Ulrich Dirnagl

Einsichten eines Wissenschaftsnarren

So urteilt ChatGPT 5:

„Über neun Jahre hinweg schrieb ein leidenschaftlicher, zugleich zutiefst skeptischer Beobachter des biomedizinischen Wissenschaftssystems die Kolumne '*Einsichten eines Wissenschaftsnarren*' für das Laborjournal. Nun erscheinen alle 77 Texte erstmals in einer Gesamtausgabe – unverändert, ungeschönt, „unplugged".

Diese Kolumnen sind keine gefälligen Randbemerkungen, sondern pointierte Kommentare und kritische Miniaturen, die den etablierten Wissenschaftsbetrieb mit scharfem Blick durchleuchten. Mit der doppelten Haltung von Enthusiast und Skeptiker feiert der Autor die Wissenschaft als Idee – und hält ihr zugleich schonungslos den Spiegel vor, wenn Systemzwänge, Eitelkeiten, Absurditäten und institutionelle Routinen ihre Lebendigkeit bedrohen.

So entsteht über die Jahre ein vielschichtiges Panorama: eine profunde Kritik der gegenwärtigen Forschungswelt, die aus Liebe zum Gegenstand formuliert ist. Mal ironisch, mal bissig, mal nachdenklich – immer aber mit dem unbestechlichen Blick dessen, der Wissenschaft ernst nimmt, indem er sie nicht unkritisch hinnimmt.

Wer die Wissenschaft wirklich schätzt, wird in diesen Texten keine Häresie finden, sondern eine Einladung: zum Nachdenken über das, was Wissenschaft sein kann – und darüber, was sie oft geworden ist."

„77 Kolumnen eines skeptischen Enthusiasten – voller Scharfsinn, Ironie und Leidenschaft."

Einsichten eines Wissenschaftsnarren

2017-2025

- nicht lektorierte Gesamtausgabe -

Ulrich Dirnagl

Impressum – Lizenzen und Bildnachweis / Links

Alle **Abbildungen** in diesem Buch wurden mit der KI-Bildgenerierungssoftware *Stable Diffusion* erstellt und nachbearbeitet. Prompt-Entwicklung: Ulrich Dirnagl. Stable Diffusion ist ein Open-Source-Projekt, lizenziert unter der CreativeML Open RAIL-M-Lizenz (https://stability.ai). Link zu einer digitalen Version und allen Abbildungen: https://dirnagl.com/narr

Druck und Bindung: pinguindruck.de; Gedruckt in Berlin

Alle Artikel wie gedruckt im **Laborjournal** (Archiv) hier online verfügbar:
https://www.laborjournal.de/rubric/narr/

Digitale Version dieser Gesamtausgabe (PDF), Zitate und weiterführende Literatur zu den einzelnen Artikeln: https://dirnagl.com/narr

A man in science past 60 does more harm than good.

Thomas Henry Huxley

Inhalt

Vorwort

Auf Einladung von Ralf Neumann, dem Chefredakteur meiner liebsten biomedizinischen Publikation – dem Laborjournal – durfte ich von 2017 bis 2025 dort einmal im Monat den Hofnarren spielen. Gemäß dem Motto *„Aufhören, wenn's am schönsten ist"* ist nach 77 Ausgaben jetzt Schluss mit den Schelmereien. Außerdem: Mein akademisches Haltbarkeitsdatum ist auch erreicht.

Aus diesem Anlass gibt's jetzt eine einmalige Collectors Edition mit allen Beiträgen. Im Laborjournal wurden die Texte stets exzellent lektoriert. Allerdings: Das Laborjournal muss seinen Spitzenjournalismus ja irgendwie finanzieren – es gibt dort reichlich Werbung. Und natürlich fehlt ein Inhaltsverzeichnis, von einem Index ganz zu schweigen.

Genau das hole ich hier nach – und präsentiere alle Beiträge in ihrer rohen, unlektorierten Originalform – *unplugged* sozusagen, wie ich sie damals beim Laborjournal eingereicht habe.

Als kleinen Bonus und zur Auflockerung gibt's zu jedem Text eine „Originalabbildung", entstanden im intensiven Dialog zwischen mir und der texttoimage KI Stable Diffusion, der ich ein bisschen Michael Sowa, Edward Hopper, Vilhelm Hammershøi und Winslow Homer beigebracht habe. Ich habe die Bilder auch ins Internet gestellt (https://dirnagl.com/lj/narr).

Unter dieser Adresse finden sich, neben der digitalen PDF Version dieses Buches auch die zitierte sowie weiterführende Literatur, da sie nicht wie bei einer wissenschaftlichen Publikation im Text Platz fanden.

Breitbrunn, 15.8.2025

Wie originell sind eigentlich Ihre wissenschaftlichen Hypothesen?

LJ 4/2017

Schon mal darüber nachgedacht, wie hoch im Schnitt der Prozentsatz ist, mit dem sie in Ihren wissenschaftlichen Hypothesen richtig liegen? Ich meine nicht den Anteil der statistisch signifikanten Ergebnisse, wenn Sie sich in neue Experimente stürzen. Es geht vielmehr um die Rate der Hypothesen, die von anderen bestätigt wurden, oder ein tatsächlich wirksames Medikament postuliert hatten. Leider werden heutzutage die wenigsten Resultate unabhängig überprüft, davon gleich mehr. Und selbst etablierte Therapien werden oft noch nach Jahren als unwirksam oder gar schädlich aus dem Verkehr gezogen. Man kann sich so einer ‚Erfolgsquote, wenn überhaupt, also nur annähern, was ich im Folgenden einmal tun will. Sie wundern sich, warum ich Ihnen diese scheinbar esoterische Frage stelle? Weil die Antwort darauf, wie hoch in etwa der Prozentsatz ist, mit dem sich Hypothesen als tatsächlich richtig erweisen, weitreichende Konsequenzen für die Bewertung von Forschungsergebnissen hätte. Den eigenen, und denen von Anderen. Und diese Frage einen überraschenden, aber direkten Bezug zur gegenwärtigen Krise der biomedizinischen Wissenschaften hat. Es geht nämlich ein Gespenst um!

Es verdichtet sich die Gewissheit, dass die meisten Studienergebnisse in Biomedizin und Psychologie sich nicht lassen reproduzieren. Nach einer kürzlich von Nature durchgeführten Umfrage glauben mittlerweile 90% der Wissenschaftler, dass wir uns mitten in einer 'Reproduzierbarkeitskrise' befinden. Davon bin auch ich überzeugt! Aber was bedeutet Reproduzierbarkeit in diesem Kontext eigentlich? Replikation des p-Werts, der Effektgröße, subjektive Einschätzung von Experten ob eine ‚Replikation' gelungen ist? Wie viel kann überhaupt reproduzierbar sein?

Ausgangspunkt der ‚Krise' waren zwei Artikel aus der pharmazeutischen Industrie. Nur in 10-20 % der Studien, welche Wissenschaftler von Amgen bzw. Bayer nachgekocht hatten, konnten sie die meist hochrangig publizierten Ergebnisse aus universitären Laboren replizieren. Nicht ganz zu Unrecht wurden die Autoren dafür kritisiert, dass sie weder die Kriterien für eine erfolgreiche Replikation preisgegeben hatten, noch welche Studien sie wiederholt hatten. Dazu kam, dass hier die Industrie ein Problem mit Resultaten aus den Universitäten hatte. Es war bei manchem in Academia deshalb schnell klar, warum die Replikationen scheitern mussten: Postdocs, die nicht gut genug für ein Karriere in Academia sind, wandern in die Industrie ab. Klar, dass die es dort letztlich auch nicht bringen: Nicht-Replikation als Folge von Kompetenzmangel. Mittlerweile konnten allerdings eine Reihe von sehr gut geplanten, systematischen Initiativen in Academia (z.B. in der Psychologie oder der Krebsforschung) ebenfalls nur einen enttäuschend geringen Teil der Resultate von ausgewählten, hochrangig publizierten Arbeiten nachvollziehen. Jetzt liest man sogar häufiger in der Zeitung, dass die Wissenschaft in einer Krise sei.

Der Kommentar eines hochrangigen Mitarbeiters der DFG hierzu: Na klar, 80 % der Befunde sind nicht reproduzierbar, aber die restlichen 20 % wurden durch uns gefördert! Wenn es nur so einfach wäre.

Denn wie viele Ergebnisse müssten eigentlich reproduzierbar sein, damit wir zufrieden wären? 80, 90, oder gar 100%? Und genau hier wird die Sache spannend, und leider auch ein bisschen kompliziert. Denn ohne Statistik und eine Prise Erkenntnistheorie kommt man hier nicht weiter! Schon 2005 hatte John Ioannidis die unerhörte (und bisher unwiderlegte) Behauptung aufgestellt, dass die meisten publizierten Ergebnisse der Biomedizin falsch sein müssten. Ergo auch nicht reproduzierbar. Seine beiden Argumente: Zum einen niedrige Qualität in Studiendesign, Analyse, und Berichterstattung. Dadurch Verzerrung (Bias) der Ergebnisse. Die Liste der Probleme ist lang, und schließt unter anderem fehlende Verblindung und Randomisierung, selektive Datenauswahl, sowie fehlende Publikation von negativen Resultaten ein. Sein anderes Argument: Zu niedrige statistische Power durch zu geringen Fallzahlen. Zur Erinnerung, Power beschreibt die Wahrscheinlichkeit, mit der tatsächlich richtige Hypothesen im Experiment bestätigt werden können. Dass beides, Bias und zu niedrige Power weit verbreitet sind, und zu einer Inflation von falsch positiven Resultaten und aufgeblähten Effektgrößen führen, ist mittlerweile durch Meta-Research gut belegt. Ich bin überzeugt davon, dass dies sehr wichtige systematische (und systemische) Ursachen für die mangelnde Reproduzierbarkeit sind. Und übrigens auch für die großen Schwierigkeiten bei der Übertragung fantastischer neuer Behandlungsstrategien im Tiermodell in am Menschen wirksame Therapien.

Aber wie steht es eigentlich um die Eingangs gestellte, fundamentale Frage: Wie viele Resultate sollten denn reproduzierbar sein? Wären in einem wissenschaftlichen Utopia, in dem Bias komplett beseitigt, und statistische Power bei 100 % liegt, alle Studien reproduzierbar? Ganz sicher nicht! Schreitet Erkenntnis nicht durch zumindest teilweise Widerlegung von bisher Anerkanntem fort? Hans-Jörg Rheinberger spricht von der ‚differentiellen Replikation von Experimentalsystemen' als wesentlichem Moment des Fortschrittes in der Wissenschaft. Danach wird im Laufe der Zeit jedes Ergebnis nur ‚differentiell', d.h. teilweise, replizierbar sein. Auch können sich Wissenschaftler irren, sie sind nicht unfehlbar.

Das bringt uns zurück zur Ausgangsfrage: Wie steht es bei eigentlich bei Ihnen? Wie viele Ihrer Hypothesen stellen sich im Rahmen Ihrer Studien als ‚richtig' heraus, und sollten damit auch replizierbar sein? Nach kurzem Zögern antworten die meisten Kollegen, denen ich diese Frage gestellt habe, mit einer Prozentzahl weit über 50 %. Man ist ja schließlich ein guter Wissenschaftler. Aber wäre das nicht tragisch, wenn sich ein hoher Prozentsatz unserer Hypothesen als richtig herausstellte? Dann läge der Verdacht nahe, dass man überwiegend triviale Hypothesen untersuchte! Dass vorher schon so viel bekannt war, dass der nächste, kleine Erkenntnisschritt schon mit großer Sicherheit vorhergesagt werden konnte. Wie langweilig!

Zum Glück sind wir aber mit unseren Hypothesen wohl weit weniger treffsicher. Wo dies formal untersucht wurde, lag die Quote eher bei 10 %. Dies hätte weitreichende Konsequenzen. Es würde zum Beispiel bedeuten, dass beim gängigen Signifikanzniveau von 5 % ($p \leq 0.05$), und einer statistischen Power wie sie in klinischen Studien gefordert (80 %) aber in den meisten präklinisch – experimentellen Studien nicht annähernd erreicht wird, mehr als ein Drittel aller statistisch signifikanten Befunde falsch positiv sind! Die meisten Experimentatoren wiegen sich in dieser Situation allerdings in einer trügerischen Sicherheit, denn sie glauben in nur maximal 5 % der Fälle falsch zu liegen. Was sie oft nicht wissen: Ein p-Wert sagt gar nichts über die Wahrscheinlichkeit aus, nach der

ein Resultat eine Hypothese bestätigt. Diese hängt nämlich nicht nur vom Signifikanzniveau ab, sondern auch von der Power, und ganz wesentlich von der Apriori Wahrscheinlichkeit der Hypothese. Nun kennen wir die a priori - Wahrscheinlichkeit unserer Hypothese aber gar nicht, und sie ist ganz sicher deutlich unter 100 %, denn wir sind ja keine unfehlbaren Langweiler. Erhöht wird die Zahl der falsch positiven Resultate noch dadurch, dass die meisten Experimente in der Biomedizin mit deutlich geringerer Power als 80% durchgeführt werden. Deshalb liegt die falsch-positiven Rate vermutlich deutlich über 50 %. John Ioannidis lässt grüßen! Und was hat das mit Reproduktion zu tun? Eben dass man falsch positive Befunde auch nicht reproduzieren kann – es sei denn durch einen weiteren falsch positiven!

Damit wird auch klar, dass explorative Forschung, es sei denn sie befasst sich mit Banalitäten, dem Wesen nach, und ganz ohne Bias und mit ausreichender Power, notwendig nicht-replizierbare Befunde erzeugen muss. Könnte man dann die Treffsicherheit durch Wiederholung des ‚positiven' Experimentes deutlich verbessern? Leider nein, es sei denn, die Fallzahl wird deutlich erhöht. Auch diese unangenehme Wahrheit ist den Wenigsten bekannt: Die Wahrscheinlichkeit, einen auf 5 % Niveau signifikanten Befund (z.B. p=0.049) der auf einer richtigen Hypothese beruht mit demselben experimentellen Setup und Fallzahl nochmals auf dem gleichen Niveau statistisch signifikant zu finden, liegt bei 50 %. Wer das verstanden hat, muss zu dem nur scheinbar verrückten Schluss kommen, dass es unter diesen Umständen besser ist, zur ‚Reproduktion ' eines Befundes eine Münze zu werfen, statt Mäuse und Ratten zu töten!

Könnte es also paradoxerweise so sein, dass insbesondere dort, wo Experimente ins wahrlich Unbekannte vordringen, vielleicht sogar bisheriges Wissen in Frage stellen, eine niedrige Replikationsrate ein Zeichen für besonders ‚heiße' und spannende Wissenschaft wäre? Es also so etwas wie notwendige, oder ‚benigne Nicht-Reproduktion' gibt? Ich denke schon. Es ist nur schwer, diese in der gegenwärtigen Literatur, in der Bias und niedrige Power ihr Unwesen treiben, von der ‚malignen Nicht-Reproduktion' zu unterscheiden. Um das zu ändern, müssen wir Experimenten mit zu geringen Fallzahlen, mangelnder Verblindung und Randomisierung, selektiver Datenauswahl, fehlerhafter Statistik, und fehlender Publikation von neutralen oder negativen Ergebnissen den Garaus machen.

Aus dem oben Gesagten lassen sich ein paar einfache Schlussfolgerungen ziehen, welche bei Umsetzung recht dramatische Wirkung hätten. Zum einen, dass es höchste Zeit ist, die ‚maligne Nicht-Replikation' zu minimieren. Da ist noch viel zu tun. Wie steht es in Ihrem Umfeld? Achten Sie als Reviewer auf Maßnahmen zur Verminderung von Bias, spektakuläre Resultate mit geringen Fallzahlen oder unerklärt asymmetrische Gruppengrößen? Veröffentlichen Sie Ergebnisse, die Ihre Hypothesen nicht bestätigt haben?

Außerdem würde es bedeuten, dass wir mit einer gewissen Rate von Nicht-Replikation leben müssen. Diese sogar dort am höchsten wäre, wo Wissenschaft so richtig ‚cutting edge' ist. Das heißt aber im Nebenschluss und ganz notwendig, dass wir mehr Augenmerk auf unabhängige Bestätigung (Konfirmation) legen müssen, welche der Exploration folgt. Und zwar um die aufregenden neuen Befunde von den in der Exploration unvermeidbar anfallenden falsch positiven Befunden zu befreien. In einem kürzlich in Nature erschienenen Kommentar schlagen Jeffrey Mogil und Malcolm Macleod vor, präklinische Studien in Top – Journalen nur noch zu veröffentlichen, wenn sie zum spektakulären und medizinisch wichtigen Grundlagenbefund gleich die Konfirmation mitliefern!

Der neben der notwendigen Verbesserung der Qualität unserer Forschung vielleicht wichtigste Schluss aus dem Gesagten ist deshalb, dass Konfirmation nicht als

zweitklassige Forschung stigmatisiert, sondern vielmehr gefördert und honoriert werden muss. Im Review Prozess, in Auswahl- und Berufungskommissionen, usw. Sie darf nicht als langweilige Fleißarbeit abgetan werden. Qualitativ hochwertige Konfirmation, und nur diese wird uns aus der Reproduktionskrise herausführen, ist eine intellektuelle Herausforderung, methodisch aufwendig, sowie Ressourcen-intensiv.

Zu Risiken und Nebenwirkungen fragen Sie Ihren Bibliothekar

LJ 5/2017

Weitgehend unbeachtet von der Wissenschaft findet derzeit Unerhörtes in deutschen Landen statt: Die Allianz der deutschen Wissenschaftsorganisationen, angeführt von der Hochschulrektorenkonferenz (DEAL-Konsortium), probt den Aufstand gegen die Verlage. Es geht um nicht weniger als den Einstieg in den Ausstieg aus dem gegenwärtigen Geschäftsmodell im wissenschaftlichen Verlagswesen! Raus aus den institutionellen Bibliotheks-Subskriptionen der Journale. Rein in den offenen Zugang zur wissenschaftlichen Literatur für alle (Open Access, OA). Finanziert durch einmalige Gebühren pro publizierte Artikel, der sogenannten Article Processing Charge (APC). Die Motive für diese Aktivitäten sind überzeugend: Das von der Gesellschaft finanzierte Wissen muss für diese auch frei zugänglich sein. Und: Die Kosten für den Zugang zu wissenschaftlichen Publikationen sind immens gestiegen. Und sie steigen jährlich weiter um über 5 %, fressen dabei den Unis ihr ohnehin schon prekäres Budget auf.

Zur Freude der großen Verlagshäuser realisieren sie mit unserer, vom Steuerzahler finanzierten und von uns produzierten, kuratierten, formatierten, und peer-gereviewten Forschung fantastische Renditen. Diese liegen bei satten 25 bis 40%, was vermutlich in keinem anderen legalen Geschäftsbereich möglich wäre. Dem Ganzen liegt ein bizarrer Tauschhandel zugrunde: Wir kaufen mit Steuermitteln unser eigenes Produkt zurück, also wissenschaftliche Erkenntnis in Manuskriptform. Nachdem wir diese den Verlagen vorab kostenlos übergeben haben. Es kommt noch toller: die Verlage geben uns unser Produkt nur leihweise, mit beschränktem Zugang, ohne Rechte auf die Artikel, und auf Widerruf zurück. Der Steuerzahler, der alles bezahlt hat, kommt nicht ran. Also nicht nur Lieschen Müller bleibt draußen, sondern auch niedergelassene Ärzte, oder Kliniker und Wissenschaftler außerhalb der Universitäten.

Nach vielen Jahren Chief-Editorschaft eines recht angesehenen Journals in meinem Forschungsbereich wundere ich mich allerdings, wieso die Rendite im wissenschaftlichen Verlagsgeschäft eigentlich nur 40 % ist. Denn sie müsste noch um einiges höher liegen! Was die Verlage tun müssen, um ein Journal zu verlegen, läuft inklusive Editoral Management System und Endherstellung der monatlichen Hefte alles nach ‚Schema F'. Die Arbeit wird ja ohnehin im Wesentlichen durch die Wissenschaft besorgt: Forschen und was rausfinden, Artikel formatiert hochladen, Editor, im Editorial board, oder Reviewer

sein, usw. Einmal für ein Journal etabliert, kann ohne zusätzliche Kosten alles, bis auf den Zitierstil und das Layout-Template, in beliebig viele andere Journale eines Verlages ‚geklont' werden. Gedruckt und per Post verschickt wird ja heute auch nichts mehr. Artikel - Downloads verursachen nur vernachlässigbare Kosten. Es wollen aber nicht nur die Verlage ordentlich verdienen, sondern oft noch jemand, daher ‚nur'30 % Rendite: Die Fachgesellschaften, denen viele der Journale gehören. Für bestimmte Zeit treten sie die Rechte an ihren Journalen an Verlage ab, wenn sie nicht sogar selbst als Verleger auftreten. Dafür kassieren sie in der Regel 5 – 6-stellige Summen. Was vielen nicht klar ist: Das System ernährt also nicht nur die Verlage, sondern auch die Fachgesellschaften, für die meisten davon ist es sogar deren Haupteinnahmequelle.

Was ändert sich aber nun, wenn das Geschäftsmodell zu OA und APCs wechselt? Welche Probleme löst das? Ganz klar, wir Autoren behalten dann das Recht auf die Wiederverwertung unserer Artikel (zumindest im Creative Commons Modell), und jedermann mit Zugang zum Internet kann sie lesen. Das wäre in der Tat ein Riesenfortschritt! Allerdings hat sich der eigentümliche Tauschhandel, bei dem wir produzieren, das Produkt verschenken, um es gleichzeitig wieder zurück zu kaufen, überhaupt nicht geändert. Statt über Subskriptionsgebühren für Journale läuft das Geschäft dann halt über APCs.

Warum sträuben sich die Verlage dann eigentlich gegen OA? Warum hat Elsevier das DEAL Konsortium auflaufen lassen, und uns für eine Weile vom Zugang zu Elsevier Journalen abgehängt? Nun, die Platzhirsche wie Elsevier, Springer Nature, SAGE etc. denken sich wohl: ‚Never change a winning horse'. Warum aufhören wenn's am schönsten ist? Auch können sie es sogar noch auf die Spitze treiben, und in sogenannten Hybrid-Modellen das tun, was man auch als ‚double dipping bezeichnet: Nämlich für das sofortige Freischalten eines Manuskriptes in einem Journal das bereits von Bibliotheken subskribiert wird, noch oben drauf APCs verlangen. Um für den Ernstfall gerüstet zu sein, testen die meisten Verlage derzeit außerdem mit ausgewählten Journalen, ob und wie sie im OA – Modell die gleichen Profite machen können.

Aber genau da liegt auch der Hase im Pfeffer: Sollten die Verlage gezwungen werden, alle ihre Journale, inklusive deren high impact Flaggschiffe wie Nature, Cell, NEJM, Lancet in OA Journale zu überführen, kann man sich jetzt schon ausrechnen, welche APCs dafür fällig sein werden! Die Verlage werden zweifelsohne die APCs so titrieren, dass wieder die alten Renditen dabei herauskommen. Sehr eindrucksvoll erkennt man die sinisteren Pläne großer Verlage bezüglich OA aus einem vor kurzem von Elsevier geleckten Dokument. Ich hoffe die Verantwortlichen des DEAL Konsortiums haben da mitgelesen (wenn nicht, siehe http://dirnagl.com/lj)!

Und noch etwas wird sich im angestrebten OA Modell nicht ändern: Wir publizieren im Zeitalter des Internets dann immer noch ziemlich genau so, wie vor mehr als 100 Jahren. Der wesentliche Fortschritt besteht heute lediglich darin, dass wir Artikel als PDFs aus dem Drucker ziehen, statt sie in der Bibliothek zu kopieren.

Nun zwingt uns aber niemand, in Journalen zu veröffentlichen, von denen wir unsere Arbeit zurückkaufen müssen. Es gäbe eine radikale, auf der Hand liegende und technisch sofort realisierbare Alternative: Unsere Artikel in von uns kuratierten Repositorien frei zugänglich zu veröffentlichen. Wir formatieren ja auch jetzt schon die Artikel professionell, wir reviewen sie, usw. Daran würde sich gar nichts ändern. Es sei denn, wir wollen auch da Neues wagen mit so sinnvollen Dingen wie open review, post – publication review, Präregistrierung, Open Data, usw. Außerdem hätten wir ja professionelle Assistenz, wir haben ja schließlich Bibliotheken mit entsprechendem Personal, und würden auch noch unglaubliche Summen freisetzen durch Einsparung der Subskriptionsgebühren und der APCs. Davon könnte man forschen, oder von einem kleinen Teil davon sogar

wieder Profis engagieren, das für uns zu organisieren, wie das z.B. der Wellcome Trust in England mit F1000Research praktiziert.

Aber halt. Ausgeträumt! Da gibt es nämlich einen Haken. Und der heißt Journal Impact Factor (JIF). Meine Beschreibung des Systems war nämlich unscharf. In Wirklichkeit verkaufen die Verlage uns gar keine Zeitschriften. Sondern den JIF! Wir tauschen JIF gegen Geld! Denn in Academia lautet die wichtigste Währung JIF. Mit Geld kann man keine Professur kaufen, oder einen Antrag bewilligt bekommen. Mit dem JIF schon! Unter http://pipredictor.com kann jede junge Wissenschaftlerin ihre persönlichen Chancen ausrechnen lassen, mit der sie sich in Academia durchsetzen wird. Wesentlicher Prädiktor: die Anzahl und der JIF der Journale, in denen sie publiziert hat. Da passt es gut dazu, dass der JIF selbst sehr viel wert ist. Er wurde letztes Jahr verkauft, von Thomson Reuters an einen Chinesischen Venture Kapitalisten, im Paket für 3,55 Milliarden US$. Warum die wohl glauben, dass sie diese ungeheure Summe sogar mit Profit wieder reinholen? Auch sollte man sich fragen, wer wohl am Ende die Zeche zahlt?

Aus Bequemlichkeit haben wir uns dafür entschieden, wissenschaftliche Leistungen nicht an deren Qualität und Relevanz festzumachen. Dazu müsste man ja Artikel lesen und diese beurteilen. Sondern wir verwenden ein Surrogat, das mit der individuellen wissenschaftlichen Leistung erstmal rein gar nichts zu tun hat. Statt wissenschaftlichem oder gesellschaftlichem Nutzen von Publikationen bewerten wir sie nach der über 2 Jahre gemittelten Zitierhäufigkeit von Artikeln in dem Journal, in dem sie veröffentlicht wurden. Als numerische Variable mit 3 Nachkommastellen. Das vereinfacht die Arbeit von Berufungskommissionen ganz ungemein, ist allerdings teuer erkauft. Zum einen bildet es die Grundlage einer ‚publish or perish' Kultur, mit all ihren Konsequenzen für die Robustheit, Transparenz, und Wahrhaftigkeit von Publikationen. Zum anderen hat es zu einem Kosmos von Journalen mit unterschiedlichen JIFs geführt, ohne die das gegenwärtige akademische Belohnungssystem kollabieren würde. Das eigentliche Kapital der Verlagshäuser besteht also gar nicht mehr in ihrer professionellen Publikationsmaschinerie. Die steht heute jedermann auch so zur Verfügung.

Seit mindestens 10 Jahren kann jeder ein Journal im Internet aufmachen – was auch einen unerwarteten und unerwünschten Nebeneffekt hat, das „Predatory Publishing". Also Journale die gegen eine APC einen Artikel ohne weitere editorische Bearbeitung veröffentlichen, und damit dem OA unverdient einen schlechten Ruf verschaffen. Mein Favorit bei den Prädatoren ist das „International Journal of Science and Nature", das mich kürzlich herzlich einlud, darin zu publizieren. Da erhält man für 2200 indische Rupien (entspricht etwa 40 €) sogar ein Nature und Science Paper im Doppelpack, ganz ohne den Ärger mit Gutachern und Editoren, bei garantierter Akzeptanz. Eindeutig das attraktivste Angebot unter 7 ähnlichen Angeboten, die mich an diesem Tag per email erreichten.

Bin ich also gegen die Umstellung auf OA? Nein, denn schlimmer als jetzt kann es nicht werden, und die freie Verfügbarkeit unserer Forschung ist ein wichtiges Ziel. Außerdem wird auf dem Weg dorthin die Zahl der Wissenschaftler grösser, welche beginnen, die Mechanismen des derzeitigen wissenschaftlichen Verlagssystems zu verstehen und sich kritisch dazu stellen. Ich fürchte, viele von uns glauben, es sei doch eh schon alles ‚OA'. Weil die meisten Wissenschaftler an den Universitäten noch ohne Nutzung ihrer Kreditkarte auf die Literatur zugreifen können, einfach per Klick auf den button ‚PDF'. Deshalb Hut ab vor dem DEAL Konsortium, gegenüber Elsevier nicht eingeknickt zu sein! Der Übergang zu OA ist aber nur dann sinnvoll, wenn wir gleichzeitig das Diktat des JIF brechen. Nur dann wird damit auch wirklich Geld gespart, uns zwar in gigantischem Ausmaß. In Deutschland allein mehrere hundert Millionen € pro Jahr. Damit könnte man

eine Menge Forschung fördern – und bei der Publikation von deren Ergebnissen die Möglichkeiten des elektronischen Publizierens endlich voll nutzen.

Werden Sie Forschungsförderer!

LJ 06/2017

Es ist Samstagmittag, die Sonne scheint. Ich begutachte gerade 9 Anträge einer Ausschreibung eines deutschen Ministeriums (je etwa 50 Seiten). Zum Glück konnte ich die Begutachtung von 4 Anträgen einer internationalen Stiftung (je etwa 60 Seiten) bereits letzte Woche abschließen. Zur Entspannung schreibe ich zwischendurch an einem eigenen Antrag für die DFG, und an einem für die EU. Ich habe den Überblick verloren, wie viele Artikel Begutachtungen ich zugesagt aber noch nicht abgeliefert habe. Aber morgen ist Sonntag, da kann ich noch was wegschaffen.

Kommt Ihnen dies Programm bekannt vor? Liegen Sie auch im Durchschnitt der Wissenschaftler, die nach verschiedenen unabhängigen Statistiken etwa 40 % ihrer Arbeitszeit mit Begutachten oder Antragschreiben zubringen? Kein Problem, denn der Tag hat 24 Stunden, und dann bleibt ja noch die Nacht für die Forschung.

Ich will aber nicht klagen. Sondern Ihnen einen Vorschlag machen, wie man viel mehr Zeit fürs Forschen erhalten würde. Interessiert? Ist aber was für starke Nerven! Ich werde nämlich eine Lanze für das Gießkannenprinzip brechen, und Ihnen eine Idee schmackhaft machen, nach der Forschungsgelder nicht mehr auf Antrag, sondern als Grundförderung für alle vergeben wird. Mit der kleinen Modifikation, dass die erhaltene Förderung zum Teil an andere Forscher weitergegeben werden muss. Klingt total verrückt, nach NFG (Nordkoreanischer Forschungsgemeinschaft)?

Erst mal ein Blick auf das gegenwärtige System – Forschungsgelder werden auf Antrag vergeben, entschieden wird per peer-review. Hat sich irgendwie bewährt. Aber wir alle kennen die Schwächen. Die wichtigste, eben schon angedeutet: Es verschlingt unglaublich Ressourcen. Beim Schreiben der Anträge, beim Reviewen derselben, in der Administration des Prozesses. Selbst unter günstigsten Bedingungen werden ja weniger als die Hälfte aller Anträge bewilligt, häufig liegt die Quote deutlich unter 10%. Bei Ablehnung sind aber die eingesetzten Ressourcen verschwendet. Vorbei sind die Zeiten, als ein Otto Warburg an die Notgemeinschaft der Deutschen Wissenschaft, dem Vorläufer der DFG, seinen legendären Einzeller schreiben konnte: ‚Benötige 10.000 Reichsmark'. Fertig war der Antrag– und natürlich wurde er auch bewilligt. Heute schreiben wir Wochen bis Monate lang an Anträgen, in denen wir taktisch vorgehen, und teils bereits Durchgeführtes beantragen, und dies mit etwas Originellem aufhübschen. Gutachterkollegen (‚peers') machen sich dann darüber. Diese lenkt das erheblich von der eigenen Arbeit ab. Und das Ganze ist von recht unvorhersehbarem Ausgang. Vor allem wenn was wirklich Originelles beantragt wurde, es also ‚risikoreich' zuging, bleiben Antrag und

damit di Innovation leicht auf der Strecke. Gefördert wird das Machbare, nicht das Mögliche. Also bevorzugt das Mittelmaß, das Konventionelle, das Inkrementelle. Und das auch noch unter Berücksichtigung des Matthäus-Prinzips („Denn wer da hat, dem wird gegeben…" – Mt 25,29). Weshalb z.B. auch das Durchschnittsalter der Antragsteller bei der DFG seit vielen Jahren immer weiter ansteigt. Oder fast jeder Nobelpreisträger den Spruch los wird: ‚Heute würde meine Forschung, die mich nach Stockholm gebracht hat, nicht mehr gefördert'. Einsteiger, und Leute denen was wirklich Neues einfällt, haben es schwer – Vierzehnender mit großen Arbeitsgruppen dagegen viel leichter. Von Interessenkonflikten, professionellen Seilschaften oder gar Fehden, die bei der Begutachtung eine Rolle spielen könnten, will ich gar nicht reden. Wir alle kennen diese Probleme, schwadronieren auch gerne beim Bier mit Kollegen darüber, insbesondere wenn uns mal wieder ein Antrag abgelehnt wurde. Das zugehörige Bauchgefühl wird durch eine umfangreiche Literatur mit empirischer Evidenz untermauert, welche unsere schlimmsten Befürchtungen bestätigt. Aber ginge es denn überhaupt anders?

Seit einiger Zeit wird ein Verfahren zur Allokation von Forschungsmitteln diskutiert, das so radikal anders funktioniert, dass man es zunächst für einen Scherz halten könnte. Und die Schellen des Narren klingeln hört. Wenn man allerdings ein bisschen darüber nachdenkt, erkennt man den immensen Charme, der in der Sache liegt.

Vom Kern her lehnt sich das Verfahren an den berühmten ‚Page rank' Algorithmus von Page and Brin an, mit dem bei Google die Webseiten bewertet werden. Die Idee geht folgendermaßen: Jeder Wissenschaftler im System erhält ein ‚Grundförderung, z.B. 100.000 € pro Jahr. Ohne weitere Bedingungen. In den USA könnte das Geld vom NIH kommen, in Deutschland von der DFG. Allerdings muss man von der Fördersumme einen bestimmten Teil, sagen wir die Hälfte, an einen oder mehrere Wissenschaftler im System weitergeben. Natürlich anonym, über die zentrale Instanz die auch die Grundförderung vergibt. An wen würde man die Mittel weiterreichen? Die Kriterien dafür setzt man sich selbst, aber naheliegend sind Originalität, Qualität, Relevanz, etc. All die Dinge, die wir ja auch beim Peer Review zugrunde legen (sollten). Natürlich würden die üblichen Regeln gelten, d.h. es dürfte niemand aus der eigenen Institution sein, mit dem man Co-Autorschaften hat, usw. Wer auf diese Weise zusätzliche Fördermittel erhält, muss auch hiervon einen bestimmten Anteil (es könnten wiederum 50%) sein, an andere weitergeben.

Wie würden sich die Mittel in einem solchen System verteilen: Analog zum Google'schen Page ranking würden bei den Wissenschaftlern mehr Mittel akkumulieren, welche von den Peers für die vielversprechendsten, tollsten, besten, etc. gehalten werden. Die eingesparten Mittel könnten zumindest teilweise an die Institutionen der Geförderten weitergegeben werden, die davon Core-Facilities finanzieren müssten. Dadurch würde für die ‚Grundgeförderten' und ihre Arbeitsgruppen eine Forschungsinfrastruktur entstehen, von der man heute nur träumen könnte. Vor allem wenn man an einer Universität arbeitet…

Das Ganze wäre eine peer-to-peer Förderung, es werden Wissenschaftler und nicht Projekte gefördert. Dieses System wurde am anschaulichsten und ausführlichsten von Bollen und Kollegen beschrieben. Diese Autoren haben das System auch in einer Simulation ‚getestet', und zwar durch Anwendung des Prinzips auf die Datenbank aller vom NIH Geförderten. In der Datenbank finden sich natürlich keine Angaben, wer wem wieviel an Fördermitteln übertragen würde. Dafür haben die Autoren als Surrogat Zitationen gewählt. Prinzip: Wen man viel zitiert, den findet man wichtig, dem würde man auch Geld geben. Sie wählten eine Grundförderung von 100.000 $, dies entspricht etwa der durchschnittlichen Förderung des NIH pro Forscher und Jahr. Resultat: Peer-to-peer

Förderung führte in der Simulation zu einer sehr ähnlichen Verteilung der Mittel wie die des NIH per Peer Review! Aber das Ganze ohne Antragsschreiberei, Review-Prozess, und der ganzen Last der Organisation des jetzigen Systems. Und wenn man die Mittel eben nicht per Zitationen, sondern auf Basis der Wertschätzung der Forschung Anderer verteilt, würde das System nicht nur massiv Ressourcen sparen, sondern wahrscheinlich auch noch innovativere und relevantere Forschung fördern. Das Verfahren ist darüber hinaus in hohem Masse steuerbar, vor allem durch die Höhe der Grundförderung, sowie die Höhe der Quote für die Weitergabe an Andere. Aber lädt das System nicht zu Absprachen, Seilschaften, usw. ein? Na ja, als ob es das nicht jetzt auch schon gäbe. Im Übrigen ist es viel einfacher in diesem System ein ‚Gaming' zu entdecken, auffällige Muster in den Mittelflüssen wären leicht identifizierbar.

Ist die Einführung von peer-to-peer Forschungsförderung realistisch? Würde es ein großer Fördergeber wagen, das gegenwärtige System, das zwar mängelbehaftet und extrem Ressourcen-intensiv ist, aber doch funktioniert, gegen etwas Unerprobtes auszutauschen? Anders ausgedrückt: Meint es der Wissenschaftsnarr wirklich ernst? Ja, in der Tat, denn man könnte so ein System parallel zum Existierenden Modell-haft erproben. Mit den Parametern spielen, klein anfangen und es langsam hoch skalieren. Das wäre doch mal ein Antrag an die DFG! Vorher muss ich heute aber noch die 9 Anträge fertig reviewen.

Exzellenztheater: Zeit für einen Wechsel im Spielplan?

LJ 9/2017

In der letzten Ausgabe dieser Kolumne (LJ 6/2017) hat der Wissenschaftsnarr allen Ernstes vorgeschlagen, Forschungsgelder nicht mehr auf Antrag, sondern als Grundförderung für alle zu vergeben. Mit der kleinen Modifikation, dass die erhaltene Förderung zum Teil an andere Forscher weitergegeben werden muss. Damit hat er sich den Boden bereitet für einen weiteren Frontalangriff auf Altbewährtes: Das Mantra von der Exzellenz. Ist der Ruf erst ruiniert, schreibt sich's ungeniert!

Viel ist schon zum Thema geschrieben worden, nicht zuletzt in der diesjährigen Sommeressay-Ausgabe des LJ (7-8/2017), ein Plädoyer von Jürgen Mittelstraß aus dem Jahr 2000. In ihm wirbt er, horribile dictu, für mehr Mittelmaß in der Wissenschaft, dafür weniger Exzellenz und Evaluation. Oder die fast 500 Seiten füllende Abrechnung mit der „akademischen Elite" von Richard Münch. In ihr charakterisiert er den Exzellenzbegriff als soziale Konstruktion zur Verteilung von Forschungsmitteln, geißelt die mit diesem Begriff vergesellschafteten Sprechblasen, und kritisiert von DFG bis zum Prinzip der außer universitären Forschung sämtliche heiligen Kühe der deutschen Wissenschaftslandschaft. Im Jahr 2007, kurz nach dem Auftakt der Exzellenzinitiative auf fast 500 Seiten bei Suhrkamp erschienen, empörte sein Buch die Vertreter der von

ihm gegeißelten „Kartelle, Monopole und Oligarchien“ und führte zu gewaltigem Rauschen in den Feuilletons der Republik.

Am Vorabend der 3. Runde der Exzellenzinitiative (jetzt: Exzellenzstrategie) erinnern sich wohl nur noch die Älteren daran. Gerade deswegen will ich mich dem Thema noch einmal ganz grundsätzlich nähern. Und weil es möglicherweise einen, für Herrn Mittelstraß damals noch nicht fassbaren direkten Zusammenhang zwischen der derzeit allenthalben beklagten Krise der (Lebens)Wissenschaften und der Exzellenzrhetorik gibt.

Was ist mit „Exzellenz“ eigentlich gemeint? Ist doch eigentlich ganz einfach, oder? Die Spitze, das Außerordentliche, die Elite, etwas Hervorragendes, usw. Bei näherem Hinsehen fällt allerdings auf, dass der Begriff keinen Inhalt hat. In der Wissenschaft gibt es exzellente Biologen, Physiker, Germanisten, Soziologen. Dass sie exzellent sind, oder hervorragend, bedeutet nur, dass sie im Vergleich zu anderen sehr viel besser dastehen, aber woran gemessen? Wir erfahren nur, dass es um die Wenigen am linken Rand einer Gauss’schen Verteilung geht. Diese Wenigen werden für Wert erachtet, belohnt zu werden, durch Professuren, mehr Forschungsmittel, ja ganze Initiativen. Und das ist beileibe kein deutsches Phänomen. Die Engländer z.B. haben ihr Research Excellence Framework (REF). Ganze Universitäten erhalten ihre Mittel relativ zu ihrer wissenschaftlichen Exzellenz. Und sie werden sagen: Aber das ist doch gut so! Und ich sage: Täuschen Sie sich da nicht!

Es stellt sich zunächst die Frage, wer eigentlich die Forscher, Projekte und Universitäten nach exzellenten und nicht-exzellenten sortiert. Und nach welchen Kriterien dies geschehen könnte. Jack Stilgoe formulierte das im Guardian (2014) so: „Exzellenz ist ein altmodisches Wort, das ein altmodisches Ideal anspricht. ‚Exzellenz‘ sagt uns nichts darüber, wie wichtig die Wissenschaft ist, aber alles darüber, wer die Auswahl trifft“. Denn es ist ganz einfach so: Die Suche nach Exzellenz wird bei den Kriterien fündig, die hierfür aufgestellt wurden. In der Biomedizin sind dies Publikationen in einer Handvoll ausgewählter Journale. Oder noch praktischer, die abstrakteste aller Metriken, der Journal Impact Factor (JIF). Was ist exzellent? Publikationen in Journalen mit sehr hohem Impact Factor. Wie wählen wir exzellente Forscher und deren Projekte aus? Durch Zählen von Publikationen mit hohem JIF. Worin zeigt sich die Exzellenz im geförderten Projekt: Durch Publikationen mit hohem JIF. Wem diese selbstreferentielle Schleife zu simplistisch ist: na klar, da kann man noch ein paar Kriterien dazunehmen, und die Schleife damit nur vergrößern. Was ist exzellent? Viele Drittmittel, bevorzugt von der DFG. Wie bekommt man viele Drittmittel? Durch Publikation in Journalen mit hohem JIF, und so weiter und so fort.

Aber ist nicht die Spitzenpublikation ein guter Prädiktor für künftige bahnbrechende Ergebnisse? Leider nein, denn wir Wissenschaftler, die wir die Arbeit im Peer Review als publikabel eingestuft haben, tun uns schwer darin, die Bedeutung und die künftige Relevanz von Forschung zu beurteilen. Dies belegen viele Studien, wie zum Beispiel diese: Die Bewertung von NIH Anträgen (genauer: der „percentile“ score) korreliert sehr schlecht mit der auf Basis von Zitationen extrapolierten Relevanz der geförderten Projekte. Hier nur als Fußnote: Für DFG Anträge könnte man so einen Zusammenhang gar nicht untersuchen, den die DFG stellt die relevanten Informationen gar nicht zur Verfügung. Am plakativsten zeigt sich unsere Unfähigkeit, Projekte oder Publikationen mit hoher Relevanz zu erkennen, in der ‚Ablehnungshistorie‘ einer Vielzahl von Arbeiten, die dann Jahre oder Jahrzehnte später mit dem Nobelpreis gekürt wurden. „Breakthrough findings“ werden nicht über Förderprogramme ausgeschrieben oder durch die Beschwörung von Exzellenz herbeigeredet. Sie „passieren“ einfach, meist wenn „Zufall begünstigt wird durch den vorbereiteten Geist“, wie es Louis Pasteur formulierte.

Die Sensitivität und Spezifität der Begutachtung von Spitzenforschung ist also ausgesprochen unbefriedigend. Von den falsch negativen werden vielleicht manche noch Jahre später entdeckt, die falsch positiven ziehen bloß Ressourcen aus dem System. Darüber hinaus hat die Rhetorik der Exzellenz aber noch weitere, korrosive Effekte. Sie fördert Narrative der übertriebenen Wichtigkeit und Effektgrößen der eigenen Ergebnisse. Sie belohnt „Abkürzungen" in Form von „flexibler" Analyse und Publikation auf dem Weg zum vermeintlich spektakulären Resultat. Dies erklärt die Inflation der signifikanten p-Werte und Effektgrößen, der behaupteten unmittelbar bevorstehenden Durchbrüche in der Therapie von Krankheiten, usw. Mancher Forscher erliegt gar der Verlockung, durch Wissenschaftsbetrug garantiert und schnell exzellente Resultate zu erhalten. Während der Drang zur Exzellenz also viele Anreize für fragwürdige wissenschaftliche Praxis bietet, ist er ein Hemmnis für „normale Wissenschaft". Normale Wissenschaft meint nach Thomas Samuel Kuhn das alltägliche, unspektakuläre Theoretisieren, Beobachten, und Experimentieren von Forschern, womit sie Wissen schaffen und konsolidieren. Normale Wissenschaft wird nur sehr gelegentlich durch ‚Paradigmenwechsel' durchgeschüttelt und dabei neu aufgestellt. Normale Wissenschaft führt nicht zu spektakulären Befunden („Stories"), sie basiert auf kompetenter Methodik, hohem Rigor und Transparenz, sie ist replizierbar. Also all das, was bei der Suche nach Exzellenz unter den Tisch fällt. Gleichzeitig ist normale Wissenschaft das Substrat für ‚breakthrough science', eben den Paradigmenwechsel. Dieser ist aber nicht steuerbar, erfolgt per Zufall, und ist auch nicht per Ausschreibung zu erzwingen. Daher, auch wenn es paradox klingt: Wer Spitzenforschung will, muss normale Wissenschaft fördern! Wer dagegen Exzellenz fördert, erhält Exzellenz, mit alle ihren Wirkungen und Nebenwirkungen. Dazu zählen natürlich Top-Publikationen, welche für sich ja keinen Wert darstellen. Außer Forscher, Initiativen, Unis, Länder in Exzellenz-Rankings nach oben zu bringen. Zudem führt die Auswahl nach Exzellenzkriterien zu Förder-Homophilie, also der Tendenz von Gutachern, Wissenschaftler zu fördern, wenn diese ihrer ähnlichen Forschung machen. Auch kommt es zur Konzentration von Ressourcen (Matthäus Effekt, „Wer hat dem wird gegeben"), zumeist auf Kosten nicht-exzellenter Bereiche, also der normalen Wissenschaft. Die Exzellenzrhetorik ist vom Wesen her rückwärtsgewandt: Sie urteilt auf Basis vorrausgegangener Exzellenz. Dadurch verringert sich die Chance, wirklich Neues zu fördern, während Rigor, Kreativität, Diversität durchs Raster fallen.

Die Rhetorik der Exzellenz hat allerdings noch eine wesentliche, auf den ersten Blick kaum zu ersetzender Funktion. Sie liefert der Wissenschaft vor der Politik ein einfaches und jedermann einleuchtendes Kriterium für die Verteilung oder gar den Aufwuchs von Forschungsmitteln. Schaut her: Bei uns fördert ihr Exzellenz! Und mit exzellenten Forschern wollen auch Politiker aufs Foto. Wie unattraktiv wäre dagegen der Ruf nach Förderung von „normaler Wissenschaft"!

Wir spielen also Exzellenztheater. Wäre es nicht an der Zeit, das Stück zu wechseln, oder zumindest das Bühnenbild? Man könnte, ganz sachte, der Rhetorik von der Exzellenz eine Rhetorik der „fundierten Wissenschaft" beistellen. Im Englischen eignet sich hierfür der Begriff „soundness". Denn soundness bedeutet Schlüssigkeit, Stichhaltigkeit, Fundiertheit und Zuverlässigkeit. Förderung von Sound Science wäre ein pluralistischer Ansatz zur Verteilung von Ressourcen. Er schließt die vielen Qualitäten, welche (gute) Wissenschaft ausmachen, mit ein. Kann man „soundness" bewerten, oder ist das nicht genauso ein „empty signifier" (inhaltsleerer Bedeutungsträger) wie „Exzellenz"? Team science und Kooperation, Open Science, Transparenz, Adhärenz zu wissenschaftlichen und ethischen Standards, Replizierbarkeit, all dies und noch eine Menge mehr lässt sich nicht nur benennen, sondern sogar bis zu einem gewissen Grad quantifizieren. Dies wären dann Kriterien für eine Förderung in der Breite. Dafür braucht es keine zusätzlichen

Mittel, denn es würde weniger Exzellenz gefördert. Als Nebeneffekt kaufen die Förderer damit auch mehr „Tickets“ in der Lotterie, welche die Forschungsförderung in Ermangelung von prädiktiven Kriterien für breakthrough science nämlich ist. Und wer mehr Tickets hat, gewinnt häufiger. Die Spitzenforschung, die neuen Therapien, die Paradigmenwechsel erwachsen dann aus einer größeren Anzahl von qualitativ hochwertigen Projekten normaler Wissenschaft. Wohl nur ein Narr hält das für machbar, oder?

Von den Gefahren allzu schöner Geschichten

LJ 10/2017

Wir Wissenschaftler sind ganz schön smart. Wir stellen Hypothesen auf, und bestätigen diese dann in einer Reihe von logisch aufeinander folgenden Experimenten. Erwünschtes Resultat folgt auf erwünschtes Resultat, mit jedem Schritt wird unsere Hypothese mehr zur Gewissheit. Nahezu ausnahmslos alle Resultate sind statistisch signifikant, manchmal auf dem 5 % Niveau, manchmal hat der p-Wert auch ganz viele Nullen. Einige unserer Experimente sind unabhängig voneinander, manche abhängig, weil sie dasselbe ‚Material‘ nutzen, z.B. für Molekularbiologie und Histologie. Nun machen wir uns ermattet aber glücklich an die Illustration und Verschriftlichung unserer Ergebnisse. Nicht nur hatten wir ein gutes Händchen bei der nun bestätigten Ausgangshypothese. Das Glück war uns umso mehr hold, da es die Kette der signifikanten p-Werte nicht abreißen ließ. Vergleichbar dem Kauf von vielen Losen einer Lotterie, bei der sich ein Los nach dem anderen als Gewinner erweist. Wenn wir dann noch die Reviewer überzeugen konnten, wird es so gedruckt.

Übertreibe ich? Ein informelles Durchblättern der führenden Journale (Nature, Cell, Science, etc.) belegt, dass die überwiegende Zahl der dort publizierten Originalartikel diesem Muster folgt. Besonders deutlich wird die von keinen Abbrüchen oder Nebenwegen getrübte Linearität dieses Musters an der Formel ‚Next we…‘, welche in vielen Artikeln mehr als 10-mal Paragraphen einleitet. Ein weiterer Hinweis besteht im fast vollständigen Fehlen von nicht signifikanten Resultaten. Dort wo man mal ein ‚n.s.‘ findet, gehört es auch in der Regel auch hin. Wenn es dort zu einer Signifikanz gekommen wäre, hätte es die Hypothese gefährdet. Wie bei einer Gruppe, die sich nicht unterscheiden sollte von einer Kontrolle, in der z.B. dasselbe Gen mit verschiedenen experimentellen Strategien manipuliert wurde.

Ein naiver Beobachter müsste zu dem Schluss kommen, dass die Autoren solcher Studien nicht nur unglaublich smart sind, sondern auch unwahrscheinlich viel Glück haben. Er könnte sie gar für Aufschneider oder Betrüger halten. Nach ein paar Jahren in der Wissenschaft wissen wir aber alle, dass da etwas ganz anderes dahintersteckt. Wir erzählen uns nämlich gegenseitig Geschichten (‚Stories‘). Die jahrelange Arbeit an der ‚Story‘ im Labor verlief ganz anders. Vieles ging schief, manches war uneindeutig, oder die Resultate passten nicht zur Hypothese. Strategien wurden gewechselt. Die

Hypothese revidiert. Und so fort. Die ‚glatte' Geschichte wurde also ex-post entwickelt und erzählt. Ist also eigentlich tatsächlich eine ‚Story'.
Aber ist das ein Problem? Wir wissen doch alle, dass es nicht so verlief wie erzählt. Außerdem interessieren wir uns auch gutem Grund nicht für all die Probleme und Holzwege, in die wir bei unserer wissenschaftlichen Exploration geraten. Die lesen sich nicht gut, würden uns mit unnützer Information überfluten. Auf der anderen Seite aber öffnet das Geschichtenerzählen einer Reihe von Untugenden Tür und Tor. Zum Beispiel dem ‚Outcome switching' und der selektiven Verwendung von Resultaten. Dies wurde verglichen mit dem ungerichteten Abfeuern eines Schusses auf eine Holzwand, auf der man dann um das Einschussloch eine Zielscheibe malt. Mit dem Loch in der Mitte. Blattschuss! So kann man nämlich jede beliebige Hypothese ‚beweisen'! Auch erfahren wir nichts über Resultate, die es nicht in die Story geschafft haben, uns aber zu anderen Hypothesen und neuen Erkenntnissen führen würden.
Lassen Sie uns deshalb an dieser Stelle die Frage stellen, woher es eigentlich kommt, dass sich die Berichterstattung über wissenschaftliche Entdeckungen fast vollständig von den Prozessen im Labor abgelöst hat, welche diesen zugrunde liegen? Ist das ein Produkt unserer Vorliebe für aalglatte, möglichst spektakuläre Stories? Unseres akademischen Belohnungssystems, welche diese belohnt, insbesondere wenn in Journalen mit hohem Impact Factor publiziert? Überraschenderweise nein. Die Rhetorik einer linearen, ununterbrochenen und fehlerlos, logisch und zeitlich mit Notwendigkeit zum Beleg der Ausgangshypothese fortschreitenden Kette von Experimenten ist mehrere hundert Jahre alt. Im ausgehenden 17. Jahrhundert wurden Experimente noch kaum publiziert, sondern vor Publikum, also quasi vor Zeugen vorgeführt. Die Ausweitung und Internationalisierung der ‚wissenschaftlichen Community', die zunächst vorwiegend von privatisierenden Gentlemen durchgeführt wurde und dann mehr und mehr von ‚Professionals', brachte die Notwendigkeit zur weithin sichtbaren Publikation. Diese Veröffentlichungen entwickelten sich unter der Schirmherrschaft der in dieser Zeit gegründeten wissenschaftlichen Gesellschaften. Federführend war hier die Royal Society in England, mit ihren noch heute publizierten ‚Proceedings'. Insofern die Experimente jetzt ohne ‚Zeugen' durchgeführt wurden, und ein sehr gemischtes und noch wenig spezialisiertes Publikum angesprochen wurde, mussten die Leser für den Gegenstand interessiert, und von der Güte der Experimente und deren Resultaten überzeugt werden. Der Rest ist im wahrsten Sinne ‚Geschichte'. Die Dissoziation von der tatsächlichen Logik und Praxis einer Studie zu deren Repräsentation in der zugehörigen wissenschaftlicher Veröffentlichung zugunsten einer ‚Story' ist heute Standard, und das nicht nur in der Biomedizin.
Eine lange Tradition, wir haben uns daran gewöhnt, und nur so werden Publikationen überhaupt von den Journalen akzeptiert – also alles gut? Ich denke nein. Zum einen, weil heute viel mehr Studien veröffentlicht, und diese wesentlich mehr Informationen in Form von Substudien enthalten, und diese methodisch wie konzeptionell wesentlich komplexer sind. Dies bedeutet, dass die Zahl der ‚Freiheitsgrade' massiv zugenommen haben, welche es den Autoren ermöglichen, durch Selektion von ‚erwünschten' Ergebnissen praktisch jede beliebige Hypothese zu ‚belegen'. Und weil es heute üblich geworden ist, die Generierung von Hypothesen durch Exploration und deren Konfirmation in einer einzigen Studie zu vermengen. Und dann wird es für den Leser ganz unübersichtlich. Wie viele Experimente wurden durchgeführt welche es nicht in die Publikation geschafft haben? Und warum nicht? Wurde die Hypothese ‚unbiased' mittels explorativer Experimente generiert, und dann in darauffolgenden unabhängigen Experimenten bestätigt? Wurde für die Konfirmation die Hypothese eindeutig formuliert, die dafür nötige Fallzahl bestimmt, und Bias soweit als möglich ausgeschlossen? Also z.B. die Experimente randomisiert und verblindet durchgeführt?

Wie aber könnte man das Risiko minimieren, uns selbst und unseren Lesern durch die selektive Verwendung von Ergebnissen zum Zwecke des ‚Storytelling' in die Irre zu führen? Wie die Ergebnisse robuster machen und in ihrer Gesamtheit der wissenschaftlichen Community zur Verfügung stellen?
Eigentlich ganz einfach. Zunächst müssten wir Exploration und Konfirmation klarer voneinander Trennen. In der Exploration suchen wir nach neuen Phänomenen. Da kann man nicht alles vorausplanen, z.B. Fallzahlen vorher abschätzen. Man kann die Richtung, welche die Experimente nehmen, aufgrund der eingehenden Befunde ändern. Man muss dem Zufall (‚Serendipity') eine Chance geben. Man braucht keine Teststatistiken, muss die erhobenen Daten nur sehr gut in ihrer Verteilung beschreiben (z.B. Konfidenzintervalle). Wie man überhaupt alles sehr genau beschreiben muss, um die Resultate nachvollziehbar und wiederholbar zu machen. Das Ergebnis solcher Discovery-Phasen sind Hypothesen. Notwendigerweise werden sich, wegen der Originalität der so gewonnenen Hypothesen sowie der niedrigen Fallzahlen in solchen Experimenten, viele falsch positive Ergebnisse einstellen. Auch werden die Effektstärken überschätzt (siehe LJ 4/2017; ‚Wie originell sind eigentlich Ihre Hypothesen'). In einer darauffolgenden Phase müssen die Ergebnisse und Hypothesen dann, sofern man sie für interessant und wichtig genug hält, in einer separaten Studie konfirmiert werden. Hier geht es darum, die falsch positiven auszusortieren, und die wahren Effektstärken zu ergründen. Nun muss die Hypothese vorab formuliert werden, die Fallzahlen so abgeschätzt werden, dass man akzeptable Typ I und II Fehlerraten erhält, usw. Man wird vor Beginn der Experimente einen detailierten Analyseplan erstellen, und von diesem und den darin niedergelegten Teststatistiken nicht mehr abweichen. Sollten wider erwarten Abweichungen vom Studien- und Analysenplan nötig geworden sein im Lauf der Untersuchung, wird man diese begründen und berichten. Idealerweise sollte man so eine konfirmatorische Studie vor Beginn registrieren (z.B. mit time stamp und bis zur Veröffentlichung mit Embargo beim Open Science Framework), um bei Publikation belegen zu können, dass man eben keine ‚Geschichte' erzählt hat. Es kommt damit zur klaren Trennung von explorativen und konfirmatorischen Studien, die natürlich nach Abschluss auch in einer Veröffentlichung publiziert werden könnte, wie dies Mogil und Macleod kürzlich in Nature für alle experimentellen Studien in hochrangigen Journalen gefordert haben.

Eine solche einfache Trennung im Design, der Analyse und Publikation von explorativen und konfirmatorischen Studien könnte die Transparenz, Validität und Reproduzierbarkeit in der experimentellen biomedizinischen Forschung deutlich erhöhen. Einziger Nachteil: Wir müssten auf etliche spektakuläre (aber dann nicht reproduzierbare) Studien verzichten.

Frage nicht, was das Experiment für Dich tun kann – frage was Du für das Experiment tun kannst!

LJ 11/2017

Eigentlich wollte ich diesmal die Physik als Vorbild herausstellen. Als Champion einer Publikationskultur und von Team-Science, von dem wir in den Lebenswissenschaften viel lernen könnten. Und dann das: Der Nobelpreis für Physik an Rainer Weiss, Barry Barish und Kip Thorne, für den ‚experimentellen Beleg' für die von Albert Einstein 1919 vorausgesagten Gravitationswellen. Publiziert in einer Arbeit mit mehr als 3000 Autoren!

Viel ist wieder kritisiert worden am Nobelpreis. Dass ihn immer die ‚alten weißen Männer welche an amerikanischen Universitäten lehren' kriegen. Oder dass der gute Herr Nobel eigentlich bestimmt hatte, dass nur einer ausgezeichnet wird pro Gebiet, und auch nur für eine Entdeckung im zurückliegenden Jahr. Geschenkt! Denn wirklich schwer wiegt, dass der Nobelpreis damit abermals ein absolut antiquiertes Bild von Wissenschaft perpetuiert: Die einsamen, genialen Forscher, von denen es nur wenige, genauer gesagt maximal drei pro Gebiet (Medizin, Chemie, Physik) gibt, welche für die „Menschheit den größten Nutzen geleistet haben". Verliehen mit einem Spektakel, das einem Eurovision Song Contest oder der Oskar-Verleihung alle Ehre macht. Es wundert mich ja nicht, dass dies von der Öffentlichkeit begeistert aufgenommen wird. Dort existiert diese Cartoon-hafte Vorstellung von Wissenschaft spätestens seitdem bereits erwähnten Albert Einstein. Und dieses Bild von Wissenschaft hatte von Newton bis zum 2.Weltkrieg, also vor der Industrialisierung und Professionalisierung von Forschung durchaus Berechtigung. Beunruhigend finde ich aber, dass sich die wissenschaftliche Community in große Stile auf diesen Anachronismus einlässt. Nun werden Sie nun fragen, warum regt sich der Narr jetzt schon wieder auf? Ist doch bestenfalls harmlos, die Preisträger fast immer preiswürdig. Und kann der Wissenschaft nicht ein wenig PR schaden in heutigen postfaktischen Zeiten, in der Impfgegner und Klimawandel-Leugner fröhliche Urstände feiern?

Ich meine, es lohnt sich tatsächlich, den Nobelpreis zu hinterfragen, weil dessen Bild von der Wissenschaft als Sache von vereinzelten Genies komplett an der Sache vorbeigeht. Und rückständig ist, letztlich sogar wissenschaftsfeindlich. Natürlich gibt es diese Ausnahmewissenschaftler. Ihr Beitrag ist wichtig. Aber der Fortschritt der Wissenschaften basiert doch wesentlich auf der Leistung vieler, ebenso origineller wie fleißiger ForscherInnen. Die dann am effektivsten vorankommen, wenn sie zusammenarbeiten. Und ‚normale Wissenschaft' im Kuhn'sche Sinn betreiben (siehe LJ 17/2009). Die internationale LIGO-Kollaboration, welche die Gravitationwellen nachgewiesen hat, ist doch das glatte Gegenteil einer three man show. Sie publizierte ihre Ergebnisse als ‚LIGO Scientific Collaboration', mit jeweils über 1000, manchmal 3000 Autoren und Hunderten von Institutionen. Und das ist gar nichts Besonderes in der Physik, insbesondere der Teilchen- und Astrophysik.

Dort erkannte man erstmals im Manhattan Project, dass große, komplizierte, die Grenzen des momentanen Machbaren überwindenden Projekte nur durch große kollaborierende Teams zu lösen sind. Denn das Projekt zur kriegstauglichen Nutzbarmachung der Kernspaltung stand noch dazu unter massivem Zeitdruck. Heute gilt der Large Hadron Collider des CERN in Genf als Mustereinrichtung einer multinationalen Forschung zu den großen physikalischen Fragen der Menschheit. Von dort kommen Publikationen mit mehr als 5000 Autoren, in alphabetischer Reihenfolge. Publiziert wird übrigens nur noch selten in dem Prestige – reichen Journalen wie Physical Review Letters oder Nature.

Die Physik-Community hat sich nämlich mit ArXiv schon in den frühen 90er Jahren des vorigen Jahrhunderts einen Dokumentenserver für Preprints geschaffen, der heute weltweit das wesentliche Forum der wissenschaftlichen Kommunikation in Physik und Mathematik darstellt. Völlig kostenlos für Autoren und Leser, und ganz ohne Review. In Peer Review Journale wird heutzutage nur noch ein geringer Anteil der Artikel submittiert, und wenn, dann häufig schon mit dem Feedback der Fachwelt aus der Preprint-Phase. Die Publikationslisten der Physiker sind voll dieser Arbeiten, auch die ‚Top 5' - Auswahl. Häufig ist man da auch nicht Koautor. Wie auch, bei Listen mit 1000 Autoren? Professor werden oder Anträge durchbringen kann man dort auch mit ArXiv-Papers. Die werden nämlich, wenn sie einen relevanten Beitrag leisten, auch gelesen. Und die Qualität eines Forschers misst sich ganz wesentlich am Beitrag zur Lösung der gemeinsamen Fragestellung.

Man vergleiche dies mit den Lebenswissenschaften. Die dort bearbeiteten Fragen sind doch in ihrer Komplexität denen der Physik absolut ebenbürtig: Krebs, Demenz, Altern.... Auch dies sind große Fragen der Menschheit, sogar unter größerem Zeitdruck als die Suche nach dem Higgs Boson, oder einer Gravitationswelle. Schreit dies nicht nach Manhattan-artiger Kollaboration? Stattdessen arbeiten wir in Gruppen von im Schnitt 8 Forschern (inklusiv Studenten), welche durchweg Arbeitsverträge im Bereich weniger Jahre haben, mit Fördergeldern welche für 3 Jahre gesichert sind. Die Forschungsstrategien und Ergebnisse werden bis zur Publikation unter Verschluss gehalten, man könnte ja gescoopt werden. Data sharing? Um Himmelswillen. Hab ich ja nichts davon, könnte im Zweifelsfall sogar schaden. Wie verrückt ist die Frage, ob sich die grundlegenden Fragen der Biowissenschaften und der Medizin nicht besser in multinationalen, koordinierten, ausreichend und langfristig alimentierten Kooperationsprojekten aufklären ließen? Sollte man das nicht mal wenigstens ausprobieren?

Sie verdrehen die Augen? Sie denken an EU-Antrags Bürokratie, an AZA-Formulare, endlose Listen von Milestones und Deliverables? Bei den derzeitigen Kollaborationen handelt es sich aber gar nicht um Zusammenarbeit im CERN-Stil. Es sind vielmehr meist Beutegemeinschaften, welche lokale Projekte finanzieren, die aus anderen Quellen keine Förderung erhalten haben, oder so auf finanziert werden. Selbst die Genetik kann hier leider kaum Vorbildfunktion haben. Das Human Genome Project war nicht das Manhattan – Projekt der Medizin, auch wenn manch einer sowas behauptet hat. Dessen Kern war die Verteilung von Sequenzierarbeit über viele Labore. Letztlich war das Auftragsarbeit im industriellen Maßstab, und konnte deshalb auch von einer Firma gescoopt werden. Echte Kollaboration im Stile des CERN funktioniert anders. Hier werden Projekte individuell oder gemeinsam entwickelt, dann in einem wissenschaftlichen Diskurs priorisiert, permanent optimiert und im Team durchgeführt. Und das Team bekommt den ‚Credit'.

Aber warum funktionieren die Lebenswissenschaften ganz anders? Und muss das so sein? Liegt es daran, dass die erwähnten Projekte der Physik nur an Maschinen

durchgeführt werden können, deren Anschaffung im Haushalt von nationalen Volkswirtschaften sichtbar werden? Der Zwang zur Kooperation spielt hier sicher eine große Rolle. Es hat aber auch sehr viel mit der Wissenschaftskultur der einzelnen Disziplinen zu tun, die natürlich wiederum von Infrastrukturfragen beeinflusst wird. Karin Knorr-Cetina hat in ‚Epistemic Cultures' (How the Sciences Make Knowledge, Harvard Press), die Organisation und Durchführung der Forschung in der Hochernergiephysik (HEP) und der Molekularbiologie verglichen. Wie sind jeweils Labore strukturiert, wie werden Gruppen geleitet, wie findet ‚Wettbewerb' statt und auf welchem Level, wie wird kooperiert, usw.?

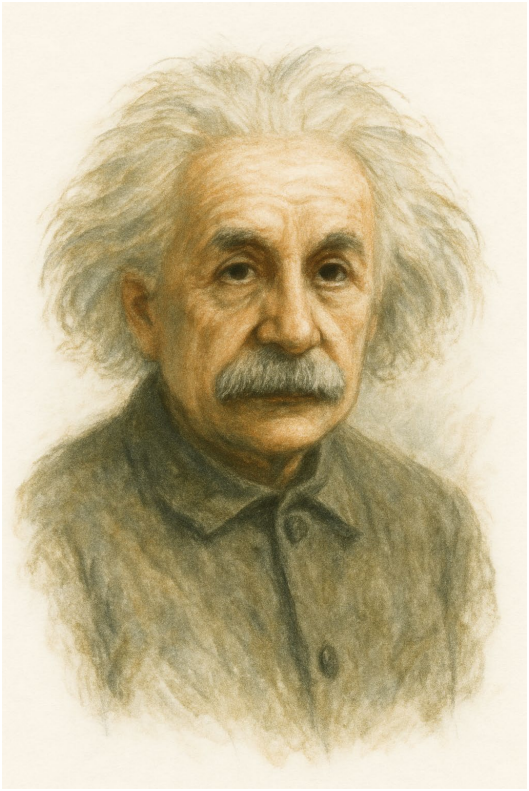
Die von ihren ausgemachten Unterschieden könnten nicht drastischer sein. Sie kommt zu dem Schluss, dass die Hochenergiephysik (HEP) das ‚epistemic subject', also den individuellen Wissenschaftler, zugunsten des bedingungslosen Austausches von Wissen aus dem Zentrum des gemeinschaftlichen Forschungsprozesses verdrängt hat. Frei nach dem Motto: ‚Frage nicht, was das Experiment für Dich tun kann – frage was Du für das Experiment tun kannst'. Umgekehrt regieren ihrer Analyse nach in der Molekularbiologie Individuen, deren Forschung auf einer Logik des Austausches beruht. Für jede Handlung wird im Gegenzug eine Leistung erwartet. Weil auch in der Molekularbiologie der Fortschritt individueller Forscher von Kollaboration abhängt, entstehen uns allen wohlbekannte Konflikte. Wenn sich ein Beitrag zum Fortschritt des Faches nicht auf ein Individuum zurückführen lässt (z.B. durch Erst/Letztautorschaft), ist es für dieses ‚vergeudet'. Daraus resultieren dann auch Probleme innerhalb eines Labors. Diese Konflikte werden durch die Hierarchien in den Laboren im Schach gehalten, die ja in der Regel mit dem Namen des Leiters assoziiert sind. Im Labor regiert ebenfalls die Logik des Austausches: Der Doktorand als Wasserträger des Gruppenleiters, der Gruppenleiter als Wasserträger des Einrichtungsleiters (dies vor allem in der Medizin). All dies existiert in der HEP entweder gar nicht, oder nur rudimentär. Man erkennt, dass hinter der Frage, wieviel in einer Wissenschaftsdomäne kollaboriert wird, nicht nur der Anschaffungspreis von Forschungssinfrastruktur, sondern auch gravierende kulturelle Unterschiede der Forschungsorganisation lauern.

Was bedeutet all dies, wenn wir das Potential von Kollaborationen im Geiste von CERN (wenn auch nicht gleich in deren Maßstab) entwickeln wollen? Dass es um wesentlich mehr geht als die Anschaffung gemeinsamer Geräte, oder die Verteilung von Workpackages a la EU-Projekt. Unsere Wertschätzung des gemeinsamen Arbeitens an Hypothesen und deren Bestätigung, der Austausch von Material und Rohdaten, das Publizieren der Ergebnisse muss sich grundlegend ändern. Solange das Fortkommen des individuellen Forschers vom Studenten zum Professor einzig und allein an deren individuellen Leistung, und nicht auch am Beitrag für das Fortkommen des Gebiets abhängt, wird das nicht passieren.

Und da wären wir wieder, beim Nobelpreis – dem ultimativen Argument für den Individualismus und gegen die Kollaboration. Dieser Physik-Nobelpreis ist ein Atavismus, weil Physik so schon lange nicht mehr funktioniert. Man könnte natürlich, aber dreht Wissenschaftsnarr wieder völlig frei, die Nobelpreise in Physik, Chemie, und Medizin auch an Kollaborationen vergeben, so wie beim Friedensnobelpreis. Vielleicht wäre das ein schönes Signal, auch und gerade für die Lebenswissenschaften.

Und die Moral von der Geschicht‘: Glaube Deinem p-Wert nicht!

LJ 12/2017



In der letzten Ausgabe des Laborjournals hat sich der Narr die Wissenschaftskultur in der Physik vorgenommen, und dort einiges gefunden, was wir Lebenswissenschaftler von denen abschauen könnten. Überhaupt ist die Physik, und da insbesondere die Teilchenphysik, eine Fundgrube von Lehrstücken. Zwei sehr aktuelle davon will ich heute mit Ihnen diskutieren.

Manch einer wird sich erinnern: Im Jahr 2011 erschütterte das Resultat eines großen, internationalen Experiments nicht nur die Physik, sondern die ganze Welt. Am 22. September titelte die New York Times auf Seite 1 „Einstein, roll over? Tiny neutrinos may have broken cosmic speed limit”! Was war geschehen? Ein sehr komplexer Versuchsaufbau war aufgeboten worden, um die Geschwindigkeit von Neutrinos zu messen. Sie wurden vom Teilchenbeschleuniger des CERN in Genf produziert, auf eine 730 km lange Reise geschickt. Dann wurde deren Ankunft von einem Detektor registriert, der durch Tausende von Metern Stein in die Dolomiten gesprengt wurde. Siehe da: Die Neutrinos kamen schneller an, als dies Photonen auf derselben Strecke gelungen wäre! Auch dem Nichtphysiker wird sofort klar, was da alles auf dem Spiel steht (Spezielle Relativitätstheorie) oder dann vielleicht möglich würde (z.B. Zeitreisen). Das hatten natürlich auch die Physiker gleich begriffen, weshalb sie ausgesprochen vorsichtig waren: Zum einen erhöhten sie das in der Teilchenphysik für die Entdeckung neuer Elementarteilchen geforderte Signifikanzniveau von sagenhaften 5 Sigma (entspricht $p< 3*10^{-7}$!) auf 6 Sigma. Außerdem wiederholten sie das Experiment mehrmals. Trotzdem, kein Zweifel, die Neutrinos machten sich nichts aus der Lichtgeschwindigkeit, und das Signifikanzniveau war unerreichte 6.2 Sigma. Daher wurde flugs die Weltpresse informiert, und ein Paper geschrieben. Allerdings hatten die Autoren da bereits trotz rekordverdächtigem p-Wert Zweifel am eigenen Befund, weshalb der Artikel endet: “The potentially great impact of the result motivates the continuation of our studies in order to investigate possible still unknown systematic effects that could explain the observed anomaly”.

Wir alle wissen, dass wir beim Zeitreisen nicht über das Kino-Stadium herausgekommen sind. Und Photonen immer noch den absoluten Geschwindigkeitsrekord halten. Was also war passiert? In den auf den Medienrummel folgenden Wochen haben die Physiker sich ihren Versuchsaufbau nochmals genau vorgenommen. Und fanden, dass das zur Entfernungsmessung genutzte GPS nicht korrekt synchronisiert war. Außerdem, man glaubt es kaum: ein Kabel war locker!

Und die Moral von der Geschicht: Glaube Deinem p-Wert nicht!

Die Physiker hatten zwar gut daran getan, für eine sehr unwahrscheinliche Hypothese ein radikal niedriges Signifikanzniveau anzusetzen. Aber, und das scheint trivial, wenn der Versuchsaufbau einen systematischen Fehler beinhaltet, nutzt weder ein noch so niedriger p-Wert, noch eine Replikation am selben Versuchsaufbau. Was lernen wir für die Lebenswissenschaften: Ein p-Wert kann einem bei der Beantwortung der Frage, ob unsere Hypothese richtig ist, eine Droge wirkt, etc., recht wenig, und oft gar nichts nützen. Und: Eine Replikation eines Experiments im selben Labor ist von sehr bedingtem Wert (Siehe auch Wissenschaftsnarr Laborjournal 4/2017).

Die ganze Sache ist nun deshalb so aktuell, weil ein All Star Team aus Statistik, Epidemiologie, und Psychologie in Nature Human Behaviour gerade einen Aufsehen erregenden Vorschlag gemacht hat. Nämlich das von Fisher in den 1920er Jahren eingeführte Signifikanzniveau um eine Größenordnung abzusenken. Von dem von uns fast wie eine Naturkonstante behandelten $p<0.05$ auf $p<0.005$! Die Autoren haben natürlich recht, das könnte die Rate der falsch positiven Resultate, unter der wir alle zu leiden haben, deutlich reduzieren. Und damit auch die Anzahl publizierter Studien, denn an der Hürde 0.005 würden viele Publikationen scheitern.

Ich halte den Vorschlag, auch mit Blick auf die OPERA-Schlappe, dennoch für einen Fehler. Den Experten, welche diese Absenkung vorschlagen, ist klar was ein p-Wert ist, und was nicht. So wissen sie, dass nicht nur Alpha, also der Typ-I Fehler für die Frage wichtig ist, ob ein Ergebnis falsch positiv ist. Das hängt nämlich auch von Beta (also dem Typ II-Fehler, bzw. der Power), und der Wahrscheinlichkeit ab, mit welcher die Hypothese richtig ist. Sie verwechseln also den p-Wert nicht mit dem positiv prädiktiven Wert, wie so viele von uns. In dem sie aber die Aufmerksamkeit in dieser Weise auf den p-Wert, ja einen bestimmten p-Wert, lenken, adeln sie ihn. Sie erwecken damit den Anschein, dass der p-Wert eben doch geeignet ist, zwischen richtigen und falschen Hypothesen zu unterscheiden, er muss nur den richtigen Wert haben. Wer den Artikel aufmerksam liest, wird alles Richtige dazu erfahren. In der Berichterstattung zu diesem Vorschlag ging es aber einzig um die neue Schwelle, und damit die ‚Rettung' des p-Werts.

Und deshalb hier gleich noch ein für uns Lebenswissenschaftler lehrreiches Beispiel aus der Physik. Bei OPERA ging es um eine sehr unwahrscheinliche Hypothese, am Ende war das Resultat trotz exorbitant niedrigem p-Wert falsch positiv. Im experimentellen Aufbau steckte ein systematischer Fehler. Beim LIGO Experiment, mit dem man kürzlich die lange gesuchten Gravitationswellen endlich nachweisen konnte, war es umgekehrt: Hier glaubte man das Ergebnis schon vorher zu kennen. Die von Einstein 1919 vorausgesagten Gravitationswellen muss es geben, denn alle Voraussagen der allgemeinen Relativitätstheorie hatten sich bisher experimentell belegen lassen. Und es gab kein ernsthaftes Argument, warum es Gravitationswellen nicht geben sollte. Das Problem hierbei war nur, dass man praktisch seit 1919 non Stopp versucht hatte, sie nachzuweisen. Aber erfolglos. Mit anderen Worten, die Experimentalphysiker fuhren ein NULL-Resultat nach dem anderen ein. Sie haben aber trotzdem nicht aufgegeben, es zu Recht auf die mangelnde Sensitivität ihres Experimentes geschoben, und an deren Verbesserung gearbeitet. Und die Moral von der Geschicht: Traue deinem p-Wert nicht! Die nicht-signifikanten p-Werte (d.h. die NULL-Resultate) bedeuteten eben nicht, dass es das untersuchte Phänomen nicht gibt. Auch hier war letztlich der experimentelle Aufbau der LIGO-Vorläufer systematisch ‚fehlerhaft'.

Was lernen wir aus diesen scheinbar exotischen Beispielen aus dem Reich der Physik, also der wohl härtesten aller Naturwissenschaften? Statistische Signifikanz, oder die Abwesenheit derselben, ist wenig hilfreich, wenn es um die Frage geht, ob unsere Hypothesen richtig oder falsch sind. Statistische Signifikanz wird überschätzt, von uns

Wissenschaftlern, genauso wie von Journal Editoren und Reviewern. Deshalb kann auch nur ein Narr dazu raten, sich bei der Beurteilung oder der Publikation von wissenschaftlichen Resultaten mehr auf die Effektstärken, die Varianzen und vor allem die Güte des experimentellen Designs zu stützen, als auf p-Werte uns statistische Signifikanz.

Wer's glaubt, wird selig!

LJ 1-2/2018

Die Medizin ist voller Mythen. Manchmal hat man sogar den Eindruck, dass sie hauptsächlich auf Mythen beruht. Viele dieser Mythen sind so plausibel, dass man ein Narr sein muss, nicht daran zu glauben. Deshalb möchte ich mich heute dem Placebo-Effekt zuwenden. Dabei werden wir uns auch einem weithin unbekannten Phänomen zuwenden, der Regression zum Mittelwert. Und die ist auch für Experimentatoren wichtig.

Kaum einer zweifelt an der geradezu magischen Effektivität des Placebo-Effektes. Es wird Sie deshalb vielleicht verwundern, dass es recht wenig Evidenz für seine Existenz gibt. Und einige gewichtige Argumente gegen ihn. Cochrane-Reviews, immerhin der goldene Standard des systematischen Reviews, konnten keine überzeugenden Belege für seine Effektivität finden. Möglicherweise sind Placebos wirksam bei Therapieresultaten, die Patienten selbst berichten (‚patient reported outcomes, PROMS'), insbesondere bei Schmerz und Übelkeit. Allerdings sind die Effekte, sollten sie exisitieren, wohl recht gering. Keine Wirksamkeit zeigte sich bei sog. ‚observer reported outcomes', also immer, wenn die Studienärzte etwas gemessen hatten.

Weil Sie den Placebo-Effekt für eine der Grundfesten der Medizin halten, und mich für einen Narren, werden Sie jetzt möglicherweise diesen Artikel kopfschüttelnd beiseitelegen. Oder Sie geben mir die Chance, Ihnen ein paar Argumente zu liefern, warum es sich hierbei vielleicht tatsächlich um einen Mythos, in jedem Fall aber um ein deutlich überschätztes Phänomen handelt. Sie würden dann auch etwas über die Regression zum Mittelwert erfahren. Dies könnte vielleicht sogar für Ihre eigene Forschung Relevanz haben.

Ein zufällig über oder unter dem Mittelwert ausfallender Messwert wird tendenziell gefolgt vom Resultat einer Messung, welche näher am Durchschnitt liegt. Trivial nicht? Noch simpler ausgedrückt: Je weiter ein Messwert vom Mittelwert abweicht, desto unwahrscheinlicher ist er. Der Naturforscher und wissenschaftliche Tausendsassa Francis Galton (1822-1911) hat dies erster erkannt, und dem Phänomen auch seinen Namen gegeben: Regression zum Mittelwert. Er nutzte 1886 Bevölkerungsregister um die Körpergröße von Eltern und deren ausgewachsenen Kindern (also in deren Erwachsenenalter) zu vergleichen. Dabei fand er, dass ausgewachsene Kinder im Schnitt näher an der Durchschnittsgröße liegen, als deren Eltern. Und nur scheinbar paradoxerweise, dass ein großes Kind in der Regel Eltern hat, die kleiner sind, als es selbst (mehr dazu bei Senn 2011). Aber was hat das denn nun mit dem Placebo-Effekt zu tun?

Patient wird man, wenn man Krankheitssymptome hat. Zum Arzt geht man, wenn diese nicht mehr ertragen möchte oder kann. Der tut dann irgendwas, und zum Glück geht es einem nach einer Weile (scheinbar) auf Grund ärztlicher Kunst häufig besser. Oder man geht nicht zum Arzt, sondern weiß selber oder aus der Apotheken Rundschau, welche Medizin am besten für einen ist (z.B. Bachblüten oder Ibuprofen). Nachdem man die Medizin genommen hat, wird es nach einigen Tagen meist besser, und nach einigen Wochen ist der Spuk vorbei. Voltaire (1694-1778) hat das so formuliert: „Die Kunst der Medizin besteht darin, den Kranken solange abzulenken, bis die Natur die Krankheit geheilt hat". Neben der Erklärung der scheinbaren Wirksamkeit von Homöopathie liegt genau hier auch der Hase im Pfeffer beim Placebo-Effekt. Als solchen bezeichnen wir die Verbesserung der Symptome mit einem Scheinmedikament oder einer Scheinprozedur. Die wird, in den guten Studien, randomisiert kontrolliert und verblindet mit dem echten Wirkstoff oder Prinzip („Verum") verglichen. Dummerweise fehlt aber in fast allen randomisiert kontrollierten Studien eine echte Kontrollgruppe! Nämlich eine, die überhaupt keine Behandlung erhält. Nur im Vergleich zu dieser könnte man überhaupt von einem Placebo-Effekt sprechen. Eine solche Untersuchungsgruppe könnte klären, wie sich die Krankheit natürlich, also ohne Behandlung entwickelt, und ob Verum und Placebogruppe einen davon abweichenden Verlauf nehmen.

Zum Glück gibt es aber auch solche Studien. Und aus diesen wissen wir, dass der natürliche Verlauf der meisten Erkrankungen fluktuierend ist. Und meist am Höhepunkt der Symptome behandelt wird. An dem Punkt, wo es natürlicherweise wieder besser wird. Und Placebo in der Regel nicht oder kaum (Schmerz, Übelkeit, Stimmung) besser ist als der natürliche Verlauf. Wenig Psychosomatik, viel statistisches Artefakt. Etwas allgemeiner ausgedrückt kann ein Vergleich innerhalb einer Gruppe zeigen, ob es einem Patienten besser oder schlechter geht, aber nicht ob und in welchem Ausmaß das auf die Behandlung zurückzuführen ist.

Es kommt aber noch dicker. Die Regression zum Mittelwert versteckt sich in fast allen klinischen Studien, und führt dort zur Überschätzung des Behandlungseffektes, egal ob Verum oder Placebo. Nehmen wir als Beispiel eine Studie, die ein Blutdruck-senkendes Medikament testet. In die Studie wird man aufgenommen, wenn man einen Blutdruck hat, der einen gewissen Wert überschreitet. Rein auf Grund der statistischen Fluktuation werden beim Blutdruckmessen in einer Gruppe von Menschen immer welche dabei sein, die bei der Messung einen erhöhten Blutdruck haben, aber keine Hypertoniker sind. Schon bei der nächsten Messung wäre der Wert wieder normal: Regression zum Mittelwert! Diese Menschen würden aber als Studienteilnehmer aufgenommen, ihr Blutdruckwert vor Behandlung in die Bestimmung des Mittelwerts der Gesamtgruppe eingehen. Nun wird behandelt, die Messung wird wiederholt, und der Mittelwert in der Gesamtgruppe ist jetzt niedriger als vor Gabe des Medikaments. Der Effekt des Medikaments wird aber überschätzt werden, da ja auch die ‚Patienten' wieder mitgemessen werden, welche gar keinen Hypertonus haben, und deren Mittelwert jetzt regrediert ist (ausführliches Beispiel mit Zahlen bei Senn 2011). Wäre alles kein Problem, wenn man jetzt eine echte Kontrollgruppe hätte, also Unbehandelte! Denn auch dort würde man den erniedrigten Blutdruck finden, aber vielleicht nicht so stark wie in der Verum-Gruppe. Leider ist dieser Vergleich aber bei den wenigsten Studien möglich, denn es fehlt die unbehandelte Gruppe. In den Studien, in denen eine unbehandelte Gruppe mitgeführt wurde, fand man keinen Placeboeffekt, oder nur in geringer Ausprägung gering bei subjektiven Symptomen wie Schmerz oder Befinden! In unserem Beispiel: Nicht im Blutdruck.

Vielleicht arbeiten Sie selbst ja mit Zellkulturen, oder mit Ratten, und werden deshalb sagen: interessant, aber glücklicherweise führe ich ja immer eine Gruppe ohne Behandlung mit. Geht mich also gar nichts an. Ich halte dagegen: Vorsicht! Denn die Regression

zum Mittelwert gilt natürlich nicht nur für individuelle Werte, sondern auch für die Ergebnisse von ganzen Studien. Insbesondere wenn sie auf kleinen Fallzahlen beruhen und daher eine hohe Varianz haben, sowie niedrige statistische Power und wenig stringente Signifikanzniveaus von 5 %. Also die meisten Studien.

Stellen Sie sich vor, sie machen ein Experiment. Wie immer mit n=8. Sie finden einen Effekt, und der war statistisch signifikant, sagen wir $p<0.03$. Sie sind glücklich. Machen noch ein paar andere Experimente für die Studie, und dann schreiben Sie das Paper, das den Effekt beschreibt. Wir gratulieren! Was aber wäre, wenn der eben signifikante Effekt ein falsch positiver gewesen wäre? Und eine Wiederholung des Experiments den Mittelwert in Richtung eines Null-Effektes korrigiert hätte? Also zum Mittelwert regrediert wäre? Durch unsere Fetischisierung von positiven, und insbesondere spektakulären (d.h. a priori unwahrscheinlichen) Befunden ist die Wahrscheinlichkeit hoch, dass wir häufig falsch positiven Befunden aufsitzen und diese in die Welt hinaustragen. Das Problem wäre leicht lösbar, aber die Lösung leider wenig populär: Größere Fallzahlen, ausreichende Power, stringente Signifikanzniveaus, Replikationen, und Publikation auch der negativen und neutralen Resultate. Good bye Nature Paper! (Siehe hierzu auch LJ 4/2017)

Von Mäusen, Makaken und Menschen

LJ 3/2018

Tuberkulose tötet weit über eine Million Menschen pro Jahr weltweit, problematisch ist vor allem die Situation im südlichen Afrika sowie in Osteuropa und Zentralasien. Ein sicher wirksamer Impfschutz gegen Tuberkulose (TB) fehlt, allerdings wird in Ländern mit hoher Inzidenz eine Lebendimpfung mit dem abgeschwächten Mykobakterien-Impfstamm Bacillus Calmette-Guérin (BCG) durchgeführt. BCG schützt aber kaum gegen Lungen TB, in jedem Fall ist der Impfschutz hochgradig variabel und unvorhersehbar. Weltweit sucht man daher seit Jahren nach einer verbesserten TB Impfung.

Der Narr interessiert sich für TB? Erst seit das British Medical Journal vor ein paar Wochen eine Untersuchung veröffentlicht hat (BMJ 2018;360:j5845), in der schwerwiegende Vorwürfe gegen Forscher und deren Universität erhoben werden: Interessenkonflikte, Tierexperimente von fraglicher Qualität, selektive Verwendung von Daten, Täuschung der Fördergeber und Ethikkommissionen, bis hin zur Gefährdung von Studienteilnehmern. Es gab auch einen Whistleblower, er musste seine Koffer packen. Das Ganze spielt in Oxford, an einem der angesehensten virologischen Institute der Erde, und die Studie am Menschen wurde an Säuglingen der ärmsten Bevölkerungsschichten Südafrikas durchgeführt. Eine explosive Mischung, die

ich hier näher beleuchten möchte, da wir daraus viel lernen können. Über die ethische Dimension präklinischer Forschung und die verheerenden Folgen, welche Qualitätsmängel bei Tierexperimenten und deren selektiver Veröffentlichung haben können, über die wichtige Rolle von systematischen Reviews von präklinischer Forschung, und letztlich auch über das Versagen von Kommissionen und Behörden, informierte Entscheidungen zu klinischen Studien zu treffen.

Am Anfang stand, wie es sich für eine spektakuläre Story gehört, eine Toppublikation: Eine Phase I Studie, publiziert in Nature Medicine. Die Autoren vom Jenner Institute in Oxford berichteten darin, dass man die unbefriedigende Wirkung der herkömmlichen BCG Vakzine auf das Immunsystem deutlich steigern kann. Und zwar durch eine gleichzeitige Impfung (‚Booster Impfung') mit einem anderen Antigen (Ag85A) des Tuberkulosebakteriums, exprimiert von einem modifizierten Vacciniavirus (dann MVA85a genannt). Dies war ein Durchbruch, und eine effizientere Tuberkuloseimpfung schien in greifbarer Nähe. Es wurden darauf Tierexperimente in verschiedensten Spezies durchgeführt, von der Maus über das Rind bis zum Primaten (Makake). Die Ergebnisse, soweit veröffentlicht, nährten die Hoffnung auf eine neue Ära der TB-Prophylaxe weiter. Die Universität Oxford schloss einen Vertrag mit einer Biotech-Firma zur weiteren Entwicklung und Vermarktung. Die Universität und Mitglieder des Forscherteams wurden Shareholder. Fördergeber von Wellcome Trust bis Paul and Melinda Gates Foundation zeigten sich spendabel, es flossen insgesamt Fördermittel von über 40 Millionen Pfund. Nun war noch die Sicherheit und Effizienz im Menschen zu zeigen, und man wählte logischerweise eine Weltregion mit hoher TB Inzidenz. Diese fand man in Südafrika, und führte die Studie an 2900 Säuglingen dort durch, wo 2- 3 % der Kinder eine klinisch manifeste Tuberkulose entwickeln. Die Studie wurde mit allen Genehmigungen und nach allen Regeln der Kunst durchgeführt: Genehmigung durch alle Behörden inklusive Ethik, randomisiert, kontrolliert, verblindet. Aber: sie verlief negativ! MVA85a reduzierte nicht die TB Rate der geimpften Kinder.

Nun sind negative (eigentlich besser: ‚neutrale') klinische Studien leider keine Seltenheit. Aber hier war die Enttäuschung besonders groß, und hatte dramatische Folgen. War doch die Ausgangslage aus den Tierversuchen (vier Spezies, inklusive Primaten!) scheinbar so vielversprechend wie selten zuvor. Die Paul and Melinda Gates Foundation, der weltweit größte Förderer im Bereich Infektionskrankheiten in den Entwicklungsländern, fasste daraufhin den Entschluss, sich aus dieser Form der translationalen Forschung ganz zurückzuziehen! Denn Tierexperimente sind ja offensichtlich nicht prädiktiv für den Menschen!

Aber stimmt das wirklich? Ein methodisch hochwertiger systematischer Review der tierexperimentellen Evidenz, auf der die Studie in Südafrika basierte, kam zu einem verheerenden Urteil. Die Qualität der diversen Studien war niedrig (keine Randomisierung, Verblindung, etc.), die Fallzahlen zu niedrig. Die Meta-Analyse fand keinen Hinweis auf eine Wirksamkeit von MVA85a im Tier. Schlimmer, bei den Primaten sah es sogar so aus als ob könnte die Booster Impfung schädlich sein. Also keine Rede von mangelnder Übertragbarkeit von Tier auf Mensch: Keine Wirkung beim Tier, keine beim Menschen! Die Autoren der Meta-Analyse stellten deshalb die naheliegende Frage, wieso es überhaupt zu der Studie an den Säuglingen kommen konnte, und diese einem Risiko bei unklarem Nutzen ausgesetzt wurden.

Was wir nun allerdings erst durch die Recherche des BMJ wissen, ist dass bereits von Anfang an Zweifel an den Tierstudien bestanden. Ganz offensichtlich wurden negative Befunde, welche eine höhere Sterblichkeit der Affen mit Booster Impfung gezeigt hatten, unterdrückt. Einem aufmerksamen Virologen, der in räumlicher Nähe und auf

ähnlichem Gebiet forschte, war dies aufgefallen. Er hatte dies der Universität gemeldet, es gab mehrere Untersuchungskommissionen, die keine Probleme finden konnten. Wer allerdings Probleme bekam, war dann allerdings der Whistleblower. Ihm wurde von der Universität mitgeteilt, dass er in den Räumlichkeiten des Instituts in Zukunft keine Forschung mehr durchführen dürfe. Die Recherche des BMJ zeigt diese leider nicht ganz untypischen Vorgänge minutiös auf. Und belegt auch, dass sowohl der Ethikkommission als auch den Genehmigungsbehörden in Südafrika selektiv nur die positiven Studienergebnisse aus den Tierexperimenten vorgelegt wurden.

Hieraus ergeben sich einige drängende Fragen. Passiert so etwas häufiger? Wie hoch ist die Qualität der präklinischen Forschung in anderen Feldern? Wie wird die Qualität gesichert? Wie häufig werden negative oder neutrale Daten nicht publiziert, oder gar verschwiegen? Die vorliegende Literatur, welche sich in letzter Zeit diesen Fragen angenommen hat, weist auf große Probleme hin. Es werden ja fast nur positive präklinische Studienergebnisse veröffentlicht, die mittleren Gruppengrößen sind in aller Regel unter 10, und Maßnahmen zur Verhinderung von Bias (z.B. Verblindung, Randomisierung) werden ebenfalls häufig keine angegeben. Wie gut ist also die präklinische Evidenz, bevor Studien am Menschendurchgeführt werden? Und wird auf Ebene der Genehmigungsverfahren (Ethik, FDA/EMA) sichergestellt, dass alle verfügbare Evidenz in den Entscheidungsprozess einfließt?

Auf letzteres gibt es bereits eine belastbare Antwort, das Manuskript hierzu ist im Review. Die Gruppe von Daniel Strech von der Medizinischen Hochschule Hannover hat eine große Zahl von Ethik-Anträgen zu klinischen Phase I oder II Studien an drei deutschen Universitäten systematisch durchforstet. Das Team suchte danach, ob die Anträge Informationen zur präklinischen Evidenz der beantragten Studien am Menschen beinhalten. Das Ergebnis war ernüchternd. Die überwiegende Anzahl dieser Anträge zitieren überhaupt keine publizierten Studien zur präklinischen Wirksamkeit des Studienmedikaments. Dort wo sich auf präklinische Daten bezogen wird, fehlt fast immer der Hinweis auf Maßnahmen zur Verhinderung von Bias sowie Fallzahlabschätzungen. Außerdem werden praktisch exklusiv positive Resultate angeführt, auch wenn es in der Literatur neutrale oder negative gibt. Wie können solche Gremien dann eine informierte Nutzen/Risiko Abschätzung durchführen?

Der im British Medical Journal aufgedeckte Fall aus Oxford ist hoffentlich extrem. Allerdings müssen wir befürchten, dass auch anderswo klinische Studien auf wackeliger präklinischer Evidenz durchgeführt werden. Ich vermute, dass ein nicht unwesentlicher Grund für die Schwierigkeiten in der Übertragung von tierexperimentellen Ergebnissen auf den Menschen darin liegt, dass die präklinische Evidenz selektiv berichtet wird und qualitativ auf wackeligen Füssen steht. Ethikkommissionen und regulatorische Behörden sollten sicherstellen, dass ihnen die Totalität der Evidenz bei der Entscheidungsfindung vorliegt, und zwar in hoher Qualität und in einer beurteilbaren Form. Manch eine klinische Studie, die enttäuschend verlief, wäre dann vermutlich gar nicht durchgeführt worden, und Studienteilnehmer nicht unnötigen Risiken ausgesetzt worden.

Wenn Du auf eine Weggabelung triffst - nimm sie!

4/2018

Zu Recht werden Forscher beneidet. Wenn sie nicht durch so lästige Dinge wie Antragschreiben, Vorlesungen, oder Formularkram aufgehalten werden, werden sie dafür bezahlt ihren tollsten Ideen nachzuspüren! To boldly go where no man has gone before! Man stöbert durch die wissenschaftliche Literatur, macht Pilotexperimente, die erstaunlicherweise ja fast immer erfolgreich sind. Dann führt dann eine Serie von wohlgeplanten und aufwendigen Experimenten durch. Diese klappen manchmal, öfters auch nicht, aber führen immer weiter ins Unbekannte. Auf diesem Weg wird aus einer Idee eine Hypothese, auf eine Hypothese folgen weitere. Die Hypothesen bestätigen sich! Am Ende, manchmal erst nach mehreren Jahren und unter erheblichem Verschleiß von Personal und Material, gelingt es, all dies zu einer ‚Story' zu verbinden (siehe dazu auch LJ 10/2017). Basierend auf einer komplexen Kette von Resultaten schließt die Geschichte mit einem ‚happy end'. In Form eines neuen biologischen Mechanismus, oder zumindest eines Puzzlesteinches dazu Und immer in die Welt gebracht mittels einer Publikation. Manchmal sogar in einer der Top-Zeitschriften.

In seiner Kurzgeschichte ‚Im Garten der Pfade, die sich verzweigen' (1944) beschreibt Jorge Luis Borges (1899 – 1986) das mysteriöse Werk des fiktiven chinesischen Schriftstellers Ts'ui Pen. Wenn im Handlungsstrang von Ts'ui Pen's Erzählung mehrere Verläufe möglich sind, geschehen diese nicht alternativ, sondern gleichzeitig! Hierdurch verästelt sich die Geschichte in ein Universum von vielfachen, möglichen Handlungen, die sich selbst wieder verzweigen, aber auch wieder zusammenführen können. Borges' Metapher vom Garten der sich verzweigenden Pfade, einem unendlichen Labyrinth, hat eine Vielzahl von Künstlern, insbesondere im Bereich der Hyperfiction inspiriert. Und die Statistiker Gelman und Loken haben sie kürzlich in die Methodenkritik psychologischer und biomedizinischer Forschung eingeführt. Sie vergleichen das Vorgehen von Wissenschaftlern mit Ts'ui Pen's Garten: diese bewegen sich in ihrer Forschung auf verzweigenden Pfaden durch einen Garten der Erkenntnis. Und so poetisch diese Wanderung auch anmutet, so Gelman und Loken, birgt sie gewisse Gefahren. Und diesen will ich mich heute zuwenden. Denn diese Gefahren sind den wenigsten Experimentatoren bewusst.

Folgen wir doch einmal einem fiktiven Wissenschaftler in den Garten seiner Forschung. Dort existiert ein veritables Labyrinth von Pfaden. Abhängig von seinen Ergebnissen, den daraus sich ergebenden Analysen, sowie der verfügbaren Evidenz anderer Forscher sucht er (oder natürlich auch sie!) sich einen Weg. Er betritt das Labyrinth mit einer Idee. Er wird sagen: Mit einer Hypothese. Sogleich führt er ein erstes Experiment zu deren Überprüfung durch. Und freut sich über das statistisch signifikante Ergebnis, eine Bande an der richtigen Stelle im Western Blot! Er biegt deshalb links ab. Bei einem weiteren, darauffolgenden Experiment ist ihm der p-Wert nicht mehr hold, er nimmt

deshalb den Pfad nach rechts. Während der Wanderung liest er ein aktuelles Paper, das ihn in seinen bisherigen Überlegungen bestätigt, und auf eine neue Idee für das nächste Experiment bringt: Schon biegt er in einen Pfad nach links ein. Dort findet das folgende Experiment wieder einen statistisch signifikanten Unterschied, von hier geht es weiter geradeaus. Der darauf verfolgte, naheliegende Ansatz bringt leider kein verwertbares Ergebnis. Unser Forscher läuft erstmal wieder zurück zur letzten Gabelung. Hier hellt sich seine Stimmung auf: das Resultat aus der knock-out Maus kann im pharmakologischen Ansatz repliziert werden! Zwei Wege führen also wieder zusammen, der Pfad wird breiter, ein Ausgang aus dem Labyrinth zeichnet sich bereits in der Ferne ab. Und auch das nächste Experiment gelingt, ein im Signalweg vermutetes Protein kann mittels Immunhistochemie nachgewiesen werden. Und: Dessen Blockade führt zu einem statistisch signifikanten Unterschied zur Kontrollgruppe! Das Literaturstudium ergibt nun, dass der Signalweg schon in einem anderen Krankheitsmodell beschrieben wurde, auch dies eine gute Nachricht. Er biegt darauf links ab und es ist geschafft: Er kann das Labyrinth verlassen. Nach vielen kompetent ausgeführten Experimenten, einer Vielzahl von statistisch signifikanten Vergleichen und ganz ohne p-Hacking (multiple statistische Tests bis einer davon signifikant wird) oder HARKING (hypothesizing after the results are known) wartet der Preis auf ihn: ein Artikel in einer angesehenen Zeitschrift.

Gute Forschung führt uns also durch das Labyrinth komplexer Biologie! Will der Narr nun wieder den Spielverderber geben? Nun, zumindest möchte ich auf ein vertracktes Problem hinweisen. Auf seinem Weg durch das Labyrinth geht der Forscher induktiv deterministisch vor. Er bemerkt gar nicht die vielen Freiheitsgrade die ihm zur Verfügung stehen. Diese ergeben sich zum Beispiel durch alternative Analysen (oder Interpretationen) der Experimente. Oder durch sich zufällig einstellende falsch positive oder falsch negative Ergebnisse. Oder auch durch die Auswahl eines anderen Artikels als Basis weiterer Experimente und Interpretationen. Das Labyrinth ist nämlich unendlich groß! Es gibt nicht nur einen Weg hindurch, sondern viele, und auch sehr viele Ausgänge. Und da unser Forscher explorativ vorgeht, hat er auch vorab keine Regeln aufgestellt, nach denen er seine Analysen durchführt, oder weitere Experimente plant. Er merkt also nichts von den vielen möglichen anderen Ergebnissen, denn er folgt ja einer Spur, die er selber legt. Nur überschätzt er dadurch die Stärke der Evidenz, die er generiert! Insbesondere überschätzt er, was ein signifikanter p-Wert auf seiner explorativen Wanderung bedeutet. Er müsste nämlich seine Resultate eigentlich mit allen anderen möglichen Analysen und Interpretationen vergleichen, welche er alternativ hätte durchführen können. Ein absurder Vorschlag, das geht natürlich nicht. Frei nach dem amerikanischen Baseball-Philosophen der New York Yankees, Yogi Berra (1925-2015), müsste der Forscher, wenn er an die Gabelung kommt, diese nehmen! In Borges‘ Garten der sich verzweigenden Pfade hieße dies, immer gleichzeitig nach links und nach rechts abzubiegen!

Deshalb gilt im Garten der sich verzweigenden Pfade nicht mehr die klassische Definition der statistischen Signifikanz (z.B. $p<0.05$). Diese lautet da: Die Wahrscheinlichkeit, rein zufällig und in Abwesenheit eines Effekts ein ähnlich extremes oder noch extremeres Ergebnis zu beobachten, ist kleiner als 5%! Man müsste nämlich über alle Daten und Analysen mitteln, welche im Garten der sich verzweigenden Pfade möglich gewesen wären. Jeder dieser anderen Wege hätte ja auch zu statistisch signifikanten Ergebnissen führen können. So ein Vergleich ist aber bei explorativer Forschung unmöglich. Wenn man trotzdem p-Werte generiert, erhält man nach Gelman und Lokens eine ‚Maschine zur Produktion und Veröffentlichung von Zufallsmustern‘. Und dies wohlgemerkt obwohl die publizierten Analysen der Forscher absolut kongruent sind mit den Hypothesen, welche deren Experimente motiviert hatten.

Was folgt aus diesen nur scheinbar esoterischen Überlegungen? Keinesfalls sprechen sie gegen Exploration, das lustvolle Wandern durch den Garten der sich verzweigenden Pfade! Allerdings folgt daraus, dass die auf dieser Wanderung gepflückten Früchte unserer Erkenntnis weniger robust sind, als uns die Kette von statistisch signifikanten Ergebnissen glauben macht. In Konsequenz bedeutet dies auch, dass die Verwendung von Teststatistiken bei der Exploration wenig hilfreich ist. Und daher eigentlich überflüssig, wenn nicht sogar irreführend. Auf eine Reihe von weiteren gewichtigen Argumenten für mehr Skepsis in unsere eigenen Ergebnisse und die Irrungen und Wirrungen der Verwendung von statistischen Tests hat der Narr bereits früher an dieser Stelle hingewiesen (LJ 4/2017). Und noch was. Ein guter Führer durch das Labyrinth ist die Konfirmation, also eine geplantes, in Vorgehen und Analyse vorbestimmtes Experiment mit ausreichender Fallzahl.

Kann denn (Nicht-)Replikation Schande sein?

LJ 5/2018

Die Ergebnisse Deiner Arbeit ließen sich nicht reproduzieren! So eine Schreckensmeldung fürchtet in letzter Zeit so mancher. Reproduzierbarkeit, Replizierbarkeit, Reliabilität und Robustheit der Forschung werden von den wissenschaftlichen Akademien, den Journalen, und mittlerweile auch von den Fördergebern allenthalben angemahnt. Es ist eine Bewegung für „reproduzierbare Wissenschaft“ entstanden. Förderprogramme für die Reproduktion von Forschungsarbeiten sind derzeit in Vorbereitung. In einigen Wissenschaftszweigen, allen voran der Psychologie, aber auch in Feldern wie der Krebsforschung werden Forschungsarbeiten nun auch systematisch repliziert. Oder eben nicht – deshalb erleben wir eine „Reproduzierbarkeits-Krise“.

Mit Daniel Fanelli hat nun kürzlich ein Wissenschaftler, den man bisher auf der Seite der Befürworter von solchen Aktivitäten vermutete, seine Stimme mahnend erhoben. In den ehrwürdigen Proceedings of the National Academy of Sciences frägt er rhetorisch: “Is science really facing a reproducibility crisis, and do we need it to?” Ich möchte mich daher heute, vielleicht am Vorabend einer aufkeimenden Gegenbewegung, mit einigen Einwänden gegen das derzeitige Mantra von der ‚Reproduzierbaren Wissenschaft‘ auseinandersetzen.

Ist Reproduzierbarkeit von Ergebnissen wirklich das Fundament der wissenschaftlichen Methode? Oder hat nicht, wie Chris Drummond anmerkt, schon Thomas Kuhn in seinem berühmten Werk ‚Die Struktur wissenschaftlicher Revolutionen‘ festgestellt, dass der wissenschaftliche Fortschritt ganz wesentlich nicht in der inkrementell fortschreitenden ‚normalen Wissenschaft‘ stattfindet, sondern durch periodisch wiederkehrende „Paradigmenwechsel“? Und der Paradigmenwechsel ist doch alles andere als die Reproduktion von bisher Dagewesenem! Ein verwandtes Argument ist das von der ‚Trivialität‘ reproduzierter wissenschaftlicher Ergebnisse. Danach sind gerade die Befunde, welche

inkrementell auf bereits sattsam Bekannten beruhen die garantiert reproduzierbarsten. Und umgekehrt, bedeutet erfolgreiche Reproduktion, dass es sich um ‚richtige' Resultate handelt'? Was, wenn Originalresultat und Reproduktion demselben systematischen Fehler aufsitzen, oder beide falsch positive Befunde sind, ganz einfach zufällig?

Und es wird noch philosophischer. Manch Kritiker der Betonung von Reproduzierbarkeit als Ziel von Wissenschaft bezieht sich gar auf Karl Popper: Nach ihm lassen sich Hypothesen nicht beweisen, sondern nur Falsifizieren. Am Beispiel des berühmten schwarzen Schwans, der die Hypothese „alle Schwäne sind weiß" widerlegt: Eine Studie, die eine vorausgegangene Untersuchung, welche an einem See nur weiße Schwäne vorfand, dahingehend reproduziert, dass sie an einem anderen See auch nur welche mit weißen Federn fand, hätte diese zwar erfolgreich repliziert. Die Hypothese wäre aber trotzdem falsch. Was sich insbesondere dann zeigen würde, wenn der schwarze Schwan vorbeifliegt. Dies ist es, was Jason Mitchel die ‚Leere der misslungenen Replikation' bezeichnet. Das tolle an Wissenschaft ist doch die Entdeckung von Neuem, nicht die langweilige Wiederholung. Reproduzieren ist also keine Wissenschaft, lautet hier das Verdikt!

Ohne theoretische Umschweife zur Sache gehen dagegen jene Kritiker, welche Replikations-Experimente für grundsätzlich problematisch halten: Weil sie Zweifel an der Kompetenz der Replizierer hegen. Man verweist dann meistens auf die Heerscharen von Doktoranden und Postdocs, die aufgerieben wurden, um eine bestimmte Technik im eigenen Labor zu etablieren. Dort würde natürlich alles replizierbar sein, von den besagten Experten. Aber das Vorhandensein von implizitem Wissen, das nicht im Methodenteil von Artikeln wiedergegeben werden kann, verhindere die Wiederholbarkeit. Die nicht-Wiederholbarkeit der Ergebnisse durch Andere beweist demnach nur eins, nämlich deren Unfähigkeit!

Und noch etwas sehr Ernstzunehmendes führen die Kritiker ins Feld: Durch die moralische Überhöhung der Replikation als Goldstandard werden Wissenschaftler, deren Ergebnisse nicht wiederholt werden können, stigmatisiert. Ganz unabhängig von den Details und Umständen der Replikation, das Resultat der Replikation gilt als das Richtige. Es steht dann auch gleich der Verdacht im Raum, dass hier jemand nicht sauber gearbeitet, ja vielleicht sogar gegen die Regeln der guten wissenschaftlichen Praxis verstoßen hat! Gute Wissenschaft MUSS replizierbar sein!

Haben die Kritiker also recht – ist es ein Fehler, Reproduzierbarkeit von Forschung aufs Schild zu heben, zu belohnen, gar Fördermittel dafür auszureichen? Ganz sicher nicht. Trotzdem empfiehlt der Narr, die Argumente ernst zu nehmen, und sich mit dem nicht ganz trivialen Thema auseinanderzusetzen.

Zunächst einmal geht es unter dem Stichwort ‚Reproduzierbarkeit' begrifflich häufig drunter und drüber. Reproduzierbarkeit der Methoden, der Resultate, der aus den Ergebnissen abgeleiteten Schlüsse (inferentielle Reproduzierbarkeit), strikte Replikation, usw., das muss man sehr wohl auseinanderhalten. Meinen wir eine Wiederholung der Effektgröße, des p-Wertes, von statistischer Signifikanz überhaupt? Und natürlich ist Reproduzierbarkeit Kontext-abhängig. Da steckt das „implizite Wissen" drin, aber noch viel wichtiger, die Robustheit der Ergebnisse, also ihre externe Validität. Hanno Würbel hat auf das Paradox des Standardisierungs-Irrtums hingewiesen: Der Wunsch nach mehr Reproduzierbarkeit führt häufig zum Ruf nach mehr Standardisierung. Dies ist aber, und darin steckt das Paradox, ein Holzweg, denn mit höherer Standardisierung werden Ergebnisse schlechter reproduzierbar! Schon der Urvater der von uns so verehrten frequentistischen Wahrscheinlichkeitstheorie Ronald Fisher, hat es 1935 so formuliert: *„Ein hoch standardisiertes Experiment liefert nur direkte Informationen in Bezug*

auf den engen Bereich der Bedingungen welche durch die Standardisierung erreicht wurden. Standardisierung stärkt daher nicht, sondern schwächt sogar unsere Schlussfolgerungen aus den Ergebnissen verglichen zur Praxis der Variation der Bedingungen". Gerade in der biomedizinischen Wissenschaft ist diese durch Verzicht auf Standardisierung, ja sogar bewusste oder unbewusste Variation verbesserte externe Validität aber sehr wichtig: Wenn sich ein Ergebnis aus einer Maus in Boston nicht in der genetisch identischen Maus in Berlin wiederholen lässt, spricht das erstmal nicht gegen die Richtigkeit und Qualität der Befunde aus Boston. Lässt aber sehr wohl Zweifel an deren Übertragbarkeit auf den Menschen aufkommen.

Das Replizieren von eigenen Befunden und solchen anderer ist Wissenschaft. Zum einen, und hier liegt das Missverständnis bei der Interpretation von Thomas Kuhn, beruhen sowohl die ‚normale Wissenschaft' (also das was die meisten von uns machen), als auch die Forschung, welche zu Paradigmenwechseln führt (also das was der Zufall und geniale Wissenschaftler so bewerkstelligen), entscheidend auf Ergebnissen die wiederholbar sein müssen. Dabei führt sowohl die Reproduktion, als auch eine mögliche nicht-Reproduktion zu wissenschaftlich relevanten Ergebnissen. Eine kompetente Reproduktion kann eine Hypothese stärken, insbesondere wenn sie auch unter Variation von methodischen Details erfolgreich war. Wird das Design der Reproduktion so verändert, dass bewusst alternative methodische Ansätze gewählt werden (z.B. statt einer Knockout Maus die Manipulation des interessierenden Gens mittels einer RNA Interferenz), spricht man von Triangulation und erhält potentiell noch robustere Resultate. Andererseits kann eine nicht-Reproduktion über die Einsicht in modifizierende Faktoren zu neuen Erkenntnissen führen.

In keinem Fall darf nicht-Reproduktion zur Stigmatisierung führen. Unzählige Faktoren können diese verursachen, die Wesentlichen habe ich oben den Replikations-Kritikern in den Mund gelegt. Und hier gleich noch eine Warnung an diejenigen, welche die Debatte irrelevant finden, da sie ‚ja schon immer ihre eigenen Ergebnisse repliziert haben'. Ein Effekt, der auf einem Niveau von gerade eben 0.05 signifikant war, und ein tatsächlich „wahres" Ergebnis darstellt, lässt sich bei strikter Replikation (gleiches Experiment, gleiche Fallzahl, etc.) nur mit 50% Wahrscheinlichkeit als signifikant wiederholen. Ein Würfelspiel also! (wer das Nachlesen will, siehe LJ 4/2017).

Der Narr meint daher: Replizierbarkeit ist zwar nicht der Zweck von Wissenschaft. Da geht es um neues Wissen. Aber neues Wissen muss reproduzierbar sein. Karl Popper hierzu: „*Alle Ereignisse, die nicht reproduzierbar sind, sind aus der Wissenschaft ausgeschlossen.*" Die Untersuchung von aufregenden Hypothesen an der Vorderfront der Wissenschaft erzeugt notwendig eine Menge falsch positiver Befunde, auch bei Forschung von höchster Qualität. Diese falsch Positiven müssen aber durch nachfolgende, kompetente Experimente wieder „ausgemerzt" werden. Reproduktion ist daher eine vornehme, hoch wissenschaftliche Tätigkeit. Die dichotome Frage: Ist ein Ergebnis reproduziert worden, ist unangebracht, die Reproduzierbarkeit folgt nicht einem einfachen ja/nein Schema. Es ehrt den Wissenschaftler, wenn andere sich an der Reproduktion seiner Ergebnisse versuchen, denn dies bedeutet, dass sie wichtig sind. Wenn sie nicht-Reproduziert werden, fängt die Wissenschaft erst so richtig an, denn dann stellen sich viele Fragen: Stimmte die Richtung, nur der p-Wert nicht? Was passiert, wenn man Originalexperiment und Replikation in einer Meta-Analyse kombiniert? Steckt dahinter gar interessante Biologie? Oder ein bisher unerkannter Fehler? usw.

Belohnt gehören also diejenigen deren Ergebnisse eines Reproduktionsversuchs würdig sind, genauso wie die Wissenschaftler, die solche Experimente durchführen. Und das

geht nur, wenn die Methoden und Ergebnisse von Studien so umfassend beschrieben werden, dass man sie auch nachkochen kann!

Bildet euch fort, ihr Etablierten!

LJ 6/2018

Viel wird derzeit nachgedacht und geschrieben, auch vom Wissenschaftsnarr, wie man Wissenschaft effizienter, robuster, reproduzierbarer, ja insgesamt werthaltiger machen kann. Ganz oben auf der Liste stehen Maßnahmen wie die Verbesserung der internen Validität (z.B. durch Randomisierung und Verblindung, Ein- und Ausschlusskriterien usw.), die Erhöhung der Fallzahlen und damit statistischen Power, die Beendigung der Fetischisierung des p-Werts, sowie die Verfügbarmachung der Originaldaten (Open Science). Fördergeber und Journale beginnen dies umzusetzen, und formulieren erstmals entsprechende Passagen in ihre Förderbedingungen oder die Anleitungen für die Antragsteller und Reviewer. Es bewegt sich also was!

Das merke ich auch bei den Studenten. Ich unterrichte unter anderem Statistik, gute wissenschaftliche Praxis, und experimentelles Design. Dabei beeindruckt mich jedes Mal der Enthusiasmus der Promotionsstudenten und jungen Postdocs, sich ins Abenteuer ihrer wissenschaftlichen Projekte zu stürzen. Und dabei ihr unbedingter Wille, dabei ‚alles richtig zu machen'. Vorschläge zur Verbesserung der Reproduzierbarkeit und Robustheit ihrer Forschungsprojekte saugen sie auf wie ein trockener Schwamm das Wasser. Oft endet die Diskussion allerdings unbefriedigend. Insbesondere wenn wir eigene Experimente und Forschungsansätze der Studenten besprechen. Ich werde dann häufig darauf hingewiesen, dass das ja alles schön und gut sei, aber in der Umsetzung am Arbeitsgruppenleiter scheitern werde. Dabei äußert sich der Widerstand ihrer Betreuer typischerweise so: „Wir haben das schon immer so gemacht, und sind damit in Nature und Science gekommen", „Wenn wir das so machen würden, kriegen wir das nie durch den Review Prozess", „Das dauert dann ja viel länger, und wir könnten gescoopt werden", „Das könnten wir doch nur in PLOS One (oder Peer J, F1000 Research etc.) publizieren, das Paper kontaminiert dann deinen CV", usw. Oft wünsche ich mir deshalb, dass nicht nur die Studenten im Seminarraum sitzen würden, sondern auch deren Betreuer!

Mir fällt dabei auf, dass sich der Berufsstand der Wissenschaftler auf erstaunliche Weise abhebt von anderen Berufen, wie Ärzten, Rechtsanwälten, Krankenpflegern, ja sogar Bundesligaschiedsrichtern. Den Wissenschaftlern fehlt eine Körperschaft, ein ethischer Code, und eine Pflichtfortbildung! Jawohl, Immobilienmakler und Bundesligaschiedsrichter haben das alles! Wissenschaftler dagegen wird und bleibt man ganz einfach durch die Ausübung der Tätigkeit, der fast immer ein spezialisierter Abschluss vorausgeht (z.B. Diplom, oder Promotion). Piloten müssen dagegen jedes Jahr soundso viele Flugstunden nachweisen, und dazu noch einen Flug mit einem Instruktor absolvieren. Ärzte

müssen unabhängig davon, ob sie niedergelassen, ermächtigt oder angestellt sind innerhalb von fünf Jahren mindestens 250 Fortbildungspunkte bei ihrer Kassenärztlichen Vereinigung nachweisen. Die Idee hinter all dem: Sicherstellen, dass man wesensmäßig auf der Höhe der Zeit ist, und seinen Beruf gemäß den aktuellen Standards ausführt. Warum brauchen Wissenschaftler das eigentlich nicht? Sind sie gesellschaftlich nicht wichtig genug? Sodass es nichts ausmacht, wenn was anbrennt in der Forschung, weil sie wichtige Entwicklungen verschlafen haben? Wird man durch seine Tätigkeit, d.h. Forschung quasi automatisch fortgebildet? Forschung also als Fortbildung? Oder hat es etwas mit der Wissenschaftsfreiheit zu tun – der Angst vor jeder Form der Regulation als Einschränkung der Kreativität im Elfenbeinturm?

Ruft der Narr nun nach einer weiteren Behörde, oder Standesorganisation? Fortbildungspunkte bei Kongressen? Eine jährliche Prüfung für promovierte und habilitierte Wissenschaftler, sonst droht Titelentzug oder Zwangs-Retraktion wissenschaftlicher Artikel? Natürlich nicht. Dennoch halte ich es für bedenklich, wenn Studenten ihre Betreuer fortbilden müssen – das kann nicht funktionieren. Auch dass Fördergeber anfangen, ihre Gutachter zu schulen, sollte einem zu denken geben. Die wichtigsten Fördergeber in England, wie z.B. Wellcome Trust, MRC, British Cancer Foundation tut dies bereits. Oder Journale ihre Reviewer fortbilden, wie das renommierte British Medical Journal es vormacht.

Wohl sind die meisten Wissenschaftler absolute Experten im unmittelbaren Gegenstand ihrer Forschung. Da lesen sie (oder schreiben gar) die aktuellsten Papers, hören auf den Kongressen die relevanten Vorträge, und diskutieren am Poster. Aber in Statistik, experimentellem Design, Reviewen etc., also zentralen Fähigkeiten ihrer Tätigkeit, wurden die wenigsten ausgebildet. Über die Jahre haben sie sich, im Wesentlichen durch einfache Assimilation gängiger Praxis im System als Wissenschaftler etabliert. Der Erfolg (Gruppenleiter, Abteilungsleiter, Professur) gibt ihnen dabei recht. Was der p-Wert wirklich bedeutet (oder auch nicht), dass die Wahrscheinlichkeit ihrer Hypothesen ganz entscheidenden Einfluss darauf hat, ob ihre Ergebnisse falsch positiv sind, was Pseudoreplikation und Nesting sind, was Regression zum Mittelwert, usw., das haben sie nie wirklich wissen müssen. Weil es ihre Kollegen, inklusive der Fördergeber und Journale, auch nicht wussten! Jetzt, wo dies ans Tageslicht kommt und als wichtige Ursache für mangelnde Replizierbarkeit, Robustheit, und Prädiktivität identifiziert wurde, ist der Nachwuchs häufig besser informiert als die Etablierten. Wenn dann noch ganz Neues dazu kommt, wie Open Data (mit so Dingen wie Common Data Elements und Metadata, Repositorien, etc.), oder elektronischen Laborbüchern, wird's ganz problematisch.

Also: Kein ‚Registered Scientist' (https://sciencecouncil.org/), sondern periodische Fortbildung zu aktuellen Themen von großer allgemeiner Relevanz. Gemeinsam mit den Studenten und Post-docs! Wie setzt man das durch: Mittels attraktiver Formate mit geringem zeitlichem Aufwand. Als Material die jeweils konkreten Forschungsprojekte. Dem könnte man noch ein bisschen nachhelfen, indem man die Teilnahme z.B. ein bisschen mit der Verteilung der Leistungsorientierten Mittelvergabe verknüpft. Aber keine Angst, aus dieser närrischen Idee wird nichts werden! Denn das Totschlägerargument ‚Wissenschaftsfreiheit' und die Angst vor Bürokratismus und gelangweiltem Absitzen von Pflichtstunden wird das verhindern. Schade eigentlich, es könnte einiges zur Professionalisierung der Wissenschaft, und zur Verbesserung der Robustheit ihrer Ergebnisse beitragen.

Im (Paper)Wald, da sind die Räuber

9/2018

Ende Juli war es wieder soweit. Ein Wissenschaftsskandal erschütterte die Republik. Recherchen von NDR, WDR und Süddeutscher Zeitung ergaben, dass deutsche Wissenschaftler in einen ‚weltweiten Skandal' verwickelt seien. Mehr als 5000 Wissenschaftlerinnen und Wissenschaftler deutscher Hochschulen, Institute und Bundesbehörden haben mit öffentlichen Geldern finanzierte Forschungsbeiträge in Online-Fachzeitschriften scheinwissenschaftlicher Verlage veröffentlicht, die grundlegende Regeln der wissenschaftlichen Qualitätssicherung nicht beachten. Die Öffentlichkeit, und nicht wenige Wissenschaftler, erfuhren so zum ersten Mal, dass es ‚Raubverlage' und ‚predatory journals' gibt. An der ganzen Sache ist einiges bemerkenswert, vieles davon stand allerdings nicht in den Zeitungen.

Raubverlage, die in ihren Phishing-Mails recht seriös auftreten, bieten Wissenschaftlern die Open Access (OA) Veröffentlichung ihrer wissenschaftlichen Studien gegen Bezahlung. Sie suggerieren dabei, dass ein ‚Peer Review' stattfindet. Der findet dann nicht statt, und die Artikel erscheinen auf den Webseiten dieser ‚Verlage', diese sind aber nicht in den gängigen Datenbanken wie PubMed gelistet. Jeder Wissenschaftler in Deutschland findet mehrere solcher Einladungen pro Tag im Email-Postfach. Sollten Sie keine bekommen und sind aber einer, sollten Sie sich jetzt Sorgen machen.

Nun könnte man meinen, mit Raubverlagen seien Elsevier und Konsorten gemeint. Diese realisieren Umsatzrenditen von über 30 %, in dem sie uns die Früchte unserer eigenen Arbeit verkaufen. Und wir noch darum gezittert haben, dass sie dies Geschenk auch annehmen, also den Artikel akzeptieren! Steuerzahler, die all dies finanziert haben und andere Unglückliche, welche nicht über einen teuren institutionellen Bibliothekszugang verfügen, kommen nur mit ihrer Kreditkarte an die Früchte unserer Erkenntnis (siehe LJ 5/2017). Aber halt, nicht so schnell! Elsevier ist ja kein Raubverlag? Denn dort gibt es (zumeist) einen ordentlichen Review Prozesses.

Und hier beginnt es interessant, ja kompliziert zu werden. Auch ich bin der Überzeugung, dass ein guter Review Prozess wissenschaftliche Studien verbessern kann. Häufig ist dies aber gar nicht der Fall. Er frisst massiv Ressourcen, doch es gibt keine wissenschaftliche Evidenz dafür, dass er ‚funktioniert'. Der Review Prozess ist langsam, teuer, erratisch, und schlecht im Detektieren von Fehlern. Er wird häufig missbraucht, seine Resultate sind potentiell anti-innovativ. Wir alle kennen das Problem. Artikel werden oft schon mit Blick auf potentielle Reviewer geschrieben, häufig wird in den Revisionen nichts anderes erreicht, als einen bestimmten Gutachter ruhig zu stellen. Statt sich auf die Suche nach neuer Erkenntnis zu machen, verbringen Arbeitsgruppen viel Zeit damit, genau die Ergebnisse zu liefern, welche die Gutachter noch gerne sehen würden. Sicherlich wird auch manche unrettbar schlechte Arbeit aus dem Verkehr gezogen. Aber nur vorübergehend, denn sie wird, nach einer Kaskade von Einreichungen in Journalen mit

abnehmenden Impact Factor irgendwo anders publiziert werden. Wenn es sein muss eben in einem räuberischen Journal! Damit der Review Prozess fair und produktiv wird, braucht es schon einigen Aufwand, und innovative Ansätze, wie z.B. bei den OA Journalen von EMBO oder F1000Research.

Wieso publizieren also gestandene Wissenschaftler in Journalen, deren Namen sie vorher noch nie gehört haben? Häufig, weil sie nach einer Reihe von absolut entnervenden erfolglosen Einreichungen die Nerven verlieren. Die Arbeit war vielleicht sogar richtig gut, aber zu wenig spektakulär, ein negativer Befund oder NULL-Resultat gar. Oder die Reviewer Forderungen waren nicht zu erfüllen, da zu aufwendig, oder der Doktorand schon über alle Berge. Die Autoren sind dann der Verlockung erlegen, gegen Zahlung einer Gebühr die Früchte Ihrer Arbeit in einem Journal mit toll klingendem Namen zu sehen. Ebenso wie eine weitere Arbeit auf ihrer Literaturliste. Und hier liegt der Kern des Problems: Ein an simplen quantitativen Indikatoren orientiertes Belohungs- und Karrieresystem. Mindestens 10 Originalarbeiten mit Peer Review als Erst- oder Letztautor werden z.B. von Habilitanden der Charité gefordert. Ähnliches gilt an den meisten deutschen Fakultäten.

Raubverlage nutzen also ein systemisches Problem unseres akademischen Systems aus. Die Opfer der Raubverlage sind gleichzeitig auch Täter! Wir beurteilen Wissenschaftler häufig nach quantitativen, leicht messbaren Größen (Zahl und Impact Factor der Publikationen, eingeworbene Drittmittel), die oft genug wenig mit der Qualität der Wissenschaft oder deren wissenschaftlicher oder gesellschaftlicher Relevanz zu tun haben. Der Blick auf den Inhalt und die Bedeutung der Forschung kommt aus Zeitgründen zu kurz. Außerdem werden Studien im Wesentlichen danach beurteilt, ob sie ein positives Resultat haben, aber nicht ob sie gut gemacht waren und ein verlässliches Ergebnis haben. Deshalb lesen wir auch jeden Tag in der Zeitung über die demnächst bevorstehende Heilung von Alzheimer, Krebs usw., ohne dass diese bisher eingetreten wären.

Apropos Heilungsversprechen: Es wird ja auch behauptet, dass von den Predatoren die Gefahr ausgeht, dass ‚fake science' publiziert und dann in klinische Anwendungen gebracht wird, welche Patienten gefährdet. Dies mag in Einzelfällen tatsächlich passiert sein. Allerdings wird auch in seriösen wissenschaftlichen Zeitschriften fake science publiziert. Und wenn das dann im Lancet geschieht, werden gleich Hunderttausende geschädigt: Andrew Wakefield und die Anti-Vaxxer lassen grüßen. Auch sollten wir nicht vergessen, dass etwa 50 % aller abgeschlossenen klinischen Studien gar nicht publiziert werden. Auch weil die Ergebnisse, wenn sie nicht eindeutig waren oder die Studienhypothese nicht belegen schwieriger zu publizieren sind. Wir also ‚Positives' und Spektakuläres fetischisieren. So gesehen könnte man den Predatoren ja sogar zu Gute halten, dass sie für die Öffentlichmachung sonst unzugänglicher Evidenz sorgen!

Klar ist dabei auch: Es wird viel zu viel (Positives) publiziert, egal ob bei Elsevier et al., oder bei Predatoren. Warum? Ganz einfach, weil wir Publikationen belohnen, aber nicht, ob sie eine wichtige Frage untersuchten und methodisch gut gemacht waren. Kommissionen können all unsere Publikationen gar nicht mehr überprüfen, geschweige denn lesen. Das gilt übrigens auch für die Wissenschaftler selbst. Allein PubMed listet für 2017 fast 1.3 Mio erschienene Artikel. Mehr als 90 % der publizierten Literatur (nicht in Raubverlagen, wohlgemerkt) werden dabei gar nicht gelesen. Trotzdem werden etwa 50 % davon mindestens einmal zitiert, häufig also ungelesen! Also zählen wir einfach Publikationen, egal was drinsteht. Oder wir addieren deren Impact Factor. Egal, dass der nichts über den jeweiligen Artikel aussagt, denn er misst nur die durchschnittliche Zitierhäufigkeit des Journals.

Tragisch auch, dass die Affäre um die Raubverlage sehr gute Open Access Zeitschriften und das Prinzip dahinter in Verruf bringt. Das ‚pay per article' Prinzip wird mit schlechter Qualität gleichgesetzt, obwohl zwischen beiden kein Zusammenhang besteht. Zur Stigmatisierung von OA führt auch, dass viele Verlage Autoren nach Ablehnung im ‚normalem' Journal anbieten, den Artikel an ein OA-Journal aus demselben Hause (mit meist niedrigerem Impact Factor) ‚weiterzureichen'. Noch komplizierter wird alles dadurch, dass es bei einigen Verlagen gar nicht so einfach zu sagen ist, ob sie Prädatoren sind. Manche der Zeitschriften, die heute als Prädatoren gelten, waren früher in allen relevanten Datenbanken gelistet und hatten ordentliche Impact Factoren (z.B. Oncotarget).

Es geht also gar nicht um Raubverlage, diese nutzen nur unsere Systemfehler. Was können, was müssen wir tun? Eine Stigmatisierung derer, welche ordentliche Studien in Raubverlagen veröffentlicht haben, bringt uns nicht weiter. Kurzfristig müssen wir aufklären, denn vielen Kollegen ist ja gar nicht klar, was ein ‚räuberisches Journal' ist und wie man es erkennt. Aber auch darüber, wie das institutionelle Subskriptionsmodell finanziert wird. Manch ein Kollege glaubt immer noch, dass das scheinbar ‚kostenlose' Herunterladen von beliebigen wissenschaftlichen Artikeln schon Open Access ist! Entscheidend wird aber sein, das akademische Anreiz- und Karrieresystem so zu verändern, dass neben quantitativen Faktoren wieder mehr qualitative, qualitätsorientierte Indikatoren Eingang finden. Neben der Frage, wie innovativ Forschung ist, müssen auch Kriterien wie z.B. wissenschaftliche Sorgfalt, Transparenz, zur Verfügungstellung der Daten für andere Wissenschaftler, oder Einbeziehung von Patienten in die Planung klinischer Studien (‚Partizipation') bewertet und ‚belohnt' werden. Dann werden wir auch herausfinden, ob innovative Formate wie Preprint Server (z.B. BioArXiv), Post-Publication-Review, und registered reports nicht viel effektiver qualitativ hochwertigere Veröffentlichungen ermöglichen.

Wenn Du ins Labor gehst, vergiss den Patienten nicht

LJ 10/2018

Der bis dahin unbekannte israelische Radiologe Dr. Yehonatan Turner schaffte es mit einem halbseitigen Artikel über ihn und seine Studie in die New York Times. Er hatte Radiologen computertomographische Aufnahmen zur Befundung vorgelegt – und dazu jeweils ein aktuelles Portraitfoto vom Patienten. Dabei nutzte er ein Cross-over Studiendesign: Einer Gruppe von Radiologen zeigte er zunächst das CT und das Portraitfoto. Drei Monate später präsentierte er ihnen dann dasselbe CT, aber ohne Foto. Andere Radiologen erhielten erst einmal nur das CT, und nach 3 Monaten CT mit Portrait. Eine weitere Kontrollgruppe befundete wie in der Radiologie üblich nur die CTs. Die Die Idee: Weil der befundende Arzt dem Patienten quasi in die Augen blickt, und nicht

nur auf eine anatomische Schichtaufnahme, würde er sich seiner Verantwortung bewusster, und damit seine Befundung gründlicher, und die Diagnose akkurater. Und so war es denn auch: Die Radiologen berichteten, durch das Foto stärkere Empathie für die Patienten zu haben, und sich mehr wie ‚Ärzte zu fühlen'. Und sie fanden statistisch signifikant mehr Auffälligkeiten und pathologische Befunde, wenn sie CT und Foto vor Augen hatten, verglichen mit CT alleine.

Wie wäre es nun, wenn man biomedizinischen Grundlagenforschern Fotos von Patienten mit der Erkrankung zeigen würde, welche diese im Labor beforschen? Ein Fotoaufsteller neben dem Computer, dem PCR Cycler, oder dem Mikroskop? Mit einer Schlaganfallpatientin für den Schlaganfallforscher, oder einem Diabetiker für die Diabetesforscherin. Ein bisschen wie die Schockbilder auf den Zigarettenschachteln, nur grösser. Es gibt ja kaum einen wissenschaftlichen Artikel, auch von den hartgesottensten Grundlagenforschern, der nicht mit Sätzen beginnt wie: „Krankheit X ist die häufigste Ursache für Y…", oder „Weltweit erkranken X Menschen an …", und die Diskussion endet „…unsere Ergebnisse könnten die Grundlage bilden für eine effektivere Behandlung von …" usw. Die Referenz auf die Wichtigkeit der eigenen Forschung für bestimmte Erkrankungen ist in der biomedizinischen Forschung allgegenwärtig. Glaubt man den Publikationen und Anträgen bei Forschungsförderern, ist dies eine Haupttriebfeder für die eigene Arbeit. Und dass trotzdem wenige Forscher direkten Kontakt zu tatsächlich Erkrankten haben, es sei denn zufällig im Bekannten- oder Verwandtenkreis. Ausnahmen bilden natürlich Ärzte in Unikliniken, welche häufig nach Feierabend, von Station ins Labor wechseln. Wo den einen aber möglicherweise der Patientenkontakt fehlt, haben die letzteren häufiger Defizite in der wissenschaftlichen Methodik.

Könnte es sein, dass die biomedizinische Wissenschaft gründlicher und robuster wird, wenn dem Forscher klar wird, dass die eigene Forschung direkte Konsequenzen für Patienten haben kann? Es also nicht nur um eine tolle Publikation, also ein paar Zeilen im Lebenslauf geht, bei denen es nicht so genau darauf ankommt, wie belastbar die Methodik, Resultate und Interpretation waren? Dass es nämlich auch um Nutzen oder die Abwendung von Schaden beim Menschen geht? Auch wenn dieser Effekt erst weit hinten in einer Kette von verschiedenen Untersuchungen und klinischer Studien eintritt.

Der tolle Befund in einem Mausmodell, den man sogleich als möglichen wichtigen neuen Mechanismus bei einer Erkrankung wie Krebs etikettiert hat. Nun war die Gruppengröße im Mausexperiment mit n=8 vielleicht recht klein, diese nicht alle verblindet, und ein paar Mal kam auch nicht das raus, was man wollte – das lag aber an den falschen Antikörpern, man hat es deshalb weggelassen. Aber statistisch signifikant mit p=0.045 war das Ergebnis allemal. Weshalb es auch ordentlich publiziert wurde, mit Peer Review und allem was so dazugehört. Der Patient ist hier scheinbar ganz ‚weit weg', trotz der Referenz auf ihn in der Einleitung des Artikels und den Hinweis auf die nun in Aussicht stehende Behandlung einer bisher nicht heilbaren Erkrankung in der Diskussion. Die translationale Mantra eben.

Aber vielleicht bildet dieser Befund die Basis weiterer Studien, möglicherweise auch diese mit ein paar Schwächen im Design und der Analyse behaftet. Und schon steht die etwas wackelige Grundlage für eine Phase I Studie am Menschen! Den wenigsten Forschern ist im Übrigen bewusst, dass häufig sog. ‚individuelle Heilversuche', also die Behandlung von einzelnen, ‚austherapierten' Patienten jenseits des Lehrbuches ohne Zulassung eines Medikament für die Erkrankung, und außerhalb von klinischen Studien, häufig auf Grundlage von ein paar gut publizierten Zellkultur- oder Mäuseexperimenten durchgeführt wird. Plausibel muss es sein, und weil andere therapeutische Optionen bei diesen unglücklichen Patienten schon ausgereizt wurden, steht der behandelnde Arzt

mit dem Rücken zur Wand. Kliniker verlassen sich nämlcich auf die Robustheit der Resultate in der experimentellen Literatur und die dort erzählten ‚Geschichten' (siehe LJ 1/2018).

Was ich damit sagen will ist, dass Grundlagen- und präklinische Forschung eine ethische Dimension haben, welche vielen Forschern nicht bewusst ist. Beim Stichwort Ethik denkt man in diesem Zusammenhang meist an die Tierethik – ist es gerechtfertigt Tieren Leid zuzufügen im Namen der Heilung von menschlichen Krankheiten? Oder daran, dass Betrug, Ideenklau, Plagiarismus etc. ‚unethisch' sind. Es gibt aber eine weitere ethische Dimension, eine die aus den mittelbaren oder unmittelbaren Konsequenzen unseres wissenschaftlichen Handelns für Patienten erwächst. Denn unsere Grundlagenforschung Auswirkungen hat auf Menschen. Dies im positiven Sinne, wenn wir einen Krankheitsmechanismus entdecken, der schließlich therapierbar wird – das hoffen wir alle. Oder aber im negativen Sinne, und das eventuell viel häufiger als uns bewusst ist: Wenn im Labor Fehler gemacht werden, unsauber gearbeitet wird, nur die positiven Befunde berichtet und die negativen in die Schublade wandern, die klinische Relevanz unserer Ergebnisse übertrieben dargestellt wird, usw., alles im Namen einer guten Publikation. Dann werden Ressourcen verschleudert, und es können Menschen zu Schaden kommen. Und dies ohne, dass wir das überhaupt merken, denn es spielt sich mehrere Schritte nach der Grundlagen- und präklinischen Forschung ab, scheinbare losgelöst davon. Und regulatorische Behörden und Ethikkommissionen bei der Zulassung von Studien sich auch nicht für die Qualität der präklinischen Evidenz zu interessieren scheinen, wie die Arbeitsgruppe von Daniel Strech kürzlich herausfand.

Kann das Patientenfoto hier helfen? Natürlich nicht. Nach dem dritten Mal hingucken ist der Effekt verschwunden, auch wenn er existieren würde. Und das ist nicht so sicher. Denn es haben sich Radiologen aus Ottawa daran gemacht, die Studie aus Israel zu wiederholen, über die zwar die New York Times berichtete die aber nur als Abstract publiziert wurde. Mit präspezifizierter Hypothese und sehr sauberer Methodik. Und sie konnten den Effekt nicht replizieren. Dies wurde 2015 ordentlich in einem Fachjournal publiziert. Ein typisches Szenario: Spektakulärer präliminärer Befund schafft es in die Zeitung, trotzdem nur ein Vortrag und ein Abstract existiert. Wenn bei dem Versuch, so etwas zu wiederholen der Effekt verschwunden ist, trotzdem (oder weil) alles lege artis und mit höherer Fallzahl gemacht wurde, ist das keine Nachricht mehr wert. Und das spektakuläre Abstract wird sechs Mal häufiger zitiert als der negative Originalartikel.

Und die Moral von der Geschicht': Vergesse den Patienten nicht! Wer sich vertieft mit dieser Problematik auseinandersetzen will, dem sei unser kürzlich erschienener Artikel hierzu wärmstens anempfohlen (PLoS Biol. 2018 Jun 6;16(6):e2006343).

Mit schlichten Wetten die Wissenschaft retten?

LJ 11/2018

Diese Frage stellte der US-Ökonom Robin Hanson im Titel seines im Jahre 1995 veröffentlichten Artikels. Er schlug darin vor, den klassischen Review Prozess durch eine Markt-basierte Alternative zu ersetzen. Statt Peer Review könnten Wetten darüber entscheiden, welche Projekte gefördert werden, oder welche wissenschaftlichen Fragestellungen priorisiert werden. In diesen sogenannten ‚Prediction Markets' (Prognosemärkten) handeln Individuen mit ‚Wetten', die auf ein bestimmtes Resultat oder Ergebnis setzen. Desto mehr auf einem solchen Marktplatz handeln, umso genauer wird die auf der aggregieren Information der Teilnehmer basierende Vorhersage des Ergebnisses. Der Prognosemarkt bedient sich also der Schwarmintelligenz. Bei Sportwetten und Wahlprognosen kennt man das. Aber in der Wissenschaft? Klingt total verrückt. Ist es aber nicht, es findet gerade Eingang in einigen Sparten der Wissenschaft. Wie funktioniert' s, und was ist dran?

Ausgangspunkt von Hanson's Überlegungen ist die gängige und ernstzunehmende Kritik an der Art und Weise, wie wir Forschungsarbeiten beurteilen und Förderentscheidungen treffen. Peer Review führt zu Risiko-averser Mainstream-Forschung und ist innovationsfeindlich, er regt zum ‚story telling' an, diskriminiert Außenseiter, und favorisiert bereits Erfolgreiche (‚Matthäus-Effekt'). Für Hanson stellt sich die Frage, wie Konsens über die Wertigkeit oder Innovation von Forschung gefunden werden kann, ohne sich auf ein paar Meinungen von ‚Peers', mit all deren Bias und persönlichen Interessen, zu verlassen. Seine Antwort: Analog zu Sport-Wetten könnte man doch eine viel breitere Basis von Experten auf den Erfolg eines geplanten Projektes oder die Richtigkeit eines Befundes wetten lassen. Je nach Fragestellung könnte man auch nicht-Experten mitspielen lassen, und damit für Partizipation und gesellschaftlichen Konsens sorgen. Letztlich schlägt er damit eine Art Crowd-sourcing von Konsensus Bildung vor, und dies mit einem spielerischen Element!

Dass sowas in der Wissenschaft funktionieren kann, zeigen einige neuere Arbeiten, in denen sogenannte ‚Prediction Markets' eingesetzt wurden. Dabei ging es zum Beispiel um die Vorhersage, ob eine Studie replizierbar sein wird, oder nicht. Das geht so: Die teilnehmenden Wissenschaftler erhalten 100 Spielmarken (manchmal auch echtes Geld, z.B. 100 $), mit denen sie dann auf den Replikationserfolg setzen. Sie würden mehr auf diejenigen setzen (d.h. Lose kaufen), von denen sie glauben, dass sie erfolgreich wiederholt werden können. Umgekehrt würden sie die meiden, welchen sie nicht trauen. Durch Kauf und Verkauf von Losen stellt sich auf diesem Markt ein Preis (in Spielmarken, oder echter Währung) für die Lose her. Dieser Preis reflektiert die Wahrscheinlichkeit, mit der die Marktteilnehmer an die Replizierbarkeit der Studie glauben. Dieses Verfahren wurde in einer eben veröffentlichten Publikation eingesetzt. Darin wurden 21 Studien aus der Psychologie, welche zwischen 2010 und 2015 in Nature oder Science veröffentlicht worden waren, wiederholt. Gleichzeitig konnte ein Gruppe von nicht an der

Replikation beteiligten Studenten und Wissenschaftlern, die nicht einmal notwendig als Experten in den jeweiligen Feldern ausgewiesen waren, durch Setzen ihrer Spielmarken Lose kaufen, und damit auf die Replizierbarkeit jeder der 21 Studien wetten. Das Ergebnis: Eine fast perfekte Vorhersage derjenigen Studien, welche dann erfolgreich repliziert wurden, und derer, die nicht wiederholbar waren. Mit einer Vergleichsgruppe wurde eine klassische Befragung durchgeführt: Glauben Sie, dass Studie X oder Y wiederholbar sein wird? Die Umfrageergebnisse waren aber nicht besser als das, was man durch pures Raten erreicht könnte. In einer anderen Studie wurde ein Prediction Market benutzt, um die Ergebnisse der sog. REF für das Fach Chemie vorherzusagen. Mit der REF (Research Excellence Framework), einem hochkomplexen und teuren Verfahren, evaluiert der englische Staat seine Universitäten, um auf Basis der Resultate (Exzellenz!) die Mittel zu verteilen. Ein einfacher Prediction Market mit nur 13 Chemikern, vom Studenten zum Professor, hat ziemlich akkurat das Ergebnis der REF 2014 für alle 33 Chemie-Departments Englands vorausgesagt. Hätte man die in der deutschen Exzellenzinitiative ausgewählten Cluster auch durch ein einfaches Wettspiel erzielen können? Ohne daß Tausende von Wissenschaftlern statt zu forschen, Anträge schreiben und ein Heer von internationalen Gutachtern durch die Republik reisen mussten?

Woran liegt es, dass Wetten bessere Vorhersagen liefern können als Umfragen oder Peer Review? Dass dies so ist, ist keineswegs neu. Wir wissen dies z.B. aus Sportwetten, oder Wetten auf Wahlausgänge. Diese erreichen erstaunliche Vorhersagekraft, fast immer deutlich besser als Umfragen. Ein Faktor hierbei ist vermutlich, dass der Anreiz grösser ist, sich mit dem Umfragegegenstand auseinanderzusetzen, wenn man dabei etwas gewinnen kann. Und wenn es nur ‚Spielmarken' sind, und kein reales Geld, wie in manchen der Prediction Market Studien in der Wissenschaft. In Deutschland könnten diese wohl ohnehin nicht mit echtem Geld durchgeführt werden, denn das wäre unerlaubtes Glücksspiel! Ein weiterer Faktor könnte sein, dass man in einem Prediction Market seine Wetten über längere Zeit durch Kauf oder Verkauf von Losen adjustieren kann. Man kann also die Kandidaten wechseln, vielleicht auch durch den Blick auf den Kurswert der Wette, der sich ja durch Käufe und Verkäufe anderer verändern kann. Vielleicht noch wichtiger ist aber die Schwarmintelligenz, die sich einstellt, wenn Viele mit unterschiedlichem Wissen oder Perspektiven sich einer Fragestellung annehmen.

Ist das aber nun alles bloße Spielerei, oder gibt es für Prediction Markets in unserem Wissenschaftssystem ernsthafte Anwendungen? Eine offensichtliche Einschränkung ist natürlich, dass sich auf diese Weise vor allem dichotomisierbare Fragestellungen bearbeiten lassen. Ist eine Studie replizierbar, oder nicht? Ist ein bestimmtes Resultat richtig oder falsch? Sollte eine bestimmte Fragestellung untersucht werden, oder nicht? Ist der Antrag förderwürdig, oder nicht? Die Ergebnisse des Marktes sind dann Wahrscheinlichkeiten, ein Prediction Market liefert kein inhaltliches Feedback. Dennoch zeigt das oben erwähnte Beispiel aus der Psychologie, dass ein solcher Prognosemarkt eine einfache, und zudem offensichtlich sehr präzise Konsensus Bildung in einem Forschungsfeld ermöglichen kann. Forschungsförderer könnten dies nutzen, wenn es um die Entscheidung geht, welche Forschungsprogramme priorisiert und dann etabliert werden sollen. Die Entscheidung würde dann von der Community, vielleicht aber auch von anderen Stakeholdern (z.B. Patienten) informiert. In der translationalen Medizin stellt sich häufig die Frage, ob eine in Modellsystemen wirksame Substanz in eine teure, und potentiell Patienten gefährdende klinische Entwicklung gebracht werden soll. Ein Prognosemarkt wäre ein simples Verfahren, die von der Community als vielversprechend angesehenen Substanzen auszuwählen, sie ließen sich sogar nach Erfolgswahrscheinlichkeiten sortieren. Natürlich garantiert das keinen Erfolg, aber in Ermangelung objektivierbarer Kriterien wird ja auch heute bereits auf die Expertenmeinung gesetzt. Allerdings nicht die des

Schwarms, sondern weniger ausgewählter Individuen, welche häufig eigene Interessen verfolgen.

Nach meinem Vorstoß zur Einführung einer peer-to-peer Forschungsgrundförderung (Laborjournal 6/2017) und zur Fortbildung für Gruppenleiter und Professoren (Laborjournal 6/2018) also wieder mal eine völlig närrische Idee – Wetten auf die Wissenschaft! Auch wenn es nie dazu kommen wird, liefert uns die Beschäftigung mit radikal von der gängigen Praxis abweichenden Lösungen einen informativen Blick in den Spiegel auf eben diesen Status quo – mit all seinen Stärken und Schwächen. Und wir erkennen: Es läuft nicht optimal, und ginge auch anders!

Mikrobiom-Manie mit melancholischen Mikroben

LJ 12/2018

In einer Bahnhofsbuchhandlung stand ich kürzlich staunend vor einem thematisch hochspezialisierten Büchertisch. Es war der Darm-Hirn Tisch. Die sich darauf stapelnden Bücher versprachen Aufklärung darüber, wie der Darm, und insbesondere sein Inhalt, uns beeinflussen, ja gar emotional fernsteuern. Eine Auswahl von Titeln: „Scheissschlau – Wie eine gesunde Darmflora unser Hirn fit hält“, „Darm heilt Hirn heilt Körper“, „Glück beginnt im Darm“, oder „Das zweite Gehirn – Wie der Darm unsere Stimmung, unsere Entscheidungen und unser Wohlbefinden beeinflusst“. Auch Zeitungen, Magazine, und das Internet verkünden solche Botschaften. Die falschen Darmbakterien machen uns depressiv, die richtigen glücklich, Jogurt hilft deshalb gegen Depressionen.

Woher kommt diese plötzliche Begeisterung für die richtige Darmflora? Hoffähig gemacht hat den Darm, seinen Inhalt, und das Gespräch darüber selbst am Esstisch ganz sicher Giulia Enders. Als Medizinstudentin hat sie 2014 mit ‚Darm mit Charm‘ einen Millionenbestseller aufgelegt, der mittlerweile in viele Sprachen übersetzt wurde. Der kam aber nicht aus dem Nichts, denn letztlich lieferte die Wissenschaft das Fundament für die Mikrobiom Manie, und bedient diese auch heute noch mit anhaltend „spektakulären“ Erkenntnissen. Alle bestens geeignet, die Aufmerksamkeit eines breiten Publikums zu erheischen. Studien in denen gezeigt wird, dass man aus den Windeln eines Einjährigen seine kognitive Entwicklung vorhersagen kann. Ängstliche Mäuse, die nach Transplantation der Fäkalien von draufgängerischen Nagern plötzlich wagemutig werden. Ein Mausmodell in dem die Tiere nur dann Symptome entwickeln, wenn sie Stuhl von Parkinson-Patienten erhalten. Autistische Kinder, die durch Präbiotikatherapie wieder sozial werden. Und so weiter und so weiter. Erinnert uns das alles ein bisschen an die Stammzell-Hype?

Eigentlich ist das mit dem Darm ja alles ein alter Hut. Schon Hippokrates wusste: „Alle Krankheiten beginnen im Darm“. Der Militärarzt Nissle kultivierte 1917 einen ganz

besonderen Bakterienstamm, den er auch gleich unbescheiden nach sich benannte: E. coli *Nissle*. Er fand ihn im Stuhlgang eines Soldaten, der besonders resistent war gegen die brutalen Bedingungen in den Schützengräben des ersten Weltkriegs am Balkan. Adolf Hitler hat das so beeindruckt, dass er sich (und seinem Schäferhund Blondi) täglich zwei Kapseln E. coli Nissle verordnete. Die Bakterien aus dem Stuhl des unbekannten Soldaten kann übrigens auch heute noch jeder schlucken, der möchte: Sie werden als Probiotikum in jeder Apotheke unter dem Namen *Mutaflor* verkauft. Der Soldat wir allerdings nicht auf dem Beipackzettel erwähnt. Aber weder Hippokrates noch Nissle können den kometen-haften Aufstieg des Mikrobioms erklären. Dieser begründet sich nämlich vor allem in den Fortschritten der Gensequenzierungstechnologie. Diese machten es möglich, im Hochdurchsatz Bakterien genetisch zu typisieren. Und zwar ohne sie vorher kultiviert zu haben. In der Kulturschale macht der Sauerstoff bei der Entnahme und Probenverarbeitung den Anaerobiern den Garaus, sie werden dann nicht detektiert. Erst die Sequenzierung der Bakteriengenome förderte die große Vielfalt des Mikrobioms in Mensch und Maus zu Tage. Der Rest ist Geschichte.

Mittlerweile spuckt Pubmed 30 neue Mikrobiom Artikel pro Tag aus. Kein Fach, das nicht von der Mikrobiom Manie erfasst wurde. Die Neurologie und Psychiatrie bilden da nur die Spitze des Eisbergs. Kardiologie, Endokrinologie, Nephrologie sind dicht dran. Aus den DFG-Fachkollegien ist zu hören, dass mittlerweile ein substantieller Anteil der Anträge einen Mikrobiomteil enthält. Und dies in der Regel von Wissenschaftlern, die weder Erfahrung in Mikrobiologie und Prokaryoten, noch Bioinformatik von bakteriellen Genomen haben. Und hier fängt‘s an problematisch zu werden.

Dazu kommt, dass viele der Mikrobiom Tierstudien die gleichen Mängel haben, die wir von aus der Präklinik kennen: Häufig nicht randomisiert, unverblindet, ohne vordefinierte Ein- und Ausschlusskriterien, geringe Fallzahlen, nicht A priori formulierte und präregistrierte Hypothesen, und so fort. Damit ist der selektiven Nutzung von Daten, einschließlich des Weglassens von unbequemen Befunden oder der Veränderung der Hypothese im Verlauf des Experiments Tür und Tor geöffnet. Befördert wird all dies dadurch, dass viele wissenschaftliche Journale derzeit ganz scharf darauf sind, Mikrobiom Studien zu publizieren. Allerdings nur solche, welche Positives bieten. Nicht solche, die Vorbefunden widersprechen, oder schlicht keinen Effekt des Mikrobioms ausmachen konnten.

Bei den klinischen Mikrobiom-Studien sieht es leider nicht viel besser aus. Diese haben in der Regel sehr geringe Fallzahlen, sind schlecht kontrolliert, häufig nicht präregistriert, schließen sehr heterogene Probanden oder Patientengruppen ein, und analysieren die Daten auf fragwürdige Weise, insbesondere was Bioinformatik uns Statistik betrifft. Aufgrund der Vielzahl der publizierten Studien konnten mittlerweile eine Reihe von Meta-Analysen durchgeführt werden. Diese schlussfolgern in der Regel: Eigentlich kann man gar nichts sagen, da die vorliegenden Studien Mindestanforderung an die Qualität nicht erreichen, zu klein sind, und ein massiver Publikationsbias vorliegt. Dort wo größere, gut gemachte Studien vorliegen, sind diese entweder neutral, oder sogar negativ, wie kürzlich eine relativ große Studie an Patienten mit Reizdarm. Den Patienten in der Stuhltransplantationsgruppe ging es schlechter als den mit Placebo behandelten. Das größte Problem der klinischen Mikrobiomliteratur ist allerdings, dass viele Beobachtungsstudien Korrelation und Kausation verwechseln. Depressive Patienten haben ein anderes Mikrobiom als glückliche Probanden. Ergo muss es das Mikrobiom sein, das depressiv macht. Dass man mit einer Depression andere Essgewohnheiten und einen anderen Lebensstil hat, was selbstredend auch aufs Mikrobiom durchschlägt, wird dabei geflissentlich übersehen. Depression kann man hier mit Autismus, Parkinson, usw. ersetzen.

Wir erleben also, wie vorher im Stammzellfeld, einen klassischen sogenannten *Gartner* Hype-Zyklus: Nach der Einführung einer neuen Technologie und einsetzender Euphorie steigen die Erwartungen immer höher. Alle Studien finden, wozu sie angetreten waren. Dieser Aufschwung wird erst gebremst, wenn sich sehr viele im Feld mit dem Phänomen befassen, und die Forschung und deren Ergebnisse nichts mehr aufregend Neues bietet. Dann ist auch der ‚Noise' so hoch, und die Datenlage mangels Qualität so instabil, dass die Hoffnungen einer allgemeinen Enttäuschung weichen. In dieser Phase ist die Mikrobiom Manie wohl noch nicht, obwohl sich erste Anzeichen dafür mehren. Einige davon habe ich oben erwähnt. Am Tiefpunkt der Enttäuschung trennt sich dann die Spreu vom Weizen, die soliden Ergebnisse verdichten sich zu einem zunehmenden Verständnis der Mechanismen, es wird ein stabiles Niveau erreicht, das „Plateau der Produktivität". Da sind die Mikrobiomforscher aber leider noch weit davon entfernt.

Muss Forschung notwendig durch solche Zyklen aus übertriebenen Erwartungen und darauffolgender Enttäuschung gehen? Gegen Enthusiasmus und Hoffnung auf neue Erkenntnis wäre ja wirklich nichts einzuwenden – das sind nämlich wesentliche Triebfedern erfolgreicher Forschung. Das Plateau der Produktivität könnte man allerdings viel schneller und Ressourcen-schonender erreichen. Man müsste dazu nur aus den bisher durchgelaufenen Zyklen lernen. Forscher, Journale, Fördergeber, und die Medien sollten sich einfach an einige Grundregeln der guten Wissenschaft erinnern. Welche sind das? Verminderung von Verzerrungen („Bias") durch Maßnahmen wie Randomisierung, Verblindung, Präregistrierung. Reduktion der Rate der falsch positiven Resultate durch höhere Fallzahlen (damit statistischer Power). Ergebnisse, welche unseren Hypothesen widersprechen sollten publiziert werden, und wichtige Befunde unabhängig reproduziert werden.

Und wenn der Narr liest, dass Stammzellen querschnittsgelähmte Mäuse heilen, oder Laktobazillen Depressive fröhlich machen, erinnert er sich an Carl Sagan's Diktum: Außergewöhnliche Behauptungen erfordern außergewöhnliche Evidenz!

Es irrt der Mensch, solang´ er strebt

LJ 1-2/2019

Unter dem Titel „Growth in a Time of Debt" erschien 2010 ein Artikel der hoch angesehenen Harvard-Ökonomen Carmen Reinhart und Kenneth Rogoff. Es ging um den Zusammenhang von nationalem Wirtschaftswachstum und Staatsverschuldung. Darin berichteten sie von ihrer Entdeckung eines erstaunlichen, weltweit zu beobachtendem Zusammenhang: Bei steigender Staatsverschuldung steigt zunächst das ökonomische Wachstum einer Nation an. Wenn allerdings die Staatsverschuldung 90 % überschreite, kehre sich dies Verhältnis recht abrupt um. Aus dem Wachstum wird eine Kontraktion, die Wirtschaftsleistung sinkt dann mit weiter steigender Verschuldung. Die Entdeckung einer „90 %

Schuldenschwelle“ schlug ein wie eine Bombe. Manche vermuten, der Artikel hätte nach der Finanzkrise von 2008 die europäische Sparpolitik mitbegründet. Sicher ist jedenfalls, dass das Paper von westlichen Politikern begeistert zur Rechtfertigung ihrer restriktiven Fiskalpolitik genutzt wurde. Im Rahmen einer Semesterarbeit und nichts Böses ahnend nahm sich dann 2013 der Student Thomas Herndon die Daten vor, auf denen das Reinhart-Rogoff Paper basierte. Nach einigem hin und her hatten ihm die Autoren das Excel-Originalspreadsheet überlassen. Und siehe da, in wenigen Minuten fand er eine Reihe von gravierenden Fehlern in der Tabellenkalkulation! Nach Korrektur verschwand die Schuldenschwelle, ja die Daten belegten nun das Gegenteil, einen stetigen, positiven Zusammenhang von Staatsverschuldung und Wachstum über den gesamten untersuchten Bereich! Was lernen wir daraus? Abgesehen davon, dass der Grundfehler von Reinhart und Rogoff natürlich in der Verwechslung von Korrelation mit Kausation steckt: Excel eignet sich nicht zur Analyse komplexer wissenschaftlicher Daten. Noch wichtiger aber: Wissenschaftler machen Fehler, und diese können fatale Konsequenzen haben.

Irren ist menschlich, so heißt es jedenfalls. Fehler treten demnach überall auf wo Menschen tätig werden. In vielen Bereichen der Gesellschaft hat man das erkannt, insbesondere dort wo Fehler unmittelbar zu kleineren oder größeren Katastrophen führen können. Zum Beispiel in Atomkraftwerken, in der Flugsicherung, oder auch im Krankenhaus. Ein professioneller Umgang mit Fehlern zum Zwecke deren Verhinderung oder gar Wiederholung ist dort in Form von Fehlermanagement-Systemen gesetzlich vorgeschrieben. In der biomedizinischen Forschung kennt man sowas interessanterweise nicht. Von Fehlern hören wir dort nur aus den „Errata“, die sich manchmal in PubMed verirren. Meist bestand der „Fehler“ dann darin, dass der Name des Instituts einer der Autoren falsch geschrieben wurde, oder, horrible dictu, ein Autor an falscher Stelle aufgeführt wurde!

Waren Sie schon mal auf einer Institutsbesprechung, in der eine Wissenschaftlerin oder ein technischer Assistent einen Fehler vorgestellt hat, den sie oder er kürzlich gemacht haben? Und dieser dann gemeinsam analysiert und diskutiert wurde? Vermutlich nicht. Gibt’s nämlich kaum. In der Klinik ist so etwas dagegen Standard. In sogenannten „Morbiditäts- und Mortalitäts-Konferenzen“ werden besondere Behandlungsverläufe, unerwünschte Ereignisse, Todesfälle usw. systematisch aufgearbeitet. Das Ziel dabei ist, in einer multiprofessionellen Umgebung Fehler und Schwachstellen – vor allem in klinischen Prozessen – zu identifizieren und daraus Verbesserungsmaßnahmen abzuleiten und umzusetzen.

Könnte es sein, dass die biomedizinische Wissenschaft so etwas nicht nötig hat? Weil kaum Fehler gemacht werden? Und wenn ja, sie ohne Einfluss auf die Ergebnisse oder deren Interpretationen sind? Sich Fehler in der Wissenschaft ohnehin nicht wiederholen? Aber natürlich machen wir Fehler, eine Menge sogar, und folgenschwere sind auch dabei! Und unsere Fehler wiederholen sich auch manchmal. Dabei können Fehler in der biomedizinischen Grundlagenwissenschaft mittelbar Patienten schädigen (siehe dazu der LJ 10-2018) , Doktoranden um Jahre Ihrer Jugend berauben, zum unnötigen Tod oder Leid von Versuchstieren beitragen, oder ganz allgemein zu massiver Ressourcenverschwendung führen.

Was gibt es nicht alles für Fehlerquellen in der komplexen Arbeitswelt des Labors! Systematische Fehler von Geräten, wie Pipetten, Waagen, Plattenreadern, etc. Geräte gibt es im Labor ja wahrlich genug, und die meisten davon sind komplexer in der Anwendung als eine Pipette. Und die kann man überdrehen, dann ist sie nicht mehr kalibriert, und das torpediert die Experimente der Nachnutzer. Überhaupt Geräte-Kalibration: Wenn

diese nicht, oder falsch durchgeführt wurde, steht alles was daraufhin mit dem Gerät gemacht wird, unter einem schlechten Stern. Dann die Fehler durch Abweichungen von Protokollen. Am wichtigsten aber wohl ganz einfach ‚menschliche Fehler', wie die offen gelassene Tiefkühlertür, die falsche Etikettierung eine Medienflasche, der Rechenfehler in der Verdünnungsreihe, eine falsch abgelesene oder niedergeschriebene Dokumentation, ein Fehler in der Benutzung Analyse-Software (Excel!), und so weiter und so fort. Damit aber nicht genug, denn die Fehler können sich ja in die Protokolle oder sogar Publikationen einschleichen, und dort bei anderen zu Problemen führen. Erschwerend kommt noch dazu, dass in den meisten Laboren eine Menge von Leuten unterwegs sind, von sehr unterschiedlichem Hierarchie-, Ausbildungs-, Motivations-, oder Einarbeitungsgrad.

Und hier liegt möglicherweise ein wichtiger Grund, warum in der biomedizinischen Forschung scheinbar kaum Fehler gemacht werden. Ist es die Angst davor, einen Fehler zu berichten, ihn zuzugeben? Man könnte ja zur Verantwortung gezogen werden, als Anfänger dastehen, oder sich den Zorn derer zuziehen, die von dem Fehler möglicherweise betroffen sind oder waren. Es sind diffuse Ängste, die da wirken. Vermutlich wird man in den meisten Laboren sagen: „Bei uns wird da ganz offen drüber gesprochen, bei uns wird keiner für seine Fehler bestraft". Aber wissen Sie, wie viele Fehler bei Ihnen wirklich gemacht werden, und wie viele davon auch berichtet? Wie werden Fehler in Ihrer Laborumgebung kommuniziert? Wie können andere aus Fehlern lernen? Wie wird sichergestellt, dass sich Fehler nicht wiederholen?

Dies kann nur in einem Umfeld funktionieren, in dem eine „Fehlerkultur" existiert. Und die ist eine komplexe Sache. Sie hat viel mit Einstellung zu tun, aber auch mit sachlichen Voraussetzungen. Die Einstellung besteht darin, Fehler als Chance zu begreifen. Es braucht eine kristallklare Losung der Labor- oder Gruppenleitung, dass Fehler, außer sie geschehen mit Vorsatz, nie zu wie auch immer gearteter „Bestrafung" oder Benachteiligung führen dürfen. Zu den sachlichen Voraussetzungen zählt es, Formate für das Berichten von Fehlern vorzuhalten, welche auch anonym genutzt werden können. Im einfachsten Fall also sowas wie ein „Fehlerkasten", in den man einen Zettel werfen kann. Das ist aber nur die halbe Miete, denn der Fehler muss ja auch analysiert, an andere kommuniziert, und evtl. Maßnahmen eingeleitet werden, damit er sich nicht wiederholen kann. Dies könnte zum Beispiel ein regelmäßiger Tagesordnungspunkt beim Labmeeting sein. Wer sich für einen etwas systematischeren Umgang mit Fehlern im Labor interessiert, dem sei unser Artikel (PLoS Biol. 2016 Dec 1;14(12):e2000705) und die dort vorgestellte open source Software LabCIRS empfohlen. Auch im Laborjournal haben wir das schon mal vorgestellt („Aus Fehlern wird man klug" LJ 1-2/2017). Das Laboratory Critical Incidence Reporting System (LabCIRS) erleichtert den Umgang mit Fehlern insbesondere in Kontext von Grundlagenforschung. Fehler können darin auch anonymisiert gemeldet werden, was insbesondere in der Anfangsphase einer proaktiven Auseinandersetzung mit Fehlern wichtig ist.

Allerdings ist es in einer überschaubar großen Arbeitsgruppe oft gar nicht möglich, einen Fehler anonymisiert zu melden, weil schon seine Beschreibung den Urheber verrät. Es geht also ganz wesentlich um das „Mind set": Alle machen Fehler, wir können von den eigenen und denen anderer lernen, und Fehler einzugestehen sowie dafür zu sorgen, dass sie sich nicht wiederholen, ist eine Ausdruck von Professionalität. Weil aber Fehlermachen tabuisiert und Angst-besetzt ist, sollte man versuchen, das Thema positiv zu wenden. Klingt komisch, aber warum nicht jährlich die ‚besten' Fehler prämieren? Also diejenigen, aus deren Aufarbeitung der größte Nutzen gezogen wurde. Auf medizinischen Konferenzen gibt es manchmal eine Session ‚Mein schlimmster Fehler'! Häufig ist

das die bestbesuchte Sitzung der ganzen Konferenz. Wäre das nicht auch was für die Grundlagenforscher?

Zu guter Letzt: Sollten Fehler erst nach Publikation einer Studie entdeckt werden, Muss gehandelt werden. Dies gilt sowohl für falsche Angaben im Methodenteil (falsche Dosierung, falscher Referenzwert etc.) als auch für zu korrigierende Resultate (Fehler in der Auswertung, in der Graphik etc.), oder auch fehlerhafte Schlussfolgerungen. Klingt irgendwie selbstverständlich, aber wenn man systematisch Errata in den gängigen Journalen durchsucht, findet sich sehr selten etwas in dieser Richtung. Liegt das daran, dass so selten Fehler im Nachhinein entdeckt werden, welche korrekturbedürftig wären? Meine Vermutung ist vielmehr, dass es daran liegt, dass die meisten Autoren Große Angst vor dem Stigma eines Erratums oder einer Retraktion haben. Befragen Sie sich selbst: Ist ihr eigenes Gewissen rein in dieser Kategorie? Waren Sie noch nie in der Situation, dass Sie von Ihnen bereits Publiziertes eigentlich nochmal hätten richtigstellen müssen?

Der vielgelesene und von mir geschätzte Blog retractionwatch.com, der ja leider auch ein bisschen von unserer Häme und unserem Voyeurismus lebt, hebt deshalb dankenswerter Weise auch immer wieder ‚gut gemachte' Retraktionen nach ‚honest error' positiv hervor. Man hat dafür eigens die Kategorie „doing the right thing" geschaffen. Ich jedenfalls habe größten Respekt vor Wissenschaftlern, die einen ‚honest error' zum Anlass nehmen, Ihre Publikation zu korrigieren!

Vom Triangulieren beim Experimentieren

LJ 3/2019

Triangulation! Die Ägypter bauten damit ihre Pyramiden. Die Griechen haben einen Zweig der Mathematik daraus entwickelt. Noch bis ins 19. Jahrhundert wurden ganze Länder so vermessen. Weit ins 20. Jahrhundert hinein haben Schiffe ihre Position damit bestimmt. Man braucht nur ein Geodreieck und einem Winkelmesser, den die Vermessungskundler einen Theodoliten nennen, und die Koordinaten von zwei sichtbaren Landmarken, um seine eigene Position durch Triangulation auf einer Karte zu bestimmen. So einfach ist das!

Was schwärmt der Narr da von der Landvermessung, der Geodäsie? Könnte es vielleicht sein, dass die Triangulation auch ein wichtiger methodischer Ansatz in der Biologie ist? Ein Heilmittel gar für die Replikationskrise? Munafo und Smith haben das vor kurzem in einem Kommentar in Nature so postuliert. Die Soziologen nennen es Triangulation, wenn sie zwei oder mehr unterschiedliche Methoden einsetzen, um einen Sachverhalt zu untersuchen. Wenn die Resultate an einem Punkt konvergieren, d.h. zum selben Ergebnis führen, erhöht dies die Validität und die Glaubwürdigkeit derselben. Machen wir das nicht auch routinemäßig in den experimentellen

Lebenswissenschaften? Hat die knock-out Maus denselben Phänotyp wie eine, bei welcher der Signalweg pharmakologisch geblockt wurde? Korrelieren Transkript und Proteinexpression mit dem Phänotyp? Auch die gute alte Dosis-Wirkungskurve hat was davon – ‚peilen' wir mit ihr doch auf verschiedene Konzentrationen.

Die biologisch-medizinische Grundlagenforschung ist es also gewohnt, von bereits etabliertem Wissen (die Landmarken des Vermessers!) mit unterschiedlichen Methoden ein Ziel ‚anzupeilen'. Konvergieren die Resultate? Bingo, schon haben wir den biologischen Mechanismus sicher verortet! Deshalb lässt es viele von uns kalt, wenn Spaßverderber mit einfacher Oberstufen-Statistik nachweisen, dass die meisten Studien in der Biomedizin trotz signifikantem p-Wert falsch positiv sein müssen (siehe der Wissenschaftsnarr LJ 4/2017). Weil wir ja nicht nur auf EIN Resultat setzen. Sondern mittels verschiedener Ansätze triangulieren! Zur Absicherung von Ergebnissen sollte das sogar besser sein als zu Reproduzieren. Wenn etwas einfach nur wiederholt wird, ist es nicht unwahrscheinlich, dass ein systematischer Fehler mit wiederholt wird. Das macht das Ergebnis vielleicht reproduzierbar, aber immer noch nicht richtig.

Lagen die Skeptiker also falsch, welche sich Sorgen machten um die Reproduzierbarkeit der Ergebnisse der Biomedizin? Schön wär's, aber leider ist die Sache trotz munteren Triangulierens in vielen Laboren nicht so einfach. Denn wie jeder Geodät bestätigen wird, das Prinzip der Triangulation ist zwar einfach, aber exakte Ortsbestimmung durch Triangulieren kein Kinderspiel. Auf die biomedizinische Grundlagenforschung bezogen ergeben sich da einige Spielregeln, die leider häufig nicht eingehalten werden. Zunächst einmal müssen wir uns sicher sein, dass unsere ‚Landmarken' biologisch fundiert sind, und nicht selbst Resultat falsch positiver Ergebnisse, einer Überinterpretation der Ergebnisse, oder ein Artefakt experimenteller Bedingungen. Der Geodät tut sich da leichter: Die Position der Referenzlandmarken findet er Dezimalsekunden-genau auf der Landkarte. Wenn wir den Winkelmesser ansetzen, das heißt verschiedene Methoden auf eine Hypothese ansetzen, müssen wir zudem sauber ablesen. Dies bedeutet: Verblindung, keine Flexibilität in der Auswahl der zu verwendenden Datenpunkte, usw. Und: Wenn die Peilung einen Winkel ergibt, der uns nicht ins Konzept passt, dürfen wir nicht einfach dessen Wert ignorieren und den Theodoliten ein wenig versetzen, um nochmals anzulegen. Nach dem Motto: Probieren wir doch einfach mal einen anderen Antikörper! Oder einen anderen pharmakologischen Blocker! Und wenn wir sowas schon machen, müssten wir dies begründen und in der Publikation berichten. Der abgelesene Winkel muss von hoher Präzision sein. Mit niedrigen Fallzahlen ist dies aber leider meist nicht zu haben, denn die biologische Varianz ist enorm. Und der gemessene Winkel muss tatsächlich existieren, darf also kein falsch positiver Befund aufgrund niedriger Fallzahl oder unwahrscheinlicher Hypothese sein.

Sie ahnen, worauf ich hinauswill: Wenn Landvermesser so ‚Triangulieren' würden, wie wir experimentieren, würden sie Landkarten erzeugen, welche sich zwar sehr plausibel zeichnen lassen. Man könnte Sie auch drucken, und sie würden hübsch aussehen. Aber wenn sich ein Wanderer danach richten würde, müsste er sich arg verlaufen.

Bei richtiger Anwendung kann Triangulation aber tatsächlich der Schlüssel zu robusteren Ergebnissen sein. Soll heißen durch Experimente mit ausreichender Fallzahl, mit Verminderung von Bias (Verblindung, Randomisierung, etc.), mit vorbestimmten Ein/Ausschlusskriterien, und Veröffentlichung der Ergebnisse unabhängig von den Resultaten. Dann ist Triangulation außerdem sehr effektiv: Die kumulativen Fallzahlen der experimentellen Serien verschiedener methodischer Ansätze können tatsächlich niedriger sein als die einer einzigen Serie mit nur einem Ansatz. Und dies sogar bei gleicher oder höherer Power und außerdem höherer externer Validität. Das ist schwer in Zahlen

zu fassen, denn hierfür lässt sich keine ‚Power' im statistischen Sinne berechnen. Und auch externe Validität, also die Generalisierbarkeit und Repräsentativität von Ergebnissen ist nicht wirklich quantifizierbar.

Wie geht es nun weiter, nachdem man durch Triangulation ein biologisches Phänomen vorläufig verortet hat? Natürlich wird man es der Welt in einer Publikation kundtun wollen. Dabei sollte man sich aber über die weiterhin existierenden Limitationen der so gewonnenen Befunde im Klaren sein. Das sollte sich schon im Titel bemerkbar machen, indem man die Studie als explorativ kennzeichnet. Und in den Conclusions sollte man sich zurückhalten. Der Verweis auf nun mögliche Therapien am Menschen, oder notwendige Überarbeitungen von Textbüchern ist da in den wenigsten Fällen angebracht. Erst eine Konfirmation durch Replikation in anderen Laboren kann Gewissheit über die Existenz und das wahre Ausmaß eines Effektes schaffen. Das benötigt in der Regel höhere Fallzahlen als im Originalexperiment, und die Studie sollte präregistriert sein. Es ist völlig klar, dass dies nur bei einer geringen Anzahl von Befunden überhaupt machbar, sinnvoll und praktikabel ist. Wenn es aber z.B. darum geht, zu entscheiden, ob man vom Tierexperiment zu einer Studie am Menschen übergeht, sollte dies selbstverständlich sein. Schön, dass das Bundesministerium für Bildung und Forschung (BMBF) dies auch so sieht, und vor kurzem eine Ausschreibung für präklinische konfirmatorische Studien veröffentlich hat. Das ist ein revolutionärer Vorstoß, der hoffentlich Schule machen wird: Bei anderen Fördergebern, aber auch bei uns Wissenschaftlern.

Liebe DFG, verlost doch Eure Fördergelder!

4/2019

Kennen Sie den schon: ‚Leider muss ich Ihnen mitteilen, dass die Deutsche Forschungsgemeinschaft nach eingehender Prüfung durch die zuständigen Ausschüsse Ihrem Antrag auf Gewährung einer Sachbeihilfe nicht entsprechen konnte.' Vermutlich ja, denn das ist der Standardsatz in den Ablehnungsschreiben der DFG. So oder so ähnlich kriegen wir ihn auch von anderen Fördergebern. Und rein statistisch gesehen passiert uns das leider recht häufig. In der Biomedizin liegen die Förderquoten zwischen 5 und 30%. Ablehnungen von Anträgen empfinden wir, nicht ganz zu Unrecht, häufig auch als persönliche Kränkungen. Haben wir doch unsere besten Ideen reingeschrieben, meist auch schon einiges an Ergebnissen verarbeitet, die wir ‚schon im Kasten hatten', das Ganze gar mit viel Prosa aufgehübscht, dazu den wichtigsten möglichen Gutachtern durch strategisch platzierte Zitate geschmeichelt usw. Und dann die Ablehnung! Also nochmal von vorne, alles umschreiben, nochmals einreichen, vielleicht bei einem anderen Fördergeber. Womit man als Forscher halt so seine Zeit verbringt. Wenn man nicht gerade selber die Anträge

Anderer begutachtet. Im Schnitt 40 % Prozent ihrer oder seiner Zeit verbringen Wissenschaftler heutzutage mit dem Schreiben oder Begutachten von Anträgen.

Dabei kennen wir alle die Misere des Antragswesens: Hoher Aufwand für alle Beteiligten (auch bei den Förderinstitutionen); häufig marginale Expertise im Review Prozess; letztlich Förderung von Mittelmaß, wohingegen Innovation und Risikoreiches auf der Strecke bleiben; fehlende Kriterien für zukünftigen Erfolg von Projekten; ‚Gutacherseilschaften' und Interessenkonflikte; Matthäus-Effekt; Bevorzugung etablierter Forscher und Mangel an Fairness, um nur einige von denen zu nennen, welche die Wenigsten von uns abstreiten würden. Aber wir Wissenschaftler scheinen eine Schafsnatur zu haben. Trotz allgegenwärtiger Kritik, insbesondere unter befreundeten Kollegen und nach ein paar Bier, drehen wir unbeirrt weiter das Hamsterrad. So ist das System halt, und ein anderes haben wir nicht.

Dabei gäbe es recht naheliegende Alternativen zum Peer Review von Anträgen. Sie sind sehr plausibel, nur wurde keine davon im großen Maßstab getestet. Aber viel ineffektiver und unbefriedigender als das gegenwärtige Prozedere kann es doch kaum werden. Das böte jede Menge Raum fürs Experimentieren! Wissenschaftler neigen doch zu dieser Tätigkeit, warum nicht auch mal in Sachen Antragswesen? Über eine solche Alternative hat der Narr schon vor einiger Zeit berichtet: Grundfinanzierung von Wissenschaftlern, kombiniert mit Peer-to-Peer Förderung. Dabei müssen Wissenschaftler einen bestimmten Anteil ihrer Grundfinanzierung an andere Forscher ihrer Wahl weitergeben. Wem das Spanisch vorkommt, der sei eingeladen, sich das nochmals anzukucken (LJ 6/2017). Dies System macht viel Sinn, ist aber recht radikal, und dürfte es im konservativen Wissenschaftsbetrieb schwer haben, je ernsthaft auf den Prüfstand zu kommen. Dagegen hat eine andere Idee größere Chancen umgesetzt zu werden, und auch die klingt hinreichend verrückt: die Förderlotterie!

Mindestens drei große Fördergeber weltweit experimentieren derzeit mit Systemen, bei denen der Zufall eine wichtige Rolle in der Förderentscheidung spielt: die ‚Explorer Grants' des Health Research Council of New Zealand, die ‚Seed Projects' von New Zealand's Science for Technology Innovation und, man höre und staune, in Deutschland die Volkswagen Stiftung mit der "Experiment!" Förderlinie. Frei nach dem Motto: ‚Forschungsförderung hat ohnehin schon Lotterie-Charakter, dann lasst uns doch gleich eine ordentliche daraus machen'.

Das Ganze ist weniger verrückt als es klingt, hat eine Historie, die bis ins 19. Jahrhundert zurückreicht, und basiert auf einer Reihe von soliden theoretischen Arbeiten und Simulationen. Wie funktioniert so eine Förderlotterie? In ihrer reinsten Form ganz einfach: Wissenschaftler stellen Anträge. Diese werden einem initialen Check unterworfen: Was ganz offensichtlich keinen Sinn macht, nicht den formalen Vorgaben entspricht, etc. wird triagiert. Dann kommt alles was übrig blieb in einen Topf, und man zieht so viele Anträge raus, wie gefördert werden können. Das lässt sich natürlich weiter verfeinern: Zum Beispiel bei den Kriterien der Vorselektion. Hier kann ein Minimalset von Anforderungen Anwendung finden, dass sowohl das wissenschaftliche Oeuvre des Antragstellers als auch die verfolgten Hypothesen und angewandten Methoden einbezieht. Diese sollten aber relativ breit und offen gehalten werden, sonst eliminiert man ja schon am Anfang die unorthodoxen Ideen und verborgenen Fertigkeiten der Antragsteller. Man könnte auch die top Anträge des initialen Review direkt fördern, und nur den ‚mittleren Bereich' losen. Also die Anträge zwischen den eindeutigen Tops und Flops. Das müsste nicht einmal in solch eindeutigen Kategorien geschehen. Auch eine gewichtete Lotterie ist denkbar, in welcher der Zufall dosiert eingespeist wird: Je besser die Noten im initialen Review waren, desto mehr ‚Lose' erhält der Antrag, damit erhöhen sich die Chancen

gezogen zu werden. In jedem Fall wird man sich in den Lotterieverfahren mit relativ kurzen Anträgen begnügen, die Prosa wird ja ohnehin nicht evaluiert. Dazu sollte die Lotterie öffentlich sein, die Details des Zufallsprozess transparent, und die Kriterien sowie die Ergebnisse des initialen Reviews offen zugänglich.

Aber was bringt das Ganze? Paradoxerweise eine fairere Selektion von Geförderten als im Peer Review! Denn der zieht willkürlich eine Grenze zwischen Geförderten und Abgelehnten. Jeder der schon mal in Review Panels war kennt den Prozess. Man reiht die Anträge nach den Noten bei der Begutachtung, und zieht dann eine Linie unterhalb derer keine Fördermittel mehr zur Verfügung stehen, weil sie für die Projekte oberhalb der Linie ausgegeben wurden. Haben die Projekte unterhalb und oberhalb der Linie wirklich unterschiedliches Potential erfolgreich zu sein? Meist reden sich dann die Mitglieder des Review Panels die Köpfe heiß, und schieben mal Anträge nach oben, mal nach unten. Völlig evidenzbefreit, dass dies was bringt. Denn wenn eines wirklich klar hervorgeht aus der Wissenschaftsforschung, dann dies: Der Peer Review Prozess, ganz egal welcher Art, ist weder prädiktiv noch konsistent. Er ist nicht prädiktiv für das Potential und den zukünftigen Erfolg von geförderten Projekten. Dies weniger, weil die Gutachter sich häufig gar nicht genug mit den Anträgen auseinandersetzen, oder genug Expertise zu deren Beurteilung haben, oder Vorurteile haben, oder einen Interessenkonflikt. Sondern hauptsächlich deshalb, weil es für zukünftigen Erfolg von Anträgen gar keine belastbaren Kriterien gibt. Der Peer Review Prozess ist zudem nicht ausreichend konsistent, da eine Wiederholung mit anderen, aber gleich qualifizierten Gutachern nicht annähernd zu den gleichen Resultaten führt. Auch hierfür gibt es solide Evidenz. Die Lotterie ist daher fairer, weil sie Antragsteller gleichbehandelt, für deren Unterscheidung es keine gesicherten Kriterien gibt. Fair ist sie auch darin, dass sie allen Qualifizierten eine Chance gibt, ob Frau ob Mann, ob jung ob alt, und auch wenn diese nicht im Windschatten einer Großen Institution oder eines etablierten Netzwerkes fahren.

Zusätzlich reduziert die Lotterie massiv den Aufwand, sowohl auf Seiten der Antragsteller (kürzere Anträge), als auch natürlich besonders bei den Gutachtern und den Administratoren der Fördergeber. Sie ist damit effizienter, und setzt Zeit und Ressourcen frei für echte Wissenschaft.

Der interessanteste Vorteil der Lotterie liegt allerdings darin, dass sie Diversität und Innovation fördert. In der Lotterie würden sicher auch viele Mainstream-Projekte gelost. Denn rein statistisch schreiben die meisten von uns Mainstream-Anträge, sonst gäbe es den Mainstream ja gar nicht. Aber weil der Mainstream durch das Zufallsprinzip nicht positiv selektiert wird, würde mehr ‚Breakthrough', ‚Disruption' oder wie immer man das kategorisieren mag, gefördert. Auch deshalb, weil mehr Anträge im System wären, die wir momentan gar nicht sehen, weil die Antragsteller in Antizipation einer Ablehnung ihn gar nicht stellen.

Jetzt werden Sie sicher eine Reihe von Bedenken haben. Würde die Einführung eines solchen Systems nicht zu einem Aufschrei führen? Wenn schon nicht bei den Wissenschaftlern, so doch in der Öffentlichkeit: Sieh her, die Wissenschaftler verlosen unser Steuergeld, jetzt sind sie völlig verrückt geworden! Und würde so ein System nicht zu einer massiven Erhöhung von Anträgen niedriger Qualität führen? Quick and dirty, einfach um ein Los in der Lotterie zu haben? Und dann wird das gewonnene Fördergeld für unsinnige Aktivitäten verprasst! Zunächst einmal sollten wir uns daran erinnern, dass auch im gegenwärtigen System viel Müll produziert wird, gefördert aus öffentlichen Mitteln. Und das mit erheblichem Aufwand in der Selektion des Mülls. Außerdem würde sich das Problem ja dadurch beheben lassen, dass Wissenschaftler, die das System mit minderwertigen Anträgen fluten, aus dem System ausgeschlossen werden könnten.

Und wo bleibt der Aufschrei? Er ist ausgeblieben: Die Volkswagen Stiftung hat 2017 begonnen, eine Antragslotterie zu testen. Dafür verdient sie höchstes Lob! Sie tut genau das, was wir Wissenschaftler auch tun: Wenn wir eine plausible und relevante Hypothese haben, führen wir ein Experiment durch um diese zu widerlegen, oder aber Evidenz für ihre Richtigkeit zu gewinnen. Genau das macht die Volkswagen Stiftung im Experiment!-Programm. Dort werden Anträge in einem teil-randomisierten Verfahren über eine Lotterie vergeben. Und das ganze mittels Begleitforschung evaluiert. Man kuckt also, ob sich klassisch per Peer Review ausgewählte Projekt in ihrem Verlauf und Erfolg unterscheiden von denen, welche erlost wurden. Die DFG dagegen gibt ihre Milliarden in die Projektförderung und setzt seit ihrem Bestehen exklusiv auf das Peer Review Verfahren. Ohne sich um Evidenz für dessen Effizienz zu bemühen, und trotz der auch dort diskutierten offensichtlichen Mängel des gegenwärtigen Verfahrens. Warum testet die DFG nicht alternative Auswahlverfahren und Förderformate, auch wenn es nur weniger als ein Promille ihres Fördervolumens beträfe? Ganz einfach: Die DFG ist die zentrale Selbstverwaltungsorganisation der Wissenschaft in Deutschland, wir Wissenschaftler ‚sind' also die DFG. Und wir haben eine Schafsnatur!

Liebe Dein Null-Resultat nicht weniger als Dein statistisch signifikantestes...

LJ 5/2019

Saublöd. Ein Riesenaufwand. Knockout-Mauslinie herstellen lassen. In Background-Linie und dann in 10 Generationen von Littermates gekreuzt. Die vielen Genotypisierungen. Und dann erst die Experimente im Krankheitsmodell: Magnetresonanztomographie, Histologie, Verhaltensuntersuchungen. Am Ende: Kein Phänotyp! Die Knockout Maus scheint eine Maus wie jede andere. Selber Outcome. Kein Unterschied zum Wildtyp. Aber halt! Es muss natürlich heißen: Kein statistisch signifikanter Unterschied zum Wildtyp. Wir können also nicht mal sagen, dass Wildtyp gleich Knockout, sondern nur: Wenn da ein Unterschied wäre, ist er wohl kleiner als die detektierbare Effektgröße, abhängig von Stichprobengröße, Fehlerniveau (also Alpha und Beta) und der Varianz unserer Ergebnisse. Denn wir hatten die Serie von Experimenten gut vorbereitet: Die Fallzahl wurde A priori bestimmt, und so gewählt, dass wir einen Unterschied von einer Standardabweichung hätten finden können. Statistiker sagen dazu Cohen's d=1, und dies gilt als substantieller Effekt. Mehr Tiere (34!) waren nicht drin, alles zu aufwendig, und das Ganze würde auch zu lange dauern, für die Doktorandin, und auch den DFG - Antrag.

Was nun? Publizieren? Ist doch ein NULL-Resultat! Wie sieht das im Lebenslauf aus, außerdem, wen interessiert das schon, und welches reputierliche Journal würde das überhaupt publizieren?

So etwas, nicht notwendig mit Knockout-Mäusen, spielt sich vermutlich in vielen Laboren weltweit ab sehr häufig ab. Resultate von Experimenten, die sauber durchgeführt werden, aber nicht zur Ablehnung der NULL-Hypothese taugen, und deshalb in der Schublade verschwinden.

Ein Riesenfehler, denn wir sollten unsere NULL-Resultate lieben wie unsere hoch signifikanten! Aber ist das nicht Blödsinn? Ein Resultat, das uns einen Schritt näher zur Heilung der Alzheimer Erkrankung oder des Brustkrebses führt, ist doch viel toller als eine NULL? Wo wir mit der NULL nicht mal sagen können: ‚Da ist KEIN Effekt'!

Vergleichen wir das ganze doch mal mit den Entdeckungsreisen von Christoph Columbus. Amerika zu entdecken, das ist doch mal ein signifikantes Ergebnis, viel toller als auf dem Ozean rumzueiern und bloß endloses Wasser zu sehen. Aber halt: Um eine Seekarte zu erzeugen, mit der man sich aufmacht, fremde Länder zu entdecken, muss man wissen, wo keine Inseln und keine Untiefen sind. Ohne so eine Karte, die Seefahrer vor ihm angelegt hatten, wäre Columbus gar nicht losgesegelt. Im Übrigen wollte er ja den Seeweg nach Indien entdecken! Und so gesehen war er nicht erfolgreich und sein Ergebnis falsch positiv, denn er dachte bis zu seinem Tod, den Seeweg nach Indien entdeckt zu haben.

Nochmal zurück zum Experiment, das den Schlüssel zur Heilung der Alzheimer Erkrankung bringen könnte, verglichen mit einem Experiment ohne statistisch signifikantes Ergebnis. Mal ganz ehrlich: Wie viele dieser Welt-verändernden Resultate kann es überhaupt geben? Und wie wahrscheinlich ist es, dass es dann auch noch wir sind, die diesen Jackpot gewinnen? Nicht Null, aber gering. Ist es nicht beruhigend, wenigstens dazu beigetragen zu haben, dass die ‚Karte' der Biologie sowie das, was da so alles schiefgehen kann (wir nennen das Krankheitsmechanismen) etwas genauer geworden ist? Und wir nun alle ein bisschen besser ‚navigieren' können?

Außerdem überschätzen wir das statistisch ‚signifikante' Resultat in der Regel in seiner Signifikanz! Der p-Wert, der signifikante, kann uns nicht nämlich nicht sagen, wie wahrscheinlich es ist, dass wir mit unserer Hypothese richtig lagen. Genauso wenig wie uns das NULL-Resultat etwas darüber sagt, ob die Hypothese falsch war. Dies liegt daran, dass wir nicht wissen, wie wahrscheinlich die Hypothese war. Und an der meist zu geringen statistischen Power. Mit genügend groß angelegten Experimenten kann man nämlich jeden Vergleich statistisch signifikant werden lassen, oder umgekehrt, mit zu kleinen Fallzahlen die NULL-Hypothese nie ablehnen. Außerdem sind viele unserer Hypothesen (hoffentlich!) recht unwahrscheinlich. Denn sonst wären wir langweilige Wissenschaftler. Denn wenn die Hypothesen unwahrscheinlich sind, nimmt die Rate der falsch positiven Resultate rasant zu, trotz statistischer Signifikanz. (Wem das spanisch vorkommt, dem sei mein Beitrag ‚Wie originell sind eigentlich Ihre Hypothesen' Laborjournal 4/2017 empfohlen.)

Experimentelle Studien müssen also so angelegt sein, dass die Ergebnisse auch dann interessant, d.h. informativ sein müssen, auch wenn die NULL-Hypothese nicht abgelehnt wird. Der Fokus sollte dabei nicht die statistische Signifikanz des Resultats sein - stattdessen die Fragestellung und die dazu passende Methodik sowie Analyse. Nämlich nur diese kann der Wissenschaftler beeinflussen, die Ergebnisse nicht! Außer er schummelt.

Wir sind zu Recht stolz darauf, dass Wissenschaft sich selbst korrigiert, falsche Schlüsse also durch darauffolgende Experimente wieder ausgemerzt werden. Das funktioniert

aber nicht richtig, wenn Resultate, welche nicht die von uns erwünschten Ergebnisse erbringen, in der Schublade verschwinden (‚File drawer Effekt').

Wann aber sind NULL-Resultate informativ: Wenn Sie nach den Regeln der Wissenschaft geplant und durchgeführt werden und ausreichend statistische Power haben. Wenn sie zum gegenwärtigen Stand der Forschung etwas beitragen. Wenn sie potentiell nützlich sind für die Community der Forscher. Wenn sie uns vor Irrwegen oder unnötigen Experimenten abhalten, oder wir die Ergebnisse in Meta-Analysen aggregieren können.

NULL-Resultate haben eine Vielzahl von tollen Eigenschaften: Sie sind robuster als statistisch signifikante. Mit anderen Worten: So komisch das auch klingt, ein NULL-Resultat ist mit viel höherer Wahrscheinlichkeit richtig, als ein statistisch signifikantes. NULL-Resultate können unsere Kollegen davon abhalten, unnötig Sackgassen zu betreten. NULL-Resultate, wenn veröffentlicht, machen Evidenzsynthesen in Form von Meta-Analysen erst aussagekräftig. NULL-Resultate erzeugen einen ‚Korridor' von Wissen, sie erzeugen Wegmarken und Grenzen, in denen statistisch signifikante Ergebnisse erst ihre volle Kraft entfalten.

Und was ist von dem Argument zu halten, dass sie sich schlechter veröffentlichen lassen? Das mag vor einer Reihe von Jahren tatsächlich so gewesen sein. Richtig ist sicher, dass sie sich tatsächlich kaum in Top-Journalen veröffentlichen lassen. Außer es handelt sich um ein NULL-Resultat, das an einem Dogma oder Textbuch-Wissen kratzt und aus einem prominenten Labor stammt. Aber das Wissen um die Nützlichkeit von NULL-Resultaten, und der Schaden, den das Schielen auf und Selektieren von statistisch signifikanten Resultaten erzeugt hat (‚Replikationskrise'), hat bei vielen Journalen zu einem Paradigmenwechsel geführt. Und damit neue Journale aufs Tapet gebracht. Etablierte Journal haben mittlerweile ‚NULL and negative result sections'. Und PLOS One, Peer J, oder F1000Research publizieren Studien ganz unabhängig von deren statistischem Ausgang. Fragestellung, Methodik und Analyse müssen stimmig sein, dann wird veröffentlicht. Das Webtool FIDDLE des QUEST - Centers kann Ihnen helfen, den richtigen Veröffentlichungsweg für NULL-Resultate zu finden (siehe Linksammlung bei https:/dirnagl.com/lj).

Sind NULL-Resultate schlecht für die Karriere, kontaminieren sie den Lebenslauf? Die Charité z.B. belohnt die Veröffentlichung von NULL-Resultaten mit zusätzlichen Forschungsmitteln. Auch frägt sie Bewerber auf Professuren danach, ob sie schon mal NULL-Resultate veröffentlicht haben, und dies dann auch weiterhin vorhaben. Ein zarter Anfang, aber immerhin ein Hinweis, dass sich auch im Karrieresystem ganz langsam was ändert.

Zum Schluss mein Kalenderspruch für den Monat Mai: ‚In der Wissenschaft ist ein Experiment nur dann gescheitert, wenn es zu keinem Ergebnis geführt hat'.

Schweine wollt ihr ewig leben?

LJ 6/2019

Aus aktuellem Anlass widmet sich der Narr diesmal den letzten Dingen. Am 17. April titelte Nature anlässlich eines Artikels, bei dem es um wiederbelebte Schweinehirne ging: ‚Turning back time'! Nicht nur die Grundfeste der Biologie, gar der Physik schienen aus den Fugen geraten. Wenn schon Nature den Verstand verliert, gibt's im Boulevard natürlich kein Halten mehr. Bild wusste, obzwar in dem Artikel gar kein Alzheimer vorkommt: ‚Durchbruch in der Alzheimer Forschung: US-Forscher reaktivieren totes Schweine-Hirn'. Der Schweizer Rundfunk, sonst eher konservativ in der Berichterstattung, dichtete: ‚Frankenschwein lebt'. Wegen der Nähe zum Osterfest sprach die Süddeutsche davon, dass Nature ein eigenes ‚Auferstehungsfest' feiere. Weltweit schallte aus Zeitungen, Fernsehen und Internet die Frohbotschaft, dass es Wissenschaftlern gelungen war, Schweinehirne wieder lebendig werden zu lassen. Der Clou dabei: Man hatte die Hirne vom Schlachthof geholt. Stimmungsmäßig oszillierte die Berichterstattung zwischen Grusel und Ekstase, denn es stand die Frage im Raum: Ist der Tot umkehrbar?

Die Studie von Forschern der US-amerikanischen Yale Universität hätte natürlich nicht nur für Metzger und Intensivmediziner große Bedeutung: Man denke nur an die ethischen Implikationen, und das mitten in der Organspende Diskussion. Wenn man Schweinhirne Stunden nach dem Tod wiederbeleben kann, können wir dann überhaupt noch von Hirntod sprechen? Und wenn man sich in dieser Frage nicht sicher sein kann, nach welchen Kriterien dürfen dann überhaupt noch Organe zur Transplantation entnommen werden? Deshalb hatte Nature dem Artikel gleich zwei mehrseitige Kommentare von prominenten Ethikern zur Seite gestellt und damit Öl ins selbstgelegte Feuer gegossen.

Bei soviel Aufregung empfiehl sich das Motto: Außergewöhnliche Behauptungen erfordern außergewöhnliche Evidenz! Fragen wir also, was die Wissenschaftler da berichtet hatten in Nature, und was eigentlich davon zu halten ist. Folgendes war geschehen: Vrselja und Kollegen hatten Schweinehirne vom Schlachthof vier Stunden nach Tötung der Tiere an eine selbstgebastelte Maschine angeschlossen. Diese Maschine, effektheischend ‚BrainEx' genannt, durchströmte das Hirngewebe für maximal 6 Stunden mit einer Lösung, welche einen künstlichen Sauerstoffträger und eine Reihe von zytoprotektiven Substanzen enthielt. In den so behandelten Hirnen konnten die Forscher die Erhaltung einiger zellulärer Strukturen, sowie rudimentäre Zellfunktionen beobachten. Dies bis maximal 10 Stunden post mortem und im Vergleich zu nicht an BrainEx angeschlossene Gehirne. Zum Beispiel reagierten die Hirngefässe auf topisch applizierte vasoaktive Substanzen. Gliazellen ließen sich mit bakteriellen Zellwandbestandteilen stimulieren.

Die strukturelle Morphologie von Neuronen, z.B. im Hippokampus, war gut erhalten. Ableitungen von einzelnen Neuronen zeigten relativ normale elektrophysiologische Eigenschaften. Synchronisierte Netzwerkaktivität war nicht vorhanden, denn ein kortikales EEG ließ sich nicht ableiten.

Das alles ist durchaus bemerkenswert, und war auch methodisch recht sauber gemacht und präsentiert. Trotzdem hat die Arbeit zwei entscheidende Mängel: Zum einen präsentiert sie nichts grundsätzlich Neues. Und die Behauptung, dass hier ein relevanter Schritt in Richtung Wiederherstellung von Hirnfunktion nach längerer Unterbrechung der Hirndurchblutung getan wurde, ist nicht nicht nur übertrieben, sondern schlichtweg falsch.

Warum ist es nichts Neues, wenn Gehirne nach Unterbrechung der Blutzufuhr und erloschenem EEG wieder Funktionszeichen entwickeln? Schon 1970 hatte Konstantin Alexander Hossmann gezeigt, dass Katzenhirne nach einer Stunde komplettem Durchblutungsstop wieder ein EEG entwickeln und evozierte Potentiale abgeleitet werden können. Die Arbeit, damals in Science publiziert, stimulierte eine ganze Generation von Neurowissenschaftlern, sich auf die Suche nach neuroprotektiven Substanzen zu machen. Bis dato war man davon ausgegangen, dass praktisch unmittelbar mit Durchblutungsstop Hirngewebe dem Untergang geweiht ist, jede therapeutische Maßnahme danach damit zwecklos wäre. Frecher Weise wird die Hossmann Arbeit, ohne Kontext und begraben in einer Liste von anderen Zitaten, von Vrjselja en passant erwähnt.

Neu ist der Befund vom ‚wiederbelebten Hirn' auch deshalb nicht, weil wir schon lange und mit Sicherheit wissen, dass sich ein EEG wieder einstellen kann auch wenn es einmal weg war – was bei den Schweinhirnen ja gar nicht passierte. Und wir wissen das sogar vom Menschen. Patienten die in tiefer Hypothermie und induziertem Kreislaufstillstand operiert werden, haben während der chirurgischen Ausschaltung großer Gefässmißbildungen im Gehirn keine messbare hirnelektrische Aktivität bzw. keine evozierten Potentiale mehr, dies durchaus auch für eine volle Stunde. Dennoch gehen glücklicherweise viele dieser Patienten nach dem Eingriff wieder ihrem Beruf nach. Analoges kann für Patienten nach geglückter Reanimation bei Herzstillstand gelten. Es sind Stillstandzeiten von bis zu 20 Minuten dokumentiert, nach denen die Patienten reanimiert wurden und danach weitgehend ‚normal' weiterlebten, allenfalls mit leichten neuropsychologischen Defiziten. Diese betreffen vor allem Störungen des Gedächtnisses, und hat mit Zelluntergängen im Hippokampus zu tun, doch davon gleich weiter unten.

Des Weiteren gibt eine Vielzahl von Studien, insbesondere aus den 80er und 90er Jahren des letzten Jahrhunderts, zur ‚isolierten Hirnperfusion' als Modell zur Untersuchung von Gehirnfunktionen. Dieses sogenannte ‚isolated brain' wurde in verschiedenen Spezies eingesetzt, darunter Schwein aber vor allem Ratte. Wie im Nature Artikel wurden die Hirne mit künstlichen Lösungen perfundiert, pulsatil oder nicht pulsatil, immer mit synthetischen Sauerstoffträgern, und häufig unter Zusatz hirnprotektiver Substanzen. So recht durchgesetzt haben sich diese Modelle ganz offensichtlich gibt, obwohl auch rezente Publikationen davon berichten.

Nun aber zu des Pudels Kern: Vrselja und Kollegen zeigen ja gar keine dauerhafte Restauration von Hirnfunktion! Die Erhaltung und Stimulierbarkeit von Gefäßen und Glialzellen verwundert niemand, zumindest nicht diejenigen, welche um die Robustheit dieser Zellen wissen. Vor allem aus der Zellkultur ist bekannt, dass diese Zellen sehr resistent sind gegen Sauerstoffmangel. Endothelzellen und Gliazellen können in Zellkultur mehr als 24 h ohne jeden Sauerstoff und Glukose unbeschädigt überstehen. Dass einzelne Neuronen nach längerdauernder Hypoxie noch elektrische Aktivität zeigten ist auch nicht verwunderlich, und ebenfalls schon lange bekannt. Was den Autoren und den

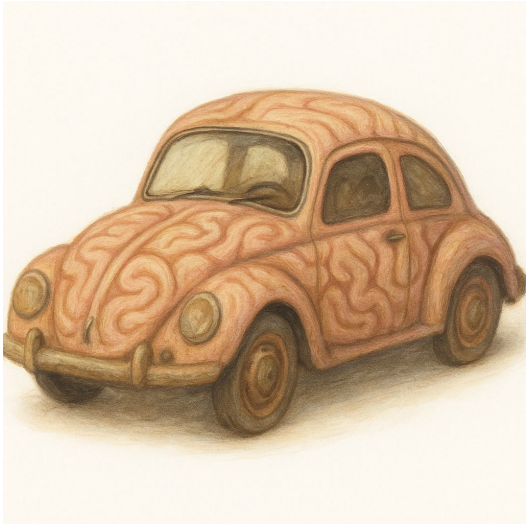
Reviewern allerdings offensichtlich nicht bekannt war, ist dass die meisten dieser Neuronen nach etwa einem Tag beginnen abzusterben, und nach drei Tagen tot sind. Nach 6 Stunden, wie bei Vrselja et al., sehen sie selbst ultrastrukturell normal aus. Dies Phänomen nennt man ‚delayed neuronal vulnerability', die Erstbeschreiber nannten es ‚maturation phenomenon'. Das war in den 60er und 70er Jahren des vorigen Jahrhunderts. Dieser verzögerte neuronale Zelltod befällt prominenter Weise die CA1 Region des Hippocampus, aber auch andere Neuronentypen, z.B. auch in Schichten des Neokortex. Über 30 Jahre haben sich ganze Hundertschaften von Neurowissenschaftlern die Zähne daran ausgebissen, dies Phänomen zu erklären. Bis man im neuen Jahrtausend ermattet aufgab und sich anderen Dingen zuwendete. Oder, wie hier geschehen, in Unkenntnis der Literatur von vorne anfing. Groundhog day!

Das ‚maturation phenomenon' ist eine der Gründe, warum ein ‚reperfundiertes Hirn' nach 4 Stunden ohne Hirndurchblutung nicht sinnvoll, das heisst nachhaltig wiederbelebt werden kann. Platt ausgedrückt: Weil die Zellen verzögert sterben, wenn man nur ein paar Stunden wartet, kriegt man das halt nicht mit.

Nun kann man aber aus der Affäre einiges lernen. Zum einen, dass ganz basales und relevantes Wissen sehr kurzlebig sein kann. Aus den Augen, aus dem Sinn. Was nur noch grauhaarige Neuropathologen kennen, in Pubmed älter als 10 Jahre ist, und nicht im Lehrbuch steht, fällt häufig der Hyperspezialisierung und Fetischisierung des frisch Publizierten zum Opfer. Sprechen Sie mit den Emeriti in deren Abstellkammer gegenüber! Zum anderen lernen wir, ein ums andere mal, dass der vielgerühmte Peer Review Prozess, in kritischen Momenten oft nicht funktioniert. Auch und gerade in Top-Journalen. Die wie Nature in ihrer Jagd nach spektakulären, öffentlichkeitswirksamen Stories bereit sind, akademische Grundregeln zu ignorieren. Dass die Tagespresse sich bei sowas nicht zweimal bitten lässt und den Verstand komplett verliert, ist da geradezu entschuldbar. Ein Hoffnungsschimmer zum Schluss: Die Süddeutsche und die FAZ (von der ich den Titel dieses Beitrags geklaut habe) haben sich nicht anstecken lassen, und sehr vernünftig über den Artikel und seine Schwächen berichtet.

Wenn Autobauer Hirne hacken

LJ 9/2019



Hektisch geschnittene Filmclips von jungdynamischen Wissenschaftlern in biomedizinischen Hightech-Laboren, Nerds am Lötkolben und Oszilloskop, Achterbahnfahrten durch Animationen eines Gewirrs aus Nervenzellen. Und dazwischen verkündet der Auto- und Raketenbauer Elon Musk in messianischer Pose seine neueste Vision: Die Symbiose des menschlichen Gehirns mit künstlicher Intelligenz (KI)! Realisiert werden wird dieser die Menschheit rettende Plan durch das revolutionäre Brain Machine Interface (BMI) seiner Firma Neuralink. Ich hätte die Narreteien meines Kollegen getrost ignoriert, wenn diese nicht in den Räumen der California Academy of

Sciences zur Aufführung gekommen wären, und das Video hiervon auf der ganzen Welt eine ungeheure Medienhype ausgelöst hätte. Einhelliges Urteil in der Presse und im Netz: ‚Ein typischer Musk, den Mund wieder etwas vollgenommen, aber wenn der sowas ankündigt, wird schon was dran sein. Aber ist das nicht auch gefährlich, brauchen wir eine neue Ethik?'

Nur: Rein gar nichts ist dran! Nicht weil es mit dem BMI einfach noch ein bisserl dauert bis wir damit unsere Gedanken runter und neue Inhalte ins Gehirn hochladen können. Wir dadurch endlich hyperintelligent werden. Auch nicht, weil Musk in seiner Präsentation maßlos übertreibt, was er mit seinem BMI schon alles erreicht hätte. Was er selbstredend tut. Nein, das wird deshalb nichts mit der Gehirn-KI Symbiose, weil die Musk'sche Vision auf gleich drei fundamentalen Fehlern gleichzeitig beruht. Einer betrifft das Prinzip von BMI, ein zweiter ist der vom Gehirn als Computer, und der dritte Fehler ist eine falsche Vorstellung davon, was KI ist. Diese Fehler sind leider sehr populär, auch bei Wissenschaftlern. Umso mehr lohnt sich hier ein närrischer Blick.

Herr Musk schreibt ein Paper: Der über Twitter gestreute Internetauftritt strotzt vor bunten Bildern und spektakulären Ankündigungen. Sogar ein Neurochirurg steht auf der Bühne, in voller OP-Montur. Inhaltlich gibt die Präsentation aber leider wenig her, Herr Musk verweist uns bezüglich technischer Details deshalb auf einen von ihm als ‚single author' publizierten ‚wissenschaftlichen Artikel' in BioRxiv. Schauen wir uns den also erstmal an. Es werden darin Komponenten eines BMI beschrieben, also chirurgisch ins Gehirn implantierter Elektroden welche elektrische Hirnaktivität ableiten. Nach Training kann das Gehirn hierüber mit einem Computer kommunizieren und so eine ‚Maschine' steuern. Dies könnte ein Roboterarm sein, oder das Bewegen eines Mauscursors. Der Artikel beschreibt oberflächlich einige Elemente eines nicht-funktionalen BMI-Prototypen: Elektroden zur Ableitung von Hirnaktivität, ein Chip zur Verarbeitung und Übertragung der Signale, und ein OP-Roboter zum Einsetzen der Elektroden in Gehirn. Es fehlt allerdings alles, was nach der Ableitung der Signale noch benötigt wird, also etwas das gesteuert wird. Herr Musk beschreibt hier also kein BMI, sondern nur Teile davon.

Bemerkenswert ist zunächst einmal seine Alleinautorschaft, obwohl völlig klar ist, dass er den Artikel gar nicht verfasst haben kann. Das ist zwar das geringste Problem des Artikels, aber festzuhalten bleibt (auch in Blick auf den von mir geschätzten Preprint-Server BioRxiv), dass dies ein klarer Verstoß gegen die international akzeptierte Publikationsethik ist. Sehr bedenklich auch, allerdings in USA lege artis aber in Deutschland zum Glück undenkbar, dass die im Artikel nur angedeuteten Tierexperimente durch ein Firmen-internes Review board ‚genehmigt' wurden. Man hat also selber entschieden, dass die Experimente ethisch vertretbar sind. Zudem weckt der Artikel unbegründete Hoffnungen bei verzweifelten Patienten Querschnittslähmung, all dies ist unethisch.

Im Westen nichts Neues: Der grundsätzliche Aufbau des BMI und das Funktionsprinzip der im Neuralink Paper beschriebenen Komponenten ist nicht neu. Verschiedene Gruppen weltweit haben sie bereits in ähnlicher Form entwickelt und bei ausgewählten Patienten mit Rückenmarksverletzungen eingesetzt worden. Diese konnten damit, nach langem Training, wieder sehr einfache Funktionen ausführen, wie z.B. das Greifen einer Tasse. Der Artikel beschreibt allerdings eine Reihe von technischen Verbesserungen, welche potentiell die Funktionalität von BMIs verbessern könnten. Dazu gehören sehr dünne Elektroden, die Möglichkeit von einigen Tausend (statt von einigen Hundert) Elektroden ableiten zu können, und der OP-Roboter zur Implantation. Trotzdem bleibt es bei der puren Versprechung, dass all dies ein BMI verbessern könne, gezeigt wird es nicht. Häufig finden sich stattdessen Formulierungen wie: ‚It is plausible to imagine….'.

Im Artikel angedeutete Funktionen des BMI, z.B. dass man damit das Hirn stimulieren könne, oder wie von Herrn Musk behauptet, eine Verschmelzung des menschlichen Gehirns mit KI möglich wird, bleiben durch nichts belegte Spekulation. Der Artikel kommt im Gewand einer wissenschaftlichen Publikation daher, ist aber nichts als ein oberflächlicher Werbeprospekt der Firma Neuralink.

Ein BMI liest keinen Gehirncode: Die mit dem Artikel und der Präsentation von Herrn Musk bewusst generierte Hype suggeriert Fähigkeiten eines BMI, welche von keinem ernstzunehmenden BMI-Forscher so für möglich gehalten werden. ‚Spikes' von Neuronen, auch wenn diese von mehreren Tausend Orten im Gehirn kommen, erlauben nicht das Auslesen von Gedanken, Vorstellungen und Gefühlen. Es ist nämlich umgekehrt: Das Prinzip eines BMI zur Steuerung einer Maschine durch Hirnaktivität besteht darin, das Hirn darauf zu trainieren, willentlich elektrische Aktivität in großen Ensembles von Nervenzellen zu generieren. Und zwar solche Aktivitäten, die vorher an dieser Stelle so noch gar nicht aufgetreten waren. Um damit dann eine spezifische, für das Gehirn bisher fremde Reaktion der Maschine auszulösen. Das Gehirn ist so plastisch, dass es so etwas lernen kann. Deshalb dauert es auch lange, bis selbst einfachste Steuer-Funktionen (Cursor rauf, Cursor runter) nur halbwegs zuverlässig klappen. Bei einem nicht geringen Anteil der Patienten, bei denen verschiedene Forschergruppen so etwas versucht haben, funktionierte es überhaupt nicht. Ob, wie von Herrn Musk einfach mal so behauptet, eine Erhöhung der Anzahl von Elektroden die Fähigkeit der Steuerung wesentlich verbessern kann, ist unklar. Forscher im Feld zweifeln dies durchaus auch an.

Das Gehirn ist kein Computer: Das Missverständnis von Herrn Musk bezüglich der Funktionsweise seines BMI beruht ganz wesentlich auf dem Glauben, dass das Gehirn wie ein Computer funktioniert. Diese Computer-Metapher vom Gehirn ist weit verbreitet, auch das milliardenschwere Human Brain Project beruht darauf. Dadurch wird die Sache aber nicht richtiger, nur noch teurer. Ein Computer ist ein Automat, der mittels programmierter Instruktionen Eingaben in strikt determinierte Ausgaben verwandelt. Von der Nutzer - (Eingabe) ebene bis hinunter zum binären Maschinencode werden mit den Instruktionen Symbole manipuliert. Diese sind total abstrakt, sie haben keinen inhaltlichen oder physischen Bezug zur Leistung, die der Computer für den Anwender erbringt. Nur unser Gehirn ordnet den Symbolen (Zeichen) Inhalte zu. Die Vorstellung vom Programm-Code im Gehirn ist deshalb unhaltbar, weil ein Code vom Wesen her nichts anderes ist als eine Abbildungsvorschrift. Ein Zeichen wird einem anderen zugeordnet: symbolische Repräsentation. Eine solche Zuordnung via Zeichen im Gehirn könnte Bewusstsein und Denken, die sich ja mit Inhalten befassen, nicht erklären, sondern würde das Problem nur verschieben: Von welchen Inhalten sollten denn die Zeichen des Codes ihre Bedeutung erhalten? Das Gehirn hat Programmierung und Codes auch gar nicht nötig. Gefühle, Gedanken, Absichten usw. ***sind*** die koordinierte Aktivität von Milliarden von Nervenzellen und Fantasillionen von Verbindungen zwischen ihnen, man kann auch sagen, Kognition ist ‚verkörpert' (embodied). Der Gedanke an einen Baum ***ist*** die elektrische Aktivität und neuronale Konnektivität wie sie beim Betrachten dieses Baumes auftritt. Die Erinnerung an diesen Baum ist die Wiederherstellung dieses elektrochemischen Zustandes.

Dabei kann das Gehirn durchaus mit Codes umgehen. Nicht nur extern, beim Programmieren. Auch intern, beim Sprechen und Schreiben. Sprache ist nämlich eine Art Code, also symbolische Repräsentation. Aber Sprache braucht es nicht für Fühlen, Denken, Handeln, sie ist nur ein Mittel dafür. Da Kognition sich keines Codes oder Programmes bedient, gibt es auch nichts auszulesen, oder einzuspielen ins Gehirn. Man könnte zwar versuchen, z.B. beim Blick auf einen Baum, die Aktivität jeder Einzelnen der 80 Milliarden Nervenzellen gleichzeitig zu messen. Und dazu den Zustand der Hunderte von

Trillionen Verbindungen zwischen ihnen. Aber dann wäre man immer noch nicht weiter. Denn dann hätte man zwar ein Abbild des elektrischen Gewitters *dieses* Gehirns beim Blick auf den Baum. Aber die Nervenzellen eines anderen Menschen erzeugen andere Verbindungen und andere Aktivitäten beim Blick auf denselben Baum. Auch weil verschiedene Gehirne eine über viele Jahre zurückreichende unterschiedliche Geschichte haben, die wiederum zu dieser spezifischen Konnektivität und Aktivität beim Blick auf den Baum beigetragen hat. Diese Geschichte müsste man kennen, um aus dem Gewitter Sinn zu machen, also den Inhalt Baum dekodieren zu können. Also, Herr Musk, da können sie das Hirn mit Elektroden spicken bis nichts mehr davon übrig ist – down und uploads von irgendwas wird es nicht geben. Insbesondere auch keine Symbiose mit KI.

Künstliche Intelligenz ist gar nicht intelligent: Wenn es ihn nicht seit mehr als 60 Jahren gäbe, könnte ‚Künstliche Intelligenz' ein genialer Begriff aus der Marketing-Schatulle von Herrn Musk sein. Es ist fast Orwell'scher Neusprech, denn KI in seiner jetzt allenthalben praktisch eingesetzten Verlaufsform hat gar nichts mit Intelligenz zu tun. Im Gegenteil, KI wird (wie andere Computersoftware auch) dort eingesetzt, wo es darum geht, aufwendige Tätigkeiten, die *keine* Intelligenz erfordern, für den Menschen zu erledigen. Das Erkennen von Katzen oder Tumoren auf digitalen Bildern. Das Übersetzen von Sprachen. Die Vorhersage von Pizzabestellungen im Feierabendgeschäft. Das autonome Fahren eines Autos. All dem hat menschliche Intelligenz die Inhalte und Kontext gegeben, Regeln geschaffen, eine Aufgabenstellung für die KI abgeleitet. Und die KI daran trainiert. Diese kann dann in dem unterschiedlichen Datenmaterial Muster erkennen, ohne zu wissen worum es geht, was der Inhalt der Daten ist, die es zu analysieren gilt, und welche Aufgabe überhaupt gelöst werden muss. Wie jede Computersoftware ist KI also völlig ignorant gegenüber den Inhalten der durch sie erledigten Aufgaben. Wenn wir behaupteten, die Katzen auf dem Foto seien Kanarienvögel, würde die KI eben ‚Kanarienvögel' finden. Die Tumoren könnten auch Würste sein. Die Verwechslung mit Intelligenz wird aber auch dadurch befördert, dass KI häufig auf Tätigkeiten angewendet wird, die für sich durchaus Intelligenz erfordern, Sprache eben, oder Tumorpathologie. Auch die Bezeichnung ‚Maschinelles Lernen', welche die Sache schon viel besser beschreibt, kann dem Missverständnis Vorschub leisten. Ist Lernen nicht eine intelligente Tätigkeit?

Durch Training mit Datensätzen, in denen Ein– und Ausgabe vorgegeben sind, erzeugt KI ein statistisches Datenmodell. Wenn alles gut geht, erzeugt das Datenmodell dann auch auf beliebige Eingaben zuverlässige Ausgaben. Die KI hat ‚gelernt', nur eben völlig begriffslos, ohne einen Funken Intelligenz. Aber sind es nicht ‚neuronale Netzwerke', die hier am Wirken sind? Also etwa doch ein künstliches Gehirn? Wieder führt uns eine Analogie in die Irre. Weil das ‚neuronale Netz' der Software einige strukturelle Gemeinsamkeiten mit der Verschaltung von Nervenzellen des Gehirns hat (viele Zellen sind in Schichten miteinander verbunden, es existieren Schwellenwerte und Verstärkungsfaktoren für die Weiterleitung eines Signales), funktioniert es noch lange nicht wie ein Gehirn. Und wenn, wir würden es gar nicht sagen können. Denn wir wissen ja überhaupt nicht, wie ein Gehirn funktioniert. Wie es Bewusstsein, Gefühle, und Gedächtnis produziert, lernt, verallgemeinert, und sich einen Begriff von der Welt macht. Wenn man für die Beschreibung der strukturellen Elemente eines künstlichen neuronalen Netzwerkes nicht suggestive Begriffe wie Neuron oder Synapse sondern Schwellenfunktion, Gewichte, Bias, Gradienten, verdeckte Schichten, Rückpropagation etc. benutzt, wird schon deutlicher, dass wir uns hier nicht im Gehirn befinden. Auf die Spitze getrieben wird die Analogismen-Logik übrigens noch durch das zirkuläre Projekt mancher Kollegen, mittels neuronaler Netzwerke im Computer rausfinden zu wollen, wie das Gehirn funktioniert: Zuerst bastelt man ein Computer Programm, das wie das Gehirn funktioniert, und dann zeigt einem das Programm, wie das Hirn funktioniert?

Der Werbeauftritt von Herrn Musk erinnert damit sehr an einen TED Talk eines gewissen Henry Markram im Juli 2009, also vor ziemlich genau 10 Jahren. Herr Markram ist der geistige Vater des Human Brain Project, das von der EU mit über einer Milliarde Euro gefördert wird. Er hat darin angekündigt, dass das Human Brain Project die Simulation des menschlichen Gehirns im Computer ermöglichen wird. Hierdurch würden wir dann Wahrnehmung und Denken, vielleicht sogar unsere physikalische Realität verstehen. Er schloss damit, dass in 10 Jahren (also heute) ein Hologramm seinen TED Talk halten wird. Passiert ist in den 10 Jahren gar nichts dergleichen, außer dass viel Geld ausgegeben wurde.

Brüder, zur Sonne, dem p-Wert ein Ende, Brüder zum Licht empor!

LJ 10/2019

„Die Wissenschaft wehrt sich gegen die p-Wert Tyrannei!“: So zumindest verkündete es vor kurzem die Financial Times. Denn international ist die Aufregung groß. Mehr als 800 Forscher, darunter viele prominente Biostatistiker haben dazu aufgerufen, sich gegen den p-Wert zu erheben. Und dies ist nur der Höhepunkt eines Aufstands, der schon letztes Jahr begonnen hatte. Eine Gruppe von Wissenschaftlern forderte damals, dass wir die Schwelle für ‚statistische Signifikanz‘ ganz neu definieren sollten. Von derzeit meist 0.05 auf 0.005. Insbesondere wenn Wissenschaftler damit behaupten wollen, etwas entdeckt zu haben. Für viele Forscher und Experten ging diese Forderung allerdings nicht weit genug, sie fordern daher nun, statistische Signifikanz gleich ganz zu beseitigen, statt nur neu zu definieren. Wieso die Aufregung? Worum geht es überhaupt? Und ist das alles wirklich neu?

Wir erinnern uns: Im Jahr 2012 gewannen Craig Bennett und Kollegen den Ig-Nobelpreis für Neurowissenschaften mit einer bemerkenswerten Studie. Sie positionierten einen toten Lachs, den sie in einem lokalen Supermarkt gekauft hatten, in einem Kernspintomographen. Dort zeigten sie dem Fisch, der eigentlich für’s Abendessen vorgesehen war, eine Reihe von Bildern. Diese zeigten Menschen in sozialen Situationen mit einer bestimmten emotionalen Aufladung, wie z.B. einem Streit, oder ein Kuss. Der Lachs musste dann entscheiden, welche Gefühle die Abgebildeten wohl durchlebt haben mussten. Die mittels funktioneller Magnetresonanztomographie durchgeführte Hirnbildgebung zeigte dabei signifikante Veränderungen der Hirnoxygenierung im toten Lachs, die auf eine Aufgaben-spezifische neuronale Verarbeitung im Fischgehirn hinwiesen.

Wie aber können ‚post-mortem neuronale Korrelate von Interspezies-Einfühlsamkeit im Lachs‘ erklärt werden, wie es der Titel des Artikels neurowissenschaftlich formuliert? Ganz einfach: Weil sich die Auswertung auf statistische Standard-Signifikanzschwellen stützte, und Mehrfachvergleiche nicht angemessen kontrollierte. Der Clou dabei: Die

Autoren zeigten in der Arbeit auch, dass in 60-70 % der damals veröffentlichten funktionellen Neuroimaging-Studien ähnlich ausgewertet wurde! Und damit die Ergebnisse eines Großteils der kognitiven Neurowissenschaften in Frage gestellt waren. Aber finden sich solche toten Fische vielleicht auch im Becken anderer Disziplinen, welche ebenfalls stark auf multiple Testungen zurückgreifen? Etwa in Genexpressions und – Assoziationsstudien?

In der Tat, auch die Genetik erkannte vor einigen Jahren, und ganz ohne Ig-Nobelpreis, dass sie ein Riesenproblem hatte. Und deshalb in einem Meer von falsch positiven Resultaten ertrank. Ein Großteil der bis dahin gefundenen differentiell exprimierten Gene und Genassoziationen waren nämlich falsch positive Befunde. Zum Glück haben die Genetiker und auch die funktionellen Hirnbildgeber mittlerweile ihre Lektion gelernt. Genetische oder Bildgebungs-Datensätze sind heute kaum noch ohne post-hoc-Korrektur für multiple Vergleiche zu veröffentlichen. Außerdem werden, zumindest in der Genetik, Validierungen mit unabhängigen Datensätzen gefordert, bevor Assoziationen akzeptiert werden. Das ist doch mal eine gute Nachricht, dass ganze Forschungsbereiche vor ihrer Haustüre gekehrt haben! Die schlechte Nachricht ist jedoch, dass andernorts unzureichende Korrektur für Mehrfachtests, laxe Schwellenwerte für Typ-I-Fehler (z.B. 5%), geringe statistische Power, sowie fehlende Validierung immer noch die Norm ist.

Mindestens so problematisch sind jedoch allgemein verbreitete falsche Vorstellungen über das, was p-Wert ist, und was das Label ‚statistisch signifikant' bedeutet. So glauben viele Forscher, dass p die Wahrscheinlichkeit ist, dass die Null-Hypothese wahr ist. Und folglich 1-p die Wahrscheinlichkeit, dass die alternative Hypothese (also ihre eigene Hypothese) richtig ist. Oder umgangssprachlich ausgedrückt: „Bei einem Alpha von 5 % laufe ich Gefahr, dass 5 % meiner Hypothesen trotz Signifikanz doch nicht richtig sind". Also eine Verwechslung mit der falsch-positiven Rate. Ein weiteres häufiges Missverständnis ist, dass der p-Wert mit der theoretischen oder praktischen Relevanz des Befunds korrelieren würde. Sowie der schwerwiegende Irrtum, dass die nicht-Ablehnung der Null-Hypothese ($p > 0.05$) belegt, dass diese richtig wäre, also kein Effekt vorliegt. Und so weiter…

Aber was ist denn dann der p-Wert, und was kann er uns über unsere Ergebnisse sagen? Wenn wir die Analyse viele Male wiederholen würden, und jedes Mal neue Daten generieren, und wenn die Nullhypothese wirklich wahr ist, würden wir sie bei $p = 0.05$ in nur 5 % (fälschlicherweise) ablehnen. Mit anderen Worten: Der p-Wert stellt die Wahrscheinlichkeit dar, Daten so extrem (oder noch extremer) als die unserer Ergebnisse zu erhalten, wenn die Nullhypothese wahr ist. Aber klingen diese Definitionen nicht vereinbar mit der Interpretation des p-Wertes als falsch-positiven Rate? Kucken wir deshalb genauer hin: In den obigen Lehrbuch-Definitionen wird die Wahrscheinlichkeit auf die Daten bezogen. Ein Irrtum ist es, sie auf die Erklärung (d.h. die Hypothese) anzuwenden. Außerdem wissen wir ja gar nicht, ob die Null wahr ist oder nicht. Und dann gibt es da noch das Problem der Wahrscheinlichkeit unserer Hypothese, die sogenannte „base rate". Und die statistische Power, d.h. die Wahrscheinlichkeit, einen Effekt zu erkennen, wenn es denn einen gibt. Dass base rate und Power für die Interpretation des p-Werts entscheidend sind, ist vielen Kollegen nicht bekannt. Und genau da liegt der sprichwörtliche Hase im Pfeffer!

Die Frage, die wir doch eigentlich gerne beantworten möchten, ist folgende: Wenn wir einen „signifikanten" p-Wert nach einem gut durchgeführten Experiment erhalten haben, mit welcher Wahrscheinlichkeit ist unser Ergebnis dann falsch positiv? Leider ist der p-Wert nur ein Teil der Gleichung, die wir lösen müssten, denn die falsch-positiven Rate hängt vom Typ I Fehler (Alpha), Typ II Fehler (bzw. Power), sowie der

Wahrscheinlichkeit der Hypothese ab, welche wir testen. Je unwahrscheinlicher nämlich unsere Hypothese und je niedriger die statistische Power, desto wahrscheinlicher ist es, dass wir ein falsch positives Ergebnis vor uns haben. Trotz eines signifikanten p-Wertes. Zur Verdeutlichung: Bei einem Typ I Fehler Niveau von 0.05, einer Power von 80 % und einer 10 % Wahrscheinlichkeit, dass die alternative Hypothese wahr ist (d.h. 10 % base rate), sind fast 40 % der statistisch signifikanten Ergebnisse falsch positiv! Und aufgemerkt: In vielen Bereichen der Biomedizin, insbesondere in der präklinischen Forschung, liegt die statistische Power oft weit unter 80%, eher bei 50 % oder darunter. Und wer sich mit explorativer Forschung in wissenschaftliches Neuland vorwagt (tun wir das nicht alle?), muss wohl auch mit base rates unter 10 % rechnen. Sonst wäre man doch nur unorigineller Mainstream-Wissenschaftler, der beforscht, was auf der Hand liegt, oder gar schon weiß! Die Kombination aus niedriger Power, laxem Typ I Fehler Niveau (Alpha = 0.05), niedriger base rate, und hoher Prävalenz von Bias (durch geringe interne Validität, z.B. durch fehlende Verblindung oder Randomisierung) erklärt, warum John Ioannidis 2005 ungestraft und seither unwiderlegt behaupten konnte, dass die meisten veröffentlichten Forschungsergebnisse falsch sein müssen.

Aber bei Alpha=0.05 ist die Wahrscheinlichkeit, einen Idioten aus sich zu machen, viel grösser als 5%. Denn der p-Wert testet nicht nur die Null-Hypothese, sondern auch alles andere im Experiment. Das schönste Beispiel hierfür ist das extrem aufwendige OPERA – Experiment, das 2011 am CERN in Genf durchgeführt wurde. Dabei gelang eine sensationelle Entdeckung: Neutrinos bewegen sich schneller als Licht! Die New York Times titelte damals, dass „winzige Neutrinos das kosmische Geschwindigkeitsbeschränkung durchbrochen haben". Das Experiment wurde mehrmals wiederholt, aber das Ergebnis blieb stabil bei einem p-Wert von kleiner 0.00000001. Leider führte dieser spektakuläre Befund nicht zu einem Nobelpreis, sondern zu einer totalen Blamage für die beteiligten Wissenschaftler. Wie sich später herausstellte, war ein Kabel im Setup lose, und ein Messinstrument war nicht richtig kalibriert. Merke: Der p-Wert bezieht sich auf die Ergebnisse eines spezifischen Experimentes, und nicht die Hypothese! Wie spezifisch ist eigentlich Ihr Antikörper?

Der p-Wert, und damit der ganze damit verknüpfte Teststatistik-Kosmos („frequentist" oder auch „Null-Hypothesis Significance Testing", NHST) führt uns also schnell auf Abwege. Der p-Wert leistet nämlich meist gar nicht das, was wir von ihm erwarten – uns zu sagen ob wir etwa Neues entdeckt haben, oder ein Effekt vorliegt. Sollten wir ihn deshalb ganz aufgeben? Einfach nicht mehr testen, wie von den 800 Kollegen gefordert?

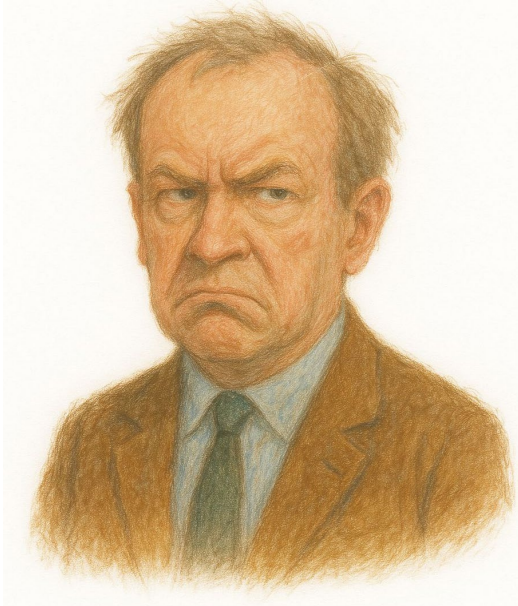
Das hiesse, das Kind mit dem Bade ausschütten! In einem kürzlich erschienenen, sehr informativen Kommentar argumentierte John Ioannidis, dass „die Signifikanz (nicht nur statistisch) sowohl für die Wissenschaft als auch für das wissenschaftsbasierte Handeln wesentlich ist, und einige Filterprozesse nützlich sind, um ein Ertrinken im Rauschen der Daten zu vermeiden". Er meint damit, dass das Aufgeben von Signifikanztests unserem Bias freien Lauf lassen würde. Jeder könnte alles behaupten, und „unwiderlegbarer Unsinn würde regieren". Wir ertrinken doch jetzt bereits in einem Meer falsch positiver Ergebnisse. Ohne irgendeine Schwelle für die Behauptung eines Zusammenhangs oder einer Entdeckung würde sich diese katastrophale Situation mit Sicherheit weiter verschärfen. Stattdessen sollten wir strengere Regeln für die Datenerfassung und -analyse festlegen, wozu die a priori Benennung und Registrierung von Hypothesen und geplanten Analyseverfahren zählen. Obwohl in den meisten Bereichen üblich, reicht eine Signifikanzgrenze von 5 % nicht aus, um das Vorhandensein eines Zusammenhangs oder eines Effektes zu beanspruchen. Ein p-Wert in dieser Region zeigt, wenn überhaupt, dass die Ergebnisse „einen Blick wert" sind und möglicherweise weitere Untersuchungen, z.B. eine Validierung mit grösserer Fallzahl rechtfertigen. Das Verkündung einer Entdeckung

oder Effektes, die nur auf $p < 0.05$ basiert, ist grundsätzlich falsch. Und ohne ausreichende Power ist sowieso jeder p-Wert unzuverlässig, während Effektgrößen (bei einem vorhandenen Effekt) überschätzt werden.

Keines der in der aktuellen Debatte zum p-Wert vorgebrachten Argumente oder vorgeschlagenen Auswege ist neu. Seit Einführung seiner Grundlagen durch R.A. Fisher, also seit fast 100 Jahren ist er zyklisch Gegenstand von hitzigen Debatten. Auch seine Abschaffung ist schon mehrfach gefordert worden, ebenso wie die Aufgabe von NHST, also frequentistischer Statistik zu Gunsten von alternativen Ansätzen, insbesondere Bayes'sche Statistik. Auffällig ist, das diese Diskussionen fast ausschließlich von Statistikern oder Statistikafficionados geführt werden. Die ohnehin wissen wie man den p-Wert (nicht) interpretiert. Und mit Bayes'scher Statistik vertraut sind. Viel wichtiger wäre es aber, dass wir, die ‚normalen' Forscher, uns vom Ritual der Hypothesentestung mit $p < 0.05$ verabschieden und die Interpretation unserer Ergebnisse nicht vom p-Wert abhängig machen sollten. Uns statt dessen auf biologisches Denken konzentrieren und das Design, die Analyse und die Veröffentlichung unserer Studien mehr Sorgfalt verwenden, und diese (prä)registieren. Methoden und Ergebnisse sollten so transparent beschrieben werden, dass Effekte und Schlussfolgerungen unabhängig bestätigt werden können. Die Angabe von statistischen Signifikanzen ist hyperinflationär und damit bedeutungslos geworden. Teststatistiken können unsere Argumentation leiten, aber nicht bestimmen.

Warum trauen WIR dem Weltklimarat, die Klimaskeptiker aber nicht?

LJ 11/2019



Es steht schlecht um die gesellschaftliche Akzeptanz unserer täglichen Arbeit als Wissenschaftler. Die Mehrheit der US-Bevölkerung erklärt sich Evolution nicht mit Darwin, sondern der heiligen Schrift. Die Masern sind weltweit wieder im kommen, weil Impfgegner eine Verschwörung der Pharmaindustrie wittern, die Kinder zu Autisten macht. Ein substantieller Anteil der Bevölkerung hält den Klimawandel nicht für vom Menschen verursacht. Sondern für Hysterie, welche interessierte Wissenschaftler aus Wichtigtuerei und in der Konkurrenz um Fördermittel schüren. Homöopathen behandeln Krankheiten mit Zuckerkügelchen, und die Kassen, also wir, müssen dafür zahlen.

Weniger Skepsis durch mehr Wissenschaft? Ein populäres Rezept gegen diese zunehmende Ablehnung relevanter Ergebnisse der Wissenschaft ist mehr und bessere Vermittlung von Wissen in Schule und Medien. Angeregt durch einen Vortrag des prominenten amerikanischen Wissenschaftssoziologen und – Historikers Steven Shapin

erlaube ich mir, hieran zu zweifeln. Denn der Kern des Problems liegt keineswegs am so naheliegenden, aber dennoch falschen Befund einer Krise in der Akzeptanz von Wissenschaft bzw. wissenschaftlicher Wahrheit. In der Kritik steht nämlich nicht die Wissenschaft, sondern Institutionen, Autoritäten, und Eliten. Den Kritikern passen konkrete Ergebnisse und Handlungsempfehlungen der Wissenschaft nicht. Die Wissenschaft und ihre Methode bleiben hingegen verschont. Die Argumente werden sogar im Namen von uns akzeptierten Wissenschaftsnormen vorgetragen: Skeptizismus und Unabhängigkeit von Erkenntnis-schädlichen Interessen. Die Kritiker sind skeptisch und reklamieren das wissenschaftliche Prinzip der Falsifikation für sich. Sie bedienen sich dabei Statistiken, Zahlen und Ergebnissen von alternativen „Experten". Die Kritiker treten häufig wie radikalisierte, „bessere" Wissenschaftler auf, die dem Mainstream den Verrat der eigenen Ideale vorwerfen.

Neben der ‚Wissenschaftlichkeit' in der Argumentation ist auffällig, dass die Liste der angezweifelten Befunde relativ kurz ist. Der in der Schule gelehrte Kanon der Wissenschaften, also die Textbuch-Wissenschaft ist nicht in der Schusslinie. Newton, Maxwell, Einstein – kein Problem. Auch die Fördergeber, wie die DFG oder die wissenschaftliche Methode kommen nicht schlecht weg, sondern gar nicht vor. Warum aber dann Ablehnung von Evolutionstheorie? Weil die Wissenschaft hier der Religion ganz grundsätzlich widerspricht! Warum Impfgegnerschaft? Weil sich Eltern ernste Sorgen um ihre Jüngsten machen! Warum Klimawandel? Weil die Leute nicht ihren Lebensstil ändern wollen, also weiter SUV fahren und mit dem Flieger nach Mallorca! Warum Homöopathie: Weil die klassische Medizin ihnen oft nicht hilft und manchmal schadet!

Könnten Sie das Klimamodell des IPCC erklären? Es geht den Kritikern nicht um wissenschaftliche ‚Wahrheit'. Es geht um ganz konkrete Dinge, die ihnen nicht passen, welche aber im Namen von Wissenschaft verordnet werden. Das Problem ist nicht wissenschaftliche Ignoranz, wie häufig behauptet wird. Natürlich herrscht auch Ignoranz bezüglich wissenschaftlicher Ergebnisse und Theorien – von Aberration bis Zellzyklus. Aber hier wird es so richtig interessant: Denn wir, die Wissenschaftler des Mainstream, akzeptieren den anthropogenen Ursprung des Klimawandels und die Evolutionstheorie nicht deshalb, weil wir die Wissenschaft dahinter verstehen. Greta Thunberg ist keine Expertin in der Modellierung komplexer Systeme. Auch unser generelles Verständnis wissenschaftlicher Ergebnisse ist ebenfalls rudimentär bis oberflächlich. Oder würden Sie behaupten, die Modelle der Klimatologen zu kennen und deren wissenschaftliche Korrektheit beurteilen zu können? Könnten Sie erklären wie ein Chip in Ihrem Handy funktioniert? Verstehen Sie die Grundlagen der allgemeinen Relativitätstheorie? Vermutlich nein. Ist auch gar nicht nötig. Weder um sich in Sachen Klimawandel zu positionieren, noch um ein Handy zu benutzen.

Wissenschaftsskepsis ist Elitenkritik. Mit welchen Argumenten beharren wir dann aber auf der Rolle des Menschen beim Klimawandel, oder Darwin's Theorie? Unsere Argumente gründen sich fast ausschließlich auf wissenschaftliche Autorität. Es ist eine Form von sozialem Wissen: Wir vertrauen den Spezialisten des IPCC, den CERN Physikern, den Virologen des Robert Koch Institutes, etc. Das sind Leute wie wir, wir sind Teil ihrer Wissenschaftskultur. Wir kennen die Strukturen, in denen sie ihre Ergebnisse erheben, veröffentlichen, diskutieren und letztendlich akzeptieren. Wir wissen, wem wir (ver)trauen können, und wem nicht. Wir gehen da alles andere als demokratisch vor. Es ist ein von uns über längere Zeit innerhalb des Wissenschaftsbetriebes durch Sozialisierung erworbenes und häufig implizites Wissen, das sich kaum operationalisieren lässt. Es ist vom Wesen her elitär, wir haben gut begründete Vorurteile. Wir berufen uns auf wissenschaftliche Autorität(en). Die Skeptiker sind deshalb auch nicht Kritiker der Wissenschaft, sondern von wissenschaftlicher Autorität und insbesondere von uns als

elitärer gesellschaftlicher Gruppe. Sie halten Wissenschaft, sofern sie Ergebnisse betrifft, die ihnen nicht in den Kram passen, für korrumpiert. Von der Politik, von der Wirtschaft, und/oder von persönlichen Interessen. Insofern unser soziales Wissen elitär ist, werden wir zur Zielscheibe rechter Elitenkritik. Diese ist im Übrigen selbst elitär, denn sie hält uns das „Wissen" alternativer „Experten" entgegen.

Die Wissenschaft hat ihre Unschuld verloren. Wie konnte es soweit kommen? Die Pioniere der Wissenschaft, wie wir sie heute betreiben, die Galileis, Boyles und Newtons, waren ‚Gentleman scientists'. Sie finanzierten sich selbst, oder forschten unter adliger Patronage. Sie waren dadurch unabhängig, nur der wissenschaftlichen Wahrheit verpflichtet. Ihre Wissenschaft war, abgesehen vom Zwecke des Erkenntnisgewinns, komplett ‚desinteressiert'. Sie hatten keinen gesellschaftlichen Auftrag, und beriefen sich nicht auf Politik, Geschäft, oder Gesellschaft. Diese Zeiten sind längst vorbei. Ein Meilenstein war zum Bespiel das Manhattan Project zur kriegerischen Nutzung der Kernenergie. Oder der Bayh-Dohle Act, mit dem amerikanischen Universitäten die Monetisierung der Erfindungen ihrer Wissenschaftler nicht nur ermöglicht, sondern ins Stammbuch geschrieben wurde. Wissenschaft hat komplett ihre Unschuld verloren, weil sie, auch an den Universitäten, voll integriert ist in die ‚Institutionen', in Geschäft, Politik, auch Militär. In weiten Teilen der Wissenschaft müssen wir, um Fördermittel zu erhalten, vorab den unmittelbaren Nutzen, die Anwendbarkeit und die Verwertbarkeit unserer Ergebnisse betonen. Wir begründen unsere eigene Wichtigkeit (und damit Forderung nach Förderung) mit dem Dienst an den Institutionen. Der Preis, den wir hierfür zahlen ist, dass Kritik an den Institutionen automatisch Kritik an der Wissenschaft mit sich bringt. Frei nach der Logik: Wenn die Politik lügt, wenn Konzerne lügen, dann lügt auch die Wissenschaft.

Dazu kommt noch, dass die Wissenschaft selbst ihr Scherflein zu diesem Vertrauensverlust beiträgt. Wissenschaftsskandale, Plagiarismus in Doktorarbeiten, Reproduzierbarkeitskrise, fragwürdige Anreizsysteme usw. sind Gegenstand öffentlicher Beobachtung und Missfallens. All dies schürt Zweifel an einer nur dem Erkenntnisgewinn verschriebenen Profession. Belegt dies denn nicht, dass man Wissenschaftlern (nicht der Wissenschaft, wohlgemerkt!) nicht trauen kann, denn sie schummeln und dienen falschen Götzen?

All dies bedeutet: Wissenschaftsskeptiker werden nicht durch mehr ‚Wissenschaft' bekehrt. Auch die Eindämmung von Falschinformation in den sozialen Netzen scheint mir wenig geeignet. Es gibt eine Menge Evidenz dafür, dass die Polarisierung und Radikalisierung in den sozialen Medien eine Folge, und nicht die Ursache des Schlamassels ist. Obskuranten treiben sich auf Seiten für Verschwörungstheoretiker um, weil sie dort die Inhalte finden, nach denen sie suchen. Impfgegner informieren sich auf Anti-vax Seiten, weil sich dort die Argumente gegen das Impfen finden. Es ist einfacher geworden, sich Gehör zu verschaffen, wozu man aber erstmal eine Botschaft braucht, die man verbreiten möchte. Es ist auch einfacher geworden Informationen zu finden, welche man vom Mainstream und dessen Publikationsmechanismen bisher ‚unterdrückt' wähnte. Im Internet sind aber doch auch jetzt schon alle Inhalte dieses Mainstreams (d.h. der Textbuch-Wissenschaft) hervorragend vertreten! Wenn irgend etwas gesichert ist, dann dass die neuen Medien eine größere Diversität in der Aneignung von Information ermöglichen. Die Kritiker kennen unsere Argumente, sie glauben uns aber nicht. Die sozialen Medien offenbaren das Problem, sie verursachen es nicht.

Was tun? Wie so häufig ist die Diagnose einfacher als die Therapie. Die wesentlich mit verantwortlichen Phänomene Populismus, Nationalismus, und Radikalisierung am rechten Rand haben erstmal gar nichts mit Wissenschaft zu tun. Da helfen keine

gestylten Aufklärungskampagnen und Wissenschaftskommunikatoren a la Hirschhausen. Wenn Wissenschaft überhaupt etwas beitragen kann, dann vielleicht mehr Zurückhaltung in der ständigen Betonung der eigenen unmittelbaren Relevanz für Geschäft und Politik. Mehr Betonung auf Erkenntnisgewinn. Der, sofern dieser robust ist, letztendlich immer relevant sein wird für die Gesellschaft. Natürlich ist auch die Vermittlung von Wissen wichtig, bei welcher Gelegenheit und in welcher Zielgruppe auch immer. Aber dies weniger in Bezug auf das unmittelbare Verständnis der komplexen Theorien und Resultate der Wissenschaft. Was ohnehin selten funktioniert, und meist zu sinnentstellender Trivialisierung im Dienste der Popularisierung führt. Vielmehr Vermittlung davon, wie Wissenschaft ganz grundsätzlich funktioniert, welche Mechanismen der Akzeptanz oder Widerlegung von Resultaten sie hat. Dass ihre Hypothesen logisch konsistent, durch Evidenz (empirisch) belegt, falsifizierbar, und ihre Ergebnisse reproduzierbar sein müssen. Aufzeigen und Vorleben der Normen von Wissenschaft. Allgemein akzeptiert sind dies (nach Robert Merton): Kommunismus (nicht erschrecken: meint gemeinsames geistiges Eigentum und kollektive Zusammenarbeit), Universalismus, Selbstlosigkeit, und organisierte Skepsis. Neumodisch auch: Transparenz (Open Science). Aber in all dem haben wir Wissenschaftler bei der Umsetzung noch eine Menge Hausaufgaben zu machen. Man könnte auch sagen, wir müssen da erst noch vor der eigenen Haustüre kehren!

Ist das Wissenschaft, oder kann das weg?

LJ 12/2019

Orangensaft?

Fleischkonsum ist schlecht für die Gesundheit. Da winken Krebs, Herzinfarkt, Schlaganfall, das volle Programm. Sagt die Ernährungswissenschaft. Und die muss es ja wissen. Ist schließlich eine Wissenschaft. Oder?

Jonathan Schoenfeld und John Ioannidis haben sich vor ein paar Jahren ein ganz normales Kochbuch vorgenommen, und daraus per Zufall 50 häufig vorkommende Zutaten ausgewählt (Zucker, Kaffee, Salz, usw.) und dann eine systematische Literaturrecherche durchgeführt. Sie gingen der Frage nach, ob es ernährungswissenschaftliche Studien gibt, welche das Krebsrisiko dieser Zutaten untersucht hatten. Und sie wurden so richtig fündig. Zu 80% der Zutaten lag eine Studie vor, häufig sogar mehrere. Von 264 dieser epidemiologischen Studien fanden 103, dass das untersuchte Lebensmittel das Krebsrisiko erhöhte, 88 erniedrigten dagegen das Krebsrisiko! Also hatte Joe Jackson doch recht: ‚Everything gives you cancer'! Aber kann das sein? Milch? Kalbfleisch?

Es kommt noch toller: 12 Haselnüsse am Tag erhöhen die Lebenserwartung um 12 Jahre, also ein Jahr pro Nuss! Alternativ kann man auch 3 Kaffeetassen am Tag trinken, denn das führt zum selben Ergebnis. Eine Mandarine am Tag ist dagegen weniger effektiv: Nur 5 Jahre Lebensverlängerung. Vorsicht ist geboten bei Eiern: Eines am Tag und man lebt 6 Jahre kürzer, 2 Scheiben Bacon kosten 10 Jahre, und das schafft man nicht mal durch Kettenrauchen. Auf solche Hochrechnungen kommt man, wenn man sich der Analyse von ernährungswissenschaftlichen Kohortenstudien anschließt, und deren kausale Rhetorik ernst nimmt. Zitate hierzu wie immer im Internet (http://dirnagl.com/lj).

Da ist es doch beruhigend, wenn man sich auf solidere Evidenz verlassen kann, die auch viel plausibler ist. Wie zum Beispiel die, dass die mediterrane Diät, also Olivenöl, Rotwein etc., so richtig gut fürs Herz ist und es nicht nur schmeckt, sondern man auch länger bei besserer Gesundheit lebt: Steht im Lehrbuch, in der Bunten, und wurde in einer Großen Studie (PREDIMED) belegt, veröffentlicht im renommierten New England Journal. Eine der wenigen interventionellen, kontrollierten und randomisierten Studien im Ernährungsbereich! Aber wussten Sie, dass diese Studie zurückgezogen werden musste? Weil es gravierende Protokollverstöße gegeben hatte, und die Daten möglicherweise manipuliert wurden. Außerdem wurde gar keine mediterrane Diät getestet, sondern Nahrungsergänzung. Schwamm drüber, der Gavi und die in Olivenöl angebratene Dorade schmecken trotzdem.

Ein ähnliches Schicksal hat in den letzten Jahren das verwandte ‚French Paradox' erlitten. Wir erinnern uns: Trotz höherem Konsum von gesättigten Fetten (Käse!) scheinen Franzosen, insbesondere im Vergleich zu Briten, ein relativ niedriges Risiko für koronare Herzerkrankungen zu haben. Wenn das mal nicht am Rotwein liegt, den die Franzosen so lieben! In der medienwirksamen Kurzform also: Rotwein schützt vor koronaren Herzerkrankungen. Endlich mal brauchbarer ärztlicher Rat! In den Jahren nach Veröffentlichung des Paradoxes und weiterer Studien explodierte der Konsum von Rotwein, insbesondere in den USA. Auch die Grundlagenforschung wurde aktiv: Legionen von Mäusen wurden betrunken gemacht, und isolierte Arterien in diversen alkoholischen Medien gebadet. Dreißig Jahre nach Erstbeschreibung und Hunderte von Studien später ist von der Euphorie leider nichts mehr übrig. Das Paradox ist vermutlich ein Artefakt der unterschiedlichen Erfassung von Herzerkrankungen in Frankreich und UK sowie einer zeitlich versetzten Änderung von Essgewohnheiten in beiden Ländern. In jedem Fall konnte letztendlich weder Rotwein noch irgendeine andere Diät dingfest gemacht werden. Vom französischen Paradox spricht man in wissenschaftlichen Kreisen daher heute diskreter Weise nicht mehr.

Auch beim Alkohol stellte sich mittlerweile heraus: Der vermeintliche protektive Effekt ist ein statistisches Artefakt (wie so häufig: ungeeignete Vergleichsgruppen und ungenügende Korrektur von Confoundern). Und die Chinesen haben dann dieses Jahr mit einer Megastudie (500.000 Teilnehmer, 10 Jahre Nachverfolgung, Genotypisierung usw.) die Story vom schützenden Effekt moderater Mengen von Alkohol endgültig abgekegelt. Jedes Tröpfchen C2 ist von einem gesundheitlichen Standpunkt aus eines zu viel. Nachrichten die zu gut klingen, um wahr zu sein, sind eben häufig genau das, nämlich nicht wahr!

Aber um neuen gesundheitlichen Rat ist die Ernährungsforschung natürlich nicht verlegen. Nun sind es die Omega-3-Fettsäuren, welche uns gesund ins hohe Alter bringen sollen! Mit Franz Beckenbauer sage ich da: Schau mer mal, dann sehn mer scho!

Aber wie steht es eigentlich um den eingangs zitierten Fleischkonsum, vor allem wenn es rot ist? Vor kurzem wurden mehrere sehr große Meta-Analysen veröffentlicht, alle in einer Ausgabe der Annals of Internal Medicine. Resultat: Der Einfluss von

Fleischkonsum auf Gesamt-Mortalität oder kardiovaskuläre Outcomes ist, wenn überhaupt vorhanden, gering!

Die Liste der Assoziationen bestimmter Diäten mit Gesundheit, Krankheit, erhöhter oder erniedrigter Lebenserwartung ist also fast endlos. Manchmal handelt es sich um dieselbe Diät, aber mit entgegengesetztem Outcome. Fast immer wird aus der behaupteten Assoziation eine kausale Beziehung gefolgert. Die Korrelation des Konsums von Lebensmittel X mit einem bestimmten Outcome Y wird dann schnell zu: Konsum von X bewirkt Krankheit Y. Dabei weiß jeder wie viele Faktoren unsere Essgewohnheiten beeinflussen, viele davon in einer unauflösbaren Wechselbeziehung. In seiner Totalität hat das, was wir essen, natürlich großen Einfluss auf unsere Gesundheit. Aber einzelne Lebensmittel spielen dabei in der Regel eine geringe Rolle. Über dem ganzen schwebt dazu außerdem der sozioökonomische Status, oder weniger verklauselt: Wieviel jemand zum Leben hat. Eine legendäre Studie hat einmal den Inhalt von Einkaufstüten an der Supermarktkasse untersucht: Nicht überraschend clustern in den Tüten Bier, Wodka, Dosenfleisch und Zigaretten auf der einen, Rotwein, Olivenöl, Salat und Müsli auf der anderen Seite. Wussten Sie, dass in Deutschland laut offizieller Gesundheitsberichtserstattung des Bundes der Unterschied in der Lebenserwartung zwischen den niedrigsten und höchsten Einkommensgruppen bei Frauen 13,3 und bei Männern 14,3 Jahre ist? Der gleiche Befund, nur noch extremer: Zwischen den Endstationen einer U-Bahn Linie in Chicago (Red Line) nimmt die Lebenserwartung von Nord (dort Wohnen die Gutsituierten) nach Süd (dort sind die ‚Problemviertel') graduell um 30 Jahre ab! Ob das wohl am Olivenöl liegt?

Für die Lebensumstände der Leute, und die Tatsache, dass diese mit Essgewohnheiten und genetischen Faktoren in nicht wirklich aufzulösender Wechselwirkung treten, können die Ernährungsforscher wahrlich nichts. Ebenso wenig sind sie schuld daran, wenn ihre Ergebnisse in den Medien übertrieben oder sogar verfälscht dargestellt werden. Fast jede größere ernährungswissenschaftliche Studie, die ein gängiges Nahrungsmittel zum Gegenstand hatte, taucht in der Laienpresse auf, oft in reißerischer Aufmachung. Auch können Ernährungswissenschaftler nichts dafür, dass es in ihrem Bereich schwierig ist, randomisiert kontrollierte, prospektive Interventionen zu untersuchen. Wofür die Ernährungswissenschaft aber schon was kann, sind methodische Mängel, einige davon habe ich oben bereits benannt. Problematisch ist auch, dass Ernährungswissenschaftler in Gremien sitzen, die häufig auf Basis schwacher Evidenz weitreichende Ernährungsempfehlungen geben. Dazu kommt, dass in der klinischen Medizin insgesamt, aber in der Ernährungswissenschaft ganz besonders, Interessenkonflikte das Design, die Analyse sowie die Interpretation von Studien stark beeinflussen. Der Einfluss der Nahrungsmittelindustrie auf die medizinische Wissenschaft ist mindestens so groß wie die der Pharmaindustrie. Und das will was heißen.

Nach Abzug von Medien-Hype, Verwechslung von Kausalität und Korrelation, Überschätzung von Effektstärken und Unterschätzung von Wechselwirkungen und Konfoundern kann man die Ergebnisse der Ernährungsforschung auf das zurückführen, was uns schon unsere Großmütter mit auf den Weg gegeben haben: Am gesündesten ist ein vielfältige und ausgewogene Diät, nicht einseitig und bloß keine Exzesse. Ein bisschen Obst und Salat, auch mal ein Stück Fleisch, nicht zu viel Fett. Was ein Omnivore halt so braucht. Und weil wir uns viel weniger bewegen als unsere Vorfahren vor ein paar Hunderttausend Jahren: Aufpassen mit den Kalorien, auch mal ins Schwitzen kommen. Oder mit Johann Wolfgang Goethe: ‚Nur durch Mäßigung erhalten wir uns'.

Aber ist das Wissenschaft?

Kaum zu glauben, wir können mehr als zehn Jahre länger leben!

LJ 1-2/2020

Wer häufiger diese Kolumne liest, muss den Eindruck gewinnen, dass der Wissenschaftsnarr ein rechter Nörgler und Misanthrop ist. Nichts und niemand scheint es ihm recht zu machen. Immer sind ihm die Fallzahlen zu gering, die Statistiken faul, die Daten handverlesen, oder die Ergebnisse zu positiv und die Schlussfolgerungen daraus überzogen. Auch scheint ihm das Peer Review System unzuverlässig, von den Fördergebern, welche hauptsächlich Mainstream fördern und Geld dort abladen, wo schon viel davon ist, gar nicht zu reden. Selbst der Nobelpreis ist ihm ein atavistisches Instrument zur Feier des einsam forschenden, natürlich männlichen und weißen Genius. Künstliche Intelligenz ist ihm zu stupide, und das akademische Karrieresystem der Kern all diesen Übels. Um nur ein paar Beispiele zu nennen.

Weit gefehlt! Der Wissenschaftsnarr ist eine Wissenschaftsenthusiast. Er ist davon überzeugt, dass Wissenschaft das beste ist, was die 1500 g Eiweiß und Fett in unserer Schädelkalotte je hervorgebracht haben. Ja, er ist vernarrt in Wissenschaft. Deshalb heute, zum Anfang der neuen Dekade, erst mal ein ordentlicher Lobgesang auf die biomedizinische Wissenschaft.

Dass wir in den sogenannten entwickelten, also industrialisierten Gesellschaften im Schnitt eine Lebenserwartung von deutlich über 80 Jahren haben, und diese Zeitspanne überwiegend gesund verbringen, hat direkt oder indirekt sehr viel mit biomedizinischer Wissenschaft zu tun. Da fallen einem natürlich sofort die Antibiotika und die auf den Erkenntnissen der Mikrobiologie fußende verbesserte Hygiene, auch im Lebensmittelbereich, ein. Die Ausrottung des Kindbettfiebers und die Verminderung der Säuglingssterblichkeit. Die Liste lässt sich beliebig fortsetzen, und beinhaltet Röntgendiagnostik, Insulin, Polio- und TBC-Vakzinierung, Organtransplantation, Antiepileptika, Antiparkinsonmittel, Antihypertensiva, Dialyse, Immunsuppressiva, und vieles mehr. Auch die letzten Dekaden brachten wieder reichlich echte ‚Durchbrüche', so z.B. Statine, Protonenpumpenblocker, HIV-Therapie, Herceptin und einige andere hochwirksame Tumortherapien. Kombiniertes Resultat all dieser Segnungen: Nicht nur die Lebenserwartung, sondern auch die Lebensqualität im Alter ist weiter kontinuierlich gestiegen. Nun benutzen wir Forscher gerne das Argument, dass uns genau diese Erfolge in Zukunft in die demographische Katastrophe führen werden. Denn wir werden – Stichwort Überalterung - demnächst alle dement oder als schwerer Pflegefall an der Schnabeltasse nibbeln. In Wirklichkeit verhalten wir uns aber nur deshalb so alarmistisch, da die meisten Erkrankungen im hohen Alter häufiger werden, wir damit die Forderung nach mehr Fördermittel für unsere Forschung begründen können.

Glücklicherweise besteht kein Grund zur Panik. Altersadjustiert sinkt nämlich die Morbidität und die Mortalität vieler Volkserkrankungen. Um nur zwei wichtige davon zu nennen: Schlaganfall und Herzinfarkt. Die Mortalität kardiovaskulärer Erkrankungen hat in den letzten 20 Jahren um mehr als 40 % abgenommen. Auch dies ganz wesentlich ein herausragender Erfolg der Medizin: Dahinter steckt die Prävention dieser Erkrankungen durch flächendeckende Behandlung von Risikofaktoren wie z.B. hohem Blutdruck. Hinzu kommen neue Therapien, wie die Behandlung von Schlaganfällen und Herzinfarkten mittels notfallmässiger Wiedereröffnung des verschlossenen Gefäßes. Wichtig für diese Erfolge waren auch noch ein paar andere Dinge, auch diese Triumphe der medizinischen Forschung. Wie zum Beispiel die Erkenntnis, dass Rauchen tödlich ist, und der Bann von Zigaretten aus dem öffentlichen Raum nicht nur die Lungenkrebsrate dramatisch reduziert, sondern auch das Auftreten kardiovaskulärer Erkrankungen.

Also: Ein dreifach Hoch auf die biomedizinische Forschung! Aber wird das so weitergehen? Der Anstieg der Lebenserwartung in den industrialisierten Ländern verlangsamt sich, in den USA ist die Lebenserwartung wieder rückläufig. Demgegenüber wird aber mehr denn je geforscht, und zumindest gemessen an Fachartikeln steigt das Wissen nach wie vor exponentiell. Sowohl die Fach- als auch die Laienpresse macht uns ordentlich Hoffnung auf bevorstehende spektakuläre Durchbrüche auf fast allen Gebieten der Medizin. Gentherapie, personalisierte Medizin, Digitalisierung, künstliche Intelligenz und Big Data sollen uns in ein neues Zeitalter führen, in dem Krebs, Alzheimer, et al. der Vergangenheit angehören werden.

Und da ist sie wieder, die Skepsis des Wissenschaftsnarren. Nicht so sehr, weil die Durchbrüche, sollten sie sich wirklich einstellen, aller Voraussicht nach erstmal nur in der Behandlung von wenigen Patienten mit sehr seltenen Erkrankungen bestehen. Und die Therapiekosten pro Patient dann locker über 1 Million Euro liegen werden. Das spricht keineswegs gegen Forschung und klinische Studien in diesen Bereichen, gemahnt uns aber zur Zurückhaltung was die Skalierbarkeit solcher ‚individualisierten' Therapien betrifft. Ganz abgesehen davon, dass noch völlig unklar ist, wie belastbare Evidenz für die Überlegenheit von derart personalisierten im Vergleich zu konventionellen Therapien zu erhalten ist. Denn randomisierte und kontrollierte Studien sind bei den geringen Fallzahlen und den kaum randomisier- oder verblindbaren Therapien nicht durchführbar. Die oben zitierten Erfolge konventioneller Therapien wurden alle in der ‚breiten Masse' erzielt, in Großen Studien konnten Risiko und Nutzen dieser Behandlungen gegeneinander abgewogen werden. Und seither helfen diese Therapien großen Kollektiven von Patienten, nicht nur wenigen Individuen.

Ich will aber meine Unkenrufe zu den Heilsversprechungen der personalisierten sowie Big Data und AI getriebenen Therapien für heute hintanstellen, und auf etwas ganz anderes hinweisen. Den vollmundigen Ankündigungen künftiger Wundertherapien sollte man die Ergebnisse des Global Burden of Disease (GBD) Projektes gegenüberstellen. GBD ‚hat sich die Quantifizierung von Todesfällen, Krankheit, Behinderung und Risikofaktoren zur Aufgabe gemacht; aufgeteilt nach Regionen und Bevölkerungsgruppen. Anhand dieser Informationen ist es möglich, wichtige Informationen abzuwägen, die von politischen Entscheidungsträgern zur Prioritätensetzung genutzt werden können' (Wikipedia). Die Ergebnisse von GBD zeigen uns, dass wir das medizinische Wissen bereits in Händen haben, Evidenz-basiert die Morbidität und Letalität bei uns und weltweit weiter massiv zu senken und die Lebensqualität dramatisch zu erhöhen. Zum einen zeigt das seit 1992 laufende GBD – für sich schon eine Glanzleistung moderner biomedizinisch-epidemiologischer Forschung, dass sich Krankheitslast dort sehr effektiv vermindern lässt, wo wir die krankheitsauslösenden Faktoren kennen. Auch verfügen wir schon über sehr effektive Therapien, sollten die Krankheiten dennoch auftreten. Die

Identifizierung vieler dieser Faktoren, sowie darauf aufbauende präventive Strategien und Therapien, gehören zu den Errungenschaften der modernen Medizin und ihrer Forschung. Allerdings liegt der vermutlich wichtigste Risikofaktor, der vielen anderen zugrunde liegt, außerhalb des Wirkungsbereiches der Medizin. Sogenannte ‚soziodemographische Indikatoren' korrelieren nämlich mit fast allen relevanten Risikofaktoren, wie zum Beispiel Rauchen, Feinstaubbelastung, Alkoholkonsum, Übergewicht. Auf Deutsch: Womit jemand sein Geld verdient und wieviel man davon hat, hat großen Einfluss darauf ob man einen Herzinfarkt, Diabetes, oder Lungenkrebs bekommt. Um die Korrelation von soziodemographischen Indikatoren mit Morbidität und Mortalität zu studieren, kann man im Lande bleiben und muss nicht südlich der Sahara forschen.

Schon in der letzten Ausgabe des Laborjournals, als es um Ernährungs'wissenschaften' ging, habe ich mir erlaubt darauf hinzuweisen, dass in Deutschland der Unterschied in der Lebenserwartung zwischen den niedrigsten und höchsten Einkommensgruppen bei Frauen 13,3 und bei Männern 14,3 Jahre ist. Aber was hat das alles mit den Versprechungen der Medizin für die nächste Dekade zu tun? Stellen Sie sich einmal vor, ein Forscher würde nächste Woche ein Medikament entdecken, das einen Gutteil der Deutschen 10 Jahre länger leben ließe! Eine Weltsensation, Ruhm und Reichtum garantiert! Aber solche ‚Therapien' sind längst bekannt, wir setzen sie nur nicht ein.

Da gäbe es viele Schätze zu heben, und alle sind sie hinlänglich bekannt: Hoher Blutdruck, hoher Nüchternblutzucker, Übergewicht (hoher Body mass index), hohes LDL Cholesterin, Alkoholkonsum, Rauchen, Feinstaub etc. sind nach GBD die führenden Risikofaktoren. In Regionen außerhalb unserer Komfortzone kommen dann noch so Dinge dazu wie unhygienische Wasserversorgung, ungeschützter Sex usw. Das Tolle an diesen ‚Risikofaktoren' ist, dass man sie vermindern oder gar verhindern kann. Es gibt sogar Maßnahmen, die alle gleichzeitig adressieren, diese laufen eine Verbesserung des Lebensstandards hinaus, und die konsequente Umsetzung von existierendem medizinischem Wissen. Ein Geschäft lässt sich allerdings damit nicht machen. Worauf ich hinaus will ist, dass wir bereits wissen, was uns krank macht, und wie wir es verhindern könn(t)en. Nach den konservativen, offiziellen Statistiken von Eurostat könnte jeder dritte Sterbefall in der EU mit dem medizinischen Kenntnisstand und den technischen Möglichkeiten von heute vermieden werden.

Diese Erkenntnis ist eigentlich trivial, aber schön ist, dass die Wissenschaft quantitative Evidenz dafür geliefert hat. Zum Beispiel würde in Europa die Bekämpfung der Hypertonie durch schon lange bekannte und mittlerweile preiswerte Medikamente, oder (sorry Raucher!) genauso eine weitere Reduktion von Tabakkonsum, gesamtgesellschaftlich um Größenordnungen wirksamer sein als jede personalisierte Tumormedizin oder Gentherapie es in Zukunft je sein könnte. Und wenn die Big data aus den Lifestyle-Trackern und der elektronischen Gesundheitsakte erst einmal mittels künstlicher Intelligenz ausgewertet werden, wird dabei wenig Überraschendes rauskommen. Nämlich dass Alkohol, Rauchen und wenig Bewegung schlecht sind, wohingegen eine ausgewogene Diät und ein bisschen Kreislaufertüchtigung hingegen gesund. Und dass Akademikerinnen gesünder und länger leben als Supermarktkassiererinnen oder Friseusen.

Nach meinem Lobgesang auf die Segnungen der biomedizinischen Forschung ist dies kein Plädoyer dafür, diese nun einzustellen, weil wir eh schon alles erreicht haben. Es ist vielmehr der Hinweis darauf, dass wir beim Blick auf die zukünftigen Segnungen personalisierter Therapien und anderer populärer medizinischer Zukunftsvisionen nicht vergessen sollten, dass das Gold sprichwörtlich bereits auf der Straße liegt. Diese Maßnahmen sind nicht so sexy, wären aber sofort umsetzbar und nach allem was wir wissen, extrem effektiv.

Wozu Tierversuche, Medikamente gibt's doch in der Apotheke?

LJ 3/2020

Tierversuche sind ein heikles Thema. Wer Tierversuche macht, so wie ich, redet ungern darüber, zumindest außerhalb unseres natürlichen Habitats, des Labors oder einer Fachkonferenz. Das gleiche gilt für Einrichtungen an denen Tierversuche durchgeführt werden. Die Max Planck Gesellschaft hat Nikos Logothetis (MPI Tübingen) im Regen stehen lassen, als er in eine unter der Gürtellinie geführte (Medien)kampagne geriet. Jetzt ist er samt Labor und Mitarbeitern auf dem Weg nach Shanghai. Auf den Webseiten der einschlägigen Forschungsinstitute findet sich alles mögliche, bunte Immunhistochemien, Weißkittel mit Pipetten in der Hand, am Computer sitzend, oder am Mikroskop. Bloß Tiere sieht man keine. Am eklatantesten ist dies bei den Auftritten der universitären Krankenhäuser. Stolz weisen sie auf ihre Forschungsaktivitäten hin, sie bewerben begeistert (künftige) wissenschaftliche Durchbrüche der medizinischen (Grundlagen)wissenschaft hin zu ganz neuen Therapien. Ein Hinweis auf Tierversuche auf dem Campus? Fehlanzeige.

Das ist bemerkenswert. Denn es gibt meiner Ansicht nach nur eine einzige Rechtfertigung dafür, Tiere in der Forschung zu züchten, zu halten, ihnen manchmal Leid zuzufügen, und zu sie zu töten: Wenn es dazu dient, unser Wissen über biologische Prozesse zu verbessern, damit daraus direkt oder indirekt neue und bessere Therapien entwickelt werden können. Das schließt wohlgemerkt die Grundlagenforschung mit ein, welche ja die Grundlage ist für die Überführung des Wissens in medizinisches Handeln. Anekdotisch wird dies von Befürwortern belegt mit Hinweis auf soundso viele Nobelpreise, welche auf Basis von Tierexperimenten vergeben wurden. Oder etwas pauschaler, der Behauptung, dass letztlich ein Großteil der Errungenschaften der modernen Medizin aus Tierversuchen hervorgegangen sind oder sich dieser bedient haben. Für die Neurologie, in der ich mich ein bisschen auskenne, erkläre ich mich hiermit satisfaktionsfähig für die Aussage, dass mindestens 50 % der mittlerweile tatsächlich fantastischen Therapien für wichtige Hirnerkrankungen (z.B. MS, Parkinson, Epilepsie) uns ohne Tierexperiment nicht zur Verfügung stünden.

Das ist aber kein Freibrief für jeden Tierversuch, der mit dem Hinweis versehen wird, dass dies möglicherweise Wissen für zukünftigen menschlichen medizinischen Nutzen bringen würde. Ein Beispiel hierfür ist eine vor kurzem in Nature veröffentlichte Studie der Stammzell- und Regenerationsforscher aus Harvard und Sao Paulo. In ihren Versuchen wurden Mäuse einer Vielzahl von Foltermethoden ausgesetzt, welche man von der CIA oder auch aus Konzentrationslagern der Nazis kennt. Mäuse wurden über viele

Stunden fixiert, ständig in neue Käfige gesetzt, isoliert, feuchtes Einstreu in die Käfige geworfen, diese gekippt, grelles Licht wechselte in kurzen Abständen mit Dunkelheit, das Ganze in unvorhersehbarer Folge oder auch in Kombination und über viele Tage. Oder es wurde den Mäusen einfach eine extrem schmerzauslösende Substanz gespritzt. Im Artikel werden diese Maßnahmen charmant als ‚Stress procedures' bezeichnet. Und wozu das alles? Um rauszufinden, dass Stress Haare ergrauen lässt, dies durch den Sympathikus vermittelt wird, und dabei Melanozyten-produzierende Stammzellen kaputt gehen, also die Zellen die den Farbstoff produzieren weniger werden. Wer das überraschend findet, hat 100 Jahre Stressforschung verschlafen. Die Tier-Ethikkommissionen aller beteiligten Institutionen haben ihren Segen zu dieser Studie gegeben. Ist das deshalb ethisch vertretbar? Nein – auch wenn das in 10 Jahren zu einem Shampoo führt, das die Bildung von grauem Haar im Alter verlangsamt. Oder Folteropfern in Guantanamo ihre Haarfarbe erhält? Der mögliche Nutzen für den Menschen muss mit dem Leid, das Tieren dafür zugeführt wird, in einem gesunden Verhältnis stehen. Dieses Verhältnis ist sicher nicht leicht zu definieren, aber es gibt klare Grenzen. Und es gibt Dinge, die so grausam sind, dass man sie Tieren, noch dazu so hoch entwickelten wie Mäusen, gar nicht antun darf. Dass wissenschaftliche Einrichtungen ethisch keine Probleme bei sowas sehen, wenn ihre Starwissenschaftler experimentieren, und dann auch noch in Nature publizieren, ist dabei wenig verwunderlich. Verwunderlich finde ich allerdings, dass die Weltpresse diesen Befund mit großem Hallo gefeiert hat, meist mit Abbildungen von putzigen Mäusen und einem Käsestückchen.

Nun wird von Vielen eingewendet, dass Tierversuche vielleicht früher nötig waren, aber heute durch Alternativen ersetzt werden können. Die Logik dieses Gedankens ist zweifelsohne korrekt: Wenn wir heute in der Lage wären, die zum Verständnis von Biologie und Krankheitsmechanismen relevanten Fragen ohne den Einsatz von Tieren zu klären, wären Tierversuche nicht nur obsolet, sondern unethisch. Dies ist jedoch, auch im Zeitalter der Organoide, der iPS – Zellen oder Computersimulationen biologischer Systeme noch nicht möglich. Insbesondere das komplexe Zusammenspiel von Blutzirkulation, Immunsystem, und Hirnaktivität welches praktisch alle Zell- und Organfunktionen moduliert, und damit fast alle Krankheiten beeinflusst, lässt sich zumindest bisher nicht in vitro modellieren. Die Betonung liegt dabei aber auf bisher, und vieles, insbesondere in weniger komplexen Organen als dem Gehirn (z.B. Leber, Lunge) lässt sich tatsächlich schon recht gut im Gläschen modellieren. Das bedeutet, dass Alternativen weiterentwickelt, und bereits vorhandene Ansätze Tierversuche ersetzen müssen.

Nun kann ich mir an dieser Stelle allerdings nicht verkneifen, auf ein paar Widersprüche und Ungereimtheiten in der Argumentation gegen Tierversuche hinzuweisen. Auch weil diese Widersprüche relevante Argumente gegen Tierversuche, und die gibt es sehr wohl, diskreditieren. Das wichtigste und als ethische Haltung nicht zu widerlegende Argument gegen Tierversuche ist die Ablehnung der Nutzung von Tieren für menschliche Zwecke aus moralischen Gründen. Noch vor einigen Jahren wäre dieser Standpunkt in der Praxis für Tierversuchsgegner nicht durchzuhalten gewesen. Denn wer so argumentiert, darf ja auch keinerlei tierische Produkte essen. Seit sogar Lidl eine vegane Abteilung hat, geht das aber ohne weiteres. Schwieriger ist sicher, über Lebensmittel hinaus ohne tierische Produkte auszukommen. Aber auch das geht, und Luxuslimousinen können heute schon gegen Aufpreis mit veganer Innenausstattung geliefert werden. Zum Arzt oder ins Krankenhaus darf man dann aber natürlich nicht, es sei denn man kommt mit Bachblüten und Globuli aus. Wer so lebt ist ein konsequenter und glaubwürdiger Tierversuchsgegner.

Aber davon gibt's ganz wenige. Viele diskreditieren ihren Standpunkt durch unlogische Argumentation und inkonsequente Lebenshaltung. So besitzen Tierversuchsgegner

häufig Haustiere. Und da fängt es an, richtig problematisch zu werden. Gar nicht mal, weil die Tiere möglicherweise nicht artgerecht vom Menschen gehalten werden. Wir denken an Hunde in der Stadt, Katzen in der Wohnung, oder viele der Züchtungen, welche die Zunge nicht mehr ins Maul kriegen, epileptisch sind oder Hüftdysplasien haben. Gravierender scheint mir, dass solche Tierversuchsgegner mit ihren Tieren fleißig zum Tierarzt gehen. Und der verschreibt dann Mittel, welche meist im Tierversuch für den Menschen entwickelt wurden, und so auf Umwegen wieder beim Tier ankommen. Das wirft auch die Frage auf: Sind Tierversuche für die Tiermedizin gerechtfertigt? Außerdem zeigt dies, dass sehr wohl eine prinzipielle Übertragung der Ergebnisse zwischen Mensch und Tier möglich ist. Noch krasser wird es, wenn man sich vor Augen hält, was manche unserer Lieben des nächtens so treiben. Streunende Hauskatzen, und dafür gibt es sehr solide Evidenz, gelten als die wichtigsten Verursacher anthropogener Mortalität von Vögeln und Säugern auf unserem Planeten. Allein in USA schätzt man, dass streunende Hauskatzen jährlich 2,5 Milliarden Vögel und 12,5 Milliarden Säuger töten. Und dabei nicht gerade zimperlich vorgehen! Zum Vergleich: In Deutschland werden etwas mehr als 2 Millionen Versuchstiere eingesetzt.

Es existiert eine in sich schlüssige, biozentrische Argumentation gegen Tierversuche. Diese kann glaubwürdig vertreten wer vegan, auch sonst Tierprodukt-frei, ohne Haustiere und ohne moderne Medizin lebt. Demgegenüber existieren sogar noch ältere anthropo- bzw. pathozentrische Argumentationen für Tierversuche. Diese sind im Alltag einfacher durchzuhalten, dies macht sie aber nicht richtiger, denn Praktikabilität ist keine ethische Kategorie. Auf der ethisch – moralischen Schiene lässt sich zwar trefflich streiten, aber das bringt uns nicht wirklich weiter. Nun gibt es da noch den Staat. Egal wie man sich als Individuum zu Tierversuchen stellt, hat der Staat, unter Berufung auf seine Bürger, durch einschlägige Gesetze, Realitäten geschaffen. Und den Tierschutz ins Grundgesetz aufgenommen. Obwohl diese Gesetze sich ebenfalls auf ethisch-moralischen Konzepte berufen, werden sie durch staatliche Gewalt und nicht durch logische Ableitung, oder Überzeugung durch überprüfbare Argumente durchgesetzt. Aus all dem ergibt sich, dass sich der Konflikt zwischen der Verpflichtung, die menschliche Gesundheit zu erhalten und zu verbessern, und dem Anliegen, Schmerzen und Leiden von Tieren zu vermeiden, weder durch rechtliche, normative, oder ethische Betrachtungen auflösen lässt.

Gibt es denn überhaupt keine von Gegnern wie Befürwortern akzeptierte ethische Prinzipien der Forschung an Tieren? Vielleicht am ehesten die 3Rs von Russell und Burch: Replacement (Ersatz), Reduction (Reduktion), Refinement (Verfeinerung). Deren recht breite Akzeptanz ist natürlich auch der Allgemeinheit, man könnte sagen Unverbindlichkeit, dieser Prinzipien geschuldet: Die Gegner können auf vollständiges Replacement pochen, die Befürworter auf Reduktion und Verfeinerung. Von Prinzipien werden ja auch keine Zahlen oder Zeiträume vorgeben. Trotzdem gibt es Preise für die Beförderung der 3R, und wer als Tierexperimentator die 3R ernst nimmt und beachtet, macht alles richtig. Alles?

Ich denke nein. Denn den 3R fehlt Entscheidendes. Die 3R sind ausschließlich auf das Tierwohl fokussiert. Was ihnen fehlt ist der Reflex auf den wissenschaftlichen Wert der Tierversuche ! Man kann nämlich ganz tolles Refinement machen, und sogar einige Reduktion erreichen, und trotzdem wertlose und damit unethische Tierexperimente durchführen. Das ist dann der Fall, wenn diese Versuche methodisch mangelhaft durchgeführt werden, z.B. durch ein falsches Studiendesign, Verzerrung durch fehlende Verblindung, falsch positiv oder falsch negativ werden durch zu niedrige Fallzahlen, die Daten selektiv verwendet werden, falsch ausgewertet wird (p-Hacking, HARKING), sie nicht publiziert

werden (wg. Null- oder negativem Resultat), oder die Beschreibung der Ergebnisse nicht ausreichend ist um sie zu wiederholen oder ihre Qualität zu beurteilen.

Hierfür braucht es gleich nochmal 3Rs, nämlich Robustness, Registration, und Reporting. Robust werden Tierversuche nämlich erst durch ausreichende interne Validität, also methodische Kompetenz und Kontrolle von Bias. Registrierung verhindert, dass Daten selektiv ausgewertet werden und Studien unter den Tisch fallen, oder Hypothesen zugrunde gelegt werden, die man erst nach Auswertung der Resultate gebildet hat. Und gutes Reporting, z.B. durch Adhärenz zu Richtlinien wie ARRIVE, sowie die zur Verfügungstellung von Originaldaten, wird eine Nach- und Weiternutzung der Ergebnisse ermöglicht (FAIR-Prinzipien). Oder dass man die Daten überhaupt veröffentlicht. Eine kürzlich veröffentlichte Studie der Gruppe um Daniel Strech legt nahe, dass weniger als 2/3 aller von den Behörden genehmigten Tierversuche überhaupt das Licht der (Fach)-Öffentlichkeit erblicken. Wer schon länger im Geschäft ist, oder die Meta-Research Literatur kennt, die all dies quantitativ untersucht, weiß dass da in vielen tierexperimentellen Studien noch erheblicher Nachholbedarf besteht. Daniel Strech und ich haben die Forderung nach Berücksichtigung dieser zusätzlichen Prinzipien kürzlich detailliert begründet (http://bit.ly/6RArtikel).

Aber noch etwas fehlt mir in der gegenwärtigen Diskussion um Tierversuche: Volle Transparenz bei denen, die Tierversuche machen. Allen voran den wissenschaftlichen Einrichtungen, und da insbesondere der Universitätsmedizin. Jedes Unikrankenhaus sollte auf seiner Internet – Präsenz, und zwar möglichst auf der Einstiegsseite und nicht irgendwo versteckt auf die Tatsache hinweisen, dass dort im Rahmen des medizinischen Erkenntnisgewinns an Tieren geforscht wird. Und diese und deren Zweck dann allgemeinverständlich beschreiben. Und auf die 6R, nicht nur die 3R achtet! Ich würde sogar noch weiter gehen. Unikliniken sollten in die Einverständniserklärung, die jeder Patient vor Behandlungsbeginn unterschreiben muss, diese oder eine ähnliche Formulierung aufnehmen: „Ich bin darüber informiert worden, dass Ärzte und Wissenschaftler des Klinikums Tierversuche zur Aufklärung von Krankheitsmechanismen und Entwicklung neuer Therapien durchführen. Viele Therapien, welche an unserem Krankenhaus zur Anwendung kommen, basieren direkt oder indirekt auf Tierversuchen."

Verrückt? Keineswegs. Es entspricht der Wahrheit, zwingt zum Nachdenken, und gibt prospektiven Patienten die Möglichkeit, sich eben doch nicht nach modernen medizinischen Standards behandeln zu lassen, da sie damit möglicherweise Nutzung von Tieren, oder sogar Tierleid billigend in Kauf nehmen.

Registered reports: Was wir von Christoph Columbus lernen könnten

LJ 4/2020

Es rauscht im Blätterwald! Nach Jahrzehnten relativer Stabilität erlebt das akademische Verlagswesen einen dramatischen Wandel. Die Geschäftsmodelle der Verlage, aber auch Schlüsselelemente des Publikationsprozesses, wie z.B. der Peer Review Prozess, stehen auf den Prüfstand. Darüber hinaus stellen Forscher und Fördergeber die Rechtfertigung der hohen Gewinne der Verlage in Frage. Der technische Fortschritt und das Internet haben die Formatierung und Verbreitung von Forschungsergebnissen erleichtert, was die Frage aufwirft, ob wir überhaupt noch Verlage brauchen.

Interessanterweise sind neben einigen wenigen Aktivisten nicht wir Wissenschaftler die Treiber dieses Wandels. Wir sind wohl zu sehr mit unserer Forschung beschäftigt - und Gefangene eines Systems, in dem unser Verbleiben und Fortkommen immer noch ganz wesentlich und sogar oft ausschließlich an die Veröffentlichung hochrangiger Publikationen geknüpft ist. Tatsächlich ist das akademische Anreizsystem (mit dem Impact Factor, oder dem Renommee eines Journals) das stärkste verbleibende Bollwerk, das die Verlagsindustrie und ihre Zeitschriftenhierarchien wie wir sie kennen aufrechterhält. Es sind die Fördergeber, die Fachgesellschaften und sogar einige Verlagshäuser selbst welche auf Veränderung setzen. Motiviert durch die aktuellen Zweifel hinsichtlich der Robustheit und der Werthaltigkeit unserer Forschung entwickeln sie neuartige Publikationsformate, um die Qualität und Zugänglichkeit der Forschungsergebnisse zu verbessern. Hierzu zählen der barrierefreie Zugang ("Open Access") zu allen öffentlich finanzierten Forschungsergebnissen und die Veröffentlichung von Artikeln vor der Überprüfung im Peer Review, die Preprints. Mathematik und Physik haben es mit ArXiv vorgemacht, BioArXiv macht es nun sehr erfolgreich in den lebenswissenschaftlichen Disziplinen nach.

Aber es gibt auch ganz neue, spannende Artikelformate welche sich derzeit verbreiten. Das Standardmodell des Publikationsprozesses: Erst Studie durchführen – Artikel schreiben und einreichen – Peer Review – dann (hoffentlich) Veröffentlichung wankt, weil dessen Nachteile immer deutlicher werden. Wir alle wissen wie anfällig der Review Prozess ist: Seilschaften oder Animositäten von Gutachtern, Inkompetenz und Zeitmangel, Homophilie – die Bevorzugung von Forschungsansätzen die unseren eigenen ähneln, anderseits Ideenklau, Intransparenz und Willkür bei der Entscheidung der Editoren, usw. Aber noch viel Grundsätzlicheres ist problematisch: Nach Abschluss der Studie ist das Pferd doch aus dem Stall - ein fehlerhaftes Design oder eine fehlerhafte Analyse kann nur selten nachträglich behoben werden. Zusätzliche Experimente, um der Kritik der Gutachter nachzukommen, sind oft durch den Wunsch der Autoren, genau die Ergebnisse zu erzielen, welche die Gutachter gefordert haben, beeinflusst. Die Nichteinhaltung der "Empfehlungen" der Gutachter führt oft zur Ablehnung, was fast immer eine Kaskade von Einreichungen auslöst. Man bewegt sich in der Hierarchie der Zeitschriften

weiter und weiter nach unten. Letztendlich werden die Manuskripte dann doch irgendwo veröffentlicht. All das verschwendet Zeit und Ressourcen von Autoren und Reviewern, ohne die Wissenschaft wesentlich zu verbessern, und führt zu einer Inflation der Literatur mit fragwürdigen Studien.

Darüber hinaus lädt die Einreichung von Studien nach ihrer Fertigstellung und Analyse ihrer Ergebnisse zur selektiven Verwendung von Daten („Rosinenpickerei") ein; zur Nichtveröffentlichung von Ergebnissen, die nicht zur Hypothese passten oder ihr sogar widersprachen; zur Hypothesenbildung nach Bekanntwerden der Ergebnisse („HARKING"); sowie zum „story telling". Diese Praktiken, in Kombination mit mangelnder interner Validität (z.B. fehlende Kontrolle von Bias durch Verblindung) und statistischen Mängeln (z.B. zu geringe Fallzahlen deshalb unzureichende Power), sind wichtige Ursachen der gegenwärtigen Reproduzierbarkeitskrise. Möglicherweise könnten all diese Probleme auf einen Schlag gelöst werden durch das neue Artikelformat der Registered Reports.

Beim Registered Report, wie er mittlerweile von vielen Journalen angeboten wird, werden vor der Durchführung der Studie zunächst die Methoden und geplanten Analysen zu Papier gebracht und beim Journal eingereicht. Dieses Protokoll wird begutachtet, und es kann, vielleicht auch erst nach Modifikation aufgrund von Kritik der Reviewer, zur vorläufigen Annahme der Studie kommen („Stage 1 Acceptance"). Sobald die Studie dann durchgeführt wurde, reichen die Autoren das vollständige Manuskript, welches nun auch die Ergebnisse enthält, zur abschließenden eher formalen Überprüfung ein. Wenn die Studie wie beschrieben durchgeführt, oder aber Abweichungen davon gut begründet wurden, wird sie publiziert (Stage 2 Acceptance). Registered Reports verhindern damit alle oben erwähnten unangemessenen Forschungspraktiken auf einen Schlag, einschließlich unzureichender statistischer Aussagekraft, selektiver Auswahl der Ergebnisse, unangemessener analytischer Flexibilität, Verzerrungen bei der Interpretation, oder nicht-Publikation von unerwarteten Ergebnissen.

Registered Reports erfordern also eine Vorab-Spezifikation der Hypothese und der geplanten Methodik und Analyse. Dadurch sind sie ideal für konfirmatorische Studien, die darauf abzielen, bereits existierende Forschungsergebnisse zu bestätigen. Eignen sie sich aber auch für explorative Forschung? Die aktuelle biomedizinische Literatur wird ja von der Erforschung und Entdeckung neuer Krankheitsmechanismen und Therapien dominiert. Es liegt auf der Hand, dass das enorme Maß an wissenschaftlicher Freiheit bei der Exploration die Forschungsarbeit in hohem Maße anfällig für die oben erwähnten unerwünschten Praktiken macht, insbesondere für Bias, geringe statistische Power und fehlerhafte Statistiken, sowie für eine nicht offen gelegte selektive Nutzung von Daten. Können Registered Reports auch helfen, diese Forschungspraktiken in exploratorischen Studien zu verhindern?

Schauen wir doch mal auf die ursprüngliche Welt der Exploration. Im goldenen Zeitalter der Entdeckung wurde Terra icognita zu Land und Wasser durchquert und kartographiert, häufig auch plündernd und mordend, meist motiviert durch die Hoffnung auf Ruhm und Reichtum. Die Forschung an den Grenzen der modernen Biologie und Medizin mag hauptsächlich von menschlicher Neugier getrieben sein, aber auch individueller und nationaler Eigennutz sind immer noch wichtige Motive. Entdecker wie Columbus oder Magellan mussten sich auf die von ihren Vorgängern beschriebenen Landmarken und angefertigten Karten verlassen. Sie wussten aber nicht, wie genau diese waren und was wirklich vor ihnen lag. Genauso segeln wir Wissenschaftler heute über einen Ozean der relativen Unwissenheit von unbekannter Größe. Wir stützen uns auf Landmarken des bereits vorhandenen Wissens - das wir dabei oft revidieren oder sogar ganz über den

Haufen werfen. Wir triangulieren unseren Weg mit verschiedensten Methoden, zwar nicht mit Kompass und Sextant, aber mit Kombinationen aus genetischen und pharmakologischen Manipulationen, Immunhistochemie oder Kernspintomographie. Auf unserer Reise gehen wir auf eine induktiv deterministische Weise vor, wobei wir uns der vielen Freiheitsgrade, die uns zur Verfügung stehen, meist gar nicht bewusst sind. Diese ergeben sich zum Beispiel aus der alternativen Analyse oder Interpretation unserer Experimente, aus falsch positiven oder falsch negativen Zwischenergebnissen oder aus der Vielzahl theoretisch möglicher methodischer Ansätze. Folglich gibt es nicht nur einen Weg, den Ozean der Biologie zu überqueren, sondern viele. Und genau wie Kolumbus könnten wir in Amerika landen, und nicht, wie geplant, in Asien. Und wären vielleicht dennoch überzeugt, dass wir die Küsten der Gewürzinseln erreicht haben.

Aber im Gegensatz zu uns haben die Entdecker ihre Reisen "präregistriert", meist bei ihren Herrschern, die ihre Expeditionen auch finanzierten. Und noch wichtiger, bereits während sie unterwegs waren kartographierten sie alles, einschließlich der Abweichungen und Widrigkeiten die ihnen widerfuhren, und schickten diese Berichte und Karten nach Hause. Dies verbesserte die Navigation für andere, die ihnen folgten, und machte künftige Expeditionen sicherer und effektiver. Analog dazu könnten wir Wissenschaftler, bevor wir Segel setzen (bzw. uns an die Bench begeben), ein Ziel in Form einer Hypothese oder eines mutmaßlichen Mechanismus (d.h. die Kernfrage unserer Forschung) festlegen und vorläufige Regeln aufstellen, nach denen wir unsere Experimente und Analysen planen. Dies könnte sich auf unsere bisherigen Reisen stützen, z.B. auf Pilotdaten, aber auch auf bereits vorhandene Karten, also die veröffentlichte Literatur. Ein solcher Plan könnte Stufe 1 eines Registered Reports werden. Während der Reise, insbesondere wenn neue Daten durch Triangulation mit verschiedenen Methoden gewonnen werden, könnten wir das Protokoll aktualisieren, und später den Gutachtern und Lesern zur Verfügung stellen. Dies würde jedes Mal geschehen, wenn Experimente abgeschlossen oder Entscheidungen über das weitere Vorgehen getroffen werden. Letztlich würde so ein Protokoll als eine Art „Logbuch“ dann den gesamten Verlauf einer Studie aufzeichnen, die Auswahl (oder Auslassung) von Daten rechtfertigen und auch Versuchslinien erfassen, welche wir nicht weiterverfolgt haben.

Was gewinnt man aber nun, wenn man exploratorische Forschung präregistrieren und die Registrierungsdatei mit einem Protokoll verknüpfen würde? Im Peer-Review der Stufe 1 würden wir auf methodische und analytische Schwächen, Konflikte mit Guidelines, übersehene oder falsch interpretierte frühere Befunde usw. aufmerksam gemacht werden. Dies würde uns helfen, die Qualität der Studien zu erhöhen, bevor wir loslegen, was möglicherweise Ressourcen und sogar Tiere einsparen würde. Mit Hilfe des „Logbuchs“ könnten die Gutachter die Arbeit während ihrer Durchführung verfolgen. Alternativ könnte das Protokoll als offenes elektronisches Laborjournal geführt werden. Das sich daraus ergebende lebende Protokoll würde zu einem echten "next we"-Narrativ führen, nicht zu den imaginären post-hoc Stories, welche derzeit üblich sind (siehe Wissenschaftsnarr LJ 10/2017 ‚Von den Gefahren allzu schöner Geschichten‘). Es würde die Verästelung unseres Forschungsprozesses und die vielen zur Verfügung stehenden Möglichkeiten erfassen und die von uns im Laufe der Studie schließlich ausgewählten Optionen rechtfertigen. Die methodische und analytische Flexibilität würde beibehalten, aber offengelegt. Es könnte trotz Präregistrierung während der Studie zu Veränderungen in Fragestellung, Studiendesign, oder Analyse kommen, die jedoch als solche begründet würden und unter den Augen der Gutachter und Leser ihr Stigma verlieren. Die Präregistrierung in Verbindung mit der Protokollierung des Studienfortgangs bewahrt die Freiheit des Forschers und nimmt uns auch nicht den Zufall („Serendipity“) als kleines Helferlein.

Die Einzelheiten der Präregistrierung und der Protokolle müssten noch genauer spezifiziert werden, aber im Prinzip handelt es sich um Varianten der "inkrementellen Registrierungen", die bereits von verschiedenen Zeitschriften eingeführt wurden. Sicherlich könnten unlautere Forscher ein solches System durch selektive Protokollierung von Experimenten aushebeln. Sie würden jedoch auf die Vorteile verzichten, vor Beginn der Studie möglicherweise wertvolle Informationen zu erhalten. Noch wichtiger ist, dass eine große Stärke von Präregistrierung und Protokollierung für die exploratorische Forschung darin bestünde, dass wir Forscher der inhärenten Grenzen der Exploration bewusst erleben. Gerade die selektive Auswahl oder Interpretation von Experimenten oder Daten bei unserer Berichterstattung führen wir uns so schlagend selbst vor Augen. Und weil schon Stage 1 (nach Wunsch mit Embargo) publiziert werden könnte, würden Doktoranden nicht Opfer der Verzögerungen, Wirrungen und Zufälligkeiten des derzeitigen Pre-Publication-Reviews. Letztendlich würden wir als Autoren und Leser von wissenschaftlichen Artikeln skeptischer, und würden wissenschaftliche Evidenz realistischer beurteilen können.

Wird das Virus die Wissenschaft verändern?

LJ 5/2020

SARS COV2 beschert der Wissenschaft derzeit den wohl größten Auftritt ihrer tausendjährigen Geschichte. Nicht nur erklärt sie bis ins letzte molekulare Detail einen Vorgang, den man noch nicht vor allzu langer Zeit als Strafe Gottes für die Sünden des Menschen erklärt hätte. Sie macht, obzwar noch mit gehöriger Ungenauigkeit, Vorhersagen über das, was geschehen könnte. Sie schafft Evidenz für die Wirksamkeit von Maßnahmen zur Eindämmung der Epidemie. Und natürlich am wichtigsten, sie entwickelt mit Hochdruck Therapien für Erkrankte, und Impfungen um uns künftig vor dem Virus zu schützen. Wissenschaft wird die Grundlage liefern, um uns gegen die nächste Pandemie zu rüsten.

Die Politik, welche unter immensem Zeitdruck, mit marginaler Expertise, und auf Basis noch recht schwacher und sich ständig wandelnder Evidenz Entscheidungen treffen muß, hat dies erkannt, und ist so Wissenschafts-hörig wie noch nie. Wissenschaftler betonen derzeit häufig ‚doch ‚nur Wissenschaftler zu sein', und die Politik ‚nur zu beraten'. Das ist aber nur die halbe Wahrheit, da Politiker den Rat einzelner Wissenschaftler mehr oder weniger ad hoc und 1:1 in Maßnahmen umsetzen, welche entweder immenses Leid und Schaden verhindern, oder aber eben diese erzeugen könnten. Die Wissenschaft ist aus dem Elfenbeinturm herabgestiegen, eine schwere Verantwortung lastet auf ihr und wenigen ihrer Vertreter.

Wie unter einem Brennglas optisch vergrößert und durch eine Zeitmaschine komprimiert exponiert das Virus gerade gnadenlos Schwächen und Stärken des gegenwärtigen Wissenschaftssystems. Das Virus ist dabei, zum Katalysator zu werden für eine Vielzahl

von Veränderungen in der Art und Weise, wie wir Wissenschaft machen. Was jetzt in den Fokus gerät, wird bereits seit etwa einem Jahrzehnt mit wachsender Intensität von den verschiedensten Stakeholdern diskutiert, und manches davon auch schon zaghaft implementiert. Der Wissenschaftsnarr hat einiges davon auf diesen Seiten aufgegriffen. Aber jetzt passiert alles auf einen Schlag.

Preprint-Server werden plötzlich zum entscheidenden Kommunikationsportal der Wissenschaft. Peer Review? Dauert viel zu lange, und hält die Wissenschaftler von der Arbeit ab – sie müssen doch forschen! Open Data – das unmittelbare zur Verfügung stellen von Originaldaten, na klar, die anderen Wissenschaftler sollen nicht nur die gemachten Versuche und Analysen nachvollziehen können, sondern durch die Daten Anderer schneller in ihrer eigenen Arbeit vorankommen. Die Publikation von Manuskripten und Offenlegung von Daten geschieht nun häufig unter bewusster Aufgabe von Ansprüchen auf Verwertung in Form von Patentanmeldungen. Das würde ja nicht nur zeitlichen Verzug bedeuten, sondern auch Ausschluss Anderer von möglicherweise wichtigem Wissen. Es wird kollaboriert wie noch nie: Gruppen die sich noch vor kurzem geradezu paranoid abschotteten aus Furcht, gescoopt zu werden, tauschen nun Protokolle, Reagenzien und Ergebnisse aus. Die Resultate der Forschung verschwinden auch nicht mehr hinter Paywalls. Covid Publikationen sind fast immer Open Access, sogar in Journalen, welche sich mit diesem Prinzip bisher noch recht nicht anfreunden konnten. Regulatorische Behörden, welche traditionell Antrags-Bearbeitungszeiten von vielen Monaten hatten, genehmigen Experimente und Studien nun innerhalb von Tagen. Das Bundesforschungsministerium (BMFT) stellt über Nacht 150 Millionen € für eine Vernetzung der Forschung der Universitätsmedizin in Deutschland zu Verfügung - Mittel welche vermutlich nicht über konventionelle, langwierige Antragsverfahren vergeben werden. Das alles ist so noch nie dagewesen, und hat etwas Rauschhaftes.

Und auch die Wissenschaftskommunikation läuft auf Hochtouren, beachtet von allen Bevölkerungsgruppen und mit bisher wenig genutzten Formaten. Allen voran der Podcast des NDR mit Christian Drosten. Wann hatte man das schon, daß Hundertausende über Wochen täglich dem Moment entgegenfiebern, an dem ein Wissenschaftler eine halbe Stunde lang die Prinzipien der PCR, die Komplexitäten des innaten und adaptiven Immunsystem, sowie infektionsepidemiologische Propädeutik wie Basisreproduktionszahl R0 und Serienlänge erklärt? Auch dabei auch ganz Prinzipielles über Wissenschaft kommuniziert wird: Daß sie etwas immer Unfertiges ist, ihre Ergebnisse von heute die von gestern über den Haufen werfen können. Das ganze noch garniert mit praktischem Ratschlag für die Zukunft, wenn Lokale wieder offen haben: Das Bier dann besser aus der Flasche trinken, wegen der Viren. Und die Anti-Vaxxer? Sind schweigsam geworden, und hoffen auf eine Impfung.

Dann diese täglichen Videokonferenzen mit Kollegen, mit Zoom, Teams, Gotomeeting, Skype. Bis vor kurzem noch gefürchtet, weil man auf Grund schlechter Audioqualität wenig mitkriegte, kaum vernünftig diskutieren konnte, weil entweder keiner oder alle gleichzeitig etwas sagen wollten. Und man selber abgelenkt war durch die Möglichkeit, nebenbei Emails zu erledigen oder Papers zu schreiben. Nun zwingt einen das Virus, sich etwas besser in die Technik und Bedienung der diversen Plattformen ein zudenken. Ein Headset zu benutzen, sich zu konzentrieren, und einer gewissen Etikette zu folgen. Und plötzlich stellt sich heraus, daß diese Videokonferenzen auch ganz hervorragend funktionieren können. Ja, daß manche davon sogar effektiver verlaufen, als wenn man im gleichen Raum säße. Teilnehmer werfen spontan Slides in die Diskussion ein, im parallellaufenden Chat werden Links und Zitate geteilt. In einem Googledoc entsteht nebenbei live das Protokoll, jeder Teilnehmer kann sich daran beteiligen. Unglaublich, wieviel Zeit und CO_2 eingespart werde kann, ohne die wie sich nun herausstellt unnötige

Reisetätigkeit. Umwelt und Produktivität lassen danken! Selbst kleinere und mittelgroße Konferenzen mit mehreren hundert Teilnehmer werden virtuell abgehalten – und siehe da es funktioniert ganz hervorragend! Vermutlich werden wir in Kürze ähnliche Effekte auch in der Lehre sehen.

Bedeutet nun all dies, daß wir nach Corona einen ‚Paradigmenwechsel' erleben werden, in dem wie wir Wissenschaft betreiben und wie diese wertgeschätzt und wahrgenommen wird? Transparenter, offener, kollaborativer, effektiver, CO_2-neutraler, immer am Puls der Öffentlichkeit? Schön wär's, aber es spricht einiges dagegen, und Evgeny Bobrov hat wichtige Argumente in seinem Blogbeitrag bei ‚Elephant in the Lab' aufgelistet (doi:10.5281/zenodo.3732948).

Schon einmal, anlässlich der Zika-Epidemie in Brasilien 2015 gab es eine gewisse Euphorie, dass danach die Wissenschaft nach der Krise viel offener sein werde. Aber vielleicht war Zika für viele von uns zu weit weg, und der Schock saß nicht tief genug. Nun geht es wieder um ein Virus. Rüttelt es uns nur deshalb stärker auf, weil wir nicht wie gewohnt weiterforschen können? Sobald wir aber wieder an der Bench stehen, und weil die Wenigsten von uns Virologen sind, werden wir uns vielleicht bald nicht mehr erinnern an all die tollen Dinge. Kehren wir dann zur Routine zurück?

Unklar ist auch, welche Auswirkungen der kometenhafte Aufstieg der Preprints haben wird. Auch der Peer Review verhindert nicht die Publikation fragwürdiger Studien. Es könnte aber vielleicht noch schlimmer kommen, als es momentan schon ist. Wenn nämlich mit den Preprints ein Tsunami problematischer Studien auf den Markt geschwemmt würde, durchmischt mit Ausgezeichnetem und Mittelmäßigem. Wie dann die Spreu vom Weizen trennen? Bei den Covid-Preprints versuchen verschieden Konsortien dies durch sog. ‚lebende systematische Reviews' zu lösen. Mittels Text mining mittels Maschinelearning Algorithmen und einer darauffolgenden menschlichen Qualitätskontrolle durch Experten bewerten und synthetisieren sie kontinuierlich die Evidenz im immer weiter anschwellenden Strom von Publikationen. Sollte dies Erfolg haben, könnte es auch auf andere Felder angewendet werden.

Und auch bei Open Data (OD) wird nicht alles Gold sein, was glänzt. Wie können wir sicherstellen, daß hier nicht Datenmassengräber entstehen, nur des Labels „OD" wegen? Die FAIRe (Findable-Usable-Interoperable-Reusable) Deposition von Daten ist alles andere als ein Kinderspiel, und überfordert schon jetzt viele Forscher.

Die ‚neue Wissenschaft' wird also nicht einfach vom Himmel fallen. Wir müssen an den neuen Formaten arbeiten, sie von tollen Ideen zu praktikablen Lösungen entwickeln. Rausfinden was funktioniert, und was nicht. All dies wird auch zusätzliche Ressourcen benötigen, z.B. in Form von Infrastruktur. Aber auch von Training und Ausbildung, und ganz wichtig: Experten, die uns im täglichen Geschäft helfen, wie zum Beispiel Data Stewards. Das wird sich rechnen, denn wir werden mit einer vertrauenswürdigeren und nützlicheren Wissenschaft belohnt werden.

Die Krise kann Katalysator sein, aber es braucht Substrat und Kofaktoren damit die Reaktionsgeschwindigkeit gesteigert wird und die Ausbeute steigt.

„Der Fall Ioannidis“ – Schlamperei beim Gralshüter wissenschaftlicher Qualitätsstandards?

LJ 6/2020

Am 17. März, just zu dem Zeitpunkt als viele Staaten drakonische Maßnahmen zur Eindämmung der SARS-COV-2 Pandemie einleiteten, meldete sich der von mir in dieser Kolumne schon häufiger erwähnte, griechisch-amerikanische Meta-Researcher und Epidemiologe John Ioannidis von der Stanford Universität mit einem provokanten Kommentar zu Wort: Möglicherweise sei ‚a fiasco in the making‘! Aber der wohl international profilierteste Kritiker von schlechter Qualität in der Wissenschaft meinte nicht das Virus, sondern die folgenschweren Maßnahmen, welche auf Grund unzureichender oder schlecht erhobener Daten eingeleitet würden und zu unabsehbaren Sekundärschäden führen könnten. Als einem der meistzitierten Forscher der Welt und lautstarkem Kritiker von Qualitätsproblemen in der Biomedizin wurden seinen Äußerungen zum Thema COVID nicht nur in der Wissenschaft, sondern ganz besonders auch in der Laienpresse große Aufmerksamkeit gewidmet.

Kurz darauf legte Ioannidis mit Daten aus zwei wissenschaftlichen Studien nach. Auf Basis der Auswertung offizieller Mortalitäts-Daten verschiedener Länder schlussfolgerte er, dass die Wahrscheinlichkeit an COVID zu sterben für die meisten Menschen etwa so niedrig sei, wie morgens auf dem Weg zur Arbeit tödlich zu verunglücken. Gemeinsam mit Kollegen von der Stanford University fand er dann in einer serologischen Studie, dass im kalifornischen Santa Clara County wohl mehr als 50 mal mehr Personen vom Virus infiziert wären als offiziell mittels PCR bestätigt.

Kein Wunder, dass er in kürzester Zeit zum wissenschaftlichen Kronzeugen für eine Lockerung oder gar Aufhebung der Eindämmungsmaßnahmen wurde. Die Ergebnisse beider Studien wurden begierig von den konservativen Medien in den USA aufgegriffen und er wurde zum begehrten Interviewpartner, insbesondere in Medien wie dem FOX News Channel. In Deutschland hat er mit Suharit Bhakdi einen neuen Anhänger gewonnen. Zeitgleich gerieten allerdings beide als Preprint veröffentlichten Studien in den sozialen Medien ins Sperrfeuer der wissenschaftlichen Methodenkritik.

Auch diejenigen, welche wie der Wissenschaftsnarr John Ioannidis als den unangefochtenen Gralshüter wissenschaftlicher Korrektheit betrachten, reagierten spätestens jetzt schockiert. Nicht so sehr wegen seiner Vereinnahmung durch reaktionäre Medien – richtige Argumente werden nicht einfach dadurch falsch, dass man sie gegenüber Obskuranten äußert oder sie von diesen zitiert werden. Auch nicht, weil Ioannidis sich mit seinen Aussagen gegen den wissenschaftlichen und politischen Mainstream stellte, dies war schon immer sein Markenzeichen. Nein, der Shitstorm, der sich zum ‚Fall Ioannidis‘ ausweitet, entzündete sich an den methodischen Schwächen, welche in der Summe Ioannidis‘ Argumente für eine Fehleinschätzung der Gefährlichkeit von SARS-COV-2 in

Frage stellen. Der Vorwurf an ihn lautete also, Studien mit verfasst zu haben, und sich mit deren Ergebnissen prominent in die öffentliche Diskussion eingemischt zu haben, die den von ihm selbst gepredigten Qualitätskriterien nicht genügen. Schadenfroh wurde darauf verwiesen, dass er nun wohl selbst den ultimativen Beweis für die Richtigkeit seines 2005 veröffentlichten, berühmtesten Artikels geliefert hätte, dessen Titel lautete: ‚Why most published research findings are false'!

Aber hat John Ioannidis, ungeachtet methodischer Schwächen der Studien, vielleicht dennoch recht? Es geht dabei im Kern um die Mütter aller Corona-Fragen: Wie gefährlich ist SARS-COV-2 tatsächlich, und könnte es sein, dass die drakonischen Maßnahmen gegen den Virus möglicherweise am Ende schädlicher als der Virus sein werden? Aber sollte nicht die Erinnerung an die Bilder aus New York oder Nord-Italien, und die Kühlcontainer voller COVID-Verstorbener, ausreichen, diese Frage eindeutig beantworten?

Ganz so einfach ist es nicht, denn die Wissenschaft hat tatsächlich noch keine eindeutigen Antworten auf den exakten Grad der Durchseuchung, die Ursachen für die Alters- und Ortsabhängigkeit der Mortalität, und vor allem nicht auf die Frage, welches Ausmaß die Kollateralschäden annehmen werden. Klar ist nur, dass man nicht nur die direkte Morbidität und Mortalität des Virus berücksichtigen darf. Sondern auch jene einbeziehen muss, welche das Resultat der Überlastung von (unvorbereiteten oder ohnehin dysfunktionalen) Gesundheitssystemen ist. Und aus der Angst vor Ansteckung im Krankenhaus resultiert, weil Hilfe bei lebensbedrohlichen Akuterkrankungen wie Schlaganfall und Herzinfarkt nicht in Anspruch genommen wurde. Auch die die psychischen Auswirkungen des Lockdowns sowie die Spätfolgen von Schulschließungen müssen berücksichtigt werden. Und dann natürlich die Auswirkungen der ökonomischen Krise oder sogar eines Kollapses der Wirtschaft in dem Gefolge des Lockdowns. Jetzt haben sich in vielen Ländern Arbeitslosigkeit und Armut bereits massiv verschärft, mit den hinreichend bekannten Folgen für Gesundheit und Lebenserwartung.

Aber wie solide war die Evidenz von Ioannidis' Studien, welche von FOX News gefeiert und von vielen Experten zerrissen wurden? In einer berechnete er auf Bevölkerungsebene das relative und absolute Risiko an COVID-19-zu sterben. Hauptkritik an der Studie: Am Anfang einer Epidemie, in der die Prävalenz einer Erkrankung ja definitionsgemäß niedrig ist, macht es wenig Sinn, das absolute (oder auch relative) Sterberisiko der Erkrankung zu berechnen. Worauf es in der frühen Phase der Pandemie viel mehr ankommt, ist nicht das absolute Sterberisiko der Infektion, sondern die Kapazität des Gesundheitssystems, und das Ausmaß der Kollateralschäden der nicht-pharmakologischen Interventionen (z.B. Social distancing, Lockdown, etc.). Ioannidis' Fokus auf die COVID-Mortalität wurde von Mark Lipsitch (Harvard), seines Zeichens einer der weltweit führenden Infektionsepidemiologen, damit verglichen, „in den ersten 3 Tagen nach einer Krebsdiagnose das absolute Sterblichkeitsrisiko zu berechnen".

Die zweite Studie, ebenfalls schon weiter oben erwähnt, fand im kalifornischen Untersuchungsgebiet eine erheblich höhere Durchseuchung mit dem Virus als dies aus den offiziellen Statistiken hervorging. Dies deshalb, weil wie überall bisher nur diejenigen getestet wurden, welche typische Krankheitssymptome hatten oder Kontakt mit Infizierten. Dadurch entgehen uns notwendig die meisten derjenigen, die trotz Infektion nicht ernsthaft erkranken, und da scheint es bei SARS-COV-2 eine Menge davon zu geben. Damit sinkt dann natürlich die errechnete Mortalität durch die Infektion, die sich aus der Division der am Virus Verstorbenen durch die Zahl der mit ihm Infizierten ergibt. Auch dieser Preprint wurden unmittelbar nach Veröffentlichung in den sozialen Medien zerrissen. Die umfassendste Kritik kam von Andrew Gelman, einem international renommierten Biostatistiker. Kritisiert wurden Selektionsbias, die statistische

Auswertung und mangelhafte Validierung des Test Kits. Hinzu kommt, dass bei nicht 100 %iger Spezifität eines Tests, wenn das untersuchte Merkmal selten ist (‚niedrige Prävalenz'), von einer hohen Falschpositivenrate ausgegangen werden muss. Dann stellte sich auch noch heraus, dass David Neeleman, der Gründer der Fluggesellschaft JetBlue einer der Geldgeber der Studie war. Und dies im Preprint nicht offengelegt wurde. Der Besitzer einer Airline hat natürlich ein massives Interesse an der Lockerung von Reisebeschränkungen. In der Studie wurde eine Seroprävalenz von um die 3% gefunden. Es ist aber aufgrund der Testspezifität und niedrigen Prävalenz nicht auszuschließen, dass die meisten davon falsch positiv waren! Deshalb, und wegen der anderen methodischen Mängel ist die Schlussfolgerung der Studie, die Infektion sei mehr als 50-fach häufiger als angenommen, und dass nun Werte vorlägen „an denen Epidemie und Mortalitätsvorhersagen kalibriert werden können" ganz sicher überzogen gewesen.

Allerdings existieren mittlerweile für beide Studien revidierte und um zusätzliche Daten und Analysen erweiterte Revisionen als Preprint. Auch ist die Pandemie weiter fortgeschritten, die nochmals aktualisierten Mortalitätsraten blieben unverändert. Ioannidis sieht alle Kritiken für ausreichend adressiert. Mittlerweile fanden eine Reihe weiterer Studien ähnliche Mortalitäts- und Seroprävalenzraten.

Momentan verbessert sich zwar die von Ioannidis' monierte unzureichende Datenbasis zur Abschätzung von Mortalität, Prävalenz, und Infektiosität von SARS-COV-2 ständig, die Unsicherheit ist aber immer noch groß. Am größten ist aber derzeit das Unwissen über die Kollateralschäden der Eindämmungsmaßnahmen. Erste Studien, ebenfalls meistens als Preprint veröffentlicht, deuten darauf hin, dass diese massiv sein werden, z.B. im Bereich Herzkreislauferkrankungen, Krebs, und psychische Störungen. Überall dort wo dies aufgrund von Registern möglich ist, interessanterweise außer in Deutschland, sehen wir jetzt zudem eine hohe Übersterblichkeit. Besonders beunruhigend ist dabei allerdings nicht nur, dass es sich dabei vorwiegend um ältere Menschen, insbesondere aus Pflegeheimen handelt. Sondern auch, dass die Übersterblichkeit einem erheblichen Anteil von ‚anderen Ursachen' zuzuschreiben ist, also nicht direkt dem Virus. Man kann an COVID sterben, ohne infiziert zu sein! Damit sollte sich der Fokus nicht mehr nur noch auf das ‚Naturereignis' einer viralen Pandemie richten, sondern auf heruntergesparte und damit dysfunktionale Gesundheitssysteme, Armut, Pflegenotstand, Zustände in Altersheimen, usw. Die teilweise deutlichen Unterschiede in der Mortalität von COVID zwischen verschiedenen Ländern sind weniger biologisch, als vielmehr gesellschaftlich bedingt.

Wie lautet also mein Urteil im Fall Ioannidis? John Ioannidis hat keine Gelegenheit ausgelassen zu betonen, auch nicht in den FOX News, dass die Politik mit sofortigen drakonischen Abwehrmaßnahmen richtig gehandelt habe im Angesicht der akuten, unklaren Lage. Die Ergebnisse seiner Studien sind mittlerweile nicht mehr wirklich kontrovers: Seroprävalenzwerte wie in der Santa Clara County Studie werden aus anderen Untersuchungsregionen berichtet, und die geringe Mortalität von SARS-COV-2 bei den unter 65-Jährigen ohne Begleiterkrankungen in Regionen mit funktionierender Gesundheitsversorgung ist mittlerweile Allgemeingut. Die Sekundärschäden durch die Eindämmungsmaßnahmen kennen wir noch nicht, aber es zeichnet sich ab, dass sie, so wie von Ioannidis bereits im März vorausgesagt, katastrophal sein werden. Trotzdem fällt ein dunkler Schatten auf den Gott der wissenschaftlichen Reform: Trotz der von ihm selbst formulierten schwachen Datenbasis hat er im selben Atemzug mit sehr drastischen Worten suggeriert, dass wir dramatisch überreagieren. Er hat sich dann, gestützt auf methodisch problematische und vermutlich durch Bias verzerrte eigene Studien unmittelbar ins Rampenlicht begeben, bevor - beziehungsweise während - eine wissenschaftliche Auseinandersetzung mit den Studien stattfand. Dass seine Auftritte politisch

instrumentalisiert würden, muss ihm bewusst gewesen sein. Die Götter der Griechen waren unsterblich, hatten aber die Gestalt von Menschen, deren negative Eigenschaften und Schwächen. Ioannidis hat uns vorgeführt, dass auch die profiliertesten Methodenkritiker Fehler machen.

Was lernen wir aus all dem? Zum einen sehen wir hier die Stärken und Schwächen von Preprints am Wirken. Deren Stärke besteht darin, dass nach ihrer Veröffentlichung eine intensive, öffentlich und in Echtzeit ausgetragene Diskussion einer wissenschaftlichen Studie stattfinden kann. Welche dann Korrekturen und Verbesserungen ermöglicht. Als Revision des Preprints sich die Arbeit dann wieder der Kritik einer Vielzahl von Experten aus den unterschiedlichsten Bereichen. Bevor der Artikel, dermaßen gestählt, in die Submission und einen regulären Peer Review bei einem Fachjournal geht. Dies ist derzeit bei ca. 70% der Preprints der Fall. Vielleicht müssen solch intensiv von der Fachöffentlichkeit vor aller Augen diskutierte Manuskripte dann auch gar nicht mehr den Weg in ein reguläres Journal finden. Die Naturwissenschaften, allen voran Mathematik und Physik, bedienen sich seit den frühen 90er Jahren des letzten Jahrhunderts dieses Publikationsverfahrens (‚arXiv'). Mit der Folge, dass die Mehrzahl der Preprints in diesen Wissenschaften gar nicht mehr bei Journalen eingereicht werden. Wozu auch?

Die Schwäche der Preprints ist natürlich ihre ‚Ungeprüftheit'. Es könnte sich um völligen Schwachsinn handeln, oder um schiere Brillanz: Nur die Fachwelt kann dies beurteilen, und selbst die tut sich da manchmal schwer. Damit liefern Preprints potentiell das Material für Obskuranten, Extremisten, oder politisches Personal mit gefährlichen Agenden. Aber mal ehrlich: Der reguläre Peer Review garantiert doch auch nicht für die Qualität und Richtigkeit der Aussage von Studien. Wir erinnern uns an Wakefield's Lancet Studie zu Vakzinierung und Autismus, oder die Science und Nature Arbeiten von Schön, Stapel, Obokata, und vielen anderen. Auf die vielfältigen Probleme des Review Verfahrens wurden vom Wissenschaftsnarren auf diesen Seiten schon mehrfach hingewiesen.

Aber vielleicht handelt es sich ja gar nicht um eine ‚Schwäche' der Preprints! Im Gegenteil: Die Preprints führen uns und den wissenschaftlichen Laien momentan ganz praktisch vor, dass Wissenschaft keine endgültigen Wahrheiten in Form von Publikationen liefert. Dass Wissenschaft schwierig ist, organisierte Skepsis eben, Fehler macht, immer in Bewegung ist, ihre eigenen Ergebnisse jederzeit in Frage stellt und diese revidiert sobald Fehler aufgedeckt oder bessere Evidenz vorhanden ist. Dies zeigen us gerade die Preprints, es gilt aber auch für begutachtete Artikel ebenso wie Lehrbücher.

Außerdem lehrt uns der ‚Fall Ioannidis', dass Wissenschaft sich in Zeiten einer universellen Krise nicht auf einen ‚research exceptionalism' berufen darf, wie es Alex London und Jonathan Kimmelman in einem lesenswerten Artikel in Science kürzlich formuliert haben. Wissenschaftliche und ethische Standards dürfen unter Zeitdruck nicht herabgesetzt werden, wie in den beiden Ioannidis Studien geschehen – sondern müssen im Gegenteil erhöht werden. Schlechte Daten sind nicht besser als keine Daten!

Vom Maus zum Mensch durchs Tal des Todes?

LJ 9/2020

‚Translation'- von der Maus zum Mensch und zurück – oh Du Mantra und ewig blaue Blume der Universitätsmedizin! Klar, wo gibt es das schon unter einem Dach: biomedizinische Grundlagenforschung und klinische Forschung, die für nötigen Studienpatienten, einen staatlichen Auftrag inklusive Finanzierung, sowie motiviertes und dafür exzellent ausgebildetes Personal! Translation ist so alt wie die akademische Medizin – nur der Begriff dafür wurde erst in den 80er Jahren des vorigen Jahrhunderts geprägt und ziert seither die Websites und Mission-Statements aller Unikliniken, und dies weltweit. Im Blick zurück ganz sicher ein Erfolgsmodell – man denke nur an Antibiotika, Epilepsiebehandlung, moderne Tumortherapie, HIV-Therapie.

Nicht nur ewige Mäkler wie der Wissenschaftsnarr, sogar DFG und Wissenschaftsrat beklagen allerdings seit geraumer Zeit, dass es nicht mehr so rund läuft mit der Translation. Allerlei poetische Metaphern werden bemüht, wie der ‚translational roadblock', oder gar das translationale ‚Tal des Todes' (‚Death valley'). In vielen Bereichen der Medizin geht es nämlich trotz internationalem massivem Forschungseinsatz nicht mehr so recht vorwärts. In meinem Fach, der Schlaganfallmedizin, forschen wir seit Jahrzehnten mit Begeisterung an pathophysiologischen Grundlagen, schreiben tolle Papers, und werden dadurch mit ein bisschen Glück auch verbeamtet – bei Patienten mit Schlaganfall ist von alledem bisher allerdings überhaupt nichts angekommen! Ist der Schlaganfall eine Ausnahme, sind Schlaganfallforscher vielleicht einfach unfähig? Was ist dann aber mit den Alzheimer Forschern? Wo bleiben die so lange versprochenen, im Tiermodell so effektiven Stammzelltherapien? Wo die wundersamen Behandlungen, die sich aus der Entschlüsselung des menschlichen Genoms ergeben sollten?

Wer sich noch nicht vollends in den schützenden Kokon der Universitätsmedizin eingesponnen hat – und deshalb den eigenen Erfolg an der Höhe der eingeworbenen Drittmittel oder dem Impact Factor von Publikationen misst, kann da schon ins Grübeln kommen. Wie erfolgreich sind wir in der Translation, gemessen am Einsatz von Ressourcen und unseren eigenen Versprechungen? Schon länger wird deshalb nach den Ursachen für die enttäuschende Bilanz translationaler Forschung gesucht. Und man ist fündig geworden. Es liegt am Tal des Todes, das es lebendig zu durchqueren gilt, und am fehlenden ‚Mindset' der beteiligten Wissenschaftler und Kliniker, also deren innerer Einstellung.

Aber schon die Metapher vom Tal des Todes führt uns auf die falsche Fährte. Es suggeriert zwei Antipoden: Hier die Grundlagenforschung, dort die klinische Forschung, in beiden läuft es ganz prima– aber die unwirtlichen Bedingungen dazwischen sind das Problem. Aus diesem Bild leiten sich dann die gängigen Strategien zur Verbesserung der Erfolgsrate des Translationsprozesses ab: Man müsse die Forscher und Kliniker an die Hand nehmen und ihnen erklären, wie sie es richtig machen sollen. Immer schön an die Patienten denken, wenn man experimentell Krankheitsmechanismen untersucht, oder

an die Mäuse, wenn man Menschen behandelt. Man müsse also nur für das richtige ‚Mindset' sorgen. Und den so Aufgeklärten dann noch ein paar Infrastrukturen zur Seite stellen, welche sie dabei unterstützen. So jedenfalls sieht das die DFG in ihren kürzlich veröffentlichten ‚Empfehlungen zur Förderung translationaler Forschung in der Universitätsmedizin'. Ich fürchte, so einfach ist das nicht – und man verpasst mit diesem Ansatz die wichtigsten Ursachen für die enttäuschende Bilanz von Translation. Das ist tragisch, denn einige davon wären recht leicht zu beseitigen.

Das vielleicht trivialste Hindernis ist natürlich die unglaubliche Komplexität der Biologie. Paradoxerweise entfernt man sich mit zunehmendem Verständnis eines Krankheitsmechanismus meist weiter von einer potentiellen Therapie, als ihr näher zu kommen. Eingriffe in Signalweg A, welche den erwünschten Effekt haben, führen oft zu schädlichen des Signalwegs B. Was hilft gegen Komplexität? Natürlich noch mehr Forschung, und zwar meist sehr Grundlagen mäßige. Mit der Komplexität zusammenhängend und ebenso unangenehm ist das Phänomen der ‚niedrig hängenden Früchte', welche wir schon gepflückt haben. Die wenigen Krankheitsmechanismen, welche einfach und nebenwirkungsarm therapierbar sind, haben wir uns bereits nutzbar gemacht – z.B. Penicillin, Insulin, Dopamin, beta-Blocker, Protonenpumpenblocker, Cyclooxygenase-Hemmer (auch selbst da konnte noch viel schiefgehen, man denke an Vioxx). Viele Volkskrankheiten können wir schon sehr erfolgreich therapieren. Den Bluthochdruck noch besser zu behandeln, oder Epilepsien, oder Multiple Sklerose, da ist sehr schwierig. Sehr zum Leidwesen übrigens der Pharmaindustrie, die nicht von Nature-Papern, sondern von profitablen Medikamenten lebt. Nachdem sie die Blockbuster 'gepflückt' hat, und ihr seit geraumer Zeit wenig wirklich Neues einfällt, lebt sie im wesentlichen von me-too Präparaten, also von vergangenen Erfolgen.

Und dann ist da noch das Problem der niedrigen internen Validität, vor allem der präklinischen Forschung. Etwas direkter ausgedrückt, deren niedrige Qualität. Die Mehrheit aller experimentellen Studien, deren Resultate ja das Fundament der darauf aufbauenden klinischen Entwicklungen darstellen, kontrollieren nicht für Verzerrungen ('Bias'), werden randomisiert noch verblindet durchgeführt. Dazu liegen die Gruppengrößen fast immer unter 10, was bei der biologisch normalen Varianz der Ergebnisse einem Würfeln gleichkommt. Allerdings mit präpariertem Würfel, denn durch das Fehlen einer Präregistierung der geplanten Experimente und Analysen hat der Wissenschaftler weitgehende Freiheit in der Auswahl erwünschter, bzw. im Weglassen unerwünschter Resultate. Unterstützt wird die selektive Datennutzung durch fehlerhafte Statistik, insbesondere das so beliebte 'p-Hacking', also die Durchführung statistischer Tests bis sich ein signifikantes Ergebnis einstellt. Steht die Story dann erstmal, heißt die Devise: 'take the paper and run'. Von der Wiederholung der Ergebnisse (Replikation), vielleicht sogar durch unabhängige Untersucher, nimmt man unter diesen Umständen besser Abstand. Gefördert dies ja ohnehin nicht, und Karriere-mäßig bringt das auch nichts – insbesondere weil dann die vorher so tolle Story im biologischen Halbschatten gar nicht mehr so eindeutig schwarz-weiß aussieht. Und weil die Null- und negativen Resultate, wo also nicht das rauskam, was man sich erhofft hatte, nur in F1000Research oder PLOS One zu veröffentlichen wären, kontaminiert man sich damit besser nicht den Lebenslauf – und archiviert es auf der eigenen Festplatte.

Wo die interne Validität niedrig ist, muss man sich da auch Sorgen um die externe Validität machen? Leider ja – denn die Mehrzahl der präklinischen Modelle ist nicht nur in ihren Spezies recht weit von den Patienten mit der untersuchten Erkrankung entfernt. Wieder ein Beispiel aus der Schlaganfallforschung: Unsere Mäuse sind genetisch praktisch identisch (Inzucht), überwiegend männlich, juvenil, und werden mit einer Vitamingeladenen Müslidiät ernährt und unter Reinraumbedingungen (SPF) gehalten. Sie

hatten also noch nie eine Infektion oder sonstige Erkrankungen, und haben damit sogar im Erwachsenenalter unreife, ja neonatale Immunsysteme. Ich erspare mir die Gegenüberstellung dieser Mäuse zu den typischen Schlaganfallpatienten. Meine einzige Erklärung dafür, warum das seit Jahrzehnten so gemacht wird, obwohl keine der in diesen Modellen so effektiven Therapien auch beim Menschen erfolgreich war? Weil wir uns daran gewöhnt haben, und sich damit tolle Publikationen erzielen lassen. Diese wiederum helfen Drittmittel zu akquirieren, mit denen man wieder tolle Publikationen schreiben kann.

An dieser Stelle der translationalen Verwertungskette, also der Beschreibung eines neuen Krankheitsmechanismus oder gar einer neuen, im Tierversuch wirksamen Therapie, ist damit das Kind meist schon mitsamt dem Wasser aus dem Bade. Sprich: Eine klinischen Entwicklung beginnt, welche auf einem unsoliden präklinischen Fundament steht. Wenn es wirklich so wäre, dass man mit Tierversuchen niedriger interner und externer Validität, geringen Fallzahlen trotz hoher Varianz, selektiver Auswahl von Daten und problematischer statistischer Auswertung bei Patienten erfolgreiche Therapien begründen könnte, bräuchte man doch gar keine Tierversuche!

Aber mal angenommen, man hat einen wirklich soliden Kandidaten für eine klinische Überprüfung gefunden, sowas gibt's ja trotz aller oben genannter Widrigkeiten manchmal? Ich überspringe hier ein paar gesetzlich nötige Zwischenschritte, welche alle auch noch zum Abbruch der translationalen Kette führen können, wie Pharmakologie/Toxikologie, und die Untersuchung von Absorption, Distribution, Metabolismus und Elimination (ADME) des Medikaments. Wie groß sind die Chancen, dass wir in einer randomisierten klinischen Studie eine wirksame Therapie finden werden? Im Durchschnitt darf sie nicht grösser als 50 % sein! Denn dies ist aus ethischen Gründen bei klinischen Studien gefordert. Man nennt es Equipoise: Möglicher Nutzen und Risiko müssen für den Patienten vor Studienbeginn im Gleichgewicht stehen. Es darf also nicht von vornherein feststehen, dass die Studienmedikation besser als Placebo ist. Es wäre ja sonst unethisch, dem Patienten diese vorzuenthalten und ein Scheinmedikament statt dessen zu geben! Darin ein weiterer Grund, warum wir gar nicht erwarten dürfen, dass Translation eine 100 % Effektivität haben kann. Klinische Studien müssen scheitern dürfen! Nur müssen sie so angelegt sein, und das gilt genauso für präklinische Experimente, dass auch ein negatives Resultat verwertbare und relevante Evidenz generiert. Zum Beispiel das Wissen um eine unwirksame Dosis, woraufhin man eine andere probieren kann, oder um eine Nebenwirkung, usw. Und die Resultate müssen deshalb auch zeitnah publiziert werden. Und schon haben wir wieder einen Grund für translationales Versagen. 60 % aller klinischen Studien der deutschen Universitätsmedizin haben 2 Jahre nach Beendigung noch keine Ergebnisse veröffentlicht, bei 40 % ist das auch noch nach 5 Jahren so. Das ist nicht nur unwissenschaftlich, sondern auch unethisch. Denn die Patienten haben an den Studien teilgenommen, weil das hieraus generierte Wissen nachfolgenden Patientengenerationen nützen soll. Sie selbst konnten, wegen Placebo und Equipoise zwar auf eigenen Nutzen hoffen, grösser als bei einem Münzwurf ist er im Schnitt tatsächlich selbst bei Erhalt der Studienmedikation nicht.

Unterstellt habe ich dabei, dass die klinischen Studien, weil durch verschiedene Behörden reguliert und kontrolliert, und unter im Sozialgesetzbuch gefordertem klinischen Qualitätsmanagement durchgeführt, robustere Ergebnisse liefern als die präklinischen Studien, auf denen sie häufig beruhen. Das gilt wohl auch in den meisten Fällen, dennoch ist falsches Studiendesign vermutlich eine häufige Ursache für translationales Scheitern. Auch hier wieder ein typisches Beispiel aus der Schlaganfallforschung: Wenn neuroprotektive, also das Hirn vor Schaden nach Schlaganfall schützende Substanzen im Nagetier nur in den ersten Stunden nach Gefäßverschluss wirksam sind, sollte man

sich eigentlich nicht wundern, wenn sie am Patienten nicht wirken, wenn man erst nach 12 Stunden therapiert. So geschehen in sehr vielen (erfolglosen) akuten Schlaganfallstudien.

Die translationale Kette kann also an ganz verschiedenen Stellen abreißen, ich habe nur ein paar genannt. Es genügt der Bruch des schwächsten Gliedes in der Kette, um Tausende von Patienten unnötigen Risiken auszusetzen und gigantische Ressourcen zu verschleudern. Denn der gesamte Prozess kostet in der Regel hunderte von Millionen Euro. Die gute Nachricht: Translationaler Erfolg muss und kann sich nicht in 100% der Fälle einstellen. Und: Ein nicht unerheblicher Teil der schwachen Glieder lässt sich relativ einfach ersetzen. Eine Erhöhung interner und externer Validität sowie ausreichende Gruppengrößen und ordentliche Statistik, sowie Präregistrierung der Studien, Publikation von Null-Resultaten, und Replikation von wichtigen Befunden würde die Translation auf ein solides Fundament stellen. Im klinischen Bereich analog: Sicherstellung, dass robuste präklinische Evidenz vorliegt, ausreichend gepowerte Studien, Studiendesigns die auch bei Verfehlung des erhofften Ergebnisses informativ sind, sowie zeitnahe Veröffentlichung der Ergebnisse. Wenn wir dies erreicht haben, können wir uns auch um das ‚translationale Mindset' der Beteiligten kümmern. Aber hier die schlechte Nachricht: Von alledem steht nichts in den Empfehlungen der DFG. Ich fürchte das liegt daran, dass viele der von mir genannten Maßnahmen nicht so recht in unser akademisches Karriere – und Fördersystem passen. Sich um die Verbesserung des Erfolges von Translation zu kümmern, heißt nämlich auch: Die Maßstäbe zu ändern, von denen das berufliche Fortkommen in der universitären Medizin abhängt!

Der Peer Review ist tot, lang lebe der Peer Review!

LJ 10/2020

Wir alle kennen das: Nach langem Warten und steigender Anspannung trifft endlich eine Antwort des Journals ein. Mit zittrigem Klick öffnet man die Email, und ließt dort dass man es sich nicht leicht gemacht habe. Aber angesichts der substantiellen Kritik der Reviewer sehe man sich nicht in der Lage, den Artikel zu veröffentlichen. Dies dürfe man bitte nicht als prinzipielles Urteil über die darin enthaltene Wissenschaft verstehen, aber man erhalte zu viele Manuskripte, und müsse deshalb priorisieren. Man wünscht weiterhin frohes Forschen und hofft, dass man dem Journal gewogen bleibe! Nach dem ersten Schock dann ein Blick auf die Reviews im Attachment. Reviewer 1 fand die Arbeit wohl ganz gut, hier ein kleines Monitum, dort ein paar wohlmeinende Vorschläge. Aber Reviewer 2! Hat er den Artikel denn überhaupt gelesen? War es vielleicht eine andere Arbeit, und er hat die Reviews verwechselt? In jedem Fall hatte der oder die Unbekannte überhaupt keine Ahnung, und erdreistete sich dennoch, auf mehreren Seiten Gülle über 3 Jahre unserer harten Arbeit und deren hochrelevante Resultate zu gießen.

Andererseits haben wir auch dies schon erlebt: Durchaus harte, aber konstruktive Kritik der Reviewer, welche die eine oder andere Problemstelle identifiziert hatten, die man selbst übersehen hatte oder nicht wahrhaben wollte. Und dann gute Hinweise gaben, wie man nach ein paar zusätzlichen Experimenten und einer textlichen Revision einen viel besseren Artikel daraus machen könnte!

Wir alle haben also vermutlich recht gemischte Erfahrungen mit dem Peer Review, akzeptieren ihn aber doch klaglos als de facto Eintrittspforte zu jeglicher wissenschaftlichen Veröffentlichung, welche uns Ansehen unter Kollegen und einer Entfristung oder Professur näher bringt. Zwar ist sich die Mehrzahl der Wissenschaftler der Vielzahl von Schwächen des Peer Review Systems bewusst. Aber wer sich von Wissenschaft ernähren will muss damit leben. Und konzentriert sich daher lieber auf die Experimente und das Schreiben der Papers, als auf vertiefte Reflexionen zu möglichen Alternativen der wissenschaftlichen Qualitätskontrolle. Würden Sie die Option ‚Send out for review' wählen, wenn sie bei Einreichung Ihres Papers auch ‚Publish immediately' ankreuzen könnten?

Aber momentan kommt Bewegung in die Diskussion. Denn in gewisser Weise bieten die derzeit so populären Preprints genau diese Option. Corona und die damit verbundende Volks- und Wissenschaftler-Aufklärungskampagne in Sachen ‚Wie funktioniert eigentlich Wissenschaft' hat das Publizieren ohne Review Prozess nicht nur den am Virus Forschenden, sondern allen Wissenschaftlern und sogar Laien nahegebracht. Und die Frage, ob der Peer Review notwendig, überflüssig, oder sogar schädlich ist, erstmals so richtig aufs Tablett gebracht.

Weil an dieser Stelle schon häufiger erwähnt, und letztlich auch im (Unter-)Bewusstsein fast aller Akteure im System präsent, nur kurz noch einmal eine kurze und unvollständige Auflistung der Schwächen des Peer Reviews: Bei Manuskripteinreichung ist das Kind bereits potentiell in den Brunnen gefallen, die Studie nämlich schon durchgeführt, substantielle Verbesserungsvorschläge kommen meist zu spät. Er verlängert die Zeit bis neue Erkenntnis auf den Markt kommt. Er fördert Mainstream Forschung. Er ist völlig intransparent, und fördert damit Seilschaften, oder Vendettas. Er begünstigt Ideenklau. Seine Ergebnisse sind nicht reproduzierbar, seine Qualität erratisch. Egal wie schlecht eine Arbeit ist, nach multiplen Submissionen bei einer Vielzahl von Journalen mit absteigender Reputation (= Impact Factor) wird sie dennoch publiziert. Er verhindert Wissenschaftsbetrug nicht. Er frisst immense Ressourcen – unsere eigenen wegen multipler Revisionen und Submissionen, unsere eigenen als Forscher, Reviewer, und als auch als Steuerzahler. Denn wir finanzieren die Lizenzgebühren der Bibliotheken oder die Open Access - Gebühren (APC), welche die Maschinerie, welche den Review Prozess der Journal aufrecht erhält. Die Liste der Probleme des Review Prozesses ließe sich noch lange fortsetzen. Und für all das gibt es solide, wissenschaftliche Evidenz.

Ein nicht unbeträchtlicher Teil der Arbeit an einem Paper besteht heute darin, den Peer Review Prozess zu beeinflussen oder sogar zu manipulieren. Da werden Arbeiten zitiert und Formulierungen benutzt, die einem erhofften Reviewer schmeicheln, endlos debattiert welche Reviewer man vorschlagen oder besser ausschließen sollte, Journale werden nach Bekanntschaften in den Editorial Boards ausgewählt, usw. Wer diese Klaviatur beherrscht hat einen klaren Karrierevorteil. ‚Meet the editor' Sessions gehören zu den bestbesuchten Veranstaltungen auf Kongressen. Man hofft dort Tipps und Tricks zu erhaschen, um demnächst im betreffenden Journal akzeptiert zu werden.

Wer das alles für normal hält, ist bereits voll im Wissenschaftsbetrieb sozialisiert. Alle anderen sollten sich allerdings die Frage stellen, was diese Umtriebe mit Wissenschaft zu tun haben. Und warum es trotz all seiner Probleme den Peer Review in der jetzigen Form überhaupt (noch) gibt. Gerüchte besagen, dass er so alt sei wie die moderne

Wissenschaft selbst, also quasi in der DNA der wissenschaftlichen Methode angelegt ist. Was aber nicht stimmt. Der Peer Review der Gentleman scientists des 17. Jahrhunderts, der Boyle's und Hooke's hatte sehr wenig zu tun mit dem heutigen Prozedere. Glauben Sie, dass ‚The Molecular Structure of Nucleic Acids' von Watson und Crick bei Nature 1953 durch einen Review Prozess gegangen ist? Natürlich nicht, das gab es damals nämlich noch nicht. Der Peer Review hat sich in seiner heutigen Form erst 20 oder 30 Jahre später richtig entwickelt. Zu Zeiten, in denen viel weniger publiziert wurde, und die verwendete Methodik viel weniger komplex war als heute. Und ‚das System' weniger kompetitiv war als heute. ‚Koryphäen' des Felds publizierten häufig wenig, und wenn, dann in den Journalen ihrer Fachgesellschaften. Jeder kannte jeden. Exzellenz maß sich nicht an der Zahl von Nature Papern. Wissenschaftliche Fehden wurden mit offenem Visier, und häufig auch mit harten Bandagen ausgetragen.

Vordergründig gilt Peer Review heute als Schlüsselelement der wissenschaftlichen Qualitätskontrolle. Wenn wieder spektakuläre Betrugsfälle sogar die Laienpresse beschäftigen reibt man sich zwar ab und an die Augen. Beruhigt sich dann aber gleich damit, dass dies ja nur die Regel bestätige, also Ausnahme gewesen sei. Trotzdem Peer Review also ganz offensichtlich eine Schlüsselrolle im wissenschaftlichen Prozess zugeschrieben wird, gibt man sich keine Mühe ihn zu professionalisieren. Obwohl es viel zu beachten, und noch mehr falsch zu machen gibt, wird gutes Reviewen nicht gelehrt. Man reviewt einfach munter drauf los sobald die erste Anfrage eines Journales kommt.

In Wirklichkeit besteht die wesentlichste Aufgabe des Peer Reviews heute nicht mehr in der Qualitätskontrolle, sondern der Operationalisierung und Quasi-Objektivierung der Hierarchie von wissenschaftlichen Journalen. Und zwar durch die Aufrechterhaltung der für diese Hierarchie notwendigen Selektivität. Die Proliferation und Quasi-Industrialisierung der Produktion von Wissen und deren Verbreitung in Form von wissenschaftlichen Artikeln, sowie die immense Zunahme der Komplexität der Themen und Methoden hat zu eine Überwältigung von Editoren, Reviewern, und Autoren geführt. Die Zeiten der Generalisten sind vorbei. Reviewer können nur noch Teilaspekte der ihnen vorgelegten Arbeiten beurteilen. Um die Qualität der Daten zu überprüfen, sofern diese und die verwendeten Analyseskripte ihnen überhaupt zugänglich gemacht werden, müssten Reviewer sich tagelang mit einem Manuskript herumschlagen. Vor einiger Zeit las ich ein Cell Paper, bei dem allein im Supplement Daten auf 15 Abbildungen verstreut waren, davon hatten einige 26 Briefmarken-große Panels, welche von a-z gelabelt waren! Wer will sich anmaßen, sowas zu beurteilen?

Der Idee vom Peer Review als inhaltlichem Diskurs unter Wissenschaftlern über konkrete Forschungsergebnisse ist edel und plausibel, und hat sich für ein paar Jahrzehnte durchaus bewährt. Peer Review hat sich dabei zum Standard entwickelt, und den Nimbus ‚Qualitätskontrollinstrument' verdient. Er ist aber heute überfordert. Dies auch wegen der mittlerweile im Wissenschaftsbetrieb vorherrschenden Hyperkompetition und der Quantifizierbarkeit des Prestiges der Journale durch den Impact Factor. Dies hat vollends zur Kommerzialisierung des Produkts ‚Wissenschaftlicher Artikel' geführt. Fachartikel sind zur wichtigsten Währung in der Konkurrenz der Wissenschaftler geworden. Die Verlage leben davon, indem sie in ihrer Konkurrenz untereinander den Wechselkurs dieser Währung festlegen. Es kommt also nicht mehr so sehr darauf an, was in einem Artikel steht, sondern vielmehr wo er erschienen ist. Das Prestige des Journals adelt den Inhalt und bürgt gleichzeitig unabhängig von diesem für seine Qualität. Der Peer Review hat in diesem Prozess die Funktion eines quasi-objektiven Steuerungs- und Selektionsinstruments.

Dass da auch mal Papers verbessert, oder großer Mist aussortiert wird, ist zur Nebensache geworden. Klar, wenn ein Artikel zur Begutachtung in die richtigen Hände gerät, gibt es konstruktive Hinweise. Und man wird davor geschützt, sich benebelt von den vermeintlich sensationellen eigenen Ergebnissen durch Publikation von methodischen Fehlern oder überzogenen Schlussfolgerungen vor der Fachwelt zum Idioten zu machen. Peer Review kann nämlich tatsächlich wissenschaftlicher Diskurs vom Feinsten sein!

Wäre es also möglich, sich der Stärken des Peer Reviews zu bedienen, seine Schwächen aber zu vermeiden? Also den Pelz zu waschen, ohne sich nass zu machen? Ich denke schon! Ein einfacher aber hochwirksamer Ansatz ist es, den review Prozess vor den Studienbeginn zu verlegen. Also der ‚Registered Report'. Der Narr hat an dieser Stelle ausführlich darüber berichtet (LJ 4/2020). Damit kombinierbar, aber auch im klassischen Review: Die Schaffung von Transparenz durch Offenlegung der gesamten Korrespondenz des Journals mit den Reviewern. Dies kann man weiter treiben bis hin zum ‚Offenen Review', also Reviews ohne den Schutz der Anonymität. Sie werden sich fragen ob das nicht zu Gefälligkeitsgutachten und Seilschaften führt? Wohl eher nein, denn sie lägen ja für jedermann sichtbar offen! Auch würden Argumentationen unter der Gürtellinie und offensichtlich inkompetente Kommentare wohl seltener, denn sie würden in den mit der Arbeit veröffentlichten Reviews den Gutacher kompromittieren. Ein potentieller Nachteil von offenen Reviews kann es aber sein, dass insbesondere junge WissenschaftlerInnen fürchten müssten, dass kritische Kommentare ihre Karriere gefährden könnten. Abgesehen davon, dass dies kein gutes Licht auf das System wirft, gäbe es auch hierfür Abhilfe, z.B. durch Co-Review mit etablierteren Wissenschaftlern. Deren Name würde dann veröffentlicht, natürlich mit dem Hinweis auf die Kollaboration, und die jungen Wissenschaftler könnten über Angebote wie Publons (https://publons.com) trotzdem Kredit für ihre Arbeit bekommen.

Die spannendsten und besten Reviews finden sich aber derzeit ohnehin erst nach Publikation. Und das sowohl von Preprints als auch von regulären Artikeln, und zwar in den sozialen Medien. Dorthin wurde mittlerweile auch die gründliche Qualitätskontrolle ausgelagert – fast alle manipulierten oder sonst wie betrügerischen Arbeiten wurden von skeptischen Lesern exponiert und dann via Twitter, PubPeer oder Blogs in den internationalen Diskurs gebracht. Viele der COVID Preprints wurden so vom ‚Schwarm' gereviewt. Die Autoren haben dann häufig die relevantesten Kommentare entweder in Revisionen ihrer Preprints berücksichtigt, oder gleich in die reguläre Submission bei einem Peer Review Journal eingebracht.

Einige Journale praktizieren bereits erfolgreich einige oder sogar alle oben genannten Modifikationen des Peer Review. Dazu zählen u.a. Elife, F1000Res, EMBOJ, PLOS Journals, BMJ und PeerJ. Mir würden auch noch weitere Verbesserungen einfallen, davon vielleicht später auf diesen Seiten. Allerdings: Damit das alles richtig durchschlägt und der Peer Review wieder Instrument des kritisch konstruktiven Austausches zwischen Wissenschaftlern wird, müssen sich gleichzeitig noch ein paar andere Dinge verändern. Vor allem: Wir müssen weniger, dafür aber bessere Artikel schreiben. Deren Inhalt und die Qualität müssten wichtigere Kriterien in der Beurteilung von Wissenschaftlern und deren Oevre werden als die Namen der Journale, in denen sie veröffentlichen. Lang lebe der Peer Review!

Wissenschaft berät Politik, oder: *Survival of the ideas that fit*

LJ 11/2020

Trotz mittlerweile wieder stark steigender Fallzahlen und der Angst vor einem zweiten Lockdown freuen wir uns in Deutschland zu Recht, dass wir bisher deutlich besser durch die Corona-Krise gekommen sind als viele unserer Nachbarn oder auch die USA. War der ‚deutsche Weg' vielleicht gar deshalb so erfolgreich, weil die Politik hierzulande ein offenes Ohr für die Wissenschaft hatte, und deshalb evidenzbasiert die richtigen Maßnahmen verordnet hat?

Das klingt plausibel, doch gibt es leider wenig Evidenz hierfür. Denn bisher hat die Wissenschaft kaum belastbare Erkenntnisse geliefert, ob und welche Maßnahmen (z.B. Lockdown) oder Szenarien (z.B. ein funktionierendes und gut vorbereitetes Gesundheitssystem) wirksam waren. Das ist tragisch, und wirft kein gutes Licht auf die Wissenschaft. Denn dieses Wissen würde uns nun wichtige Argumente liefern, was zu tun und was besser zu lassen ist, um im Herbst und Winter die Intensivstationen nicht überlaufen zu lassen und dabei gleichzeitig ein möglichst normales Leben zu gewährleisten.

Bei näherem Hinsehen wird man zudem feststellen müssen, dass es ja auch gar keine evidenzbasierte Beratung der Politik durch die Wissenschaft gegeben hat. Aber halt, haben wir nicht einen Christian Drosten, der die Politik berät und zudem noch Wissenschaft in die Breite kommuniziert wie noch keiner zuvor? Dazu eine Physikerin als Kanzlerin, welche Treffen der Ministerpräsidenten mit Impulsvorträgen von Epidemiologen einleitet. Und einen Gesundheitsminister, der zwar von der Ausbildung her Bankkaufmann ist, aber rational argumentiert und einer Beratung durch die Wissenschaft gegenüber aufgeschlossen scheint? Reicht das nicht? Ich fürchte nein.

Ohne Ausnahme betonen Politiker in allen Ländern, von Albanien bis Zypern (USA eingeschlossen), dass ihre Corona Maßnahmen auf ‚best available science' beruhen. Aber wer entscheidet denn, was diese beste verfügbare wissenschaftlich Evidenz sein soll? Natürlich die Politik selbst. Denn die Bewertung, Priorisierung und Verwendung wissenschaftlicher Evidenz folgt politischem Kalkül. Das heißt unter Einbeziehung anderer staatlicher Interessen, wie zum Beispiel dem Funktionieren der Wirtschaft. Und natürlich mit Blick auf die Wahlbarometer. Man könnte dies frei nach Darwin als ‚Political selection: The survival of the ideas that fit' bezeichnen.

Außerdem, was wäre denn 'best available science' in Zeiten, in denen durch 'Covidization' der schon vorher beeindruckende wissenschaftliche Müllberg noch weiter anschwillt? Wo durch die Inflation von hastig produzierten, teilweise per Pressekonferenz kommunizierten Ergebnissen eine Trennung von Signal und Rauschen immer schwerer wird, und Evidenzsynthese wegen zum Scheitern verurteilt ist: Wo man Müll reinsteckt, kommt auch Müll raus. Denn in kürzester Zeit hat sich *Research exceptionalism* breitgemacht, mit der Maxime: ‚In Zeiten einer Pandemie sind schlechte Daten besser als

keine Daten'. Mehr als 1000 klinische Studien testen derzeit, welche Therapien gegen COVID helfen. Tendenz exponentiell steigend. Die meisten dieser Studien werden nie brauchbare Resultate liefern – aber davon im nächsten Wissenschaftsnarren mehr.

Wie müsste denn Politikberatung durch die beste verfügbare wissenschaftliche Evidenz überhaupt aussehen, um als robuste Entscheidungsgrundlage für gesellschaftliche Interventionen gegen das Virus zu bilden? Sie müsste vier Prinzipien verpflichtet sein, von denen erschreckenderweise derzeit keine einzige Beachtung findet. Die Prinzipien lauten: Inklusivität, Gründlichkeit, Transparenz, und Zugänglichkeit.

Inklusivität bedeutet, dass alle verfügbaren Quellen der Evidenz und Expertise systematisch Berücksichtigung finden müssen. Im konkreten Fall also nicht nur aus der Virologie, sondern auch aus der Epidemiologie, der Immunologie, der Hygiene, und natürlich auch aus relevanten nicht biomedizinischen Domänen. Die Qualität der vorhandenen Evidenz aus Studien muss mittels klar formulierter Validitätskriterien objektiv bewertet werden. *Gründlich* wird die Beratung, wenn sie Limitationen, Verzerrungen (Biase) und Interessenkonflikte, welche der Wissensbasis zugrunde liegen, möglichst vollständig aufzeigt und diese minimiert. *Transparent* ist Beratung, wenn ihr Auftrag und ihre Aufgabenstellung klar formuliert wird. Annahmen, Limitation, Unsicherheiten, offene Fragen müssen klar herausgestellt werden. Potentielle Konflikte, welche persönlich, politisch, kommerziell oder organisatorisch sein können, müssen offengelegt und kontrolliert werden. *Zugänglich* sind frei für jedermann verfügbare Beratungsergebnisse, die in verständlicher Sprache formuliert sind.

Klingt eigentlich alles einfach und einleuchtend, aber geht sowas in Zeiten einer potentiellen massiven Bedrohung überhaupt? Insbesondere wenn alles ganz schnell gehen muss? Das gelänge insbesondere dann, wenn man auf einen Notstand, wie er zum Beispiel durch einen bisher unbekannten Virus ausgelöst wird, gut vorbereitet wäre. Dann würden zumindest Strukturen vorgehalten, welche in kürzester Zeit die Etablierung eines solchen Beratungsgremiums, natürlich angepasst an die Spezifika der aktuellen Bedrohung, erlauben. Außerdem könnte man mit einer akuten Evidenzsynthese beginnen, die noch nicht allen oben genannten Kriterien entspricht, aber im Lauf der Zeit weiter optimiert wird. Mittlerweile ist bei Corona deutlich mehr als ein halbes Jahr vergangen, und noch nichts ist in diese Richtung passiert.

Vielleicht werden Sie an dieser Stelle einwenden, dass wir das doch alles schon haben. Experten die sich äußern, Drosten, Streeck et al.. Die nationale Akademie der Wissenschaften (Leopoldina), welche Empfehlungen gibt. Dazu eine Vielzahl von Fachgesellschaften und Organisation mit wohlmeinenden Analysen und Vorschlägen. Ja man könnte sogar befürchten, es gäbe zu viel, und nicht zu wenig Politikberatung. Aber der momentan verfügbare Rat, und dessen Einfluss auf Entscheidungen der Politik, folgt keiner der oben genannten Prinzipien. Er ist im wesentlichen Eminenz-basiert, denn er kommt von ‚führenden Virologen', einer Nationalen Akademie', nicht aus einer systematischen Analyse. Dabei bleibt völlig intransparent, welche Experten mit welchen Argumenten gehört wurden - und welche nicht. Welcher Wissenschaftler oder Gruppierung hat wann und warum Zugang zur Politik? Welche Meinungen oder Befunde haben letztendlich Eingang gefunden in politisches Handeln? Welches Wissen fehlt ganz besonders und wo müsste systematisches Vorgehen und Studien deshalb priorisiert werden? Es kann wegen der gegenwärtigen Intransparenz des Vorgehens auch nur vermutet werden, dass die Beratung nicht inklusiv war. Es scheint aber durch, dass wesentliche Gewerke der Wissenschaft in diesem Diskurs, sollte es überhaupt einen gegeben haben, gar nicht beteiligt wurden. Auch an der Gründlichkeit muss zumindest stark gezweifelt werden.

Gerade was Robustheit, Kontrolle von Bias und Interessenkonflikten und dergleichen betrifft hat zumindest die Biomedizin ja schon im Normalbetrieb ihre Schwierigkeiten.

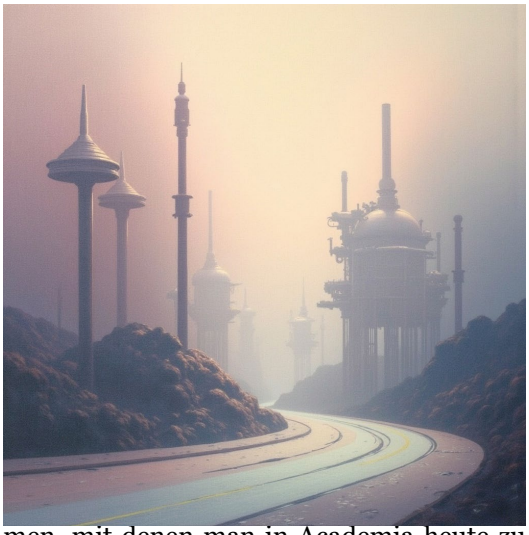
Aber belegt nicht der bisherige Verlauf der Pandemie in Deutschland, insbesondere im Vergleich zu anderen Industrienationen, dass Wissenschaft und Politik hierzulande alles richtig gemacht haben? Während der Narr über diesen Zeilen brütet, steigen die Infektionszahlen in Deutschland wie anderswo massiv. Gerade wurde bei uns eine Sperrstunde eingeführt. Auf welcher Basis? Gehen die Leute dann aus den Kneipen nach Hause und stecken sich dort bei privaten Nachfeiern an? Wird der Virus nach 23 h erst besonders gefährlich? Im den Medien treten Virologen auf, welche diese Maßnahme mit dem Brustton der Überzeugung verurteilen, kurz darauf andere welche sie verteidigen. Welche Evidenz gibt es zu solchen Maßnahmen? Wurde Sie berücksichtigt? Und welche Evidenz gibt es zu den Beherbergungsverboten? Schulschließungen? Uns so weiter…

Warum haben wir nicht systematisch versucht, die fehlende Evidenz in den zurückliegenden Monaten zu schaffen? Die Wirksamkeit einer Sperrstunde ist ein klassisches Beispiel für einen ‚Evidence gap', also eine Lücke in der Beurteilung die man versucht zu schließen, sobald man sie identifiziert hat. Das gleiche gilt zum Beispiel für die Frage, was eigentlich passiert, wenn man Patienten nicht mehr aufnimmt oder nach Hause schickt, um Betten für COVID-Erkrankte freizuhalten. Vor 8 Monaten war es noch besserwisserisch, solche Fragen zu stellen – denn wir wussten praktisch nichts über das Virus, seine Infektiosität, Morbidität und Mortalität und Ausbreitungsdynamik. Mittlerweile gibt es weltweit über 37 Millionen bestätigte Fälle und über 1 Million Tote. Eine Pubmed Suche mit dem Term „COVID" ergibt über 60.000 Treffer. (Stand 10.10.2020). Rationelle wissenschaftliche Politikberatung hätte viel früher die relevantesten Wissenslücken identifizieren und die Politik zur Bereitstellung von Mitteln für deren Überwindung durch qualitativ hochwertige Forschung drängen müssen.

Es gibt wenig Hinweise darauf, dass sich eine evidenzbasierte, inklusive, gründliche, transparente, und zugängliche wissenschaftliche Beratung der Corona-Politik in den kommenden Monaten doch noch einstellen könnte. Dass sowas möglich wäre, beweist sehr schön eine private Initiative, welche alle eben genannten Kriterien erfüllt (Thesenpapier 4.0, link siehe http://dirnagl.com/lj), aber eben keine Beachtung findet, da sie keine nationale Corona-Evidenz Task Force ist. Was bleibt ist die Hoffnung, dass spätestens nach dem Ende der Pandemie der Beschluss gefasst wird, für die nächste Krise – und die kommt bestimmt – besser gerüstet zu sein. Hierzu werden dann nicht nur genug Masken und Beatmungsgeräte gehören. Sondern eine ganz grundlegende Neuorganisation des Verhältnisses von Wissenschaft und Politik in Krisenzeiten.

Wie konnte es eigentlich soweit kommen?

LJ-12/2020



Covid? Trump? Nein, denn diesmal soll es, dies auch zur vorweihnachtlichen Entspannung, um die Frage gehen, wieso heutzutage eigentlich wissenschaftliche Karrieren ganz wesentlich vom Journal Impact Factor (JIF) abhängen. Und der Einwerbung von möglichst vielen Drittmitteln. Oder allgemeiner ausgedrückt, warum Inhalte, Originalität und Verlässlichkeit von Forschungsergebnissen oft eine Nebensache sind, wenn sich Kommissionen die Köpfe darüber heiß reden, wen man in die eigenen Reihen aufnehmen will. Und wen nicht. Oder welche Anträge es verdienen, gefördert zu werden. Kurzum, folgen Sie mir auf eine kurze und unvollständige Geschichte der Mechanismen, mit denen man in Academia heute zu etwas kommt. Vielleicht ergeben sich aus dieser historischen Perspektive auch Hinweise, wie wir dem Schlamassel, in dem wir uns befinden, wieder entkommen können. Doch ich eile voraus. Beginnen wir dort wo alles begann, bei den Gründungsvätern der modernen Wissenschaft.

Die frühen Pioniere modernen wissenschaftlichen Arbeitens wie Galileo, Hooke, Boyle, Newton waren ‚Gentlemen scientists'. Nicht nur waren die ausnahmslosen Männer, sie waren auch alle finanziell unabhängig. Entweder per Geburt, oder durch Mäzenatentum. Getrieben von der Neugier ‚wie die Welt funktioniert' war ihr Ziel natürlich nicht nur Wissen zu produzieren, sondern Ruhm und Ehre zu erlangen. Der Nutzen des so erworbenen Wissens wurde dabei nicht darin gesehen, die Grundlagen zu schaffen für eine rationellere Aneignung der Natur durch den Menschen. Weit gefehlt, es ging diesen allesamt tief religiösen Herren ganz wesentlich darum, das von Gott geschriebene Buch der Natur und damit die Ordnung der Welt zu dechiffrieren und dadurch besseren Glauben und gottesfürchtigeres Verhalten zu befördern. Wissenschaft war Gottesdienst. Von Fürsten und Königen wurden damals nicht die Wissenschaftler, sondern die Erfinder und Ingenieure gefördert, denn nur sie versprachen Hilfe dabei, sich die Welt durch Eroberung und Krieg Untertan zu machen.

Der Umgang, den Newton und Kollegen mit ihrer Konkurrenz pflegten, war häufig allerdings alles andere als Gentleman-like. Ging es doch um Primat und Posterität. Ausgangspunkt ihrer Ideen und Hypothesen war das, was Lorraine Daston 'Ground zero empiricism' nannte. Mit anderen Worten, sie schrieben auf ein fast leeres Blatt. Die Community von Forschern war sehr übersichtlich, vielleicht ein paar Hundert, maximal ein paar Tausend Gleichgesinnter weltweit. Lose organisiert in Akademien, in denen man sich gegenseitig Theorien und Experimente vorstellte und kritisierte. Publiziert wurde neben Büchern hauptsächlich in den Annalen der nationalen wissenschaftlichen Akademien. Die Royal Society Englands war dabei führend in Geschwindigkeit und Reichweite: Zweimal im Jahr wurden Exemplare gedruckt, z.B. im Jahr 1829 800 Stück, und an korrespondierende Akademien und ausgewählte Wissenschaftler versandt. Häufig verging dabei nicht mehr als ein halbes Jahr zwischen Vortrag bzw. Einreichung und

Veröffentlichung. Konkurriert wurde damals natürlich nicht um Stellen oder Forschungsförderung, sondern um Reputation und Zugang zu diesen Akademien und deren internationaler Korrespondenz. Neben der Originalität und Güte der Wissenschaft dürften hier sicher auch damals schon Hierarchien, Beziehungen und Machtspiele wichtig gewesen sein.

Mit dem zunehmenden Verständnis dessen, was die Welt im Innersten zusammenhält, begann man sich aber auch vermehrt für die Nützlichkeit der wissenschaftlichen Erkenntnisse zu interessieren. Als sich bürgerliche Gesellschaften etablierten und die Industrialisierung im 18. und 19. Jahrhundert aufblühte, begannen Staaten, Wissenschaft systematisch zu organisieren, und diese insbesondere über Universitäten zu fördern. Maxwell, Pasteur, Virchow usw. waren universitäre Brotwissenschaftler, welche staatlich alimentiert forschten. Auch ihnen ging es nicht um Reichtum, sondern immer noch vorrangig um Fortschritte im Wissen, und die darüber zu erlangende Anerkennung und Ruhm.

Gleichzeitig spezialisierten sich die Wissenschaften mehr und mehr, Fachjournale kamen auf und wurden neben Vorträgen zum wichtigsten Medium des wissenschaftlichen Diskurses. Noch kannten sich alle Wissenschaftler eines Gebietes. In Wort und Schrift focht man wissenschaftliche Kontroversen nicht anonym, sondern von Angesicht zu Angesicht aus. Neu war allerdings die akademische Konkurrenz um die Anstellung als Assistent, oder Berufung und Verstetigung als Professor. Wichtig waren dabei vor allem die Reputation unter den Kollegen, aber natürlich auch akademische Hierarchien und Zugehörigkeiten zu ‚wissenschaftlichen Schulen'. Quantitative bibliometrische Indikatoren oder Drittmittel spielten auf jeden Fall keine Rolle, denn die gab's damals ja noch nicht. Auch nahm man es auch damals mancherorts schon nicht so genau mit der guten wissenschaftlichen Praxis, wenn es nur dem akademischen Fortkommen diente. Charles Babbage, der von der Rechenmaschine, beschrieb 1830 in seinen ‚Reflections on the Decline of Science in England, and on Some of Its Causes' die wesentlichen auch heute noch praktizierten Spielarten der unsauberen Wissenschaft. Er unterschied dabei ‚Hoaxing' (Fabrizieren), ‚Forging' (Fälschen), Trimming' (selektive Datenanalyse) und ‚Cooking (unsauberer Satistik).

Im frühen 20. Jahrhundert kamen dann die Drittmittel dazu. Unmittelbar nach dem verlorenen ersten Weltkrieg hatten die deutschen Universitäten, Akademien und die Kaiser-Wilhelm Gesellschaft (heute Max Planck Gesellschaft) eine Idee, wie sie ihre durch Krieg und Krise klamme Finanzsituation aufbessern könnten. Sie gründeten die ‚Notgemeinschaft der deutschen Wissenschaft' (deren Rechtsnachfolger bekanntermaßen die DFG ist) und konnten so auf Antragsbasis individuelle Wissenschaftler fördern. Aber dies lief ganz anders ab als heute. Erhalten hat sich der Antrag von Otto Warburg bei der Notgemeinschaft. Er bestand aus einem einzigen Satz: ‚Benötige 10.000 Reichsmark' und darunter ‚gez. Otto Warburg'. Er wurde vermutlich genehmigt, aber nicht nach Begutachtung. Der Name Warburg war ausreichend. Ein paar Jahre später wurde dann auch noch das Parteiabzeichen wichtig. In den Zeiten einer ‚Deutschen Physik' war Gesinnung und Parteizugehörigkeit natürlich auch für die Einstellung oder Berufung an der Universität ein wesentliches Kriterium. Der JIF und die Drittmittel aber immer noch in weiter Ferne!

Erst durch den 2. Weltkrieg änderte sich dieses System ganz grundsätzlich, und zwar weltweit. Während des Krieges kam es nämlich zu einer bisher ungekannten Industrialisierung der Forschung, am konsequentesten in den USA. Forschungsprogramme, welche die Grundlagen lieferten zur Entwicklung von Langstreckenraketen, RADAR, Atombombe, Computern usw. wurden mit gigantischen Summen ausgestattet und

generalstabsmässig exekutiert. Am Ende des 2. Weltkrieges war der Großteil der universitären (Natur)Wissenschaft im Dienste des Militärs. Nützlichkeit der Forschung, hier zur Sicherung militärischer Überlegenheit, hatte oberstes Primat. So sehr, dass man sich damals um das Überleben der ‚Blue Skies' Grundlagenforschung ernsthaft Sorgen machen musste. Heute noch viel gelesen und zitiert wird Vannevar Bush's Bericht ‚Science – the last frontier'. Im Auftrag des amerikanischen Präsidenten 1945 erstellt, gilt es auch heute noch als Manifest des staatlichen Auftrages, Forschung auch um ihrer selbst Willen zu fördern. Denn die Grundlagenforschung liefert das Wissen für spätere, noch nicht antizipierbare Anwendungen. Auch schrieb Bush dem Staat ins Stammbuch, für wissenschaftlichen Nachwuchs zu sorgen, und sich inhaltlich bei all dem möglichst raus zu halten.

Diese Entwicklungen katalysierten durch immer weiter zunehmende Spezialisierung der verschiedenen Disziplinen sowie steigende Staatsausgaben für akademische Forschung einen steil ansteigenden Forschungs-Output. Trotzdem war das für die Forscher in ihren Spezialgebieten und sogar darüber hinaus noch immer alles recht überschaubar. Editoren entschieden auf ihren Schreibtischen über die Publikation von Manuskripten, der Peer Review wie wir in kennen, war noch nicht geboren. Pro Fach gab es nur einige wenige Journale, publiziert in den jeweiligen Landessprachen. Man tauschte sich noch vor allem auf nationaler Ebene aus, dort wurde auch entschieden, wer ‚exzellent' ist, und wer nicht.

Irgendwann, so etwa in den 80er Jahren des vorigen Jahrhunderts, erreichte die exponentielle Wissensproliferation, deren Spezialisierung, und die schiere Menge von ‚Wissensproduzenten' aber eine kritische Schwelle. Es wurde immer schwieriger, auf Basis der Kenntnis der Inhalte Qualität und Originalität von Forschern zu beurteilen und Förder- und Karriereentscheidungen zu treffen. Dazu kam wohl noch die in den späten 60ern allgemein einsetzende Auflehnung gegen verstaubte Hierarchien. Der Wunsch nach Objektivierung und Quantifizierung von Leistung, auch in der Forschung, war geboren. Mittlerweile hatte sich in der Folge dieser Entwicklungen auch eine Hierarchie der Journale etabliert, die durch Eugene Garfield's geniale Erfindung des Impact Factors 1955 quantifizierbar wurde, und von ihm (und den Verlagen) folgerichtig auch massiv kommerzialisiert wurde.

Der Rest ist Geschichte. Laut UNESCO gibt es allein in Deutschland mittlerweile mehr als 400.000 Vollzeitwissenschaftler, auf der Welt viele Millionen. Welcome to the club! Diese Wissenschaftler publizieren nun jährlich Millionen von Artikeln. Innerhalb eines Jahrhunderts ist die mittlere Anzahl von Autoren von 1 auf 6 angestiegen. In diesen hundert Jahren ist aber auch die Produktivität von Wissenschaft, definiert als das Verhältnis von Output an Wissen zu Input in die Wissenschaft, stark zurückgegangen. Es geht dennoch voran, denn die Zahl der Wissenschaftler (Input!) hat parallel etwa um denselben Faktor zugenommen, vermutlich sogar überproportional (Zitate für all dies wie immer unter http://dirnagl.com/lj). Wir wissen nämlich schon recht viel, gute Ideen sind rarer geworden, die niedrig hängenden Früchte sind gepflückt, alles wird immer komplexer – Inhalte wie Methoden. Damit es weiter vorwärts geht, braucht es immer mehr Wissenschaftler, und immer kompliziertere und teurere Apparate um der Natur ihre Geheimnisse zu entreißen.

Der anschwellende akademische Massenbetrieb der letzten Jahrzehnte bot dabei auch ein ausgezeichnetes Substrat für die Perfektionierung objektiver, einfacher, und transparenter Kriterien zur Beurteilung von Forschern und Forschung: JIF, HirscHirsch-Faktor, Drittmittel. Wozu Artikel von Bewerbern oder Antragstellern lesen, wenn man weiß, dass deren Impact Factor im Mittel bei 20,162 liegt? Oder eben ‚nur' bei 6, 531? Man

beachte auch die Genauigkeit des Indikators: In den meisten Lebensläufen und Anträgen wird er mit 3 Nachkommastellen angegeben!

Abgesehen davon, dass diese Objektivierung der Güte von individueller Wissenschaft auf falschen Prämissen beruht: Der JIF misst, wenn irgend etwas, die Popularität des jeweiligen Journals und Faches. Außerdem: 80% der Zitate in Nature (und vergleichbar) werden von 20 % der Artikel (inklusive Reviews) erwirtschaftet. Die überwiegende Mehrheit der Artikel dieser auch als ‚Glamjournals' bezeichneten Zeitschriften zieht nicht mehr Zitationen als die in einer guten Fachzeitschrift veröffentlichten. Oder eben auch gar keine. Noch korrosiver als diese Ungeeignetheit der Metriken war allerdings, dass nun zwei schon lange bekannte Phänomene wirksam werden konnten. Zum einen *Goodhart's Gesetz*, formuliert im Jahr 1975, das vorhersagt, ‚dass ein Maß, das zum Ziel wird, aufhört ein gutes Maß zu sein'. Und genau das ist passiert. Das Schürfen von Impact Factor Punkten begann das erkenntnisgeleitete Interesse zu korrumpieren. Immer mehr Papers müssen immer mehr Punkte erzeugen. Forschungsergebnisse die solche Punkte verheißen, werden priorisiert. Mit allen Konsequenzen, von der geschickten Auswahl und Überinterpretation von Ergebnissen bis hin zum Betrug. Babbage lässt grüßen. Dazu kommt dann noch der *Matthäus Effekt*, für die Wissenschaft zum ersten Mal von Robert Merton 1968 formuliert: ‚Wer hat, dem wird gegeben'. So steht's schon in der Bibel. Drittmittel erzeugen Drittmittel. Und der Mainstream feiert fröhliche Urstände. Science Paper erzeugen Nature Paper, und umgekehrt. Natürlich kann nicht jeder sowas kriegen, denn die ‚Währung', für welche die Impact Punkte den Umtauschkurs festlegen, werden von den Verlagen über Ablehnungsquoten gesteuert. Das ist ihr Geschäftsmodell. Die über 10.000 Max Planck Wissenschaftler, die deutsche Forscherelite also, schaffen es nicht mehr als 400 Artikel jährlich in Nature und Nature-brand Journalen zu platzieren!

Die besondere Attraktivität, aber auch Toxizität dieser Indikatorik besteht in ihrer scheinbaren Plausibilität, Transparenz, Simplizität und Praktikabilität. Und der Tatsache, dass die offensichtliche Alternative, die Auseinandersetzung mit wissenschaftlichen Inhalten und deren Qualität und Originalität, in Anbetracht der oben geschilderten Paper- und Wissenschaftler – Tsunami alternativlos erscheint. Sie hat sich deshalb weltweit durchgesetzt. Mindestens eine Generation von Wissenschaftlern und Administratoren wurde damit bereits sozialisiert – sie können sich andere Mechanismen oft gar nicht mehr vorstellen. Die Beurteilung der Originalität und Qualität von Wissenschaft und deren Produzenten auf Basis von Zitierraten und Reputation von Journalen, oder der Akkumulation von Drittmitteln erscheint ihnen als etwas Natürliches. Weil es, wie oben beschrieben, evolutionär als Antwort auf den Erfolg, man könnte auch sagen, die ‚Industrialisierung' von Wissenschaft entstanden ist.

Es stellt sich also die Frage, ob Wissensproduktion im 21. Jahrhundert, mit ihrer Armada von Wissenschaftlern und der schieren Masse ihrer Outputs andere Kriterien der ‚Leistungsbewertung' braucht? Und wenn ja, ob es denn überhaupt andere gäbe, und ob diese dann auch noch praktikabel wären? Wer diese Kolumne schon mehrfach gelesen hat, wird ahnen, dass der Wissenschaftsnarr hierzu klare Vorstellungen hat. Die wird er der verehrten Leserschaft dann im nächsten Heft vorstellen!

Back to the future: Von industrieller zu inhaltlicher Forschungsbewertung

LJ 1-2/2021

Wissenschaft verschlingt massiv gesellschaftliche Ressourcen, nicht nur finanzielle. Insbesondere für die akademische Forschung, welche sich selbst verwaltet und sich gerne auf die im Grundgesetz verankerte Forschungsfreiheit beruft, stellt sich damit die Frage, wie sie die ihr von der Gesellschaft zur Verfügung gestellten Mittel alloziert. Es gibt keine natürliche Begrenzung, wieviel geforscht werden könnte – aber sehr wohl eine Beschränkung der Mittel, welche die Gesellschaft für Forschung einsetzen kann und will. Welche Forschung soll also gefördert, welche Wissenschaftler ins Brot gesetzt werden?

Für die Beantwortung dieser für den akademischen Betrieb zentralen Fragen haben sich über viele Jahrzehnte Mechanismen evolutionär entwickelt. Diese Mechanismen steuern aber nicht nur die Verteilung der Mittel in und zwischen den Institutionen, sondern letztendlich auch die Inhalte und die Qualität der Forschung. Den individuellen Wissenschaftlern, welche in Academia nicht nur ihrem Forscherdrang nachgehen, sondern auch ihren Lebensunterhalt verdienen, geht die in Fleisch und Blut über. Sie halten das für so etwas wie eine natürliche Ordnung. Die Verteilungs- und Leistungsbewertungsmechanismen und die dazugehörige Indikatorik bestimmen ihren Tagesablauf und die Art und Weise wie sie forschen mehr als die tägliche Lektüre von Fachliteratur, der Blick durchs Mikroskop, oder Kongressvorträge. Auch wenn das den Wenigsten wirklich bewusst ist.

In der zurückliegenden Folge hat der Wissenschaftsnarr die Frage gestellt, wie sich das heute weltweit durchgesetzte Karriere-, Belohnungs- und Begutachtungssystem entwickeln konnte, von den Anfängen der modernen Wissenschaft im 17. Jahrhundert bis heute. Wie es dazu kommen konnte, dass quantitative Indikatoren wie Journal Impact Factor (JIF) und Höhe der Drittmitteleinwerbung bei der Beurteilung von Forschern und deren Anträgen eine wichtigere Rolle spielen als die Inhalte, Relevanz, oder Qualität der Forschung. Frei nach dem Motto: ‚Sag mir deinen JIF, und ich sage Dir, ob Du geoder befördert wirst'.

Dabei stellte sich heraus, dass dieses Beurteilungssystem nur wenige Dekaden alt ist. Vermutlich ist erst eine Generation von Wissenschaftlern komplett in ihm sozialisiert worden. Das System erwuchs im Wesentlichen aus zwei Entwicklungen: Zum einen der Industrialisierung und massiven Ausweitung von akademischer Forschung. Diese wiederum ist das Resultat ihres eigenen immensen Erfolgs, aber auch der durch diesen Erfolg gleichzeitig abnehmenden Effizienz von Forschung geschuldet. Denn weil die ‚Früchte der Erkenntnis' immer höher hängen, benötigt es immer größeren Einsatz an Forschung um den Erkenntnisgewinn pro eingesetzte Mittel konstant zu halten, wenn nicht zu steigern. Zusammengenommen führt dies zu einer immer weniger beurteilbaren Flut von Forschern, Projekten, Anträgen und Artikeln. Um da noch

durchzukommen, benötigen wir einfach und schnell zu erhebenden Bewertungskriterien. Am besten welche, die man anwenden kann, ohne sich die Mühe zu machen, die eigentlichen Inhalte der Wissenschaft zu beurteilen.

Die zweite wesentliche Triebfeder der Entstehung des heutigen Beurteilungssystems ist der nachvollziehbare Wunsch nach Verteilungsgerechtigkeit. Wir wünschen uns objektive Kriterien, welche reproduzierbar sind, und nicht der Willkür unterworfen. Und eine eindeutige Diskriminierung zwischen Bewerbern oder Anträgen erlauben, am besten sogar ein einfaches Ranking. Niemand soll gefördert werden, weil sich ein mächtiger Mentor hinter den Kulissen eingemischt hat. Sondern weil man etwas geleistet hat, das jeder nachvollziehen kann. Und schon sind wir beim JIF und den akkumulierten Drittmitteln. Einfach, objektiv, quantifizierbar, nachvollziehbar. Man muss weder einen Artikel eines Kandidaten gelesen haben, noch den ganzen Lebenslauf. Ein Blick auf die Literaturliste (und für die Doofen den mit 2 oder 3 Nachkommastellen angegebenen JIF dahinter, denn die Namen der Journale sind ja schon ausreichend) und die Aufreihung der Drittmittel reichen völlig. Wer Übung hat, und häufiger in Kommissionen sitzt oder begutachtet, schafft das locker in 3 Minuten pro Kandidaten. Und hier liegt der Hase im Pfeffer: Wie effizient und nützlich kann ein Verteilungs- und Bewertungssystem sein, dessen Mess- und Steuergrößen sich von den Inhalten, der Relevanz und der Qualität der zu steuernden Forschung verabschiedet haben? Nicht nur der Wissenschaftsnarr schlägt hier Alarm. Die Spatzen pfeifen es von den Dächern, dass es so nicht weitergehen kann.

Wie aber könnte man es besser machen? Ist ein System überhaupt reformierbar, das sich weltweit durchgesetzt hat, und doch offensichtlich funktioniert – Stichwort CRISPR, SARS-COV-2 Vakzine, et cetera? Nur ein Narr kann sich erlauben, diese Frage mit einem klaren Ja zu beantworten, und gleich noch ein paar praktische Vorschläge hierfür zu geben. Vorneweg erst mal drei Prämissen: 1. Forschung kann nur beurteilen, wer sich kompetent und konkret mit deren Inhalten, Methoden, Ergebnissen, und Interpretationen auseinandersetzt. Das ist natürlich sehr unangenehm, denn das ist aufwendig, kann nicht automatisiert werden, und ist nicht quantifizierbar. 2. Wir dürfen den gesellschaftlichen Einsatz an Ressourcen für Forschung nicht reduzieren, sondern müssen die vorhandenen Ressourcen effizienter einsetzen. Man könnte nämlich darauf kommen, einfach weniger Forschung zu fördern. Damit könnte man den Output so weit reduzieren, bis dieser wieder inhaltlich bewertbar würde. Damit würde man die Uhr allerdings mehr als 100 Jahre zurückstellen, und eine wissenschaftliche Eiszeit induzieren, das geht natürlich gar nicht. 3. Die nötigen Veränderungen im Bewertungs- und Verteilungssystem müssen von außen kommen, also von den staatlichen Fördereinrichtungen, den Hochschulen, den außeruniversitären Wissenschaftsorganisationen. Die Wissenschaftler, welche ihren Weg in Academia suchen, haben keine andere Wahl als sich den Bedingungen der Konkurrenz um Fördermittel und berufliche Positionen zu stellen. Sie sind ja das Objekt der Bewertungsmechanismen.

Womit könnte man beginnen? Um eine inhaltliche und qualitative Bewertung von Forschungsleistungen durchzusetzen, müsste man zunächst ganz einfach die Verwendung von abstrakten Indikatoren (JIF, Drittmitteleinwerbung etc.) gezielt verhindern - und nicht nur deren sparsamen und nur unterstützenden Gebrauch anmahnen! Dies bedeutet konkret: Verbot der Angabe von JIF, HirscHirsch-Faktor, etc., dafür obligatorische Verwendung von Narrativen zur Beschreibung des eigenen Beitrages. Zitationen eigener Arbeiten in Lebensläufen und Anträgen nur unter Angabe von deren Titel, Autoren, und eines Identifiers wie z.B. der PMID (Pubmed Identifier), aber nicht des Namens des Journales. Damit kann die Referenz aufgerufen und gelesen werden, ein bloßes Durchsuchen von Literaturlisten nach Journalnamen ist dann aber nicht mehr möglich. Diese

Kurznarrative führen auch ganz natürlich zu einer Beschränkung auf wenige relevante Literaturstellen. Denn wer wollte schon mehr als 10 oder mehr davon schreiben?

Eine Fokussierung auf Erst- und Letztautorpositionen ist dann auch nicht mehr nötig und sollte ganz fallen. Denn es handelt sich hierbei ohnehin um eine potentiell schädliche Fiktion: Heutzutage liefern zu jeder relevanten biomedizinischen Arbeit eine Vielzahl von Wissenschaftlern verschiedenartigste Beiträge. Diese lassen sich nicht auf 2 Positionen in der Autorenleiste reduzieren, die noch dazu gar nicht eindeutig definiert sind. Die geteilte Autorschaft mit Sternchen ist der alberne Versuch, sich um diese Erkenntnis herumzumogeln. Damit zusammenhängend müssten auch die Mindestanzahlen von Publikationen fallen, welche für Promotion, Habilitation etc. derzeit gefordert werden. All dies führt zum ‚Slicing' von Studien in kleinere Einheiten, zur Inflation von Publikationen die keiner braucht, zu unsinnigen und unnötigen Diskussionen um Autorenpositionen etc. Stattdessen sollte ein Narrativ den wissenschaftlichen Beitrag individueller Wissenschaftler darlegen. Ob dieser dann Promotions- oder Habilitations-würdig ist, müssen die zuständigen Kommissionen in einer inhaltlichen Auseinandersetzung mit dem Oevre der Kandidaten entscheiden, aber nicht wie derzeit üblich aus dem Studium eines Spreadsheet-Rankings ableiten.

Bei dieser Gelegenheit sollte man dann gleich versuchen, zu alphabetischen Autorenleisten überzugehen, wie in den Multiautorenkollaborationen der Hochenergiephysik schon lange erfolgreich praktiziert. Dafür gibt es bereits eine hervorragende Taxonomie (https://casrai.org/credit/), die sich auch für die Lebenswissenschaften eignet. Die Reputation und das Renommee von Wissenschaftlern ergeben sich doch aus ihren inhaltlichen Beiträgen und ihrem ‚Standing' in der Community. Hierbei sollten dann auch Reviews und Beurteilungen durch Peers Berücksichtigung finden, welche nach Publikation entweder bei den Journalen (z.B. nach Post-Publication-Review) oder aber auch auf sozialen Medien publiziert werden. Der ‚Science Twitter' ist in vielen Feldern bereits heute wesentlich transparenter, nachvollziehbarer, aktueller und damit wissenschaftlicher als althergebrachte Formate (Letter to the Editor, etc.) des Diskurses. Die Qualitätskontrolle wissenschaftlicher Publikationen findet ohnehin effizienter in den sozialen Medien statt als im klassischen Peer Review. Ein Beispiel hierfür ist, dass die Qualitätsprobleme, welche zur Retraktion von Papers aus hochrangigen Journalen führen, bereits seit einiger Zeit zuerst auf Twitter oder in Blogs exponiert werden. Und vorher im Peer Review regelhaft übersehen wurden.

Die oben genannten Maßnahmen würden bereits zu einer massiven Reduktion der Artikelflut führen, wodurch eine Auseinandersetzung mit deren Inhalten erleichtert wird. Inhalte und Qualität können dann auch wissenschaftliche Reputation und Renommee bestimmen, nicht Proxies wie JIF und Drittmittel. Es fehlt aber noch etwas Wesentliches: Die Karrierewege in Academia müssen sich ändern. Dreiundachtzig Prozent des wissenschaftlichen Personals sitzt auf befristeten Stellen! Der immense Konkurrenzdruck, nicht aus dem System ausgebucht, oder es gar von der Basis der Pyramide zur Spitze zu schaffen, führt zur Selektion von Eigenschaften, welche weder förderlich für Qualität noch für Kooperation in der Wissenschaft sind. Die Pyramide muss zu einem Trapez geformt werden. Dabei muss die Spitze flacher, und die Basis etwas weniger breit werden. Das bedeutet aber auch, weniger PhD-Studenten (als ‚billige' Arbeitskraft) ins System einzuschleusen als bisher. Wer den nicht unbeschwerlichen Weg in die akademische Berufswelt nimmt, muss die reelle Chance haben, durch gute Wissenschaft (und nicht nur durch 3 ‚Top-Publikationen') langfristig ein Auskommen zu haben.

Und nun setzt der Narr zum letzten Schlag an: Nach Einführung eines rein inhalts- und qualitätsorientierten Bewertungssystems und der Kappung der akademischen Karriere-

Pyramide fehlt nun noch als drittes Element der Zufall! Da echte Innovation nicht vorhersagbar ist, und jeder wie auch immer geartete Begutachtungsprozess tendenziell den Mainstream begünstigt, sollte ein Teil der Förderung in Lotterien vergeben werden. Ja, Sie haben richtig gelesen, verlost werden! Wer Näheres dazu wissen will, dem sei ein Blick in einen früheren Wissenschaftsnarren empfohlen (LJ 04-2019). So ein Vergabemodus würde uns auch mehr Zeit zum Forschen lassen, weil ein Teil der Antragschreiberei, und der Begutachtung derselben, wegfallen würde. Da würde Vieles gefördert werden, was eher mittelmäßig ist und nicht den versprochenen Durchbruch bringt. Wie jetzt auch. Die Wahrscheinlichkeit aber, dass mal was bahnbrechend Neues gefördert würde, steigt erheblich.

Aber wie realistisch sind solch närrische Gedankenspiele eigentlich? Kaum zu glauben, aber die DFG arbeitet derzeit an einem Positionspapier, das, abgesehen von der Lotterie, all das und mehr umzusetzen empfiehlt, was der Narr sich da weiter oben so zusammengesponnen hat. Und das ist nicht nur eine der üblichen ‚Denkschriften', mit denen man zeigt, dass man ein ‚Problem' erkannt hat, man sich viel vornimmt aber wenig tut, weil alles so furchtbar komplex ist. Die DFG empfiehlt sich darin selbst, gleich und ganz konkret mit der Umsetzung zu beginnen! Wenn dies verabschiedet wird, reiht sich nun endlich auch unser wichtigster Forschungsförderer ein in die weltweite Riege der Fördergeber und Institutionen (z.B. Wellcome Trust, ZonMW), die es ernst meinen damit, dass es so nicht weiter gehen kann. Sogar die Lotterie wird übrigens mancherorts schon ausprobiert, zum Beispiel bei der Volkswagenstiftung. Vielleicht ist also gar kein Fluxkompensator mehr nötig, und die Zukunft hat schon (ein bisschen) begonnen?

Im Kampf gegen Corona heißt von Botswana siegen lernen!

LJ 3/2021

Solange die Mehrheit der Bevölkerung nicht immun gegen SARS-COV-2 ist, muss das Gesundheitssystem vor dem Kollaps durch Überlastung mit COVID Patienten geschützt werden. Seit einem Jahr erproben wir daher mit einigem Erfolg Maßnahmen, welche von verstärktem Händewaschen bis hin zum totalen Lockdown reichen. Dabei werden Maßnahmen eingeführt, verschärft, gelockert, oder abgeschafft, um dann wieder eingeführt zu werden, … und so geht's dahin. Die Politik begründet ihr Vorgehen mit Inzidenzwerten, Auslastung von Krankenhäusern, Modellrechnungen und dem Rat von Experten (hierzu auch der Wissenschaftsnarr in LJ 11/2020). Unbestritten haben viele dieser (Anti)Corona-Maßnahmen enorme Plausibilität. Auch ist die Einsicht trivial, dass ein totaler Lockdown die Verbreitung eines Virus stark einschränken kann. Der ist aber nicht ewig durchzuhalten. Ungemein relevant ist deshalb die Frage, welche der Maßnahmen aus der Blackbox Lockdown Wirkung haben, und bei welchen der Schaden den Nutzen überwiegt. Man könnte mit diesem Wissen ein evidenzbasiertes Paket von Corona-Maßnahmen

schnüren, das weniger drastisch ist als der Lockdown, aber genauso effektiv. Und vielleicht so auch manchen Skeptiker zum Mitmachen bewegen. Deshalb ist die Frage, welche Evidenz wir für die Wirksamkeit einzelner Maßnahmen haben so wichtig. Aber Vorsicht! Die Frage nach der Evidenz von Corona-Maßnahmen ist mittlerweile recht gefährlich geworden. Denn man läuft Gefahr, sofort ins Lager der Virus-Leugner, Querdenker (was früher eigentlich eher ein Kompliment war) und Rechtsradikalen verortet zu werden. Oder man wird gleich als Narr abgestempelt, weshalb das Thema ja auch in mein Resort fällt. Und ich in dieser Sache unsere Aufmerksamkeit auf Botswana lenken möchte.

Wir haben eine Flut von Studien, welche die Wirksamkeit von Corona-Maßnahmen mittels statistischer Modellierung untersuchen. Und nicht ganz überraschend zu sehr unterschiedlichen Ergebnissen kommen. Denn geringfügige Veränderung von Modell-Parametern führen häufig zu ganz anderen Vorhersagen. Auch machen sich die Modellierer untereinander ihre Modelle madig. Und wer von uns würde sich zutrauen, deren Qualität, Validität und Prädiktivität einzuschätzen? Es gibt auch eine Flut von Beobachtungsstudien, welche Effekte von Corona-Maßnahmen zum Gegenstand haben. Aber solche Beobachtungsstudien liefern nur schwache Evidenz und erlauben keine kausalen Schlussfolgerungen. Was wir deshalb bräuchten, sind randomisierte, kontrollierte Studien (RCT) in denen spezifische Corona-Maßnahmen als Intervention getestet werden. RCTs sind schließlich der Goldstandard zur Überprüfung therapeutischer Interventionen in der Medizin, und deshalb ganz nebenbei auch die Grundlage für die Zulassung der Corona-Vakzinen. Nun schätzen Sie mal, wie viele RCTs zur Untersuchung der Wirksamkeit von Social Distancing Maßnahmen es bisher gegeben hat?

Ich konnte nur 3 finden, weltweit! Eine in Norwegen, in der Teilnehmer randomisiert die Nutzung von Fitnessstudios erlaubt bzw. verboten wurde. Dann die dänische Masken-Studie. Dort wurde das Tragen von Gesichtsmasken in der Öffentlichkeit untersucht. Dies zu einer Zeit in der es noch nicht Pflicht war. Die Teilnehmer wurden per Zufall in 2 Gruppen aufgeteilt, eine trug Masken, die andere nicht. Endpunkt war in beiden skandinavischen Studien, welche jeweils mehrere Tausend Teilnehmer rekrutieren konnten, das Auftreten von SARS-COV-2 Infektionen. Außerdem wurde eine sehr interessante Intervention, Sie haben es sicher schon vermutet, in Botswana durchgeführt! Auch dort wurden die Schulen wegen Corona geschlossen, und es wurde verglichen, ob sich mittels ‚low tech' Maßnahmen dennoch ein Lernerfolg erzielen lässt. Hierzu wurden Schüler in 3 Gruppen randomisiert – kein Unterricht, täglicher Kontakt mit Lehrkräften via SMS oder Anruf. Dort haben Schüler nämlich keine Smartphones oder Laptops.

Ich will nichts zu den Ergebnissen, der Qualität, oder auch zur Übertragbarkeit dieser 3 Studien auf Deutschland sagen. Es ist aber ein Skandal, dass es bisher nur Botswana geschafft hat, eine randomisierte Intervention durchzuführen, welche die Auswirkungen einer der am heftigsten diskutierten Maßnahmen untersucht, welche bisher weltweit schon 1,6 Milliarden Schüler betrifft. Und das ein Jahr nach Beginn eine weltweiten Pandemie, für die es immer noch keine spezifische Therapie gibt, nach über 100 Millionen gesicherten Infektionen und 2.3 Millionen assoziierten Toten, und einer Kakophonie von fluktuierenden, teils drastischen Maßnahmen der sozialen Distanzierung.

Statt auf randomisiert kontrollierte Studien verlassen sich unsere Modellierer und Politiker auf Beobachtungsdaten, inklusive solcher, die in der Zeit der Spanischen Grippe von 1918/19 erhoben wurden. Wäre es nicht an der Zeit zu untersuchen, ob ein totaler Lockdown von Alters- und Pflegeheimen effektiver ist als die Kombinationsstrategie negativer Virusnachweis (PCR), Schnelltests am Eingang und FFP2 Maske? Oder ob

Schulschließungen besser sind als die Kombination von Masken, Tests, und Wechselunterricht? Ich bin sicher Ihnen fallen auch noch ein paar interessante Fragen ein.

Jetzt werden Sie vermutlich sagen: So ein Narr, ein Sofa-Epidemiologe! Fordern kann man sowas natürlich – aber kontrollierte Interventionsstudien zu Social Distancing, das geht doch gar nicht! Sind sie sich da so sicher? Hat es denn jemand versucht bei uns? Und ist gescheitert, sodass wir wissen würden, wie man es besser machen könnte mit einem modifizierten Ansatz? Es sieht leider ganz so aus, als wäre das nicht der Fall.

Schon aus der Botswana Studie, und den beiden skandinavischen, sowie einer geplanten aber nie durchgeführten norwegischen Schulschließungsstudie hätten wir viel lernen können. Zunächst einmal, dass es ganz grundsätzlich machbar ist. Die Methoden für solche Studien stehen im Prinzip, sie kommen aus der ganz normalen klinisch-epidemiologischen Studienroutine. Aber auch von randomisiert kontrollierten Interventionen, welche mittlerweile auch in den Erziehungs -, Wirtschafts- und Sozialwissenschaften durchgeführt werden. Man kann solche Studien sowohl auf der Ebene von Individuen, als auch der von Gruppen machen. Letzteres z.B. in sogenannten Cluster-randomised trials, die auch in klinischen Fragestellungen häufig eingesetzt werden.

Einfach ist das natürlich nicht, ganz besonders unter den Bedingungen einer Pandemie. Um ein brauchbares Protokoll für so eine Studie aufzusetzen, muss man sich eine Menge Gedanken machen. Am Beispiel Schulschließungen: Welche ‚Dosis' und welches Timing soll die Intervention haben? Randomisiert man Schulschließung versus Wechselunterricht und verringerter Klassenstärke? Welche Klassenstufen sollen untersucht werden, wie groß dürfen die Klassen sein, wie oft wird gelüftet? Welchen primären Outcome wählt man? SARS-COV-2 Infektionen, na klar. Aber in welchem Kollektiv? Im Landkreis, in der Umgebung der Schule, nur bei Eltern und Schülern? Außerdem will man ja auch etwas über die sonstigen Auswirkungen erfahren. Führen solche Schließungen zum Beispiel später zu schlechterem Bildungsniveau und Abschlüssen? Aber nach welcher Zeit, und wie gemessen? Im Schulschließungsfall geht das auch nicht auf der Ebene von Individuen. Man kann nicht einzelne Schüler aus einer Klasse in die Studiengruppen randomisieren. Also muss man Schulen, oder Schulbezirke randomisieren. Würden Eltern bei so etwas zustimmen? In wieweit ist ihre Einwilligung überhaupt nötig? Der Staat frägt die Eltern ja auch nicht, bevor er Schulen auf oder zu macht. Ethisch ist sowas dann unproblematisch, wenn man nicht weiß, welche Maßnahme besser ist, wenn potentieller Nutzen und Risiko bei Intervention und Kontrolle gleich verteilt sind. Die Ethiker nennen das Equipoise. Bei Schulschließungen ist das der Fall. Aber selbst wenn eine Ethikkommission zustimmt, welche Schulbehörde würde bei sowas mitmachen? Würde es dann einen Aufstand von unwilligen Eltern geben? Fragen über Fragen. An diesem Beispiel sieht man, dass es ganz und gar nicht einfach ist, solche Interventionen zu planen und durchzuführen. Aber man könnte sich ja erstmal etwas einfachere und ebenso wichtige Fragestellungen vornehmen, wie z.B. welche Hygienemaßnahmen in Restaurants wirksam sind. Sollten Sie jetzt denken, das geht doch eh alles nicht, denn wir sind ja schon mitten im Lockdown, irren Sie sich. Man bräuchte die zeitliche Sequenz ja nur umzudrehen, und nicht die Einführung der Maßnahme, sondern deren Lockerung als Intervention zu testen. Außerdem wechseln wir ständig von strengeren zu weniger strengen Maßnahmen und zurück, eigentlich ideale Bedingungen für kausale Studien.

Mein Punkt ist: Solange man sich nicht auf den Weg macht und versucht, Hindernisse zu überwinden, neue Untersuchungskonzepte und Methoden zu entwickeln, ist das Argument, dass man die Blackbox Lockdown nicht knacken könne falsch und gefährlich. Und selbst wenn man jetzt keine Antworten mehr bekäme, welche in dieser Pandemie politische Entscheidungen auf eine rationelle Grundlage stellen würden: Die nächste

Pandemie kommt bestimmt. Vielleicht sind wir ja auch schon mittendrin, mit irgendeiner der Mutanten. Insbesondere wenn gegen eine davon die derzeitigen Vakzinen nicht mehr wirksam sind. Denn dann heißt es: Und ewig grüßt das Murmeltier - Lockdown, Lockerung, Lockdown usw.

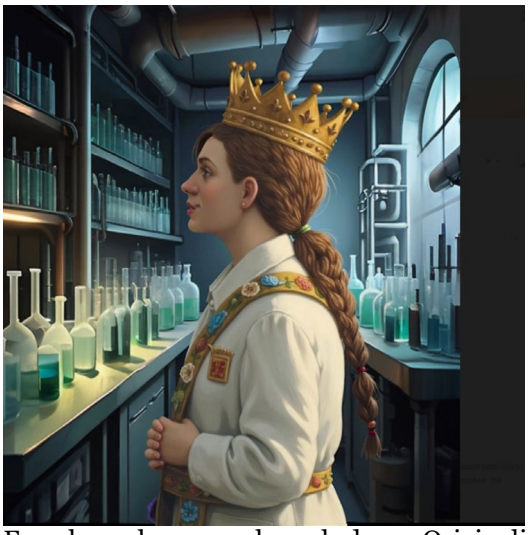
Aber was rege ich mich eigentlich über das Fehlen randomisiert kontrollierter Studien auf? Bei uns werden ja nicht einmal einfach durchzuführende und extrem aussagekräftige Observationsstudien und Datenerhebungen durchgeführt. Wissen wir, ob Pflegekräfte, Paketausfahrer, oder Supermarktkassierer häufiger SARS-COV-2 positiv sind, und häufiger symptomatisch? Wäre doch eigentlich recht gradlinig: Man müsste doch nur die Berufsgruppen melden, zusammen mit den Virustestergebnissen. Und wieso kennen wir eigentlich nicht die Dunkelziffer der Infizierten, also derer, die nicht getestet wurden, aber dennoch vom Virus befallen waren? Das ist nicht nur für die Berechnung der infektiösen Mortalität wichtig, sondern auch für die Frage, wie weit man schon in Richtung Herdenimmunität ist. Wie wäre es da mit zufällig ausgewählten Stichproben, welche in repräsentativen Regionen (Stadt, (Bundes)land) wiederholt getestet werden, sowie das in der ersten Welle in München gemacht wurde? Wieso gibt es eigentlich keine flächendeckende, systematische molekulargenetische Überwachung der Virusgenome? Wenn man erstmal ein paar hundert Milliarden für die Pandemiebekämpfung ausgegeben hat, ist dieser Neglect zumindest ökonomisch nicht mehr zu begründen.

Stattdessen starren wir wie hypnotisiert auf die tägliche Verkündung der gemeldeten Infektionszahlen, und die Vakzinierung. Vier Prozent des Forschungsoutputs der Welt im Jahr 2020 befasste sich mit Corona. Pubmed listet bereits über 100.000 Artikel zum Thema. Mehr als 4000 klinische Corona-Studien sind bei Clinicaltrials.gov registriert, mehrere Hundert davon haben die Wirkung von Chloroquin getestet. Studien, welche mittels einer randomisierten und kontrollierten Intervention versuchen, herauszufinden was nützt und was schadet bei der sozialen Distanzierung, kann man dagegen an einer Hand abzählen.

In Botswana kam übrigens heraus, dass sowohl Textnachrichten als auch Anrufe der Lehrer bei den Eltern und Schülern deren Interaktion signifikant verbessern konnte, und damit die Rechenfähigkeiten der Schüler gegenüber denen in der Kontrollgruppe erhöhte. Deshalb kriegen jetzt alle Familien mit Schulkindern Textnachrichten und Anrufe.

Boost your score: Freiwillige Selbstinszenierung in der Konkurrenz der Wissenschaftler

LJ 4/2021



Haben Sie einen Fitness-Tracker? Sind Sie auf Twitter oder Facebook und zählen ihre Likes und Followers? Kennen Sie Ihren Research Gate Score? Achten Sie bei Restaurantbesuchen auf Gault Millau Hauben und Michelin Sterne? Dann sind Sie in guter Gesellschaft, denn Sie betreiben auf verschiedensten Ebenen Reputationsmanagement mit quantitativen Indikatoren. Genau wie die Universitäten und Forschungsförderer. Nur dass Sie das privat und ganz freiwillig machen!

Kürzlich hat sich der Wissenschaftsnarr über zwei Folgen hinweg (LJ 12/2020, LJ1-2/2021) ausführlich darüber Gedanken gemacht, wie es dazu kam, dass wir heute in unserem Wissenschaftssystem Forschung kaum noch nach deren Originalität, Qualität und Einfluss beurteilen. Sondern vielmehr mit quantitativen Indikatoren wie Journal Impact Factor (JIF) oder Drittmitteleinwerbung, und darüber dann Fördermittel oder akademische Titel verteilen. Auch hatte er ein paar närrische Ideen, wie man das Rad wieder ein Stück zurückdrehen könnte, in Richtung einer inhaltlichen Bewertung von Forschungsleistungen. Bei diesen Betrachtungen blieb aber noch unberücksichtigt, dass sich die Institutionen und Fördergeber in guter Gesellschaft – nämlich unserer - befinden, wenn sie Wettbewerb und Konkurrenz mit einfachen, abstrakten Messgrößen anfeuern. Und das macht ihnen die Sache leichter. Und gleichzeitig wird der Status quo stabilisiert, es schwieriger, ihn zu verändern.

Es soll also hier nicht um die institutionelle, sondern um die individuelle Seite von quantitativer Leistungsbewertung gehen. Das wissenschaftliche s ist nämlich nur eine spezialisierte Verlaufsform und ein Spiegel eines gesamtgesellschaftlichen Quantifizierungskultes, der auch vor dem Privaten nicht halt gemacht hat. Denn nicht mehr nur im Beruf dient Quantifizierung der Herstellung eines Marktes, auf dem über den Wettbewerb mit Zahlen Leistung gemessen und gesteigert wird. Individuell geht es dabei um Status, um Reputation. Aus der Notwendigkeit, Papers in renommierten Journalen zu publizieren (oder einfacher gesagt, mit hohem JIF) um sich im akademischen System zu halten, oder gar aufzusteigen, entwickelt sich das Management von persönlichem wissenschaftlichem Status: ‚Der hat im letzten Jahr zwei Nature Paper geschrieben!', oder ‚Mein Hirsch-Faktor ist über 50', und so fort. Objektive wie auch subjektive Unsicherheit in der Konkurrenz der Wissenschaftler untereinander erhöht dabei nur noch den Wunsch nach Status und Informationen, welche diesen quantifizieren. Daraus hat sich eine Fetischisierung der Selbst- bzw. Außendarstellung entwickelt, welche unter anderem in der Hege und Pflege des Lebenslaufs (nur kein Journal vergessen, für das mal schon mal gereviewed hat!), einer eigenen professionellen Website oder Twitter-Accounts ausgelebt wird. Das Motto lautet hier ‚looking good', und nicht mehr ‚being good'. Ordentliche Graduiertenprogramme bieten ihren Studenten mittlerweile Seminare in der Kunst

dieser professionellen Selbstdarstellung und Selbstoptimierung an. Wir trainieren den Nachwuchs im Statuswettbewerb, und prämieren Statusstreber.

Der Nachwuchs lehnt sich dagegen mehrheitlich keineswegs auf, sondern wünscht sich weitere Vertiefung. All dies ist natürlich schon deshalb keineswegs verwunderlich, da Reputationsmanagment über quantitative Indikatoren auch im Privatleben mittlerweile voll durchgesetzt ist. Die Wissenschaft ist von solchen Quantifizierungsauswüchsen aber auch deshalb besonders betroffen, da Wissenschaftler möglicherweise ein gesteigertes Anerkennungsbedürfnis und Geltungssucht haben. Titel, Toppublikationen, Auszeichnungen, immer der Erste sein: Wissenschaftler sind geborene Konkurrenzler. Außerdem sind Wissenschaftler für die Quantifizierungslogik des Wettbewerbes auf natürliche Weise empfänglich. Was messbar ist und in Zahlen ausgedrückt werden kann, ist transparent, nachvollziehbar, evidenzbasiert, rational, neutral, präzise, einfach, unmittelbar und objektiv vergleichbar. Vermessung gehört zum Grundrepertoire der wissenschaftlichen Methode. So gesehen ist das Zählen beim JIF und Hirsch-Faktor, aber eben auch bei Gault Millau Hauben oder Twitter Followern nicht weit von der wissenschaftlichen Praxis.

Ist das nicht harmlos? Gar nützlich, da die sich auf diese Weise selbst und gegenseitig anstachelnden Wissenschaftler dann forscherische Großtaten vollbringen? Ich fürchte nein. Quantifizierung vereinfacht durch Abstraktion. Eine Qualität (‚Was') wird in eine Quantität (‚Wieviel') transformiert. Unvergleichbares wird plötzlich vergleichbar, sogar die sprichwörtlichen Äpfel mit den Birnen! Es gilt nun ein gemeinsamer Maßstab für unterschiedliche Dinge. Herr Dr. Maier und Frau Dr. Müller können sich jetzt direkt vergleichen. Über den kumulativen JIF, oder den Hirsch-Faktor, den Research Gate den Altmetric Score. Die letzten beiden sind wunderbare, aber auch traurige Beispiele für den Kern des Problems. Impact wird als Aufmerksamkeit verstanden. Nicht originelle Hypothesen, neue Erkenntnisse oder gar wissenschaftlicher oder gesellschaftlicher Nutzen werden hier zum Wesentlichen, sondern Sichtbarkeit und Popularität. Research Gate fordert seine Nutzer auf: ‚Boost your score'. Dabei legt Research Gate gar nicht offen, wie der Score berechnet wird. Ob er reproduzierbar ist, oder was er eigentlich aussagen soll. Das macht aber gar nichts, denn er produziert eine Zahl, und über diese kann man sich vergleichen und konkurrieren. Unser Denken und Urteilen richtet sich dadurch mehr und mehr an solcher Indikatorik aus, und verdrängt dabei professionelle Standards und Inhalte.

Wer die Hyperkompetition im System, das ‚publish or perish' kritisiert, und die Institutionen zum Umsteuern auffordert, muss sich deshalb auch an die eigene Nase fassen. Wir beteiligen uns freiwillig und mit großem Eifer an einer Vielzahl teilweise privaten Spielarten der Konkurrenz, welche mittels karger, vom Gegenstand getrennter Zahlen ausgetragen werden. Und außerdem: So richtig aufregen tun wir uns vor allem dann, wenn wir mit den eigenen Zahlen, unserem Ranking also, nicht zufrieden sind. Denn die Zufriedenheit mit einem Indikator und seiner Berechnung korreliert sehr gut mit der eigenen Platzierung. Wenn die nicht gut ist, ist der Indikator mutmaßlich ungeeignet. Und dann hört man selbst von bisher nicht als kritisch aufgefallenen Kollegen so manchen wahren Satz zu JIF oder Drittmitteln. Oder es wird am Algorithmus gemäkelt: man wünscht sich eine Formel bei deren Anwendung man besser dasteht.

Wir haben die quantitative, abstrakte Statuslogik, die uns von den Institutionen aufgemacht wird, verinnerlicht, sie zu einer wichtigen Zielgröße unseres Selbstwertgefühls gemacht. Wir haben die Indikatoren und Maßstäbe freiwillig übernommen, und weil die Institutionen und die Kollegen ihnen großen Wert beimessen, tun wir selbst dies mit umso größerer Überzeugung: Conform and perform!

Steffen Mau, der in seinem lesenswerten Buch , Das metrische Wir - Über die Quantifizierung des Sozialen‘ diese Umtriebe aus sozialwissenschaftlicher Perspektive ausführlich analysiert, weist am Anfang seines Traktates darauf hin, dass im Deutschen schon das Wort ‚vermessen‘ eine Vorahnung auf Schlimmes enthält: ‚Vermessen‘ meint ja nicht nur den Vergleich mit einem Maßstab, sondern bedeutet auch falsch messen, sowie ‚überheblich‘ bzw. ‚anmaßend‘. Der falsche Maßstab, der Reflex auf die Reputation, und letztlich das Setzen von falschen Anreizen, alles dies nimmt die deutsche Sprache da schon vorweg!

Politikberatung bis dass der Elefant mit dem Rüssel wackelt!

LJ 5/2021

Das Zeitalter der Universalgelehrten kehrt zurück! Seit etwa einem Jahr eifern Wissenschaftler da Vinci, Leibniz, und von Humboldt nach. Virologen äußern sich öffentlich und gegenüber politischen Entscheidungsträgern zur Epidemiologie, Physiker zur Infektionsbiologie, Mathematiker zu viralen Oberflächenproteinen, und so fort. Dabei war es doch bisher die Domäne der Narren, ungestraft Späße zu beliebigen Themen zu machen! Auch deshalb erlaube ich mir heute mich ungeniert der mathematischen Modellierung in Zeiten der Pandemie zuwenden.

Modellierer sind momentan ja sehr gefragt. Wir lesen ihre Arbeiten in Nature und Science, man lauscht ihnen bei Lanz und Konsorten, sie beraten Politiker, auch rechnen sie für nationale Akademien. Das ist kein Wunder, denn ihre Formeln und Modelle versprechen Aufklärung komplexer Zusammenhänge. Sie sagen uns, was passieren könnte, wenn wir gewisse Dinge tun oder lassen. Auch erklären sie uns, welche Maßnahmen zur Pandemiebekämpfung wirksam sind, und welche nicht. Häufig mahnen sie, und belegen ihre Empfehlungen mit konkreten Zahlen. Genau so wünscht man sich Handreichungen aus der Wissenschaft. Die Politik bekommt Argumente für ihre Entscheidungen, und Bürger sehen ein, warum die Schule schließen muss, oder das Geschäft die Türe wieder öffnen darf.

Modellierer sind in vielen Bereichen bereits recht erfolgreich. Ein Paradebeispiel hierfür ist der Wetterbericht. Mit im Mittel etwa 70 % Treffsicherheit gelingt es den Meteorologen, das Wetter in 7 Tagen vorherzusagen. In die Modelle, welche auf Supercomputern gerechnet werden, gehen unzählige Messungen ein, welche das atmosphärische Geschehen vom Boden bis in viele Kilometer in die Höhe abbilden. Ihre Rechnungen berücksichtigen die Temperatur- und Strömungsdynamik der großen Gewässer, und sogar die fluktuierenden Bahnen von Mond und Sonne. All dies mit höchster Meßgenauigkeit. Möglich wird eine Wettervorhersage dieser Treffsicherheit aber nur, weil die meteorologischen Zusammenhänge von Temperaturen, Drücken, Wind-, Wasser-, und Planetenbewegungen durch internationale wissenschaftliche Kooperationen bereits lange untersucht und mittlerweile recht gut verstanden werden. Ein anderes schönes Beispiel für

erfolgreiche Modellierungen kommt aus der Geophysik. Ausbrüche von Vulkanen lassen sich überraschend gut vorhersagen, wie zuletzt bewiesen beim Fagradalsfjall-Vulkan in Island. Auch diese Vorhersagen beruhen auf einer Vielzahl von exakten seismologischen und Satelliten-Messungen, zumindest teilweise verstandenen Mechanismen vulkanischer Aktivität, und jahrelanger Optimierung der Modelle. Aber selbst diese Modellierer liegen oft daneben, und wir ärgern uns, vor Regen nicht gewarnt worden zu sein. Uns so mancher Vulkan will trotz eindringlicher Warnungen einfach nicht ausbrechen.

Aber wie steht es eigentlich um die Vorhersagekraft und damit die Nützlichkeit der so allgegenwärtigen Modellierungen in der Pandemie? Leider gibt es mittlerweile eine Menge Hinweise darauf, dass es da nicht zum Besten steht. Die Modellierer sind offensichtlich so sehr mit dem Generieren neuer Modelle beschäftigt, dass sie kaum dazu kommen, die Güte und das Eintreten ihrer Vorhersagen zu analysieren. Dies hat man offensichtlich den Journalisten überlassen. So analysiert ein Artikel in der Tageszeitung *Die Welt* (Literaturzitate wie immer bei http://dirnagl.com/lj) die wichtigsten Vorhersagen aus dem Umfeld von Deutschlands prominentester Modelliererin, Viola Priesemann. Dabei zeigt sich zum einen, dass die meisten Schlussfolgerungen aus den Modellrechnungen sehr vage verfasst waren. Wie bei Horoskopen passten sie damit zu jedem Verlauf. Und dort, wo konkret Zahlen vorhergesagt wurden, sind diese sehr häufig nicht eingetroffen. Es sei denn, es handelte sich um Triviales, wie die Prädiktion eines weiteren Anstieges am Anfang eines bereits deutlich sichtbaren Verlaufes. Und sobald es darum ging, die Wirksamkeit von Pandemiemaßnahmen zu prognostizieren, wurde es richtig problematisch. Nur ein Beispiel hierfür ist die Vorhersage aus der Leopoldina- Stellungnahme vom 8. Dezember letzten Jahres. Dort wurde Folgendes vorausgesagt: *‚Wenn ab dem 14. Dezember die Maßnahmen streng verschärft werden, dann sinken die Fallzahlen in der Modellrechnung bis Januar auf unter 50 pro 1.000.000 Einwohner‘*. Wie wir alle wissen, ist dies trotz erfolgtem hartem Lockdown nicht eingetreten: Die Inzidenzraten stiegen zwar nicht weiter, verharrten aber auf hohem Niveau.

Diese Modellierung basierte auf dem im Juli 2020 in Science veröffentlichten Modell aus dem Max-Planck-Institut für Dynamik und Selbstorganisation in Göttingen. Und auf Daten aus dem Frühjahr 2020. Das Modell bezog sich damit auf eine völlig andere Umsetzung und Akzeptanz von Maßnahmen als im Vorhersagezeitraum. Wie vielen solcher Modellierungsstudien fehlten hier aber auch Kontrollen, wie wir sie in jeder biomedizinischen Arbeit erwarten würden. Zum Beispiel hätte man die Güte des Modells durch Anwendung auf anderen Datensätze, zum Beispiel aus einem anderen Land oder über einen anderen -am besten auch längeren Zeitraum hinweg – überprüfen können.

Kontrollen beim Modellieren? Ja, das geht, sogar recht einfach. Das Modell, so eine zentrale Aussage des Artikels, belegten die Wirksamkeit und damit Notwendigkeit des harten Lockdowns in Deutschland. Hätten die Autoren ihr Modell aber z.B. auch auf Schweden angesetzt, wäre dort ein ganz ähnlich gearteter Abfall der Fallzahlen herausgekommen. Nur dass es dort keinen Lockdown gab! Diese Kontrollrechnung konnte der Neurologe und Physiker Christian Meisel durchführen, denn die Priesemann Forschungsgruppe (Kudos!) stellte ihr Modell inklusive Daten ins Netz. Meisel entwickelt normalerweise Modelle, mit denen sich aus EEG – Daten epileptische Anfälle vorhersagen lassen und ist deshalb mit der Technik wohlvertraut. Ähnliches wie für das Göttinger Modell gilt auch für die Modelle des Imperial College in London (ICL). Diese hatten großen Einfluss auf die Pandemiemaßnahmen der englischen Regierung. Auch hier lagen die Vorhersagen häufig extrem daneben. Chin und Kollegen konnten außerdem zeigen, dass verschiedene publizierte Modelle des ICL zu ganz unterschiedlichen Resultaten kommen, wenn man sie auf die gleichen Länder loslässt. Was die Londoner selbst bezeichnenderweise nicht gemacht hatten.

Ist alles dies überraschend? Deutet es darauf hin, dass die Pandemie-Modellierer ihr Handwerk nicht recht verstehen? Im Gegensatz zu den Meteorologen basieren ihre Modellierungen auf schlechten oder sogar nicht vorhandenen Daten, also bloßen Annahmen. Dies gilt für Corona-Inzidenzen und noch mehr für die Auswirkungen nicht-pharmakologischer Interventionen. Außerdem hängt alles davon ab, ob und wie die Maßnahmen in der Bevölkerung dann umgesetzt werden. Bei einer höchst unsicheren Datenlage, wie sie z.B. allein schon aufgrund sich ständig ändernder Testkapazitäten und -raten, insbesondere am Anfang einer Pandemie, vorkommt, ist es unabdingbar, diese elementare Fehlerbehaftung kritisch zu berücksichtigten. Datenfehler pflanzen sich fort, das lernt man spätestens im Physikpraktikum. Und sie tun das umso mehr, wenn sie in komplexe, multiparametrische Modelle und Wachstumsverläufe eingehen. Dazu kommen jede Menge nicht antizipierbarer Einflussgrößen wie z.B. das Auftreten von Virusmutanten mit veränderter Infektiosität, Letalität, oder Effektivität von Vakzinierungen sowie Rückkoppelungs- und Selbstregulierungsmechanismen, weil die Vorhersagen ihrerseits ja Wirkungen auf das Verhalten in der Bevölkerung haben. In Anbetracht all dessen ist die oft propagierte Pseudogenauigkeit der Modellierungsergebnisse schlichtweg vermessen. Es ist, als würde man mit Kanonen, nämlich komplexen, multiparametrische Modellierungen auf Spatzen, also auf grob Fehler-behaftete und nicht-valide Datengrundlagen schießen. Ein schönes Beispiel ist hier auch der Rückgang des Autoverkehrs in der Pandemie. In den USA wurden im letzten Jahr ca. 13 % weniger Meilen gefahren. Folglich nehmen auch die Verkehrstoten ab, einer der wenigen positiven Effekte der Pandemie? Falsch. Die haben so zugenommen wie seit 1924 nicht, nämlich um 25% pro gefahrene Meile. Retrospektiv sucht man nun nach Gründen hierfür, wie z.B. vermehrtem Alkoholkonsum. Man hätte diesen überraschenden Effekt wohl kaum vor dessen Bekanntwerden in einem Modell der Gesamtmortalität während der Pandemie berücksichtigen können.

Ein weiterer wichtiger Grund für das Versagen der Modelle ist, dass deren Annahmen ja durch die in der Pandemie angeordneten Maßnahmen modifiziert werden. Dies ist sogar ein erwünschter Effekt, denn die Modellierer erheben häufig ihren Zeigefinger. Das wäre aber gerade so, als wenn sich das Wetter ändern würde, abhängig davon, ob wir einen Regenschirm aufspannen oder nicht. Dann würde auch der Wetterbericht nicht mehr funktionieren.

Hinzukommt, dass Modellierungsstudien in der Regel weder vorab Studienprotokolle veröffentlichen noch präregistriert werden, wie dies eigentlich heutzutage für qualitativ hochwertige Studien selbstverständlich sein sollte. Damit ist einem Herumprobieren ‚bis es passt' Tür und Tor geöffnet.

Auch historisch betrachtet haben Modellierungen von Epidemien keinen guten Track record, aber daran erinnert sich heute kaum noch jemand. Man denke zurück an die Schweinegrippe, oder die Bovine Spongioforme Encecpalitis (BSE). Auch damals schon lagen die prominenten Modellierer, welche übrigens heute immer noch ganz vorne dabei sind, mit ihren Vorhersagen massiv daneben. Bei der bereits erwähnten Prädiktion von epileptischen Anfällen - auch hier geht es ja um die Vorhersage zukünftiger Ereignisse aus komplexen Datensätzen - hat man übrigens aus den initialen Fehlern gelernt. Nach einer anfänglichen Euphorie mit darauffolgender kritischer Ernüchterung und Fehleranalyse ist eine etwas demütigere, aber dennoch nicht weniger relevante Wissenschaft entstanden. Mittlerweile gibt es dort rigorose Methoden, mit denen die Güte von Vorhersagen geprüft werden können. Die Pandemie-Modellierer von heute täten gut daran, mal einen Blick hierauf zu werfen.

Vielleicht besteht aber der eigentliche Nutzen der Pandemie-Modellierungen darin, worst case Szenarien wissenschaftlicher, und damit einschneidende Maßnahmen für die breite Masse einleuchtender und akzeptabler zu machen. Diese also wissenschaftlich zu bebildern. Das ist aber eine gefährliche Strategie: Zum einen, weil Vorhersagen, welche daneben liegen, ihre Überzeugungskraft verlieren. Zum anderen, weil die Modelle ja behaupten, die Nützlichkeit oder Schädlichkeit bestimmter Maßnahmen oder auch von Verhalten zu ‚objektivieren'. Wie zum Beispiel Schulschließungen, Ausgangssperren, oder Abstandsregeln. Wenn die offensichtlichen und teils schwerwiegenden Limitationen der Modelle nicht erkannt oder berücksichtigt werden, sie aber dennoch die Grundlage für unser Handeln in der Pandemie liefern, dann läuft was schief.

Aber ob und wenn ja welchen Einfluss die derzeit sehr medienpräsenten Modellierer auf die Politik überhaupt haben, oder von dieser nur benutzt werden um politisch motivierte Entscheidungen zu rechtfertigen, ist unklar. Dafür können die Modellierer natürlich erstmal nichts. Allerdings wehren sie sich gegen eine solche Instrumentalisierung auch nicht, sondern genießen die mediale Aufmerksamkeit. Der Narr hat sich über das komplette Fehlen einer evidenzbasierten, inklusiven, gründlichen, transparenten, und zugänglichen wissenschaftlichen Beratung der Corona-Politik bereits in Laborjournal 11/2020 echauffiert. Darin kommt er zum Schluss, dass das post-Darwinistische Motto ‚wissenschaftsbasierter' Pandemiepolitik derzeit ein ‚Survival of the ideas that fit' ist.

Modellierung funktioniert in der Pandemie bisher nur dort, wo sie sich auf wenig komplexe, und teilweise gut verstandene Zusammenhänge verlässt und die Datenlage einigermaßen robust ist. Das ist leider nicht häufig der Fall. Zum Beispiel liefert sie recht verlässliche und nützliche Vorhersagen, wo es um den Zusammenhang von COVID-Inzidenz, Auslastung von Intensivstationen und Todesfällen geht. Sobald die Modellierer sich aber auf komplexe, kaum oder gar nicht verstandene, dazu noch volatile Zusammenhänge stürzen, und die zugrundeliegenden Parameter auf nicht verlässlichen oder nur geschätzten Daten beruhen, und die Vorhersage Einfluss auf ihr eigenes Ergebnis hat, funktioniert es nicht mehr richtig. Die daraus resultierenden, überkomplexen Modelle werden, damit überhaupt etwas Plausibles dabei rauskommt, ‚overfitted', es wird mehr Noise als Signal modelliert. Eine vertiefte Diskussion der Limitationen und Unsicherheiten solcher Modelle und deren Aussagen, genauso wie Kontrollen, würden dabei nur stören, und in weniger öffentlicher Aufmerksamkeit resultieren. John von Neumann, Mathematiker, Physiker, und Computerpionier wird mit dem Bonmot zitiert: ‚Mit vier Parametern kann ich einen Elefanten fitten, und mit fünf ihn mit dem Rüssel wackeln lassen.' Wenn mit Rüssel-wackelnden Elefanten und dem Gestus mathematisch-physikalischer Autorität Politikberatung gemacht wird, ist das nicht ohne Risiko.

Der Wissenschaftsnarr dankt Prof.Dr. Christian Meisel und Prof. Dr. Gerd Antes für anregende Diskussionen.

Die medizinische Habilitation: Vom professoralen Herrschaftsinstrument zum Jodeldiplom für Chefärzte

LJ 6/2021

Vor nun bald 20 Jahren saß ich gemeinsam mit 5 anderen Leidensgenossen in einem Vorraum eines Hörsaals der Medizinischen Fakultät der Ludwig-Maximilians-Universität München. Drinnen tagte der hohe Fakultätsrat. Es galt für uns die letzte Hürde zur Erlangung der Habilitation zu überwinden: Ein freier Vortrag ohne Hilfsmittel zu einem Thema, das nicht Gegenstand unserer Habilitationsschrift war. Unmittelbar vor mir dran war ein gestandener Neurochirurg. Als solcher war er gewohnt unter einem Mikroskop Aneurysmen an der Hirnbasis zu klippen. Er tat dies routiniert, Leben oder Tod seiner Patienten lagen dabei in seinen Händen. Jetzt aber war er kaltschweißig, trotz vorab Einnahme eines Betablockers. Kurz bevor er an der Reihe war, wollte er sich aus dem Staub machen. Er sei zu aufgeregt, er könne weder klar denken noch sprechen. Es gelang mir, ihn in letzter Minute durch gutes Zureden von seiner Flucht abzubringen, wankend bewegte er sich in den Hörsaal.

Habilitanden wurden nämlich regelmäßig Opfer im Grabenkrieg der Ordinarien. Ihre Scharmützel trugen die Professoren mit harten Bandagen aus. Der Abschuss eines Habilitanden des Konkurrenten konnte einen Stellungsvorteil bringen, oder versprach einfach nur süße Rache für eine an anderer Stelle durch diesen Konkurrenten erlittene Schmach. Allerdings wurde dann meist in der nächsten Sitzung mit gleicher Münze heimgezahlt, und ein weiterer Habilitand geriet so ins Sperrfeuer. Wir Habilitanden waren damit Spielbälle in der ganz normalen Konkurrenz der Professoren der Fakultät. Der für mein Fach zuständige Ordinarius beruhigte mich in den Wochen vor meinem Vortrag damit, dass er nur noch mit dem Dekan Tennis spielen müsse, und dabei verlieren. Dann bräuchte ich mir keine Sorgen machen. Und tatsächlich ließ er den Dekan gewinnen, und besitze seither die Venia legendi. Wie recht hatte doch Ernst-Ludwig Winnacker, ehemals Präsident der DFG, als er die Habilitation im Jahr 2006 als ‚spätmittelalterliche Errungenschaft', und als ‚Herrschaftsinstrument altgedienter Professoren' bezeichnete. Der Vortrag vor der hohen Fakultät war somit ein letztes Initiationsritual vor Eintritt den Club derer, die auf eine Berufung zum Professor hoffen dürfen.

Heute bin auch ich gewähltes Mitglied einer solchen Fakultät, und urteile über Vorträge von Habilitanden. Vermutlich sind diese immer noch sehr aufgeregt, denn auf unerklärliche Weise hat das ganze Verfahren nach wie vor eine bedeutungsschwere akademische Aura. Außerdem wird man ja nochmal ‚geprüft', und das in einem Alter, in dem man normalerweise Anderen Noten erteilt. Von den Auseinandersetzungen der Ordinarien und dem dabei fließenden Habilitandenblut ist heute aber rein gar nichts übriggeblieben. Die Habilitanden benutzen Powerpoint und werden gelobt für ihre Vorträge, und es

folgen ein oder zwei artige Fragen. Dann wird gratuliert. Dabei ist es ziemlich egal, wie die Qualität des Vortrags und der dort dargebotenen Wissenschaft war. Der Narr sitzt dort häufiger, den Blick fremdschämend nach unten gerichtet. Er wird dort nämlich Zeuge irrgeleiteter Studiendesigns, offenem Missbrauch von Statistik, überinterpretierten Ergebnissen, und meist vollständig fehlender Erwähnungen der Limitationen der vorgestellten Studien. Viele der Vorträge würden, von Studenten in einem ordentlichen Labmeeting gehalten, in diesem nicht durchgehen. Das wissenschaftliche Niveau der Verteidigungen von PhD-Dissertationen vor der gleichen Fakultät ist im Mittel deutlich höher.

Hat man sich erst einmal auf den Weg zur Habilitation gemacht, ist es nur eine Frage der Zeit, bis man den Titel Privatdozent auf der Visitenkarte hat. Keiner fällt durch, man muss (in Berlin) mindestens 11 Originalartikel als Erst- oder Letztautor geschrieben haben, dazu Pflichtlehre absolviert und ein bisschen Didaktik geschnuppert haben. Dann lässt man das Ganze in eine schwarze Kladde binden. Die Familie kriegt davon ein Exemplar, ein paar kommen in den Bücherschrank, der Rest wird in einem Karton gelagert, denn man irgendwann entsorgen wird. Versucht man ausländischen Kollegen klarzumachen, worum es sich bei der Habil handelt, wird man meist nicht verstanden, und wird ungläubig belächelt. Manch ein Habilitand führt den Titel Privatdozent (PD) dann gleich im internationalen Lebenslauf als PhD. Hat zwar nichts miteinander zu tun, klingt aber gut und erzeugt keine Rückfragen.

Die Habilitation wird aus all den genannten Gründen seit Jahrzehnten als spätmittelalterlicher akademischer Zopf kritisiert, den man abschneiden müsse. Frau Buhlman hat es tatsächlich 2002 versucht. Sie wollte den Teufel mit dem Beelzebub der Juniorprofessur austreiben. Jetzt habilitieren bei uns die Juniorprofessoren! Das Projekt ist bekanntermaßen gescheitert. Für die Medizin kann man auch ganz einfach sagen warum: Zum einen natürlich, weil Ärzte als Juniorprofessoren sehr viel weniger verdienen als im normalen Ärztetarif. Zum anderen, vermutlich noch wichtiger, weil Habilitierte auf dem Arbeitsmarkt der Chefärzte in nicht universitären Krankenhäusern einen großen Konkurrenzvorteil haben. Wenn sie dann nach der Habil noch 5 Jahre durchgehalten und noch ein paar Artikel veröffentlicht haben, dann kriegen sie ein APL – Professur verliehen. Und das bringt dem Krankenhausträger Nimbus und zusätzliche Patienten, und dem Chefarzt ein deutlich höheres Gehalt.

Ist das Ganze also harmlose akademische Folklore und skurrile Brauchtumspflege? Ich denke nein. Es handelt sich vielmehr um Zeit- und Ressourcenverschwendung großen Stils, und gaukelt wissenschaftlichen Professionalismus vor, wo oftmals keiner ist. Viele Habilitanden forschen nämlich um zu habilitieren. Klingt harmlos - ist es aber nicht. Wenn der Titelerwerb zum primären Ziel der Forschung wird, geht es nicht mehr vorrangig um Erkenntnisgewinn. Dann ist es egal, ob eine klinische Studie zu wenig Patienten untersucht um relevante Aussagen zu generieren. Denn irgendein Paper wird schon draus zu zimmern sein, genauso wie aus einer x-beliebigen tierexperimentellen Studie. Hier werden nicht nur Ressourcen verbraten, sondern auch potentiell Patienten in Studien rekrutiert, deren Ergebnisse niemand weiterbringen. Oder Tiere für Experimente verbraucht, deren Resultate nicht reproduzierbar sind – und die im Zweifel auch gar nicht der Mühe Wert sind, repliziert zu werden. Dazu sitzen uns noch den Hintern auf in den zugehörigen Kommissionssitzungen. Habilitanden füllen Formulare aus und schreiben dicke Bücher, die keiner liest. Gutachter verfassen über diese Bücher Gutachten, die ebenfalls keiner liest – denn sie empfehlen ohnehin die Annahme. In den Strudel solcher Habilitationsforschung geraten dann auch noch häufig mäßig gut angeleitete medizinische Doktoranden, die ihrerseits häufig nur für einen Titelerwerb forschen,

nämlich dem Dr.med. Die ist ein nicht minder problematisches Unterfangen, das sich der Narr bei Gelegenheit einmal separat vornehmen wird.

Natürlich gibt es auch ganz tolle Habilitationen. Aber wenn sie auf solider und relevanter Wissenschaft beruhen, bringt der Titel aber nichts Zusätzliches. Die Ergebnisse der Arbeit stehen für sich, Habilitationsschrift und Urkunde tragen zum darin erarbeiteten Erkenntnisgewinn rein gar nichts bei. Und zum Professor kann man auch mit ‚Habilitations-äquivalenter Leistung' berufen werden, also mit einem ordentlichem wissenschaftlichen Oevre und Lehrerfahrung.

Wofür braucht es also die Habilitation? Um es mit Frau Hoppenstedt zu sagen: ‚Da hab ich was in der Hand. Da hab ich was Eigenes. Da hab ich mein Jodeldiplom.'

Tu felix Britannica – Notizen aus der deutschen Corona-Studienprovinz

LJ 9/2021

Das Virus hat biomedizinischer Forschung zu einem Allzeithoch im öffentlichen Interesse verholfen. Man ist allenthalben voll des Lobs für die Forscher, und abgesehen von einer Minderheit von Obskuranten vertraut die Bevölkerung diesen mehr denn je. Kein Wunder, denn es dauerte nur sagenhafte 9 Monate vom Beginn der Pandemie bis zu den ersten Impfungen, und die praktische Anwendung der Erkenntnisse der Medizin rettete einer Vielzahl von COVID Patienten das Leben. Aber die geneigten Leser dieser Kolumne wären enttäuscht, wenn der Narr nicht anlässlich der 4. Corona Welle seiner schelmischen Rolle gerecht würde. Lassen Sie mich also ketzerisch die Frage stellen: Verdient es die medizinische Wissenschaft wirklich, so pauschal über den grünen Klee gelobt zu werden? Ist die Wissenschaft für die nächste Welle, oder gar die nächste Pandemie, optimal aufgestellt?

Wie steht es zum Beispiel hierzulande um die randomisierten und kontrollierten Studien, welche Interventionen zur Prävention oder Therapie der SARS-COV-2 Infektion untersuchen? Diese Königsklasse von klinischen Studien generiert die bestmögliche Evidenz zum Nutzen aber auch möglichen Schaden von Impfungen, vom Tragen von Masken oder Schulschließungen, und natürlich auch von Medikamenten zur Behandlung von COVID. Es sind diese Studien, welche eigentlich die Basis für unsere Reaktion auf den gesundheitlichen Notstand liefern sollten. Nicht überraschend hat die Pandemie weltweit zu einer Tsunami solcher Studien geführt. Aber was kommt dabei raus, wie hoch ist deren Qualität?

Eine Gruppe von Autoren um den klinischen Epidemiologen Lars Hemkens (Basel) hat sich diese Frage gestellt, und alle weltweit registrierten klinischen Studien der ersten 100 Tage der Pandemie angeschaut. Es waren über 700 Studien, welche sich vornahmen,

insgesamt über 400.000 Patienten zu rekrutieren. Hemkens et al. fanden, dass die meisten dieser Studien viel zu klein konzipiert waren, auch untersuchten viele ein und dieselbe Intervention. Allein über 100 klinische Studien widmeten sich der Wirksamkeit des Anti-Malaria Mittels Hydroxychloroquin! Viele der Studien rekrutierten zu wenige oder manchmal gar keine Patienten, und produzierten daher auch keine brauchbare Evidenz.

Hemkens, der übrigens zusammen mit Gerd Antes auf sehr lesenswerte Weise in der diesjährigen Laborjournal Sommer-Ausgabe (LJ 7/2021) den Unsinn des Begriffs ‚Präventions-Paradox' aufspießt, hat sich nun aktuell die klinischen Corona Studien in Deutschland vorgenommen. Lief es bei uns besser als anderswo auf der Welt? War doch Schlimmes zu befürchten, denn der Wissenschaftsrat kam bereits 2018 zu der Einschätzung, dass ‚gemessen an einer leitenden Rolle bei herausragend publizierten klinischen Studien die deutsche Forschung im Vergleich wichtiger Referenzländern keine internationale Spitzenposition ein[nimmt]'. Wer die etwas verschwurbelten, weil diplomatischen Formulierungen des Wissenschaftsrates kennt, ahnt dass dies ein vernichtendes Urteil ist. In Schulnoten ausgedrückt hat der Wissenschaftsrat den deutschen klinischen Studien, und zwar insbesondere aus der universitären Medizin, mit ‚mangelhaft' bewertet. Aber vielleicht konnte die klinische Forschung Deutschlands in der Pandemie ja aufholen, insbesondere da allein das BMBF nach eigenen Angaben mehr als 1,6 Milliarden in die Corona-Forschung investiert hat?

Was Hemkens und sein Team fanden, war aber dann doch ernüchternd. Der Anteil der in Deutschland geplanten Corona – Studien und der in Deutschland für solche Studien rekrutierten Patienten ist im weltweiten Vergleich minimal. In einer Pandemie ist Geschwindigkeit alles – aber nur weniger als eine Handvoll der Studien, die komplett in Deutschland durchgeführt wurden, war bis zum April 2021 abgeschlossen. Die Mehrzahl der Studien erreichte nicht die geplanten Teilnehmerzahlen oder wurden gar abgebrochen. Nur jeder Hundertste in ein Krankenhaus eingewiesene Corona Patient in Deutschland wurde in eine randomisierte Studie eingeschlossen! Zudem fand sich keine einzige registrierte, randomisiert kontrollierte Studie, welche nicht-pharmakologische Interventionen untersuchte. Also Maßnahmen wie soziale Distanzierung oder Verhaltensinterventionen. Auch wurden in Deutschland keine Studien in Pflegeheimen, Kindergärten, Kindertagesstätten oder Schulen durchgeführt. Über diesen Missstand hat sich der Narr schon in einer früheren Folge aufgeregt, als er räsonierte, dass ‚im Kampf gegen Corona von Botswana lernen siegen lernen heißt' (LJ 3/2021).

Die Ohrfeige des Wissenschaftsrat hatte also keinerlei Wirkung! Woran liegt es aber, dass das mit den Studien allgemein und nun speziell bei Corona in Deutschland so schlecht läuft? Etwa daran, dass das unter pandemischen Bedingungen einfach nicht schneller und besser geht? Die Antwort hierauf ist ein klares Nein, andere Länder machen es uns vor! Die RECOVERY Studie der Universität Oxford im englischen National Health System wurde in 2 Tagen geplant und schloss dann 10.000 Patienten innerhalb von 2 Monaten ein! In kürzester Zeit wurde mit einem smarten, pragmatischen Design die Unwirksamkeit (Lopinavir und Ritonavir, beides Anti-HIV Therapeutika), bzw. Wirksamkeit (Dexamethason) gleich mehrerer Medikamente belegt, weltweit konnten mit diesem Wissen Leben gerettet werden. Es gelang in England allein für RECOVERY, jeden 6. hospitalisierten COVID Patienten in die Studie einzuschließen. Bei der großen internationalen Studien SOLIDARITY der WHO hat Deutschland erst nach einem halben Jahr mitgemacht, als die ersten Ergebnisse schon analysiert waren, welche die Unwirksamkeit von Remdesivir, Hydroxychloroquine, Lopinavir und Interferon beta-1a belegten. Beim Europäischen DisCOVery-Trial unter Federführung des französischen Instituts de la Santé et de la Recherche Médicale (INSERM) hat Deutschland gleich gar

nicht mitgemacht. In den ersten hundert Tagen der Pandemie waren China, USA, England, Spanien und Frankreich Spitzenreiter bei COVID Studien und Rekrutierung, Deutschland blieb unter ferner liefen.

Über die Gründe für die klinische Studienmisere in Deutschland kann man nur spekulieren. Es wird auch nicht gerne darüber geredet oder systematisch Ursachenforschung betrieben. Deshalb hier mal, und aus dem hohlen Bauch heraus, ein paar närrische Erklärungsversuche: In Deutschland liegt der Fokus seit je her auf von der Industrie geplanten und finanzierten Studien. Auf diesem Sektor ist Deutschland sogar die Nummer 3 hinter USA und England. Mit solchen Studien querfinanzieren Unikliniken ihre Forschung, und nicht selten auch die Patientenversorgung. Außerdem ermöglichen Industriestudien lukrative Nebenverdienste für die Beteiligten, durch Mitwirkung an diversen dazugehörigen Gremien, wie Advisory-, Data-, Safety-, Monitoring- usw. Boards. Geplant, gemonitort, und analysiert werden diese Studien von den Pharmafirmen, sogar das Schreiben der Papers wird von spezialisierten Schreibbüros erledigt. Man kriegt aber selbstverständlich trotzdem eine Co-Autorenschaft. Dazu stellt die Pharmaindustrie über Anzeigenaufträge bei den Top-Journalen sicher, dass die Papers auch dort erscheinen. Da kann es schon passieren, dass nicht mehr viel Zeit, Interesse, und vor allem keine Patienten übrig bleiben für sogenannte Investigator Initiated Trials (IITs), also die Wissenschafts-getriebenen Studien aus der akademischen Medizin.

Dazu kommt häufig ein Mangel an professioneller Infrastruktur zur Studiendurchführung. Studienzentren (Clinical Trial Centers) etabliert man an den meisten Universitätskliniken erst seit wenigen Jahren. Diese haben sich aus den Koordinierungszentren Klinische Studien (KKS) heraus entwickelt, welche zwischen 1999 und 2009 vom BMBF an 12 deutschen Unis gefördert wurden. Das Programm des BMBF war schon damals eine Kritik an der deutschen Studienmisere. Außerdem gab und gibt immer noch viel zu wenige Möglichkeiten, IITs gefördert zu bekommen. Dies trotz DFG und BMBF Förderlinien, die wie Tropfen auf einem heißen Stein verdampfen. Gefördert werden dann in der Regel zu kleine Studien, die dann aber häufig dennoch nicht oder nicht in einem vernünftigen zeitlichen Rahmen zu Ende rekrutieren. Was im Übrigen nicht nur Ressourcenverschwendung ist, sondern auch unethisch den rekrutierten Patienten gegenüber. Denn die haben ja bei den Studien mitgemacht und durchaus auch ein Risiko auf sich genommen um zukünftigen Patientengenerationen durch den zu erwartenden Wissensfortschritt zu nutzen. Der sich aber nicht einstellen kann, wenn nichts Belastbares rauskommt oder gar keine Ergebnisse publiziert werden. Den Krankenkassen, die eigentlich ein großes Interesse an Industrie-unabhängigen Studien haben sollten, ist übrigens in Deutschland die Förderung solcher Studien durch das Sozialgesetzbuch verboten!

Aber damit nicht genug: In Deutschland hat die klinische Epidemiologie, also jene Disziplin, welche die methodischen Grundlagen für klinische Studien liefert und dabei sicherstellt, dass sie gut geplant, durchgeführt und analysiert werden, keine Tradition und ist nur an wenigen Standorten nennenswert vertreten. In England oder den Niederlanden haben medizinische Unis meist Abteilungen für klinische Epidemiologie mit mehr als 50 Wissenschaftlern, und diese sind sogar spezialisiert für bestimmte medizinische Disziplinen. Aber warum ist das dort so, und bei uns nicht? Da beißt sich die Katze in den Schwanz: Es gab und gibt einfach in einem solchen Umfeld wenig Bedarf für Methodiker in Deutschlands medizinischen Fakultäten, weil eben relativ wenige Studien aus Academia durchgeführt werden. Aber dafür umso mehr für die Industrie, und die hat ihre eigene Expertise oder kauft diese von Firmen wie PAREXEL, QUINTILES, usw. ein. Das heisst auch, dass in Deutschland klinische Studien sich vor allem mit Fragen beschäftigen, welche für die Industrie interessant sind. Und diese decken sich nicht immer mit den Fragen, welche für Patienten relevant wären.

Und dann gibt es noch so etwas wie ein ‚Founder'-Effekt. In Nordamerika und England wurde von den 70er und 80er Jahren des vergangenen Jahrhunderts an die Evidenzbasierte Medizin (EBM) entwickelt. Deren Giganten Guyatt, Sackett, Cochrane, Chalmers et al. EBM gründeten Epidemiologie-Schulen, die auch heute noch in diesen Ländern das Fundament klinischer Studienexzellenz bilden. Und aus denen sich kompetenter Nachwuchs rekrutieren lässt.

Mein Fazit zur klinischen Studienkultur in Deutschland lautet also: Es gibt großen Nachholbedarf! Wir brauchen mehr Methodenkompetenz, also vor allem klinische Epidemiologie und Biostatistik, mehr Fördermittel, und vielleicht dafür ein bisschen weniger Auftragsforschung für die Industrie. Demut ist angesagt, und ein Blick nach England, das glückliche England! Da könnten wir viel lernen.

Abschließend möchte ich aber nicht versäumen, noch etwas zum ‚Wunder' der Express-Impfstoffentwicklung zu sagen. Selten wird erwähnt, dass die Technologie, auf der die mRNA Impfstoffe von BioNTech und Moderna beruhen, das Resultat von vielen Jahrzehnten Grundlagenforschung war. Abgesehen von der darin steckenden allgemeinen Molekular- und Biotechnologie, welche natürlich noch viel ältere Wurzeln hat, ging es bereits so richtig los in den frühen 90er Jahren des letzten Jahrhunderts. Damals hatte Katalin Karikó die Idee, exogen administrierte mRNA zu nutzen, um Zellen zur Synthese von Proteinen zu bewegen. Dies gelang ihr dann 2005, zusammen mit Drew Weissman, durch spezifische Modifikation der mRNAs, welche diese weniger immunogen machten. Parallel dazu, und ebenfalls über Jahrzehnte, wurden Nanopartikel entwickelt, welche die Aufnahme der mRNAs in die Zellen ermöglichten. Eine ähnlich langwierige und komplizierte Geschichte hat der virale Gentransfer, der nun die Basis für die Impfstoffe von AstraZeneca, Johnson&Johnson und anderen ist. Auch dieser war lange Zeit zunächst nur aus pur mechanistischem Interesse, dann weiter als Forschungstool entwickelt worden. Beide Technologien sind nicht das Produkt von Forschungsprogrammen zur Entwicklung von Impfstoffen! Deshalb eine weitere Moral von dieser Gschicht: Vergiss die Grundlagenforschung nicht - aber davon ein andermal mehr!

Der Wissenschaftsnarr dankt Prof.Dr. Lars Hemkens für Einblicke in seine aktuellen Studienergebnisse und anregende Diskussionen.

Wozu braucht der Doktor einen Doktor?

LJ 10/2021

Der Dr.med. ist eine Anomalie in der Welt der Dissertationen. In der Serie: ‚Unnütze oder schädliche akademische Grade in der Medizin' wollen wir uns nach der Habilitation (LJ 6/2021) diesmal mit der medizinischen Promotion befassen. Der Narr weiß, wovon er spricht. Nicht nur hat er den Titel selbst erworben, er hat sein Studium auch als Promotions-Ghostwriter teilfinanziert. Wie alle (Klein)Kriminellen kann er sich aber heute damit rechtfertigen, dass es ihm allzu leicht gemacht wurde. Womit wir mitten im Thema wären.

Mehr als 80 % der Absolventen der Medizin promovieren. Die Dissertation kann vor Abschluss des Studiums angefertigt werden, der Titel wird dann gleich nach dem letzten Staatsexamen verliehen. Und das macht die überwiegende Zahl der Mediziner so. Die Einarbeitung ins Thema, die Durchführung der Studie, die Auswertung und das Zusammenschreiben der Ergebnisse, all dies findet während des Studiums statt. Dass dies überhaupt nebenbei möglich ist, wirft natürlich nicht nur Fragen bezüglich des Umfangs und der Qualität solcher Dissertationen auf. Offensichtlich lasten die Aneignung der Inhalte und Fähigkeiten der Ausbildung zum Arzt die Studenten auch nicht sonderlich aus. Die Mehrzahl kann nebenbei nicht nur das selbständige wissenschaftliche Arbeiten erlernen, sondern an Studien teilhaben oder aufwendige Experimente durchführen, diese analysieren, interpretieren, und als Monographie oder als Fachartikel veröffentlichen.

All dies ist hinlänglich bekannt, und wird seit vielen Jahren kritisiert. Der Wissenschaftsrat hat sich mehrfach mit dem Thema befasst, und kommt zu dem Schluss, dass ‚Das wissenschaftliche Niveau der studienbegleitenden Doktorarbeiten [...] in der weit überwiegenden Zahl der Fälle nicht den Standards der Doktorarbeiten anderer naturwissenschaftlicher Fächer [entspricht].' Das European Research Council (ERC) akzeptiert Antragsteller mit Dr.med. nur, wenn diese nachweisen können, dass sie mehrere Jahre an ihrer Dissertation geforscht haben. Deutsche Mediziner/innen müssen daher zusätzlich eine Stelle nachweisen können, die eine PhD-Äquivalenz voraussetzt (z. B. Postdoc-Fellowship, Ruf auf eine Professur). Der ehemalige Vorsitzende der Wissenschaftsrats und Charité-Vorstandsvorsitzende Karl Max Einhäupl, formulierte es einmal so: „Überspitzt gesagt, untersuchen manche solcher Dissertationen irrelevante Fragestellungen mit unzulässigen Methoden und erhalten zu guter Letzt noch einen wohlklingenden Titel."

Das Ganze ist natürlich insbesondere für jene Promovenden bitter, welche aufwendige, über Jahre gehende, ausgezeichnete Forschungsarbeit geleistet haben, die sie dann oft auch erst einige Jahre nach der Approbation abschließen. Auch sie erhalten nur den Dr.med.(ioker), ein Titel der dem Dr.rer.nat./PhD nicht das Wasser reichen kann, obwohl ihre Arbeiten deren Standards entsprechen. Aber mir geht es um mehr als

Gerechtigkeit zwischen verschiedenen Disziplinen, Diskriminierung deutscher Antragsteller beim ERC, oder Spott von Naturwissenschaftlern über die mangelnden Qualitätsstandards bei medizinischen Titeln.

Denn die Kritik am Dr.med. hatte Wirkung. Die meisten deutschen Medizinischen Fakultäten haben die Diagnose akzeptiert, dass es der medizinischen Dissertation an Qualität mangelt. Sie wollten aber am Grundproblem, dass nämlich studienbegleitend promoviert wird, nicht rütteln. Sie wollen den Kuchen Essen, ihn aber trotzdem auf dem Teller behalten. Um Abhilfe zu schaffen wurden daher an vielen Fakultäten durchaus wohlmeinend erhöhte Qualitätsstandards für den Dr.med. definiert und dazu Promotionsvereinbarungen eingeführt, sowie strukturierte Ausbildungsangebote gemacht. Und als pièce de résistance das Desiderat formuliert, dass medizinische Promotionsarbeiten nach Peer Review in möglichst internationalen wissenschaftlichen Journalen erscheinen sollen. Natürlich mit den Promovenden als Erstautoren. Die Publikationspromotion soll nun also die Regel sein, und die Monographie überflüssig machen. Die letztere verstaubt ja ohnehin nur im Bücherschrank der Großeltern, die das alles mitfinanziert hatten. Und ein gehobenes Prädikat ist mit so einer Monographie mittlerweile auch gar nicht mehr zu erreichen. Ist damit nun das Ziel erreicht, und die ‚Die Promotion dient dem Nachweis der Befähigung zu vertiefter wissenschaftlicher Arbeit durch eine eigene, selbstständige und originäre Forschungsleistung, die zum Erkenntnisgewinn im Fachgebiet beiträgt'? So formuliert es z.B. die aktuelle Promotionsordnung der Charité. Ich fürchte nein, das Ganze hat das Problem sogar verschärft!

Denn am Grundübel hat sich rein gar nichts geändert: Dem studienbegleitenden Promovieren! Neben einem anspruchsvollen Studium, dessen Praktika und Prüfungen, vielleicht sogar einem Job um all das zu finanzieren, soll Kompetenz in wissenschaftlicher Methodik oder Studienauswertung erworben werden. Dabei sollen sich Studenten in diesen Methoden so einarbeiten, dass sie robuste und relevante Ergebnisse erzielen, die es wert sind, in einer Publikation der Fachwelt präsentiert zu werden. Weil dies im Peer Review passieren soll, muss die Arbeit nach dem Zusammenschreiben auch noch eingereicht, begutachtet, vermutlich ein oder mehrere Revisionen durchlaufen, bei Ablehnung nochmals bei einem anderen Journal eingereicht werden, etc. Dies bedeutet, dass ‚minimally publishable units' erzeugt werden müssen, und eine Story präsentiert, welche bei Gutachtern möglichst wenig aneckt. Und hier liegt der Hase im Pfeffer: Dadurch führen die neuen Regeln dazu, dass der Druck, in einem vorgegebenen Zeitraum publizierbare Ergebnisse zu erzeugen noch höher ist, als ohnehin schon. Damit wächst auch die Versuchung, Resultate selektiv auszuwählen, bei der Statistik durch multiple, aber nicht vorher festgelegte oder dafür korrigierte Vergleiche signifikante p-Werte zu erzeugen, Hypothesen nach Auswertung der Ergebnisse zu modifizieren oder gar auszutauschen ohne dies in der Arbeit kundzutun, etc. Das volle Programm nicht offengelegter ‚wissenschaftlicher Freiheiten' eben. Dazu darf man nicht vergessen, dass Medizindoktoranden in der Regel von Klinikern angeleitet werden, welche den größeren Teil ihres Arbeitstags am Patientenbett, im Hörsaal, oder auf Kongressen verbringen, und daher häufig nicht die optimalen Betreuer sind. Unter anderem auch deshalb, weil die meisten von ihnen auf dieselbe Weise in die Wissenschaft sozialisiert wurden – und sie sich ihre eigene Methoden- und Statistik-Kompetenz ‚on the job' angeeignet haben. Die Betreuung übernimmt deshalb auch häufig das technische Assistenzpersonal, oder gleich andere, etwas seniorere Studenten.

Auch das Problem der mangelnden Vergleichbarkeit der Promotionen wird durch die neuen Anforderungen nicht gelöst. Nach wie vor wird der Titel von Vielen im Schnelldurchgang erworben, während andere wie auch schon bisher Forschungsarbeiten

abliefern, die einem PhD oder Dr.rer.nat. absolut äquivalent sind. Auch hier wurde also nichts gewonnen.

Eine besondere Tragik der bundesweiten Versuche, den Standard des Dr.med. zu heben liegt darin, dass die ‚neue Promotionsordnung […] weltweite Anstrengungen reflektiert, die Robustheit, Reproduzierbarkeit und Werthaltigkeit biomedizinischer Forschung zu verbessern'. So jedenfalls formuliert es die Charité, typisch für viele medizinische Fakultäten in Deutschland. Nicht nur läuft man so Gefahr, das Gegenteil von robusten und reproduzierbaren Forschungsergebnissen zu erhalten. Sondern trägt auch zu dem Tsunami von wissenschaftlichen Publikationen bei, welche bestenfalls kleine inkrementelle Beiträge liefern, oft aber nur die Literatur mit Wertlosem weiter verdünnen. Diese Publikationen werden zwar zum Glück häufig gar nicht gelesen, erschweren aber auf Fall die Evidenzsynthese in systematischen Reviews und Meta-Analysen.

Aber warum wollen über 80% der Medizinstudenten den Dr.med. überhaupt erwerben? Zum einen geht es ihnen dabei häufig einfach um den Titel, sodass der Doktor auch einen Doktor hat! Zum anderen wollen manche Studenten herausfinden, ob ‚Wissenschaft' etwas für sie ist, ob sie Karriere in der akademischen Medizin machen wollen. Beide Anliegen sind nachvollziehbar und legitim. Das gegenwärtige Verfahren ist dafür aber ungeeignet. Den Titelerwerb könnte man z.B. wie in Österreich gestalten. Der akademische Grad wird dort bei Studienabschluss nach Verfassen einer einfachen Diplomarbeit verliehen. In Deutschland darf dieser Grad deshalb nur mit Zusatz als „Dr. med. univ." geführt werden. Für die Frage, ob Wissenschaft für einen als berufliche Perspektive in Frage kommt, eignen sich andere Formate viel besser, wie Hausarbeiten, Hospitationen, dafür zugeschnittene Kurse und Seminare. Die jetzt im Zuge der Reform des Dr.med. mancherorts aufgebauten strukturierten Ausbildungsangebote wären dafür bereits sehr tauglich. Außerdem sind diese eine gute Grundlage für die Ausbildung derer, die mit einer ‚echten', PhD äquivalenten Promotion nach dem Medizinstudium methodische Kompetenz erwerben wollen. Dazu gibt es an vielen Unis mit medizinischer Fakultät bereits Graduiertenkollegs und Programme, in denen Mediziner und Naturwissenschaftler in den Lebenswissenschaften ausgebildet werden und einen ‚soliden' PhD oder Dr.rer.nat. erwerben können. Das Rad muss also nicht nochmals neu erfunden werden. Obwohl dies Modell, also eine Art Dr. med. univ. fürs Praxisschild, und der PhD für die Forschungs-interessierten relativ einfach und zeitnah umzusetzen wäre, wagen sich einzelne Fakultäten dennoch nicht daran. Denn sie fürchten einen Konkurrenznachteil, wenn sie einen Dr.med. gar nicht mehr anbieten, die anderen Unis aber schon. Die Sache müsste also bundesweit umgesetzt werden, und das geht Vielen zu weit. Also machen alle weiter wie bisher.

Und dann gibt es ja noch das Argument, dass durch Promotionsphase nach dem Studium, Facharzt, und Habilitation Mediziner ihre Ausbildung erst kurz vor der Berentung abschließen könnten. Da ist sogar was dran. Aber man müsste ja nur die Habilitation abschaffen, welche der Narr in einer früheren Folge dieser Kolumne als unnützes Jodeldiplom entlarvt hat.

Die Abschaffung des Dr.med. würde allerdings auch die Ghostwriter medizinischer Promotionen brotlos machen. Googlen sie mal danach, sie werden sehen, das ist immer noch ein florierendes Geschäft!

Das Märchen von denen die auszogen, der Alzheimer'schen Krankheit den Garaus zu machen

LJ 11/2021

Es war einmal, vor gar nicht langer Zeit. Die biomedizinische Wissenschaft entschlüsselte Schritt für Schritt die Entstehungsmechanismen einer der furchtbarsten und gleichzeitig häufigsten Erkrankungen des Menschen. Wenn wir nur alt genug werden, befällt sie die meisten von uns. Forscher aus den verschiedensten Gewerken der medizinischen Forschung, Genetik, Molekularbiologie, Histopathologie machten sich gemeinsam auf, ihre Mechanismen zu verstehen und eine Therapie zu entwickeln. Mit Tiermodellen konnte man die humane Pathologie rekapitulieren. Eine elegante und dabei recht simple biochemische Theorie wurde entwickelt, welche Entstehung und Symptome der Erkrankung erklären konnte. Aus der Theorie ließ sich direkt eine Therapie ableiten, welche die Progression der Krankheit stoppen, vielleicht sogar umkehren könnte. Gleichzeitig wurden bildgebende Methoden entwickelt, welche das Krankheits-auslösende Protein im menschlichen Gehirn schon vor dem Einsetzen erster Symptome anzeigten. Und damit Diagnose und Therapiekontrolle ermöglichten. Aus dem Blut von Menschen, welche ganz offensichtlich resistent gegen diese Erkrankung waren, wurden Antikörper gegen das krankheitsauslösende, fehlgefaltete Protein isoliert. Diese konnten rekombinant industriell hergestellt werden und erwiesen sich in großen klinischen Studien als wirksam. Sie wurden deshalb im beschleunigten Verfahren von der FDA als Medikamente zugelassen. Zum ersten Mal war damit eine Therapie gefunden, welche den Verlauf dieser schrecklichen Erkrankung beeinflussen konnte. Ein Triumph der medizinischen Forschung und des Zusammenspiels von akademischer Wissenschaft, Pharmaindustrie, und Zulassungsbehörden. Die Krankheit hatte, wie von den Forschern vorhergesagt, viel von ihrem Schrecken verloren. Die Lebensqualität von Abermillionen von Patienten und deren Angehörigen verbesserte sich, die Aktienkurse der beteiligen Pharmakonzerne stiegen unaufhaltsam. Bald darauf wurde den Erstbeschreibern des Pathomechanismus ein Nobelpreis verliehen. Ein modernes, medizinisches Märchen! Und wenn sie nicht gestorben sind, dann leben sie noch heute.

Vieles davon ist tatsächlich passiert. Das auf der kanonischen Amyloid-Hypothese basierende Medikament Aducanumab der Firmen Biogen und EISAI, ein rekombinanter humaner Antikörper gegen aggregierte lösliche und unlösliche Formen des Amyloid beta (Aβ), wurde in diesem Jahr tatsächlich von der FDA zugelassen. Allerdings eine Therapie, die zwar Aβ im Gehirn reduziert, aber dennoch keine gesicherte Wirkung auf den Krankheitsverlauf hat. Gesichert ist dagegen, dass die Nebenwirkungen dieser Therapie schwer sein können und häufig sind. Und die Therapie im Jahr fast 60.000 US$ kostet, die dazu nötige sehr teure Diagnostik gar nicht eingeschlossen. Dieses Medikament hat also nicht nur medizinische, sondern auch ökonomische Toxizität. Wegen der Lehren,

die sich aus dieser Geschichte ziehen lassen, lohnt es sich genauer hinzuschauen, und die Frage zu stellen, wie es soweit kommen konnte.

Alois Alzheimer legte bereits in seiner Erstbeschreibung der nach ihm benannten Hirnerkrankung eine Fährte, welcher die Wissenschaft der kommenden 100 Jahre wie hypnotisiert folgen sollte. Alzheimer hat nämlich nicht nur die klinischen Symptome, also im Wesentlichen die Demenz als typisches Merkmal der Erkrankung beschrieben, sondern auch die dazugehörige, charakteristische Gehirn-Pathologie. Er fand in den Gehirnen der Patienten zugrunde gegangene Nervenzellen sowie Eiweißablagerungen, die sogenannten Plaques. So richtig los mit der Pathophysiologie ging es dann in den 80er Jahren des vergangenen Jahrhunderts, als man die Plaques als extrazellulär deponiertes Aβ, und in den Zellen Fibrillen das Tau-Protein identifizieren konnte. Von da an ging es Schlag auf Schlag: Die Biochemie des Amyloidstoffwechsels wurde aufgeklärt, mit all den zugehörigen Enzymen. Genetiker fanden in Familien von Patienten, welche an seltenen erblichen Demenzformen litten, Mutationen in Amyloid-Vorläuferproteinen und Amyloid-prozessierenden Enzymen. Ein bestimmtes, in Plaques konzentriertes Amyloidfragment stellte sich in Experimenten an Zellkulturen und Versuchstieren als toxisch für Nervenzellen heraus. Gentechnisch veränderte Mäuse, in deren Genom man die mutierten Gene von Patienten mit dominant vererbter Alzheimer – Erkrankung inseriert hatte, entwickelten Plaques im Gehirn. Reduziertes Amyloid $beta_{(1-42)}$ im Nervenwasser korrelierte mit der Erkrankung. Auch konnten nuklearmedizinische Kontrastmittel entwickelt werden, welche das Aβ im Gehirn von Patienten nicht-invasiv bildgebend darstellen und sogar quantifizieren konnten. Die Aβ-Bildgebung hat damit eine hohe Sensitivität für den Nachweis der Alzheimer-Pathologie. Darüber hinaus konnte sie das Fortschreiten von leichter kognitiver Beeinträchtigung zur Alzheimer'schen Erkrankung vorhersagen. Da falsch prozessiertes Amyloid, welches sich im Gehirn ablagert, Neuronen zerstören kann, kann es zu Hirnleistungsstörung und Demenz führen. Verschiedene Strategien zur Elimination des Übeltäterproteins wurden entwickelt und klinisch getestet. Das Prinzip: Aktive Immunisierung (Vakzinierung), bzw. monoklonale Antikörper, sodass letztlich das eigene Immunsystem den Eiweismüll einsammelt und beseitigt. Auch die Antikörper wurden der Natur abgeschaut, isoliert aus Menschen im hohen Lebensalter, die geistig noch überdurchschnittlich fit waren. Und siehe da, Vakzinierung und Antikörper waren in der Lage, Plaques aus den Gehirnen von Alzheimer-Patienten zu beseitigen! Auch in Mäusen, welche Gene von Patienten mit der erblichen Form der Erkrankung exprimierten und ebenfalls Plaques entwickelten, hatte das funktioniert.

Dieses eindrucksvolle Universum vielfältiger Evidenz machte die Amyloid-Hypothese der Alzheimer Erkrankung zu einer scheinbar wasserdichten, linearen Theorie, welche sich innerhalb von 30 Jahren zum absoluten Dogma von Wissenschaft und Industrie entwickeln konnte. Sie hatte einen einzigen Schönheitsfehler: Die daraus abgeleitete Therapie beseitigte Aβ, nicht aber die Demenz. Aber einer Erklärung hierfür war schnell gefunden: Die Therapie kommt zu spät, der Schaden im Gehirn war vorher schon passiert! Also muss vor Krankheitsbeginn therapiert werden! Aber seit kurzem ist klar: Auch dies scheint nicht zu funktionieren. Genträger der dominant vererbten Alzheimer Erkrankung wurden Jahre vor Ausbruch der Erkrankung behandelt. Sie litten unter den bekannten Nebenwirkungen der Therapie, aber wurden dennoch symptomatisch.

Nun ist es aber keineswegs so, dass all dies überraschend kam. Bereits während sich die Amyloid-Hypothese zum Dogma entwickelte, gab es eine Vielzahl von Hinweisen, dass alles vielleicht doch anders, auf jeden Fall aber viel komplizierter sein könnte. Zum Beispiel fand man, dass es alte Menschen gibt, deren Hirne voller Plaques sind, diese aber trotzdem geistig voll leistungsfähig. Auch die Tierversuche waren im Nachhinein prädiktiver, als es den Alzheimer-Forschern lieb sein konnte: Die Tiere mit den Plaques

entwickelten nämlich gar keine ‚Demenzäquivalente', deshalb konnte die ‚Plaqueauflösung' bei diesen auch klinisch gar nichts rechtes bewirken. Überhaupt die Tierversuche: Sehr häufig geringe Fahlzahlen, niedrige interne Valididät (z.B. fehlende Verblindung), fragwürdige Statistik, selektive Auswahl von Ergebnissen, nicht-Veröffentlichung von negativen Resultaten, usw. Also das volle Programm der Qualitätsprobleme, wie sie auch in anderen Forschungsgebieten gang und gebe waren, bzw. immer noch sind. Der Narr hat sich hierüber an dieser Stelle schon öfters aufgeregt. Auch diesmal gibt es unter http://dirnagl.com/lj Literaturhinweise, welche die gravierenden Qualitätsprobleme in der experimentellen Alzheimerforschung dokumentieren. Aber auch die großen klinischen Anti-Amyloid Antikörper- und Vakzinierungs-Studien fuhren ein negatives Ergebnis nach dem anderen ein. Fast scheint es, als hätte eine Folie à deux die akademischen Forscher und die Pharmaindustrie erfasst.

Denn schon früh traten auch Mahner auf den Plan, welche auf Probleme der Amyloid-Hypothese hinwiesen und alternative Mechanismen ins Spiel brachten. Sie wurden aber vom wissenschaftlichen Mainstream bestenfalls ignoriert, oder aber deren Arbeiten aus den Top-Journalen gegutachtet, und ihre Förderanträge abgelehnt.

Für eine gewisse Zeit gab es sogar noch so etwas wie eine wissenschaftliche Kontroverse, ja sogar richtigen Streit. Die ‚Tauisten' betonten, dass ein weiteres histopathologisches Merkmal der Erkrankung, die intraneuronalen Tau-Faser Versteifungen und Ablagerungen viel besser mit dem Krankheitsverlauf der Erkrankung korrelieren als das Amyloid. Tauisten und Baptisten - die Anhänger der Amyloid-Hypothese - führten regelrecht Krieg! Bis man um die Jahrtausendwende das Kriegsbeil begrub und sich einigte, dass die Taupathologie biochemisch downstream vom Amyloid auftritt. Die Tauisten konvertierten daraufhin in Scharen zum Baptismus!

Trotz der aufziehenden dunklen Wolken entwickelte die Pharmaindustrie ein ähnliches Skotom wie die universitären Forscher. Nachdem sich dann Eli Lilly, Pfizer, Roche und MSD negative Studien und Aktienkurseinbrüche geholt und das Feld verlassen hatten, machten die Firmen Biogen und EISAI weiter. Und waren, so scheint es, auf den ersten Blick erfolgreich. Im Juni dieses Jahrs wurde Aducanumab, der Amyloid-Antikörper aus dem oben angerissenen Wissenschaftsmärchen, als erster Krankheits-modifizierender Wirkstoff für die Therapie der Alzheimer'schen Erkrankung zugelassen. Der letzte fehlende Puzzlestein im Siegeszug der Amyloidhypothese? Biogen und EISAI gelang es tatsächlich, ein sündteures, in den eigenen Studien ziemlich unwirksames aber nebenwirkungsreiches Medikament gegen das Votum der FDA-Spezialisten und Gremien zur Zulassung zu bringen! Sekundiert wurden sie dabei von Patientenorganisationen, welche großzügig von Biogen gefördert worden waren, und auch Wissenschaftlern, welche ihre Karrieren auf Amyloid gebaut hatten. Und über die Jahre als Berater, Vortragende, und ‚Key opinion leaders' der Industrie ebenfalls gut verdient hatten. Die FDA, die ja schon immer als industrienah galt, hat sich mit diesem Verfahren komplett unmöglich gemacht. Untersuchungsausschüsse klären zurzeit, was da passiert ist. Vielleicht wird es auch den Freedom of Information Act brauchen, um Dokumente ans Licht der Öffentlichkeit zu bringen, welche diesen Skandal aufklären.

Und wer zahlt die Rechnung für all dies? Natürlich zunächst die Alzheimer Patienten und deren Familien, die sich verschulden und sich an den Strohhalm der falschen Hoffnung klammern. Wohl aber auch wir alle, nicht nur als Steuerzahler und Krankenversicherte. Sondern auch als potentielle zukünftige Alzheimer Patienten. Weil eine Forschungsmonokultur, welche über 30 Jahre abgeschottet und unbeirrbar in einer Echokammer vor sich hinforscht, bei einer komplexen Hirnerkrankung erfolglos bleiben muss. Weil es keine Alzheimer'sche Erkrankung gibt, sondern viele Alzheimer'sche

Erkrankungen. Weil wohl Aβ und Tau, aber auch Inflammation, Mikroglia, Mitochondrien, Mikrogefässe und vieles mehr in komplizierter Weise zusammenspielt. Wird das Hirn nur alt genug, hält es den konzertierten Anschlägen dieser Mechanismen nicht mehr stand und wird in seiner Funktion gestört. Dabei spielen komplexe Gen-Interaktionen genauso eine Rolle wie eine Vielzahl von Umweltfaktoren. Vieles von dem, was in den letzten Dekaden von den Alzheimer Forschern herausgefunden wurde, kann sehr wohl nützliche Beiträge in der Aufklärung der komplexen Pathologie von Demenzerkrankungen liefern, war deshalb nicht vergeudet. Aber hier hat sich ein ganzes Feld, von wenigen Ausnahmen abgesehen, mit einer letztlich naiven Theorie verzockt, und die Wissenschaft am Ende sehr viel Zeit verloren und Ressourcen verschwendet, die man besser breit gestreut hätte.

An dieser Stelle erinnern wir uns an Max Planck, der 1948 in seinen Lebenserinnerungen schrieb: „Eine neue wissenschaftliche Wahrheit pflegt sich nicht in der Weise durchzusetzen, dass ihre Gegner überzeugt werden und sich als belehrt erklären, sondern dadurch, dass die Gegner allmählich aussterben und dass die heranwachsende Generation von vornherein mit der Wahrheit vertraut gemacht ist."

Preprints – Heilsbringer oder apokalyptische Reiter des wissenschaftlichen Publizierens?

LJ 12/2021

Am 24. November wurde dem Physiker Paul Ginsparg von der Cornell Universität der mit 200.000 € dotierte *Einstein Award for the Improvement of Research Quality* verliehen. Mit diesem dieses Jahr zum ersten Mal von der Damp Stiftung und dem Land Berlin verliehenen und mit insgesamt 500.000 € dotierten Preis werden ForscherInnen, Institutionen und Nachwuchswissenschaftler aus der ganzen Welt ausgezeichnet, die einen wesentlichen Beitrag zur Verbesserung der Qualität und Robustheit von Forschungsergebnissen geleistet haben. Aber wer zum Teufel ist Paul Ginsparg, und warum hat ihn eine hochkarätig besetzte internationale Jury ausgewählt? Nur wenigen ist bekannt, dass Ginsparg die Welt des wissenschaftlichen Publizierens revolutioniert hat. Er hat nämlich 1993 den Preprintserver ArXiv gegründet, als elektronisches Austauschforum einer kleinen Community von theoretischen Physikern. Der Rest ist Geschichte, denn spätestens seit Corona werden auch in den Lebenswissenschaften immer mehr Artikel vor Einreichung bei einem Peer-Review Journal zunächst als Preprint gepostet, z.B. bei BioRxiv oder MedRxiv. Also ohne

formalen Review, für jedermann zugänglich, auch für die nicht-wissenschaftliche Öffentlichkeit. In den Wissenschaftsseiten der Tageszeitungen liest man deshalb jetzt häufig die Formulierung: ‚In einer noch nicht begutachteten Studie haben Wissenschaftler …

Mit Preprints publizieren Wissenschaftler sich selbst, also ganz ohne Verlage. Sozusagen ein YouPublish, eine Art Youtube für Wissenschaftler! Aber hoppla, schafft das nicht ein Riesenproblem? Wieso wird man für so eine ganz offensichtliche Narretei mit einem der höchst dotierten Wissenschaftspreise ausgezeichnet, noch dazu einem für die Verbesserung von Qualität? Öffnen Preprints denn nicht das Tor zur Hölle der Pseudowissenschaften, zumindest der Vorhölle der nicht durch die Fachwelt begutachteten und deshalb potentiell fehlerbehafteten Studien?

Zunächst mal das offensichtlich Positive an Preprints: Sie bringen eine absolute Beschleunigung des wissenschaftlichen Austausches. Das ist nicht nur, aber besonders in einer Pandemie wichtig. Ein Beispiel: Britische Forscher veröffentlichten im Juni 2020 einen Preprint, in der sie die Wirkung von Kortikosteroiden auf COVID-19 untersuchten und zeigen konnten, dass diese die Mortalität von hospitalisierten COVID-19 Patienten senken. Unmittelbar nach Veröffentlichung stieg der Corticosteroideinsatz bei solchen Patienten von 30 % auf 92 %. Unzählige Menschleben konnten im Intervall, welches bis zur Veröffentlichung des Peer-Review Artikels vergangen wären, auf diese Weise gerettet werden.

Bis zur Publikation eines Artikels im peer-review Verfahren dauert es nämlich fast immer mehr als ein Jahr, in der Pandemie oft etwas weniger, aber sonst häufig noch viel länger. Denn wir versuchen als Autoren, mit jedem Artikel so viel Renommee, d.h. konkret Journal Impactpunkte, zu schürfen wie möglich. Das bedeutet in der Regel, es zunächst bei Journals zu versuchen, bei denen man schon rein stochastisch wenig Chancen hat. Und dann in einer Kaskade der abnehmenden Impactfaktoren es so lange mit Resubmissionen zu versuchen, bis sich endlich ein Journal erbarmt. Jedes Mal mit neuer Formatierung, neuen Reviews, neuen Rebuttals, usw. Ein Reprint dagegen ist sofort auf dem Markt des wissenschaftlichen Diskurses – ums Renommee kann man sich danach ja immer noch bemühen. Dafür bekommt man durch den Preprint möglicherweise sehr hilfreiche Kommentare – über die Comment – Funktion des Servers, via Twitter, Email, usw. Dies kann zu neuen, besseren Versionen des Manuskripts führen, die dann gleich wieder versioniert auf dem Preprint-Server erscheinen können. Gratis gibt es dabei Schutz vor Scooping, denn man bekommt ja den Credit für sein Werk sofort nach Fertigstellung des Manuskripts. Auch ermöglichen es Preprints, problemlos NULL-Resultate zu publizieren. Also z.B. Studien in denen man die eigene Hypothese nicht bestätigen konnte, oder es sonst wie Überraschungen gab, die eine reguläre Publikation zumindest erschweren würden.

Wo viel Licht, muss auch viel Schatten sein? Zweifellos wird da auch etliches hochgeladen, was wissenschaftlich fragwürdig, wenn nicht sogar unsinnig oder betrügerisch ist. Und von Obskuranten dann als Beleg für deren abstruse Theorien genutzt wird, wie in der Pandemie tatsächlich geschehen. Z.B. ein Preprint, der nahelegte, dass die Struktur von SARS-CoV-2 "unheimliche" Ähnlichkeiten mit HIV aufweist. Aber auch Studien zu möglichen Nebenwirkungen der lange gehypten Chloroquinderivate, oder dem vermeintlichen Risiko der Behandlung von COVID Patienten mit Medikamenten, die am Angiotensinsystem ansetzen. Nur: Die letzteren beiden Preprints überstanden unbeschadet den Review Prozess in Lancet und im New England Journal of Medicine. Und mussten deshalb zurückgezogen werden. Das ist anekdotisch, aber mittlerweile gibt es eine Reihe von Studien, welche das Schicksal von Prepreprints verfolgt haben. Dabei

kommt heraus: Mindestens 70 % davon werden danach in Peer Review Journalen veröffentlicht. Das bemerkenswerteste daran: Preprint und daraus entstandener Artikel unterscheiden sich in der Regel kaum, manchmal ändern sich Tabellen und Abbildungen ein wenig, manchmal auch etwas der Spin, aber die Kerndaten und Aussagen bleiben bestehen.

Mit anderen Worten, der Peer Review ist keineswegs der Filter, für den man ihn immer hält. Am schönsten hat dies Drummond Rennie, einer der Editoren von JAMA bereits 1986 formuliert: *‚Trotz [des Begutachtungssystems] muss jeder, der die Fachzeitschriften aufmerksam und kritisch liest, feststellen, dass es kaum Hindernisse für eine Veröffentlichung gibt. Keine Studie scheint zu bruchstückhaft, keine Hypothese zu trivial, keine Literatur zu einseitig oder zu partikulär, kein Design zu verzerrt, keine Methodik zu stümperhaft, keine Ergebnisdarstellung zu ungenau, zu obskur, zu widersprüchlich, keine Analyse zu biased, kein Argument zu trivial oder zu ungerechtfertigt und keine Grammatik und Syntax zu anstößig zu sein, als dass eine Arbeit nicht in den Druck gehen könnte. Die Funktion der Peer Review kann also darin bestehen, nicht zu entscheiden, ob, sondern wo eine Arbeit veröffentlicht wird.‘* Daran hat sich nichts geändert.

Dass in der Pandemie schlechte, als Preprint veröffentlichte Studien die öffentliche Gesundheit mehr gefährden als Peer Review Artikel, ist deshalb nicht haltbar. Im Gegenteil, insofern Preprints immer den Disclaimer ‚Achtung: Nicht Peer Reviewed‘ tragen, sind sie weniger gefährlich als wenn sie fast wortgleich in Journalen erscheinen, aber das vermeintliche Gütesiegel ‚Peer Review‘ tragen. Leider ergab eine Studie, dass etwas über 40% der Artikel in der Laienpresse, die über in Preprints publizierte Forschung berichten, die Ergebnisse nicht korrekt als wissenschaftlich unsicher oder ohne Peer-Review darstellen. Aber: Medizinische Fachartikel, ob nun als Preprint oder normaler Journalartikel, sind doch nie dazu angetan, Laien Handreichungen zu deren persönlicher Gesundheitsvorsorge oder Therapieempfehlungen zu geben. Dies passiert nur wenn Journalisten ihr Handwerk nicht verstehen und ihnen die nötige Grundskepsis gegenüber publizierter Evidenz fehlt. Dem Format Preprint ist das jedenfalls nicht anzulasten.

Aber wie ist das eigentlich mit den Preprints und dem Peer Review? Per Definitionem handelt es sich ja um nicht gereviewtes Material. Aber auch dies stimmt häufig nicht mehr so recht. Zum einen schalten biomedizinische Preprint Server einen Qualitätscheck voraus, der verhindern soll, dass totaler Nonsense oder offensichtlich Gefährliches hochgeladen wird. Aber noch wichtiger: Die wissenschaftliche Community rezipiert doch die Preprints und diskutiert diese, sofern die Fragestellung und Ergebnisse eine Relevanz für das Feld haben. Darin besteht ja der ganze Sinn des Preprints – deshalb haben sie sich ja auch in der Physik so durchgesetzt. Dort sind in einigen Bereichen Preprints das Hauptmedium des wissenschaftlichen Diskurses. Diese füllen dann auch die Lebensläufe und Anträge, und sind über ihre Inhalte und deren Rezeption in der Community Reputations-relevant.

Und genau hier geht in den Lebenswissenschaften derzeit die publikatorische Post ab! Elife, mittlerweile eines der renommiertesten Journale in der Biomedizin, begutachtet nur noch was vorher als Preprint publiziert wurde. Mit Review Commons haben ASAPbio und EMBO eine Plattform geschaffen, über die man Preprints zu deren Review einreichen kann, ohne sich für eines der bisher 17 teilweise hochrenommierten Journale entschieden zu haben. Nach dem Review und möglicherweise Revision wählen die Autoren dann eines der Journale. Die Artikel werden also Journal-agnostisch, nur nach wissenschaftlichen Meriten bewertet, während sie schon als Preprint für die Community verfügbar sind. Wird der Artikel von dem ausgewählten Journal nicht angenommen,

kann er inklusive der Reviews bei einem der anderen beteiligten Journale eingereicht werden und muss nicht nochmal begutachtet werden. Die Reviews werden übrigens, wenn die Autoren zustimmen, zusammen mit dem Preprint veröffentlicht. Der ganze Prozess wird dabei von Review Commons koordiniert. Sciety ist eine andere spannende Initiative, welche die offene Evaluation und Kuratierung von Preprints innerhalb eines wachsenden Netzwerks von Wissenschaftlern strukturiert organisiert.

Manches von all dem mag noch im Erprobungsstadium oder gerade erst an der Schwelle der vollen Funktionalität und Nützlichkeit sein. Auch hängen viele dieser Initiativen noch am Tropf von Stiftungen wie Chan Zuckerberg Initiative, Helmsley Charitable Trust oder Arnold Ventures, sowie Fördergebern wie Howard Hughes Medical Institute (USA) oder MRC (UK). Sie müssen also erst noch nachhaltige Geschäftsmodelle etablieren. Die eingeschlagene Richtung scheint aber schon jetzt unumkehrbar – Preprints als primäres Disseminationsmedium wissenschaftlicher Erkenntnis und damit Dreh- und Angelpunkt einer Umwälzung des Publikationswesens. Nicht wirklich überraschend ist leider, dass die meisten deutschen Fördergeber und Institutionen gerade dabei sind, diese Entwicklungen zu verschlafen – die Musik spielt wie so häufig anderswo.

Die Tatsache, dass die Filterfunktion des Review Prozess ein unerreichtes Ideal ist, und Preprints den Erkenntisaustausch stark beschleunigen, Ressourcen schonen, etc., wirft natürlich eine Menge Fragen auf, bis hin zu der, wozu es dann überhaupt noch Journale braucht? Dass es diese in der gegenwärtigen Form gibt, begründet sich nicht unwesentlich darin, dass dem derzeitigen wissenschaftlichen Publikationssystem eine Schlüsselrolle in der ademischen Reputationsökonomie zukommt. In der Publikation geht es ja nur teilweise darum, wissenschaftliche Evidenz unter die Leute zu bringen. Mindestens so wichtig ist es, mittels der Publikation Renommee zu erwirtschaften. Die Journale verkaufen uns dies Renommee, das in seiner abstraktesten Form aus den Journal Impact Factor Punkten besteht. Diese kaufen die Wissenschaftler – bzw. die institutionellen Bibliotheken, und damit der Steuerzahler - bei den Verlagen und tauschen sie dann gegen Drittmittel, Verstetigung, oder Professuren ein. Preprints führen uns derzeit vor Augen, dass es offensichtlich möglich ist, den Mittelmann, d.h. die Verlage aus der Gleichung zu nehmen. Der Haken dabei: Solange Preprints nicht Reputations-wirksam sind, muss auf jeden Fall weiter eine klassische Publikation angestrebt werden.

Preprints haben also revolutionäres Potential, weshalb der Einstein Preis für Ginsparg absolut verdient ist. Diejenigen, denen es gelingt, das akademische Reputationssystem so zu verändern, dass Preprints ihre volle Kraft entfalten können, sollten gleich für einen der nächsten Einstein Preise nominiert werden.

Ulrich Dirnagl ist Wissenschaftlicher Sekretär des Einstein Award for the Improvement of Quality in Research.

Wie Hanna beinahe das deutsche Wissenschaftssystem reformiert hätte

LJ 1-2/2022

In den Wissenschaftssystemen vieler Länder hangeln sich Forscher von Vertrag zu Vertrag, permanente Anstellungen sind die Ausnahme. 90 % der Wissenschaftler in Deutschland arbeiten zeitlich befristet. Dies erschwert oder gar verunmöglicht ihnen eine rationale Lebensplanung. Hyperkompetition und perverse Anreize sowie steile Hierarchien führen zu einem brain drain hochkompetenter und eigentlich maximal motivierter Forscher, welche die Wissenschaft frustriert verlassen. Wer im System verbleiben darf, entscheidet sich oft an Indikatoren, welche vorgeben, Qualität und Innovation zu objektivieren, diese dabei aber negativ beeinflussen. Nicht nur die Betroffenen fordern daher eine grundlegende Reform des Wissenschaftssystems.

Und so eine Reform wurde in Deutschland tatsächlich im Jahre 2007 mit dem Wissenschaftszeitvertragsgesetz eingeleitet. Dieses Gesetz begrenzt die Möglichkeit, Wissenschaftler nach der Promotion befristet einzustellen auf 6 Jahre. Danach ist nochmals eine Qualifizierungsphase von weiteren 6 Jahren möglich (in der Medizin 9 Jahre), z.B. auf dem Weg zur Professur. Eine Weiterbeschäftigung an Universitäten ist dann nur noch auf einer Dauerstelle möglich. Dort kam es aber in der Folge nicht zu dem hierfür nötigen Aufwuchs an unbefristeten Stellen. Im Gegenteil, durch große nationale Wissenschaftsförderprogramme wie die ‚Exzellenz-Initiative' hat sich seither der Pool an befristeten Doktoranden und Postdoc Stellen sogar noch massiv erhöht, damit wurde die Konkurrenz um die weiterhin stark begrenzten Dauerstellen weiter angefacht. Damit hat das Gesetz seine intendierte Wirkung nicht nur verfehlt, ja es führt sogar zu einer verschärften ‚Aussortierung' von gut ausgebildeten und bewährten Wissenschaftlern. Diese würden zwar gerne weiter forschen, die Mittel im System wären dafür sogar vorhanden, allerdings eben nur für befristete Anstellung. Und diese muss ihnen aus rechtlichen Gründen verwehrt werden.

Seither rumort es noch mehr unter Deutschlands Early Career Researchers (ECRs). Das BMBF fühlte sich daher im Sommer 2021 bemüßigt, ein beschwichtigendes Erklär-Video ins Netz zu stellen. Die fiktive Postdoc ‚Hanna' erklärte darin graphisch wie intellektuell auf Kindergarten-Niveau anderen ECRs die akademischen Befristungsregeln. Hanna erläuterte, dass das Gesetz dazu diene, das Wissenschaftssystem vor dem ‚Verstopfen' mit Postdocs auf Dauerstellen zu bewahren. Diese Postdocs kennzeichnete sie als Teile einer ‚akademischen Wertschöpfungskette', welche nur zu Innovation führen könne, wenn ein substantieller Teil der Wissenschaftler kontinuierlich aus dem System ausgebucht würde. Tatsächlich ist Hannas Sichtweise weit verbreitet, insbesondere bei der Professorenschaft und bei den Univerwaltungen. Allerdings waren die Wortwahl und die graphische Anmutung des Videos so zynisch und offensiv, dass es einen Shitstorm in den

sozialen Medien auslöste. Und das Ministerium das Video ganz schnell wieder aus dem Netz nehmen musste. Wer es verpasst hat, auf Youtube ist es von anderen Hannas archiviert worden.

Unter dem Hashtag #IchbinHanna entlud sich der Zorn und die Frustration der vom BMBF über ihre dienende Rolle und unsichere Zukunft im deutschen Wissenschaftssystem belehrten ECRs. Dies wurde von vielen Medien, sogar international, aufgegriffen und kommentiert. Es war aber reiner Zufall, dass sich just zu dieser Zeit die Parlamentarier des Landes Berlin mit einer Novelle ihres Hochschulgesetzes befassten. Die Abgeordneten der regierenden Koalition aus SPD, Linken und Grünen zeigten sich beeindruckt von der Kampagne der ECRs und dem Medienrummel. Aber auch die Politik war seit Jahren frustriert über die Berliner Universitäten. Diese hatten in den Hochschulverträgen, welche ihre Finanzierung durch das Land regeln, immer eine Erhöhung des Anteils von unbefristeten Mittelbaustellen zugesagt. Aber diese nicht geschaffen. Also griffen die Parlamentarier wenige Monate vor dem Ende ihrer Legislaturperiode und ohne Rücksprache mit den Universitäten zu einem drastischen Mittel: Sie fügten einem bereits existierenden Gesetzes-Paragraphen einen wirkmächtigen Satz hinzu: *‚Sofern der wissenschaftliche Mitarbeiter oder die wissenschaftliche Mitarbeiterin bereits promoviert ist und es sich bei dem im Arbeitsvertrag genannten Qualifikationsziel um eine Habilitation, ein Habilitationsäquivalent, den Erwerb von Lehrerfahrung und Lehrbefähigung oder um sonstige Leistungen zum Erwerb der Berufungsfähigkeit gemäß § 100 handelt, ist eine Anschlusszusage zu vereinbaren.‘* Im Klartext: Universitäten müssen Postdocs, welche sich auf einer Stelle qualifizieren, z.B. um Professor zu werden, bereits bei der ersten Anstellung die Übernahme auf eine unbefristete Stelle zusichern. Was genau in diesem Kontext als Qualifikationsziel gilt, und was die Kriterien sein müssten, welche bei Einstellung vereinbart und bei Erreichen zur Verstetigung führen, lässt das Gesetz allerdings offen.

Obzwar das Gesetz momentan noch nicht umgesetzt wird, brach bei den Universitäten akute Panik aus. Die Präsidentin der Berliner Humboldt Universität trat zurück, die Freie Universität Berlin stoppte sofort alle Einstellungen von Postdocs, die Berlin University Alliance (ein Zusammenschluss der vier größten Berliner Universitäten und der Charité im Rahmen der Exzellenzstrategie) beendete laufende Verfahren zur Einrichtung von Nachwuchsgruppen, usw. Die Begründung der Universitäten für diese drastischen Reaktionen: Die Umsetzung des Gesetzes würde zwar eine erste Kohorte von Studenten glücklich machen, aber die Nachfolgenden komplett im unakademischen Regen stehen lassen, denn dann seien ja alle Dauerstellen schon vergeben. Zudem würden die Berliner Unis, gerade in der Exzellenz-Strategie als exzellent geadelt, im internationalen Wettbewerb zurückfallen. Nach Berlin zu berufenden Professoren könnten nun mangels disponibler Stellen keine personellen Ausstattungsangebote mehr gemacht werden können. Die zurückgetretene Präsidentin der Humboldt Universität, Sabine Kunst, formulierte es so: Das *‚Gesetz ist gut gemeint, aber schlecht gemacht‘*. Ohne zusätzliche Stellen und ohne Übergangsphase schlechterdings nicht umsetzbar.

Da ist natürlich was dran, und mit dem nachvollziehbaren Ruf nach zusätzlichen Mitteln zur Schaffung von Dauerstellen enden deshalb auch die meisten Diskussionen und Medienberichte zu dieser Gesetzesänderung. Das ist ein Fehler, denn es geht hier um sehr viel mehr als um prekäre Anstellungsbedingungen für junge Akademiker und die strukturelle Unterfinanzierung der Universitäten. Wir sollten vielmehr das Mantra in Frage stellen, nach der Dauerstellen das System verstopfen, oder verstetigte Akademiker weniger kreativ sind. Und begreifen, dass Dauerstellen und akademischer Mittelbau sehr viel mit der Qualität, der Vertrauenswürdigkeit, und damit auch der Reproduzierbarkeit von Forschung zu tun haben. Alles Dinge um die wir uns derzeit Sorgen machen müssen.

Zunächst einmal ist festzuhalten, dass dieses Mantra meist von denen rezitiert wird, die selbst auf Dauerstellen sitzen. Jene Professoren würden allerdings den Vorwurf, dass sie wegen der Verstetigung träge geworden und weniger engagiert und innovativ forschen, vehement zurückweisen. Auch ein Blick auf die Industrie, in der Entfristung nach Probezeit die Regel ist, liefert wenig Argumente für Kettenarbeitsverträge.

Die Sichtweise, nach der Verstetigung engagiertes Forschen hemmt, basiert auf einem unschönen Menschenbild: Sobald ein Wissenschaftler ein gewisse Jobsicherheit erlangt, kauft er sich eine Kaffeemaschine und zieht sich vom konfokalen Mikroskop auf die Couch zurück! Das Maß an Selbstausbeutung mit dem in der Wissenschaft, egal mit welchem Vertrag, nach sehr langer Ausbildung und zu jeder Tages- und Nachtzeit bei im Vergleich zur Wirtschaft moderaten Gehältern geforscht wird, belegt hingegen, dass in Wirklichkeit ganz andere Motive als Bequemlichkeit oder Profitstreben am Wirken sind.

Und welche Evidenz gibt es eigentlich für Hannas Argument, dass die Unsicherheit auf Weiterbeschäftigung und Bereitschaft zu ständigem Vertrags-, Projekt- oder Stellenwechsel Innovations-fördernd ist? Zunächst einmal lassen sich eine Menge Gründe finden, warum das Gegenteil der Fall sein sollte. Psychischer Stress ist keine gute Basis für kreatives Denken und Arbeiten. Gute Wissenschaft braucht Zeit, Einarbeitung in Methoden, Routine, etc. Unterbrechungen und Themenwechsel sind da kontraproduktiv. Natürlich sind auch ‚Wanderjahre' für Wissenschaftler wichtig. Man lernt neue Methoden, kommt auf neue Ideen, knüpft Netzwerke für Kollaborationen, usw. Es gibt aber keinen logischen Grund, warum dies nicht auch in einem System mit vielen Dauerstellen möglich wäre. Im Gegenteil: Ausgangspunkt ist doch gerade die Mobilität der Wissenschaftler. Sie würden doch auch von Dauerstelle zu Dauerstelle wechseln. Diese Stellen würden damit keineswegs ‚zementiert', sondern durch befristete Aufenthalte in anderen Laboren, Wechsel in andere Institute (national wie international) oder Wegberufungen ‚flüssig' gehalten. Mit anderen Worten: Wissenschaftler würden einfach von ‚Tenure' zu ‚Tenure' wechseln. Eine Dauerstelle wäre nicht immer mit derselben Person besetzt.

Die Unsicherheit der eigenen Stelle, das ständig sich mittels einer fragwürdigen Indikatorik bewähren müssen, hat daneben noch weitere korrosive Wirkungen auf die Wissenschaft. Wer nur wenig Zeit hat, und weniger an inhaltlichen Resultaten gemessen wird, als an der Anzahl und dem Impact Factor seiner Publikationen macht weniger Team Science, forscht weniger transparent, und wird auch häufiger versucht sein, ‚Abkürzungen' zu nehmen, um zum Ziel zu kommen. Und das lautet Publikation in Journalen mit hohem JIF. Dabei hilft dann die selektive Nutzung von Daten, das HARKING, also die nicht offengelegte Anpassung von Hypothesen auf Basis der Ergebnisse, fragwürdige statistische Methoden, zu geringe Fallzahlen, und viele andere Praktiken, welche die interne Validität der Forschung verschlechtern. Und dadurch mit verantwortlich sind für die in vielen Bereichen enttäuschende Übertragbarkeit von Studienergebnissen, mangelnde Reproduzierbarkeit, und eine generelle Ineffizienz und Ressourcenverschwendung in der Wissenschaft. Dauerstellen vermindert den Druck auf Wissenschaftler und sind damit ein Bollwerk gegen die Notwendigkeit der Anwendung von fragwürdigen Wissenschaftspraktiken. Zudem befreien langfristige und noch mehr permanente Stellen Wissenschaftler von Machtstrukturen, die noch in vielen Bereichen weit verbreitet sind. Frühe Selbständigkeit fördert die Kreativität und Eigeninitiative und verhindert die Appropriation von Leistungen durch andere, insbesondere hierarchisch höher Stehende.

Akademische Dauerstellen sind besonders dazu geeignet, Wissenschaftler aktiv dabei zu unterstützen, ihre Forschung vertrauenswürdiger, transparenter, und nützlicher zu machen. Ein gutes Beispiel ist hier das Forschungsdatenmanagement, einschließlich des Teilens von Forschungsdaten zur Nachnutzung. Forschungsdaten FAIR (findable,

accessible, interoperable, reusable) zu teilen ist aufwendig und erfordert spezielle Kenntnisse. Viele Wissenschaftler würden das gern tun, es fehlen ihnen aber Ressourcen und das nötige know how. Da helfen sog. Data Stewards, welche aus der Wissenschaft kommen, aber selbst nicht notwendig eigene wissenschaftliche Projekte verfolgen. Auch Core Facilities sind hier zu nennen, in denen methodische Kompetenz von Wissenschaftlern auf höchstem Niveau vorgehalten wird. Wissenschaftler in Core Facilities können durchaus an Projekten beteiligt sein, verantworten diese aber nicht selbst. Wissenschaftler auf solchen Dauerstellen helfen dabei, die Spannung zwischen Qualität und Geschwindigkeit von Forschung zu vermindern. Qualität in der Wissenschaft braucht Dauerstellen.

Wer Hannas Argument folgt, nach dem Kettenarbeitsverträge und die Aussicht auf ein frühes Ausscheiden aus Academia Innovationen fördert, ist zudem ein weiteres fragwürdiges Mantra auf den Leim gegangen: Nämlich der von Innovation als allein selig machendem Ziel wissenschaftlicher Betätigung. Dahinter steckt latent die Vorstellung, dass es wesentlich um die Erzielung spektakulärer Befunde geht, welche es in Glamourjournale wie Science, Nature und Cell, oder zumindest in die Zeitung schaffen und den Wissenschaftlern damit zu Ruhm und Ehre verhelfen. Weit gefehlt. Wissenschaft ist ganz überwiegend ‚normal' (Thomas Kuhn). Ohne normale Wissenschaft, welche das Wissen in kleinen Schritten voranbringt, gibt es keine ‚großen' Innovationen oder gar Paradigmenwechsel. Normale Wissenschaft muss belastbare Ergebnisse produzieren um nützlich zu sein. Echte Innovationen sind selten, nicht vorhersagbar, und häufig das Ergebnis von Zufällen. Innovationen entstehen auf und aus normaler Wissenschaft. Diese braucht Zeit, und funktioniert nicht unter Druck. Normale Wissenschaft braucht Dauerstellen.

Es spricht also sehr viel für, aber wenig gegen eine frühe ‚Tenurisierung' von ECRs. Modelle hierfür existieren in einer Reihe von Ländern, z.B. mit dem sog. ‚Lecturer'. In Deutschland ist die 2002 eingeführte Juniorprofessur ein Schritt in diese Richtung. Allerdings nur mit Blick auf das hierarchische professorale Modell, nicht auf den Mittelbau. Das neue Berliner Gesetz nimmt nun genau diesen in den Blick, denn es geht darum, Wissenschaftlern – natürlich abhängig von bestimmten Kriterien – eine sichere Perspektive zu bieten. Und nicht nur die vage Aussicht auf eine Professur in weiter Ferne.

Aber gäbe es denn überhaupt Evaluations-Kriterien, bei deren Erreichen den Postdocs die schon bei der ersten Anstellung in Aussicht gestellte Verstetigung auch zu gewähren wäre? Würden hier unverändert jene Kriterien angewendet werden, welche heute bereits durchgesetzt sind (und ja gerade eine wichtige Ursache für die Notwendigkeit einer Systemreform darstellen), wäre nichts gewonnen. Drittmittel, Hirsch-Faktor und Journal Impact Factor sind ja schon für die Beurteilung arrivierter Wissenschaftler ungeeignet. Für ECRs kämen sie auch aus praktischen Gründen gar nicht in Frage. Sie hatten ja noch wenig oder gar keine Gelegenheit, bei diesen Indikatoren zu punkten. Hier müssten dagegen vorwiegend qualitative Kriterien im Vordergrund stehen. Mit welcher Kompetenz, Sorgfalt, und Transparenz hat der ECR geforscht? Hat er oder sie Disseminationsformate jenseits des Peer Review Artikels genutzt, wie z.B. Präregistrierungen, Preprints? Wurden Methoden und Ergebnisse FAIR geteilt, oder auch NULL und Negativ-Resultate veröffentlicht? Welche methodischen oder inhaltlichen Beiträge wurden in den Diskurs der Forschergemeinschaft eingebracht, wie wurden diese rezipiert? Hier eignen sich z.B. Einladungen zu Vorträgen, Preise, aber auch Stellungnahmen von Peers als Indikatoren.

Wenn über eine Entfristung schon in frühen Karrierestadien entschieden werden soll, muss auch die Vorbereitung auf akademische Karriereweg oder Alternativen hierzu viel früher als heute einsetzen. Wir bereiten junge Wissenschaftler heute oft gar nicht und manchmal mehr schlecht als recht darauf vor, was ‚auf sie zukommt'. Welche alternativen Karrierewege gibt es, im akademischen System und vor allem auch außerhalb?

Welche Weichen müssen wann gestellt werden, von wem und nach welchen Kriterien? Wo gibt es vertiefende Einblicke in die vielfältigen Berufsoptionen, auch im Sinne von Hospitationen oder Praktika? Die meisten ECRs bewegen sich im Studium und dann im Postdoc wie in einem Labyrinth. Es gibt kaum Wegweiser, und Karriereentscheidungen werden häufig eher zufällig, d.h. opportunistisch getroffen. Dabei wird die Rolle des individuellen Glücks (oder aber Pechs) massiv unterschätzt: Ein PhD Student mit dem Glück, im richtigen Labor zur richtigen Zeit promoviert zu haben und dabei eine Erstautorschaft in einem tollen Journal ergattert zu haben, gilt als erfolgreich, als ‚High potential'. Eine andere wurde auf die falschen Methoden angesetzt, schlecht betreut, oder geriet zwischen die Fronten einer Gruppen-internen Konkurrenz. Dann wird man wohl eher zu der Einschätzung kommen, dass sie vielleicht für die Wissenschaft doch nicht so geeignet war! Auch dies ein Grund dafür, Wissenschaftler im System zu halten durch planbare Karriereoptionen, statt sie wie heute in einer Art Glücksspiel Lose ziehen lassen.

Es macht auch gesellschaftlich wenig Sinn, Ressourcen dafür zu verschwenden, Wissenschaftler lange auszubilden um sie dann in ganz anderen Berufssparten tätig werden zu lassen. Momentan müssen Wissenschaftler eine Professur anstreben um an der Universität eine dauernde Anstellung zu finden. Von diesen Professuren gibt es aber nur sehr wenige, und viele Wissenschaftler sind weder interessiert noch dafür geeignet, sich den Hintern in Kommissionen aufzusitzen, Machtkämpfe mit Kollegen in der Fakultät zu bestehen, große Forschungsgruppen oder gar ganze Institute zu leiten, oder sich sonst wie wichtig zu machen. Sie wollen im Labor forschen, andere Wissenschaftler trainieren, Studenten unterrichten, und Gruppen mit überschaubarer und wissenschaftlich sinnvoller Größe an der Bench anleiten.

Vieles spricht also dafür, dass Postdocs in früher Selbständigkeit eine Chance erhalten sollten, in der Wissenschaft zu verbleiben. Hierfür gibt es eine Reihe von Modellen, gemeinsam ist vielen, dass sie den ECR drei Optionen bieten. Ein Weg führt zur Professur, einer zur Daueranstellung im akademischen Mittelbau. Ein Dritter, über den am besten vor den beiden anderen entschieden werden sollte, stellt das Ausscheiden aus der universitären Forschung dar. Entscheidend dabei ist, dass die Evaluationskriterien immer transparent und verantwortungsvoll sein müssen, und dem angestrebten Weg, also Professur oder Mittelbau, angemessen.

Jetzt bleibt eigentlich nur noch die Frage, warum das neue Gesetz auf so vehementen Widerstand der Berliner Universitäten stößt? Widerstand gegen ein Gesetz, das die Umsetzung einer lange herbeigeredeten und eigentlich konsentierten Reform verspricht. Ganz einfach: Weil das Berliner Gesetz sehr viel Sinn macht, aber finanziell nicht umsetzbar ist. Denn ein großer Anteil der akademischen Wissenschaftsförderung in Deutschland läuft über Projektfinanzierung. Und Projekte sind ihrer Natur nach zeitlich begrenzt. Projekte generieren die Stellen, entlang derer sich die PhD Studenten und Postdocs derzeit hangeln, ohne Aussicht auf Verstetigung, denn die Projekte enden ja früher oder später.

Aber auch ohne zusätzliche Finanzmittel könnte das Problem gelöst werden, in dem Teile der über Projekte eingeworbenen Mittel für Dauerstellen an die Institutionen abgeführt werden. Derzeit erhalten die deutschen Unis von den großen Forschungsförderern DFG und BMBF für bewilligte Projekte einen Overhead von 20%. Das ist natürlich viel zu wenig, und das wenige geht in Administration, Infrastruktur, usw., aber nicht in Stellen für Wissenschaftler. An amerikanischen Top-Universitäten sind Overheads von über 100 % die Norm. Die meisten privaten Förderer in Deutschland bezahlen im Übrigen gar keinen Overhead. Die Unis schlucken das, aus Angst sonst gar keine Mittel mehr

aus diesen Quellen zu bekommen. Paradoxerweise bringt die Einwerbung von Drittmitteln Universitäten daher in finanzielle Bedrängnis, denn die wahren Overheadkosten, bereits ohne Bereitstellung von Dauerstellen, sind wesentlich höher. Je erfolgreicher eine Uni also ist, desto größer deren Unterfinanzierung und die Schwierigkeiten, in die sie dadurch gerät.

All dies ist aber lösbar, allerdings nicht von den Universitäten selbst, und nicht nur auf lokaler Ebene, als z.B. in Berlin. Abgesehen von einer nötigen allgemeinen Erhöhung der staatlichen Grundförderung, denn die deutschen Unis sind seit vielen Jahren strukturell unterfinanziert, müssten die Mittelflüsse in der deutschen Hochschulförderung angepasst werden. Universitäten sollten aus der Projektfinanzierung zusätzliche Mittel für Dauerstellen erhalten. Dies sollte natürlich nicht zu Lasten der Projektförderung gehen. Über neu zu etablierende Verteilungsschemen könnte den Universitäten aus den geförderten Projekten die Bereitstellung von Personalressourcen aus dem Dauerstellenpool möglich gemacht werden. Dies ist sicher nicht auf Arbeitsgruppenniveau möglich, aber wohl auf Einrichtungs-, also z.B. Institutsebene. Das erfordert sicherlich einiges an Kreativität und Bereitschaft, neue Wege zu gehen. Nicht nur in Berlin wären die Universitäten spätestens jetzt gefragt, endlich proaktiv solche Anstellungs-Modelle zu entwickeln, und diese mittels mathematischer Simulationen auf Basis der ihnen als Arbeitgeber ja vorliegenden Daten (Drittmittelaufkommen, Laufzeiten, Personalfluktuation, etc.) vor einer Pilotierung zu optimieren.

Gleichzeitig ist die Politik gefragt im Dialog mit den Universitäten die rechtlichen Grundlagen für so eine Reform zu schaffen. Da trifft es sich doch hervorragend, dass wir soeben eine neue Regierung bekommen haben. Laut deren Koalitionsvertrag hat sie sich eine Reform des Wissenschaftszeitvertragsgesetzes ebenso vorgenommen wie eine Erhöhung der Planbarkeit und Verbindlichkeit in der Postdoc Phase. Die Formulierungen hierzu klingen fast so, als wäre Hanna mit am Tisch gesessen. Dazu möchte man alternative Karrieren außerhalb der Professur schaffen, einen Ausbau und die Verstetigung des Tenure Track Programms, sowie Dauerstellen für Daueraufgaben. Eigentlich ideale Voraussetzung nun endlich schon lange erkannte Probleme im deutschen Wissenschaftssystem zu lösen, und dabei gleichzeitig die Bedingungen für den akademischen Nachwuchs zu verbessern, die Qualität der Forschung zu erhöhen, und Deutschland als Wissenschaftsstandort attraktiver zu machen. Hanna würde staunen, was sie da ins Rollen gebracht hat!

Eine englische Version dieses Artikels wird bei EMBO Reports erscheinen.

Wie die Reputationsökonomie Papiermühlen antreibt

LJ 3/2022

Gerade hatte sich die Aufregung um die Raubverlage (‚predatory publishers') etwas gelegt, schon geht ein neues Schreckgespenst um: Die Papiermühlen (‚Paper Mills'). Die Raubverlage (und dabei geht es nicht wie man meinen könnte um Elsevier und Co) offerieren ihrer Kundschaft, also uns Wissenschaftlern, einen verkürzten Weg zur Publikation. Unsere Artikel werden, natürlich gegen eine stattliche Gebühr, nach einem Fake-Review Prozess garantiert und expediert in einem Open Access Journal mit häufig wohlklingendem Namen veröffentlicht. Der Kunde fügt diesen dann seinem Lebenslauf hinzu, und rückt karrieretechnisch ein Feld vor, insbesondere dort wo Erbsen, also Publikationen gezählt werden. Der Narr hat darüber bereits ausführlich berichtet (https://www.laborjournal.de/rubric/narr/narr/n_18_09.php). Panik brach damals aus bei vielen Kollegen, denen plötzlich auffiel wie schnell das ging mit dem Review ihres Papers, und wie freundlich die Kommentare waren, wenn es überhaupt welche gegeben hatte. War man etwa unter die Räuber geraten? Habilkommissionen suchten hektisch nach Listen mit den Namen verdächtiger Journale, und begannen die Literaturlisten derer unter die Lupe zu nehmen, die nach dem Erwerb dieses spätmittelalterlichen akademischen Grades strebten.

In einer Reputationsökonomie, in der die Anzahl von Publikationen und deren Nimbus (sprich: Impact Factor) mehr zählt als deren Inhalt und Qualität, kann man, wenn man genug kriminelle Energie hat, das Geschäftsmodell der Raubverlage aber noch einen Schritt weitertreiben. Indem man nämlich aufstrebenden Wissenschaftlern ein Komplettpaket anbietet. Wozu sollen diese überhaupt noch Studien durchführen, analysieren, und dann aufwendig alles zusammenschreiben? Hierfür bieten die Betreiber der Paper Mills nun das entsprechende Produkt. Sie liefern nämlich einen ‚full service': Der Artikel, die Koautoren, und auch die Veröffentlichung in einem Journal, alles in einem Paket. Der prospektive Autor muss nur noch ein bestimmtes Fachgebiet, evtl. auch ein paar Schlüsselwörter oder Methoden angeben und ein Journal auswählen. Wem auch das zu viel ist, kann eine Koautorschaft auswählen bei einer Publikation, die gerade auf diese Weise angefertigt wird. Natürlich ist das Ganze nicht billig, und nach Recherchen von Bernhard Sabel (Magdeburg), abhängig vom Impact Factor des Zieljournals. Für einen Impact Factor über 3 können da schon mal 25.000 € oder mehr fällig werden. Aber Qualität hat eben ihren Preis, und das ist ja schließlich eine Investition in die eigene Zukunft. Ganz davon abgesehen, dass natürlich eine tatsächlich durchgeführte Studie viel teurer käme, und viel länger dauerte.

Ich muss zugeben, als ich zum ersten Mal von diesen Umtrieben hörte, habe ich die ganze Sache nicht besonders ernst genommen. Mag sein, dass man in China mit solchen Publikationen im CV Karriere machen kann, aber doch nicht bei uns! Diese Einschätzung war leider einigermaßen naiv. Zum einen war mir das Ausmaß dieser Artikelmarktes

nicht klar, und auch nicht welche Journale davon betroffen waren. Eine Reihe von sehr reputierlichen Fachjournalen mussten letztes Jahr Dutzende von Artikeln zurückziehen. Das Journal of Cellular Biochemistry (JIF 4,5) hatte im Oktober ein eigenes Supplement herausgebracht, mit 129 Retraktionen von Artikeln, welche sich als Produkt von Papiermühlen herausstellten. Naunyn-Schmiedeberg's Archives of Pharmacology (JIF 3,0), gegründet 1873 und damit die älteste noch existierende Fachzeitschrift für Pharmakologie, musste 10 Artikel zurückziehen, und hat 30 weitere abgelehnt, die ganz offensichtlich ‚gemahlen' waren. Viele davon sind dann allerdings, leicht modifiziert, in anderen Journalen erschienen. Roland Seifert, der Chief Editor des Journals hat hierüber bereits im Laborjournal berichtet). Bernhard Sabel (Chief Editor von Restaurative Neurology and Neuroscience, JIF 2,1) schätzt, dass bis zu 15 % der Artikel seines Journales aus Paper Mills stamm(t)en. In einer systematischen Analyse von neurowissenschaftlichen Journalen fand er, dass etwa 10 % aller Arbeiten hochgradig verdächtig waren, einer Papiermühle zu entstammen. Es handelt sich also um eine veritable Industrie, und die Kunden (d.h. Autoren) kommen nicht nur aus China, sondern auch aus den USA (an 2. Stelle!), Japan, Indien, Korea.

Nun könnte man einwenden, dass die meisten wissenschaftlichen Artikel in der Biomedizin ohnehin nicht gelesen werden, die Sache also gar nicht so schlimm. Dass die so ist, lässt sich ganz gut aus Zitationsanalysen ableiten, indem man nachweist, dass Zitate aus Literaturlisten übernommen werden, ohne gelesen worden zu sein. Hält sich der Schaden also in Grenzen, weil der Spuk im Rauschen des gigantischen Blätterwaldes untergeht? Weit gefehlt.

Zum einen ist ja nicht auszuschließen, dass bei der schieren Masse von gefälschten Publikationen nicht doch zu Misinformation, Ressourcenverschwendung, und vielleicht sogar einer Gefährdung von Patienten kommt, denn da sind ja auch eine Menge klinischer Fake-Studien dabei. Zum anderen werden hierdurch systematische Reviews und Meta-Analysen ad absurdum geführt. Deren Ziel ist es ja gerade, die gesamte vorhandene wissenschaftliche Evidenz zu einer Intervention zu sammeln, und gewichtet zu synthetisieren. Wenn ein substantieller Teil der eingeschlossenen Studien Fake ist, kann man sich vorstellen wie belastbar die Evidenzsynthese wird.

Die Papiermühlen werfen übrigens auch ein wenig vorteilhaftes Schlaglicht auf die Retraktionskultur und den Idealismus der ‚selbst korrigierenden Wissenschaft'. Nur durch das Engagement einiger weniger Chief-Editors tritt derzeit das ganze Ausmaß dieser Vermüllung des Publikationskorpus ans Tageslicht. Viele Journale kümmern sich entweder gar nicht darum, oder verschleppen nötige Retraktionen. Und sie fördern das Kaskadenpublizieren: Wenn ein Artikel verdächtig erscheint, wird er einfach abgelehnt. Was natürlich heißt, dass er letztendlich, vielleicht nach einer ganzen Reihe weiterer Submissionen, doch irgendwo unterkommt und publiziert wird. Wissenschaft kann sich selbst korrigieren – aber das ist ein sehr ineffektiver und unakzeptabel langwieriger Prozess.

Am beunruhigendsten erscheint mir aber, was uns die scheinbar mühelose Akzeptanz von Papiermühlen Artikeln in von uns hoch geschätzten Fachjournalen über die Qualitätskontrolle des Peer Review Prozesses sagt. Die geschätzten Peers lassen sich also allzu oft von Artikeln narren, welche von einer AI geschrieben wurden, welche auch die zugehörigen Daten und die Abbildungen gefakt hat. Trainiert wurde die AI zuvor an Millionen von Artikeln, sie weiß also wie es geht. Oder, dies eine Technik vor allem russischer Papiermühlen, es werden Artikel ins Englische übersetzt, welche zuvor schon in russischen Zeitschriften erschienen waren. Leicht modifiziert, und mit neuen Autoren versehen, die für diesen Service bezahlen. Plagiarism meets Paper Mills! Der Tsunami an Artikeln, der uns nicht nur als Leser, sondern vorher schon in unserer Rolle als Reviewer

unter sich begräbt, fordert eben sein Tribut! Wie soll man auch in ein oder zwei Stunden, denn mehr Zeit bleibt ja meist nicht, einen Artikel gründlich begutachten, sich die Originaldaten dazu anschauen, vielleicht sogar noch mit spezieller Software die Abbildungen auf Manipulationen überprüfen? Wenn es noch einen Beweis gebraucht hätte, hier ist er: Der Peer Review wird in seiner Filterfunktion massiv überschätzt.

Natürlich gäbe es eine Reihe von Gegenmaßnahmen, die man in Anschlag bringen kann, um diese Fake-Papers zu erkennen und auszusortieren. Häufig geben z.B. Papiermühlen-Paper Autoren private Email-Adressen an, ein erster Hinweis, dass da was im Argen liegen könnte. Am verrücktesten hierbei sind chinesische Autoren, welche Gmail Accounts angeben, obwohl diese schon seit 2014 von der Regierung geblockt werden. Häufig stellen Paper-Mill Autoren aus nachvollziehbaren Gründen auf Nachfrage keine Originaldaten zur Verfügung. Dies, neben der Möglichkeit der Nachnutzung, ein weiterer wichtiger Grund für das FAIRe (findable-accessible-interoperable-reusable) Teilen der Originaldaten in allen Publikationen. Das Fehlen von ORCID-IDs, also eindeutigen Identifizierungsnummern der Autorinnen und Autoren kann auch eines von vielen Indizien sein. Am tollsten aber ist der Vorschlag, AI zu nutzen um verdächtige Artikel zu identifizieren. Man trainiert hierzu eine Maschine an Fake-Artikeln, die von einer anderen AI nach Training an echten Artikeln geschrieben wurden!

Dies sind alles wohlmeinende Vorschläge, die das Problem aber nicht lösen werden. Solange wir uns gegenseitig weniger nach dem Inhalt, dem tatsächlichen Impakt, und der Qualität unserer Forschung beurteilen – mithin lesen und kritisch beurteilen was wir produzieren – wird die Artikelinflation nicht enden. Und findige Köpfe werden Mittel und Wege finden, sich daran zu bereichern. Welchen Anschlag braucht es noch auf das Publikationswesen, nach Raubverlagen, Papiermühlen, Verlagshäusern mit 30 % Profitrate, etc., bis wir verstanden haben, dass all dies nur Symptome eines viel grundsätzlicheren Grundübels sind. Papier-Mühlen werden nicht vom Wind, sondern von der Reputationsökonomie angetrieben!

Tu felix Brittania reloaded: Wie schön sich Politik in Wissenschaft einmischen kann

LJ 4/2022

Manch einer wird sich vielleicht erinnern, vor nicht allzu langer Zeit wünschte ich mich an dieser Stellen nach England, weil dort die klinische Corona-Forschung der deutschen so sehr überlegen ist (LJ 9/2021). Und schon wieder packt den Narr der Neid, da einiges in Sachen Wissenschaft auf der Insel so viel besser läuft als bei uns. Und das hat ausgerechnet mit parlamentarischer Kontrolle der Wissenschaft zu tun – eine staatliche Einmischung ins freie Forschen, allein deren Vorstellung jeden deutschen Wissenschaftler in Angstschweiß ausbrechen lässt.

Aber immer der Reihe nach. Stellen Sie sich vor, sie leiden an den Symptomen einer bisher nicht befriedigend behandelbaren Erkrankung. In einer deutschen Uniklinik eröffnet man Ihnen, dass es ein neues, aussichtsreiches Medikament gibt, Sie zu behandeln. Man bietet Ihnen an, an einer laufenden Studie teilzunehmen. Im Aufklärungsgespräch erfahren Sie, dass Sie in so einer Studie mit 50 % Wahrscheinlichkeit ein Scheinmedikament (Placebo) erhalten, und dass das Studienmedikament - von dem man ja noch nicht weiß ob es wirkt, eine Reihe von unangenehmen, teils auch gefährlichen Nebenwirkungen haben könnte. Von ihrer Teilnahme an der Studie profitieren also möglicherweise nicht sie selbst, vielleicht schadet sie ihnen sogar. Aber in jedem Fall nützen die Ergebnisse der Studie nachfolgenden Patienten mit derselben Erkrankung, die nun möglicherweise besser behandelt werden können. Sie willigen unter diesen Umständen in die Studienteilnahme ein. Es besteht ja zumindest die Möglichkeit eines persönlichen Nutzens, und der Nutzen für andere ist sogar garantiert. Aber würden Sie an der Studie auch teilnehmen, wenn sie wüssten, dass solche klinischen Studien häufig ihre Ergebnisse entweder gar nicht, oder erst viele Jahre nach Abschluss veröffentlichen? Vermutlich nein. Leider ist aber genau dies traurige Praxis, was man ihnen im Aufklärungsgespräch jedoch ganz sicher verschwiegen hätte.

Randomisierte und kontrollierte klinische Studien (RCT) wie die eben beschriebene sind der Goldstandard, wenn es darum geht, herauszufinden ob ein neues Medikament wirkt oder nicht, oder ob es gar schädlich für die Patienten ist. Ein neues Medikament kann nur zugelassen werden, wenn positive Evidenz aus einer, meist sogar 2 großen RCTs vorliegt. Studienteilnehmer werden in diesen Studien per Zufall in Gruppen eingeteilt, die entweder Scheinbehandlung (Placebo) oder das Studienmedikament erhalten. Ein Patient, der sich zu Teilnahme an einer klinischen Studie entscheidet, hat damit nicht nur keine Gewissheit, dass er oder sie das neue überhaupt Medikament erhält. Zum Zeitpunkt der Studie ist auch unklar, ob das Medikament nützt, oder vielleicht sogar schädlich ist. Studienteilnahme birgt also ein Risiko, das sich in etwa mit dem potentiellen Nutzen die Waage halten muss (sog. ‚Equipoise'). Nur dann wird die Ethikkommission grünes Licht für die Studie geben. In der Aufklärung vor Studienteilnahme wird all dies den Patienten erläutert. Die Teilnahme an der Studie dient also nicht notwendig der eigenen Gesundheit (obzwar das natürlich nicht ausgeschlossen ist), man geht ein Risiko zum Nutzen späterer Generationen von Menschen mit derselben Krankheit ein. Studienteilnahme ist damit ein altruistischer Akt.

Was aber, wenn die Ergebnisse dieser Studien gar nicht, oder erst stark verzögert veröffentlicht werden? Dann hätten die Studienverantwortlichen die Patienten getäuscht, und diese umsonst ein Risiko auf sich genommen. Deshalb fordert sowohl die Europäische Union, als auch die WHO, dass die wichtigsten Studienergebnisse innerhalb von 12 Monaten nach Abschluss der Studie veröffentlicht werden müssen. Bei Studien an Kindern ist diese Frist sogar nur 6 Monate. Deshalb ist es schockierend, dass die Mehrzahl der klinischen Studien an medizinischen Universitäten in Deutschland nicht fristgerecht, sondern oft erst viele Jahre später, nicht selten aber auch gar nicht veröffentlicht werden. Das ist unethisch, ein Verrat am Altruismus der Studienteilnehmer.

Entgegen ihrem Ruf hält sich die Pharmaindustrie in den von ihr organisierten Studien übrigens überwiegend an diese Regeln, vermutlich aus Angst vor rechtlichen und finanziellen Konsequenzen sowie einem möglichen Imageschaden. Ganz im Gegensatz zu den medizinischen Universitäten, welche in Studien die Wirksamkeit der Therapien untersuchen, welche in ihren Laboren entwickelt oder von ihren Klinikern erdacht wurden. Das wissen wir, weil Ben Goldacre (Oxford), Autor des Klassikers ‚Die Pharmalüge' (‚Bad Pharma') sogenannte ‚Trial Tracker' ins Netz gestellt hat (z.B. für das Europäische Studienregister https://eu.trialstracker.net/). Ein digitaler Pranger, an dem sich

jedermann davon überzeugen kann, wie gut oder wie schlecht eine Firma oder akademische Einrichtung im Veröffentlichen ihrer klinischen Studienergebnisse ist. Wenn man in so einem Trial tracker stöbert, fällt einem alsbald auf, dass die britischen Universitäten ausgezeichnet abschneiden, sie bleiben beim Veröffentlichen der Ergebnisse fast all ihrer Studien im vorgeschriebenen Zeitrahmen. Aber die deutschen Studien schneiden viel, viel schlechter ab. Das hat auch mein Kollege Daniel Strech in einer sehr detaillierten Studie nachgewiesen, in der er sich alle 36 deutschen universitären medizinischen Zentren vorgenommen hatte. Aber woran liegt es, dass die britischen Unis ihre wichtigsten Studienergebnisse fast immer zeitnah veröffentlichen, und die deutschen nicht? Und war das eigentlich schon immer so?

Auch die britischen Unis waren vor ein paar Jahren noch genauso ‚unethisch' wie die deutschen! Allerdings hat sich in England die Politik des Problems angenommen, und die Unis ganz einfach dazu verdonnert, ihre Studienergebnisse zeitgerecht und vollständig zu berichten. Und das ging so: Das britische Parlament hält sich eine Reihe von sogenannten ‚Select Committees'. Eines davon ist das ‚Science and Technology Committee' des Unterhauses. Dessen Aufgabe ist es, darauf zu achten, dass die Politik und die Entscheidungsfindung der Regierung auf soliden wissenschaftlichen Erkenntnissen und Ratschlägen beruhen. In England ist der Staat wie in Deutschland direkt oder indirekt ein wesentlicher Geldgeber auch der klinischen Forschung. Diese ist recht teuer, und kann viel Nutzen bringen, es kann aber auch einiges schiefgehen. Der englische Staat achtet deshalb darauf, ob seine Fördermittel effektiv, effizient und ethisch eingesetzt werden. Um 2018 herum und getriggert durch eine im British Medical Journal vom schon erwähnten Ben Goldacre veröffentlichten Studie setzte das ‚Science and Technology Committee' die skandalöse da verzögerte oder gar ganz fehlende Veröffentlichung von klinischen Studienresultaten auf seine Tagesordnung. Schon 2013 hatte sich das Committee des Themas angenommen, und schöpfte einen ersten Verdacht. Noch beließ man es damals mit der Veröffentlichung eines mahnenden Reports. Nun hörte man sich die führenden Experten abermals zum Thema an, und lud auch die wichtigsten Vertreter von Universitäten ein, welche klinische Studien durchführen. Das Ganze, wie fast alle Sitzungen dieser Unterhaus-Committees wurde live im Fernsehen übertragen und danach auf einem Youtube Kanal archiviert. Es ist durchaus sehenswert wie sachkundig, parteiübergreifend, und direkt die Parlamentarier hier zur Sache gehen. So macht selbst Parlamentsfernsehen Spaß!

Die Protokolle dieser Anhörungen wurden mit den Resultaten der Diskussionen als ‚Report' ins Netz gestellt. Und man höre und staune: Die Politik hat den Unis daraufhin das Messer an die Brust gesetzt. Entweder ihr löst das Problem innerhalb von 6 Monaten, oder wir überdenken die staatliche Förderung eurer Institutionen! Der Rest ist Geschichte. Die Unis veröffentlichten alle überfälligen Ergebnisreports in den zugehörigen, jedermann zugänglichen Studienregistern, und sind seither auch nicht mehr rückfällig geworden.

Der Narr war beeindruckt, auch durch andere Themen die sich der Ausschuss auf den Tisch gezogen hatte. Wie zum Beispiel die britischen Corona-Wissenschaft. Spätestens da wurde er zum Binge-viewer des Britischen Parlamentsfernsehens. Was für eine Transparenz! Was für ein Sachverstand! Was für eine no-nonsense Debattenkultur! Derzeit befasst sich das Committee übrigens, dies kein Scherz, mit einem Lieblingsthema dieser Kolumne, nämlich mit ‚Reproducibility and Research Integrity'. Nach einem Call, bei der jeder Brite seine Sichtweise einbringen konnte, hat das Komitee nun schon zweimal (natürlich öffentlich) getagt, und schloss dabei alle wichtigen Stakeholder, also die Unis, die Verlagshäuser, die Fördergeber, und führende Kritiker des gegenwärtigen Wissenschaftssystems ein. Manchmal taten mir die in den Sitzungen peinlich Befragten

richtig leid. Zum Beispiel die Vertreterin des Verlagshauses Wiley, die von den Parlamentariern mit lauter richtigen Fragen und Argumenten regelrecht an die Wand genagelt wurde. Auch junge Wissenschaftler kamen in den Anhörungen zur Sprache, #IchbinHanna ließ grüßen (siehe auch Laborjournal 1/2022)

Da stellt sich doch unmittelbar die Frage, wie das in Deutschland so ist mit der wissenschaftlichen Politikberatung für den Deutschen Bundestag? Kümmert es die Politik, wie Forschungsmittel eingesetzt werden? Was macht eigentlich der Ausschuss für Bildung, Forschung und Technikfolgenabschätzung des Deutschen Bundestages? Haben Sie von dem schon mal gehört? Und wüssten Sie, was der mit seinen 38 Mitgliedern so macht?

Beflügelt von den ‚englischen Verhältnissen' und unterstützt durch die Lobbyisten von Wikimedia, hatte ich es in der letzten Wahlperiode sogar zu einem Termin bei Vorsitzenden dieses Ausschusses gebracht. Mein Ziel: Aufmerksamkeit zu schaffen bei der Politik für die nicht- oder verzögerte Veröffentlichung von klinischen Studienresultaten. Die Untersuchung von Daniel Strech, und der Trial Tracker von Ben Goldacre, sowie die Aktivitäten von Till Bruckner von Transparimed hatten 2019 gerade die ganze Misere in Deutschland offengelegt. Hatte mein Vorstoß irgendwelche Konsequenzen? Etwa eine Befassung des Ausschusses mit dem Thema? Schließlich sind BMBF und DFG die Geldgeber der klinischen Studien von deutschen Universitäten, da müsste doch der Staat ein Interesse haben, das in Ordnung zu bringen. Leider komplette Fehlanzeige. Nichts ist passiert.

Wenn man sich die Protokolle der zumeist nicht öffentlichen Sitzungen dieses Bundestagsausschusses anschaut, wird einem klar, warum das so war. Im Ausschuss geht es bei marginalem Sachverstand ganz wesentlich um Parteipolitik. Partei A bringt einen Antrag ein – Partei B (Opposition) bringt diesen dann zum Fall. Mal A den von B, mal B den von A. Wieder und wieder. Das Ganze total intransparent, da die Sitzungen fast ausschließlich nicht öffentlich und die Sitzungsprotokolle wenig informativ sind. Gibt es vielleicht ein anderes parlamentarisches Gremium in Deutschland, das sich mit solchen Themen befassen würde? Mir ist keines bekannt – und wenn hätte es keinen Impact. Interessiert sich in Deutschland irgend jemand dafür, ob das Geld, das in die Forschung fließt, verantwortungsvoll eingesetzt wird? Ergebnisse von öffentlich geförderten Projekten anderen Forschern und der Öffentlichkeit zu Verfügung gestellt werden? Oder gar nicht veröffentlicht werden? Publikationen aus solchen Studien hinter Paywalls verschwinden? Ob die Corona-Maßnahmen des Bundes- und der Länder evidenzbasiert und effektiv waren? Was man in (oder vor) der nächsten Pandemie besser machen könnte?

Solange wir sowas nicht haben, bleibt nur der neidvolle Blick zu den glücklichen Briten. Immerhin können wir im Internet deren Parlamentsfernsehen kucken, und die Kommissions-Berichte runterladen. Da stehen schlaue Sachen drin, und einiges davon lässt sich auch ohne Politik umsetzen. Bei der Veröffentlichung der klinischen Studienresultate ist das gerade so. Die deutschen Unis werden langsam besser, denn sie beginnen dem englischen Beispiel zu folgen.

Mehr Handys, mehr Dicke?

LJ 5/2022

Das titelte kürzlich die BILD! Der originellen Alliteration wegen ignorierte der Reporter sogar, dass es im Artikel ums Telefonieren und nicht um Spielen mit Handys ging. BILD bezog sich in ihrem Bericht auf eine Pressemitteilung der Uni Lübeck. Diese hatte eine Studie eines Wissenschaftler-Teams der dortigen Psychoneurobiologie angepriesen, in der der ‚Einfluss von Handystrahlung auf die Nahrungsaufnahme nachgewiesen' wurde. Und das ging so: In einem ‚durchdachten Versuchsdesign' hatte man 15 junge Männer in einem Abstand von zwei Wochen insgesamt dreimal einbestellt. Im Experiment wurden die Probanden dann mit zwei verschiedenen Handys bestrahlt bzw. einer Scheinbestrahlung als Kontrolle ausgesetzt. Im Anschluss durften sich die Probanden für eine definierte Zeit an einem Buffet bedienen. Gemessen wurde die spontane Nahrungsaufnahme, der Energiestoffwechsel des Gehirns anhand von Phosphor-Magnetresonanz-Spektroskopie (MRS) sowie verschiedene Blutwerte vor und nach Bestrahlung. Und siehe da: Nach 25 Minuten Handy am Ohr nahmen die Probanden sage und schreibe etwa ein Viertel mehr Kalorien zum Frühstück im Vergleich zur Scheinbestrahlung. Das entspricht kalorisch etwa einer halben Bier, oder einem Stück Apfeltorte!

Die Pressemitteilung fand dieses Ergebnis ‚erstaunlich', und den Effekt ‚überraschend deutlich'. Dies kann man getrost als massives Understatement bezeichnen. Denn sollte dies stimmen, müssten wir uns alle nicht nur wundern, warum wir den ca. 30 Minuten nach Mobiltelefonaten einsetzenden Heißhunger noch nicht an uns selbst bemerkt haben. Und noch wichtiger: Der Befund in ‚Nutrients' veröffentlichte Befund hätte nicht abzuschätzende Implikationen für die Menschheit. Dies blieb auch den Autoren nicht verborgen, ja war sogar die Motivation für ihre Studie: Bereits im ersten Satz des Abstracts wird ein Zusammenhang zwischen gestiegener Mobilfunknutzung und der weltweiten Adipositas-Epidemie suggeriert. Laut WHO waren im Jahr 2016 mehr als 1,9 Milliarden Erwachsene ab 18 Jahren übergewichtig, und davon 650 Millionen gar fettleibig. Wir wissen, dass dies das Risiko für Diabetes mellitus, Herz-Kreislauf-Erkrankungen, Bluthochdruck und Schlaganfall sowie bestimmte Krebsarten deutlich erhöht. Sollte also die Nutzung von Mobiltelefonen ein wichtiger Grund für einen der wichtigsten vermeidbaren Gründe von weltweiter Morbidität und Mortalität sein, wäre das eine Entdeckung in der Kategorie Penizillin, Polioimpfung oder Heliobacter und Ulcus– der Nobelpreis für die Entdecker damit in greifbarer Nähe. Deshalb ist es umso erstaunlicher, dass - wie die Pressemitteilung stolz vermeldet - die Studie zwar von RTL, der Fachzeitschrift Elektrosmog und dem Portal diagnose:funk, letzteres ein Organ für Elektrosmog-Obskuranten, aber bisher von keiner der seriösen deutschen oder Internationalen Medien beachtet wurde.

Hunderte von kleinen und methodisch problematischen Studien haben in den letzten Dekaden einen Zusammenhang von elektromagnetischer Strahlung aus Handys und

Krebs, Frühgeburten, Depressionen, und jeder Menge anderer Gesundheitsschäden nahegelegt, nach deren Interpretation sogar bewiesen. Grosse, gut gemachte Studien haben aber all dies widerlegt. Dass eine sehr kleine Studie nun einen akuten Effekt von elektromagnetischer Strahlung aus dem Handy auf das Gehirn, nämlich eine massive, einfach nachzuweisende Veränderung im Essverhalten nachweisen kann, ist bemerkenswert. Der Wissenschaftsnarr hat sich den Nutrients Artikel deshalb genauer angeschaut. Auch um die Frage zu klären, was da so alles von der DFG gefördert wird. Denn diese Studie wurde letztlich von uns Steuerzahlern finanziert, und zwar über den Transregio SFB 134.

Um es vorweg zu nehmen: Diese Studie ein Lehrstück für all die Probleme, welche die Biomedizin seit Jahren plagen, und dem Narren leider unerschöpflich Stoff für diese Kolumne bieten. Nichts wurde hier ausgelassen. Die Probleme fangen beim Studiendesign an, und hören beim Reporting, also der eigentlichen Veröffentlichung auf. Aber immer der Reihe nach.

Zunächst einmal gilt in der Wissenschaft der Grundsatz: „Außergewöhnliche Aussagen benötigen außergewöhnliche Evidenz." Und die ist mit 15 jungen Männern, auch in einem cross-over Design einfach nicht zu haben. Das kann man ganz einfach statistisch begründen, das sagt einem aber auch der gesunde Menschenverstand. Ob nun ein paar Hundert oder ein paar Tausend Probanden nötig sind, und wie oft das dann an anderer Stelle repliziert werden müsste, darüber lässt sich trefflich biometrisch fachsimpeln. Klar ist nur: 15 Probanden, und ohne Konfirmation, das geht gar nicht.

Umso mehr hier eine heterodoxe, sehr unwahrscheinliche Hypothese untersucht wird. Mit 15 Studienteilnehmern kann man bestätigen, was man im Grunde eh schon weiß. Zum Beispiel dass ein neues Medikament aus einer lange untersuchten Substanzklasse, von der bekannt ist, dass sie den Blutdruck erniedrigt (z.B. Sartane), tatsächlich den Blutdruck verringert. Statistisch ausgedrückt besteht das Problem u.a. darin, dass mit abnehmender ‚prior probability' der Hypothese (‚Vortestwahrscheinlichkeit', ‚base rate') die Zahl der falsch positiven Resultate zunimmt. Bei unwahrscheinlicher Hypothese, niedriger statistischer Power (15 Probanden!), und nicht sehr stringenten Typ I Fehler - Signifikanzniveau (hier 5%), werden falsch positive Ergebnisse immer wahrscheinlicher. Nur wenn man den p-Wert mit dem positiven (bzw. negativen) Vorhersagewert verwechselt – und viele Kollegen tun dies leider - kann man sich unter diesen Bedingungen mit einem ‚statistisch signifikanten Ergebnis' wohl fühlen. Alle anderen, auch diejenigen die den Wissenschaftsnarr ‚Brüder, zur Sonne, dem p-Wert ein Ende, Brüder, zum Lichte empor!' (LJ 10/2019) zum Thema gelesen haben, werden unbeeindruckt stärkere Evidenz fordern. Ganz nebenbei sei erwähnt, dass die Physiker für eine ‚Entdeckung', und als dies dürfte man das von den Lübeckern beschriebene Phänomen dann getrost zählen, eine statistische Signifikanz auf 5-Sigma Niveau gefordert wird. Das bedeutet, einen Typ I Fehler mit einer Wahrscheinlichkeit von über 1:3.3 Millionen zu akzeptieren, ein p-Wert mit mehr als 7 Nullen. Wer sich bereits mit dem 5% Niveau begnügt, tut dies in einer von 20 Fällen. Ronald Fisher, der ‚Erfinder' des p-Werts charakterisierte Befunde auf 5% Niveau lapidar so: ‚Worth a look'.

Es geht aber weiter bei den statistischen Problemen der Handy-Studie. Es werden eine Vielzahl von Vergleichen gemacht, aber keiner davon als ‚primärer Endpunkt' definiert, und die anderen Vergleiche dann als explorativ gekennzeichnet. Deshalb hätten die Typ-1 Fehler Niveaus auch entsprechend korrigiert werden müssen (z.B. mit einer Bonferroni-Korrektur). Auch die sich im Artikel andeutende (bzw. auch unmittelbar suggerierte) Verwechslung von Korrelation und Kausalität (mehr Handys, mehr Dicke) muss hier erwähnt werden. Zeitgleich mit der Zunahme der Nutzung von Handys sind viele

Dinge passiert. Nicht nur wurde die Weltbevölkerung dicker. Auch haben Elektroautos zugenommen, und der Thunfisch im Mittelmeer hat abgenommen. Auf der sehr unterhaltsamen Website https://www.tylervigen.com/spurious-correlations sind eine Vielzahl von solchen ‚spurious correlations' gelistet. Mein Liebling darunter ist die nahezu perfekte Korrelation der US-Ausgaben für Wissenschaft, Raumfahrt und Technologie mit den Selbstmorden durch Erhängen, Strangulieren und Ersticken.

Die Studie war im Übrigen nicht präregistriert. Wir wissen also nicht, was alles in deren Verlauf angepasst wurde, welche Analysen vorgesehen waren und welche dann tatsächlich gemacht wurden, welche Daten Verwendung finden sollten und welche es dann in die Auswertung geschafft haben, usw (‚undisclosed flexibility'). Ich will hier gar nichts unterstellen, aber es macht die Interpretation einer Studie soviel eindeutiger, wenn alles auf dem Tisch liegt bevor es los geht. Selbstverständlich hätten sie dies mit einer Sperrfrist bis zur Veröffentlichung der Resultate tun können, das ist bei solchen Studien sogar die Regel. Damit ihnen niemand die tolle Idee klaut und vorher publiziert.

Das Studiendesign war laut Autoren ‚single-blinded'. Man kann nur annehmen, dass die Verblindung sich hier auf die Studienteilnehmer bezog. Diese wussten nicht einmal was die eigentliche Fragestellung der Studie war, noch welches Handy nun sendete oder nur ‚sham' war. Das ist vorbildlich. Nur waren es dann wohl die Untersucher selbst, die nicht verblindet waren bei der Auswertung der Daten. Ich denke es braucht keiner weiteren Erläuterung welche Auswirkungen unbewusster Bias in Studien haben kann.

Und dann die (nicht)Verfügbarkeit der Originaldaten. Im Artikel steht, dass diese auf Anfrage zur Verfügung gestellt werden könnten. Das ist für sich schon eine Enttäuschung. Denn wer je versucht hat, ‚data on reasonable request' von Studienautoren zu bekommen, weiß dass dies in den seltensten Fällen gelingt. Wenn man überhaupt eine Antwort auf die Anfrage bekommt, spricht meist irgendwas gegen eine Herausgabe. Manchmal ist auch die Festplatte kaputt gegangen, oder der Zuständige ist unbekannt verzogen. Auch hier will ich nichts unterstellen, fest steht aber, dass die Daten von öffentlich geförderten Studien auch der Öffentlichkeit zur Verfügung gestellt werden müssen. Sagt das nicht sogar der Fördergeber dieser Studie, die DFG? Warum können wir die Daten dieser Studie nicht einfach von ZENODO oder einem anderen Repositorium herunterladen?

Dabei hätte man sich die Originaldaten sehr gerne angesehen, denn deren Repräsentation in der deskriptiven Statistik des Artikels lässt leider zu wünschen übrig. Dort finden wir, statt Dot- und Box Plots sowie vernünftigen Varianzmassen leider nur die allgegenwärtigen - aber die wahren Varianzen beschönigenden und die Verteilungen verschleiernden - Bar Graphen mit Standard Error (SEM). Bei der Darstellung der neuroenergetischen Daten fehlen die SEMs für die „Sham"-Ergebnisse gleich ganz. Der hier wohl gezeigte Gruppen-Mittelwert führt denn auch zu einer starken Überbetonung der Effektstärke. MR-Spektroskopiker hätten sich auch über ein paar Originalspektren zur Beurteilung der Messqualität sehr gefreut. All dies Hinweise darauf, dass die Reviewer von ‚Nutrients' hier wohl nicht so genau hingeschaut haben.

MR-Spektroskopie ist aufwendig und teuer. Man hätte das Experiment aber durchaus an einer größeren Zahl von Probanden, ganz ohne MR-Spektroskopie replizieren können, nur unter Messung der Kalorienaufnahme. Da wären dann nur noch die Kosten für das Frühstücksbuffet angefallen. Damit hätte man auch gleich ausschließen können, dass es zu einer wie auch immer gearteten Wechselwirkung der Leistungsdeposition in das Gehirn beim Senden über die Mobilfunkantenne und dem Leistungseintrag durch die HF-Pulse während der Magnetresonanz-Spektroskopie gekommen ist. Vor allem wenn man von 15 Probanden auf die Weltbevölkerung schließen will, würde das viel Sinn

machen, denn die wenigsten von uns liegen vor und nach einem Handytelefonat im Tomographen.

Manch Leser wird sich wohl fragen, warum ein so sensationeller, uns potentiell weltbewegender Befund in Nutrients veröffentlicht wurde, einer Zeitschrift die schon mehrfach durch Skandale aufgefallen war (für Interessierte: googeln sie z.B. mal ‚Australian paradox'). Und nicht in Nature oder Science. Ich halte diesen Aspekt allerdings für wenig relevant. Eine Studie muss auf Basis ihres Designs, ihrer Ergebnisse und deren Interpretation bewertet werden, und nicht nach dem Journal in dem sie publiziert wurde. Klar ist aber, dass sich – viele von uns kennen das aus eigener Erfahrung – dieser Artikel in einer Kaskade von Journalen mit abnehmendem Journal Impact Factor abwärts bewegt haben muss. Für mich wieder einmal ein Hinweis darauf, dass letztlich alles irgendwo publiziert werden kann und wird, und dass der Review Prozess nicht die derzeit - auch wegen der Zunahme von Preprints - viel beschworene Filterfunktion hat. Dass ein Artikel in einem Journal mit einem Impact Factor von über 5 und damit in der ersten Quartile der unter Nutrition & Dietetics gelisteten Journale veröffentlicht wird, ist eben für sich erstmal gar kein Ausweis für Qualität. Man muss sich immer noch die Mühe machen, den Artikel zu lesen und ihn inhaltlich und methodisch bewerten.

Sollte man also hochgradig unwahrscheinliche Hypothesen erst gar nicht untersuchen? Ist die Untersuchung einer möglichen Wirkung von elektromagnetischer Strahlung auf das Gehirn und unser Verhalten grundsätzlich abzulehnen? Ist der Ansatz der Autoren total esoterisch? Natürlich nicht! ‚To boldly go where no man has gone before' ist eine der vornehmsten (und spannendsten) Aspekte der Wissenschaft. Nur müssen wir methodisch solide bleiben, und unsere Ergebnisse nicht überinterpretieren. Eine Lübecker Handy-Studie, präregistriert, mit ein paar hundert Probanden (nur Frühstück!!), doppelt verblindet, inklusive dem versuchten Nachweis einer Dosis-Wirkungsbeziehung (15 Minuten Handytelefonat vs. 30 Minuten), repliziert in einem anderen Labor, die Daten frei verfügbar bei ZENODO, das wäre einer Berichterstattung in Nature, New York Times, und Neuer Züricher Zeitung würdig. Auch wenn man nach Handybestrahlung nur ein Löffelchen Müsli mehr essen müsste.

Der Wissenschaftsnarr dankt Prof. Dr. Harald Möller vom Max-Planck-Institut für Kognitions- und Neurowissenschaften in Leipzig für seine sachdienlichen Hinweise zur MR-Spektroskopie in der Studie.

Warum es klemmt bei Open Science und der Reform des akademischen Bewertungssystems?

LJ 6/2022

Im November letzten Jahres haben die 193 Mitgliedstaaten der UNESCO, also alle von der UN anerkannten unabhängigen Länder dieser Erde, ‚Empfehlungen zur Offenen Wissenschaft (Open Science)' unterzeichnet. Darin verpflichten sich die Staaten, eine Kultur der offenen Wissenschaft zu fördern, in diese zu investieren und Anreize für sie zu schaffen. Dies war wohl nur ein vorläufiger Höhepunkt eines wahren Tsunami von Empfehlungen, Manifesten, und Aufrufen Wissenschaft transparenter, ihre Ergebnisse robuster und werthaltiger zu machen, mehr zu kollaborieren und Ergebnisse zu teilen, sowie die Gesellschaft stärker in den Forschungsprozess einzubeziehen. Die Europäische Kommission, die Leading European Research Universities LERU, die European University Alliance EUA, SCIENCE EUROPE (ein Zusammenschluss der größten europäischen Forschungsförderer, inklusive der DFG) – um nur ein paar Organisationen zu nennen – alle haben sie in den letzten Monaten und in der Regel mit länglichen Papieren nicht nur zu Open Science aufgerufen. In den Fokus dieser Aktivitäten nahmen sie eine Reform des Belohnungs- und Karrieresystem! Dieses, so steht es in all diesen Papieren, sei der wichtigste Hebel die Wissenschaft offener zu machen, und müsse bedeuten uns abzunabeln von ungeeigneten und gar schädlichen Metriken wie dem Journal Impact Factor (JIF). Viel mehr sollten wir uns entwickeln hin zu einer inhaltlich orientierten Bewertung von Forschern und deren Produkten. Die EU organisiert deshalb derzeit eine ‚Koalition der Willigen', in einem großangelegten ‚Process towards an agreement on reforming research assessment'.

So richtig bewusst wurde mir all dies, als ich im Februar auf eine Open Science Konferenz nach Paris eingeladen wurde. Veranstaltet wurde diese vom französischen Staat. Zeitungsleser wissen, dass Frankreich derzeit die Ratspräsidentschaft der EU innehat. Und ob Sie es glauben oder nicht, die Franzosen haben glatt Open Science zu einer wesentlichen Priorität ihres Vorsitzes über die europäischen Staaten gemacht. Auf dieser Konferenz sprachen und diskutierten nun die Granden der EU Wissenschaftspolitik und Forschungsförderung, wie Jean-Eric Paquet, der Generaldirektor der europäischen Kommission für Forschung und Innovation, Maria Leptin, die Präsidentin des European Research Council ERC, oder Marc Schlitz, der Präsident von Science Europe. Und sie alle argumentierten als hätten sie das Laborjournal abonniert und alle meine Kolumnen gelesen und verinnerlicht!

Das machte mich nachdenklich, ich habe deshalb alle oben erwähnten, fast 20 Aufrufe zur Durchsetzung von Open Science und einer Reform der akademischen Begutachtungssysteme studiert (eine Link-Liste hierzu finden sie unter http://dirnagl.com/lj). Da steht viel Schlaues drin, aber wenn sie nur eines davon im Original lesen wollen, empfehle ich die UNESCO Empfehlungen. Diese haben ja schließlich alle Staaten

unterzeichnet, damit natürlich auch Deutschland. Aber wenn sie die Sache weiter abkürzen wollen, hier ist meine 2-Satz Zusammenfassung aller Dokumente:

‚Wissenschaftliche Erkenntnisse sollen für jedermann offen verfügbar, zugänglich und wiederverwendbar gemacht werden, die wissenschaftliche Zusammenarbeit und der Informationsaustausch zum Nutzen von Wissenschaft und Gesellschaft gestärkt und die Prozesse der wissenschaftlichen Erkenntnisgewinnung, -bewertung und -vermittlung für gesellschaftliche Akteure über die traditionelle Wissenschaftsgemeinschaft hinaus geöffnet werden. Das System der Forschungsbewertung muss reformiert werden, damit Qualität, Leistung und Impakt von Forschung und Forschern auf der Grundlage geeigneterer Kriterien und Verfahren bewertet werden.‘

Spätestens jetzt werden Sie sich vermutlich die Augen reiben. Wenn Staaten, Forschungsförderer und die wichtigsten akademische Institutionen das so wollen, wenn deren Vertreter reden wie Open Science Aktivisten der ersten Stunde, warum wird Ihr Antrag oder Ihre Bewerbung dann immer noch nach der Reputation der Journale bewertet, in denen sie veröffentlich haben? Warum sind in Ihrer Einrichtung JIF und Drittmittelaufkommen immer noch die wesentlichen Kriterien für’s akademische Fortkommen? Warum wird, wenn Sie an einer medizinischen Fakultät arbeiten, ihre Forschungsleistung über den JIF mit einer Genauigkeit von drei Nachkommastellen und der Summe Ihrer Drittmitteleinwerbung berechnet und als LOM (Leistungsorientierte Mittelvergabe) honoriert?

Was also sind die Gründe für dieses eklatante Implementierungsdefizit? Die Problematik ist, wie sollte es anders sein, komplex. Und auch nicht wirklich neu. Vor ziemlich genau 20 Jahren forderte die Budapester Open Access Initiative, öffentlich geförderten Ergebnissen der Wissenschaft eben dieser Öffentlichkeit bitteschön auch kostenfrei und online zur Verfügung zu stellen. Mehr als 1300 Institutionen und Organisationen und mehr als 6000 teils sehr prominente Individuen haben dies bisher unterzeichnet. Mittlerweile ist Open Access (OA), das ja ein Teil des weit mehr umfassenden Konstrukts Open Science ist zwar ein etablierter Teil des wissenschaftlichen Publikationssystems. Allerdings ist ein substantieller Teil der Literatur immer noch nicht OA verfügbar, die prestigereichsten Journale (d.h. die uns den für die Karriere so wichtigen möglichst hohen JIF verkaufen) bieten gleich gar kein OA, oder die Kosten für OA sind so hoch, dass sie Autoren bzw. Institutionen finanziell ruinieren können. Dies führt auch dazu, dass Ungleichheiten im Wissenschaftssystem nicht beseitigt, sondern sogar verstärkt werden. OA muss man sich leisten können – als Land, bzw. auch als Forscher. Die OA Bewegung hat also zwei Jahrzehnte benötigt vom Manifest-Stadium bis zu einer partiellen Umsetzung in ein alternatives Geschäftsmodell der Verlage mit teils zweifelhaften Resultaten!

Aber in Open Science steckt noch viel mehr drin als offener Zugang zu wissenschaftlichen Artikeln. Dazu gehört auch die Verfügbarkeit von Daten und Code, die Einbindung gesellschaftlicher Akteure in den Forschungsprozess, Kollaboration und Team Science, oder die faire Begutachtungs- und Evaluationspraxis. Diese Breite generiert noch mehr Komplexität. Es ist eben nicht damit getan, ein Manifest zu schreiben und dann zu erwarten, dass alle darin angesprochenen Akteure sogleich überzeugt und begeistert sind und sich umgehend an dessen Umsetzung machen. Irgendwie hängt alles mit allem zusammen: Die Auflagen der Forschungsförderer, die Policies welche sich Universitäten geben und welche Anreize sie ihren Wissenschaftlern setzen, welche Ressourcen die Universitäten dafür einstellen, aber auch die gesetzlichen Vorgaben und die Berechnungsmodelle, mit welchen die Länder die Universitäten finanzieren. Und natürlich was alles so in den Köpfen der Wissenschaftler spukt.

Wenn sich zum Beispiel die Finanzierung einer Hochschule durch das Land an deren Publikationsleistung und eingeworbenen Drittmittel orientiert, wird die Leitung dieser Hochschule ihre Wissenschaftler genau hierauf verpflichten wollen. Hochschulleitungen schielen auch sehr auf ihre Positionierung in internationalen Rankings. Bei diesen spielt die offene Wissenschaft nun gar keine Rolle. Und dann natürlich die Ressourcenfrage: Zum Beispiel bräuchten Wissenschaftler zum FAIRen (findable – accessible – interoperable -reusable) Teilen von Forschungsdaten Unterstützung durch Data stewards und IT-Infrastruktur. Die deutschen Unis sind aber bekanntermaßen strukturell massiv unterfinanziert, da muss jeder Groschen zweimal umgedreht werden, bevor er dann am Ende doch für die an den JIF gekoppelte Leistungsorientierte Mittelvergabe (LOM) ausgegeben wird.

Aber auch wir Wissenschaftler tun uns mit der Open Science schwer. Vor allem weil wir in einem System sozialisiert wurden, in dem diese für unser Fortkommen bisher völlig irrelevant war. Durch die Allgegenwart und Allmacht reputationsbasierter Metriken halten wir Drittmittel und Publikationen in Journalen mit hoher Reputation für Forschungsleistungen. Diese folgenschwere Verwechslung, ist dabei nur zu verständlich: Denn es sind genau diese ‚Forschungsleistungen' welche zur Professur und zum Drittmittelerfolg führen. Auch die besagte LOM gaukelt uns die Gleichsetzung von JIF mit Forschungsleistung vor. Was soll da jetzt plötzlich das Gerede von Open Data, oder gar der Beteiligung gesellschaftlicher Akteure in unserer Forschung? Oder die Präregistrierung von Studienprotokollen, das würde ja glatt unsere nicht offen gelegte Freiheit in der (nicht-)Verwendung von Studienergebnissen einschränken und damit das Geschichtenerzählen durch ‚next we' Narrative in Publikationen erschweren. Für diejenigen, die in Kommissionen sitzen und Professoren berufen, oder Anträge begutachten, hat die Fokussierung auf JIF und Drittmittel noch ganz andere, nämlich sehr praktische Vorteile: Es erleichtert die Selektion ungemein. Man kann Kandidaten damit ganz einfach in Spreadsheets sortieren und muss sich nicht mühsam durch deren wissenschaftliches Oeuvre quälen. Und wird dazu noch mit einer auf Zahlen beruhenden, damit objektiven und ‚gerechten' Auswahl belohnt.

Nur den Fördergebern fällt es viel leichter, neue Kriterien für die Bewertung von Forschung sowie Open Science Praktiken durchzusetzen. Wer zahlt, schafft eben an. Deshalb kommt von diesen auch momentan der größte Druck auf das System. Sie können all dies bei Antragstellern ganz einfach einfordern, und in den Begutachtungsprozess integrieren. Europäische Union, Wellcome Trust, auch das BMBF beginnen hier die Daumenschrauben anzusetzen. Nur die DFG tut sich dabei noch schwer. Kein Wunder, sie ist ja das Organ der ‚Selbstverwaltung' bzw. ‚Selbstorganisation' der deutschen Wissenschaft' – oder anders ausgedrückt, wir sind die DFG! Das bedeutet, dass die arrivierten Wissenschaftler bei uns das sagen haben und nicht irgendwelche Apparatschiks – und so schallt es denn aus dem Akademikerwald zurück wie einen Absatz weiter oben beschrieben.

Wir brauchen also keine weiteren Manifeste, sondern mehr Pioniere (‚Champions') in der Umsetzung, und mehr Handreichungen wie die Implementierung von Open Science Praktiken denn praktisch geschehen soll. Bei den Wissenschaftlern brauchen wir zum Beispiel mehr Mut bei denjenigen, die es im System schon ‚zu etwas gebracht' haben, und nicht mehr das akademische Hamsterrad drehen müssen. Die jungen Wissenschaftler, welche überwiegend hinter einer grundlegenden Reform des akademischen Systems stehen, geraten durch den Auftrag der in den Manifesten an sie heran getragen wird in einen ‚double bind'. Sie sollen neue Praktiken und Kriterien umsetzten, während sie gleichzeitig nach den alten beurteilt werden. Zudem sind sie das schwächste Glied in der Kette.

Von den akademischen Institutionen braucht es mehr Unterstützung ihrer Wissenschaftler und Kommissionen, z.B. in Form von Trainings. Noch wichtiger aber: Sie müssten zusätzliche Kriterien bei der leistungsorientierten Mittelvergabe und Rekrutierung einführen. Das könnte die Belohnung von Open Data sein, oder kurze Narrative zum eigenen Beitrag in Sachen Open Science in den Lebensläufen, oder auch kurze Begründungen warum bestimmte Publikationen als die ‚besten Fünf' ausgewählt wurden. Dass sie in Cell, Nature oder Science veröffentlicht wurden, wären dabei dann keine geeigneten Antworten. Oder wie wäre es z.B. mit einer Verblindung der Namen der Journals in der Literaturliste? Und stattdessen Links bei den Titeln, die direkt in das Paper führen. Dann müssten Kommissionsmitglieder die Artikel zumindest aufrufen um rauszufinden, wo es publiziert wurde – und könnten so vielleicht auch gleich was vom Inhalt des Artikels mitbekommen.

Natürlich gäbe es viele weitere Möglichkeiten. Von keiner wissen wir, wie praktikabel sie ist, oder ob sie ihren Zweck erfüllt oder gar nicht intendierte, negative Auswirkungen hat. Aber nur in der Anwendung finden wir das heraus – durch Pilotprojekte, und maßvolle Modifikation bestehender Verfahren. Was funktioniert wird ausgebaut, was nicht funktioniert wird verbessert oder aufgegeben. Am besten würden solche Pilotprojekte gleich als ‚Interventionen' verstanden und wissenschaftlich begleitet. Evaluations- bzw. Implementierungsforschung kann solide Evidenz liefern um Institutionen, Fördergeber und Wissenschaftler bei der Umsetzung von spezifischen Maßnahmen rationell zu beraten.

Die Fördergeber sollten all dies unterstützen, indem sie für solche Maßnahmen und Pilotprojekte Förderlinien ausschreiben, aber auch für Implementierungsforschung. Auch sollten sie die Beantragung von Mitteln ermöglichen, welche Open Science fördern. Also zum Beispiel für Forschungsdatenmanagement, wissenschaftliches Qualitätsmanagement, Data stewards, Patienten – und Stakeholder Engagement, usw. Last but not least müssen auch die Geldgeber der Universitäten aktiv werden – sie sollten die Berechnung der Landeszuführungsbeiträge auch an die Umsetzung von offener Wissenschaft knüpfen, so wie sie dies ja auch schon für Open Access und Gleichstellung mit einigem Erfolg gemacht haben. Ob das alles dann weniger braucht als 2 Dekaden wage ich zu bezweifeln. Aber das sollte uns nicht abhalten, jetzt aktiv zu werden. Gut Ding will eben Weile haben.

Pimp your paper!

LJ 9/2022

Täglich ergießt sich ein Tsunami wissenschaftlicher Artikel über uns. Es gibt etwa 30.000 medizinische Journale, keiner weiß das so genau, die jährliche Wachstumsrate liegt bei über 5 %. Die MEDLINE listet jährlich mehr als 1,7 Millionen Artikel, Tendenz unaufhörlich steigend. Da lesen wir dann Triviales oder gar Obskures, und sehr häufig auch Spektakuläres. Befunde, die nach den Worten der Autoren die medizinische Praxis revolutionieren würden. Sie fragen sich vielleicht, wie auch Sie dazu beitragen können, diesen Strom biomedizinischer Evidenz nicht versiegen zu lassen, und damit gleichzeitig Ihren CV zu bereichern? Im Folgenden möchte ich Ihnen deshalb einige Tipps aus meiner langjährigen Praxis als Autor, Reviewer und Journal Editor geben. Vieles mag Ihnen selbstverständlich oder gar trivial erscheinen. Ich denke aber, dass meine Handreichungen Ihnen gerade in dieser Zusammenschau helfen können, auch aus noch so fragmentierten, irrelevanten, oder schlecht designten Experimenten oder Studien einen Artikel zu stricken, der den Peer Review übersteht und sich danach auch nicht auf der Liste Ihrer Originalarbeiten verstecken muss.

Es beginnt damit, dass sie auf keinen Fall auf Sprüche wie ‚Spektakuläre Ergebnisse oder Behauptungen erfordern außergewöhnliche Beweise!' hereinfallen dürfen. Carl Sagan hat diesen Spruch von Pierre-Simon Laplace plagiiert, der die Maxime im 18. Jahrhundert so formulierte: "Das Gewicht der Beweise für eine außergewöhnliche Behauptung muss im Verhältnis zu ihrer Seltsamkeit stehen". Himmel, das war doch zu einer Zeit als Gentleman scientists forschten, ohne Anträge schreiben zu müssen oder ihr Ansehen oder Anstellung an der Anzahl der Publikationen und deren Impact Factor gemessen wurden! Diesen Luxus können wir uns heute wahrlich nicht mehr leisten. Lassen Sie also Ihren Spekulationen von vorneherein freien Lauf und verengen sie Ihre Schlussfolgerungen nicht durch engstirnigen Blick auf die Qualität der von Ihnen generierten Evidenz. Höchstes Gut ist nach wie vor die Freiheit der Wissenschaft, außerdem ist Kreativität die Haupttriebfeder von Innovation.

Analog gilt das übrigens auch für die Formulierung der Hypothese. Sollte es Ihnen wider erwarten nicht gelungen sein, diese in der Studie zu belegen, sollten sie erwägen, diese im Lichte ihrer Ergebnisse kreativ zu modifizieren. Ich rate allerdings davor ab, dies im Manuskript zu erwähnen, es soll Gutachter geben, die das aus Unkenntnis über die aktuelle epistemische Praxis als unwissenschaftlich kritisieren. Festzuhalten bleibt, dass sich in Kenntnis der Ergebnisse ein viel überzeugenderes Narrativ aufbauen lässt. Dieses kann man noch erheblich stärken, in dem man Befunde, welche die sich abzeichnende ‚Story' stören, im Artikel nicht erwähnt. Letztlich kommt alles auf eine gute Auswahl der für die Studie ausgewählten Befunde aus dem großen Pool der Ihnen zur Verfügung stehenden Ergebnisse an. Nur durch eine ebenso umsichtige wie selektive Auswahl werden Artikel möglich, die dem Leser eine faszinierend lineare (‚Next we…') und schlüssige

(‚We have demonstrated…') Argumentation bieten, ihn dabei aber nicht auch noch mit unwesentlichen Nebenbefunden überfrachtet. Unsere Artikel sind ja selbst im von uns kuratierten Narrativ meist schon komplex genug!

Beim Design der Studie werden sie der statistischen Power, also dem Typ II Fehler keine große Aufmerksamkeit gewidmet haben. Das war eine weise Entscheidung, denn sie präsentieren ja positive Ergebnisse, wozu sollte man sich um falsch negative sorgen? Außerdem hätte eine a-priori Power-Analyse und Fallzahlabschätzung vermutlich ergeben, dass die von Ihnen und in Ihrem Forschungsfeld verwendeten Gruppengrößen viel zu gering sind. Aber zum einen haben wir alle doch schon immer so kleine Gruppen verwendet, und außerdem würden bei den tatsächlich notwendigen Fallzahlen die Ressourcen aus dem Förderantrag nicht reichen, die Genehmigungsbehörde sich beschweren, und es für einen Doktoranden zu lange dauern. Vielleicht haben sie in diesem Zusammenhang auch schon mal den sogenannten ‚Sample Size Samba' tanzen müssen. Falls nein, möchte ich Ihnen diese Technik an dieser Stelle ans Herz legen. Dazu geben Sie einfach eine unrealistisch hohe Effektstärke in das Statistikprogram ein (z.B. 1.5 Standardabweichungen), dann errechnet sich daraus schon bei den gewohnt niedrigen Fallzahlen ein ausreichendes Niveau für Typ I (5 %) und Typ II (20% - das heißt 80 % Power) Fehler. Spielen sie in der Software mit den Effektgrößen bis es passt. Dass Sie eine a-priori Power Analyse und Fallzahlabschätzung gemacht haben macht sich auf jeden Fall gut im Artikel, und die Behörde ist auch glücklich.

Bei der statistischen Analyse halten Sie sich am besten in gewohnter Weise an den seit nunmehr über 100 Jahre bewährten p-Wert und die magische 5 % Signifikanzschwelle, Ronald Fisher sei Dank! Auch hier gilt: Alle machen das schon lange so, dann kann das doch nicht falsch sein. Lassen Sie sich auch nicht durch Unkenrufe sogenannter ‚Experten' irritieren, dass ein so wenig stringentes Signifikanzniveau kombiniert mit niedriger Power zu einem sehr hohen Anteil falsch positiver und falsch negativer Ergebnisse führen muss. Und zudem noch tatsächlich existierende Effekte größenmäßig stark überschätzt werden. Meist folgt dann von diesen Besserwissern gleich noch der Hinweis, man solle sich doch bitte auf biologische Effekte und deren Ausmaß, und nicht die statistische Signifikanz fokussieren. Solche Kommentare müssen als realitätsfern zurückgewiesen werden. Wenn man sie ernst nehmen würde, könnte man viele der Effekte, die man doch gerade eben publizieren will, nicht mehr belegen und müsste die gesamte Diskussion umschreiben. Auch die für die Akzeptanz durch das Journal wichtigen spektakulären Schlussfolgerungen ließen sich nicht mehr halten. Das ganze Manuskript wäre gefährdet!

Noch ein Wort zur Wahl der Teststatistik. Nutzen Sie die grenzenlosen Möglichkeiten, welche moderne Statistikpakete bieten. Oft führt erst die Durchführung einer Reihe verschiedener Testverfahren zur gewünschten Signifikanz. Auch bei den post-hoc Tests sollten sie nicht zu schnell aufgeben, es findet sich fast immer ein wenig konservativer Kontrast der eine Signifikanz an der richtigen Stelle ergibt. Ganz klar muss ich allerdings in diesem Zusammenhang davor warnen, für multiple Vergleiche zu korrigieren. Wir machen ja häufig in einer Studie so viele verschiedene Schlüsse von der Stichprobe auf die Grundgesamtheit in Form von statistischen Tests, dass dies schon aus praktischen Gründen gar nicht mehr möglich ist. Da kommen schnell mal mehr Vergleiche zusammen als unabhängige experimentelle Einheiten vorhanden waren. Aber noch viel wichtiger: Das Adjustieren der p-Werte zerstört häufig die mühsam erarbeitete statistische Signifikanz, das geht also gar nicht.

Das bringt mich zur graphischen Darstellung der Resultate. Hier haben sich Bar-Graphen mit Standfehler des Mittelwertes (SEMs) unglaublich bewährt. Diese

Darstellungsform zeichnet sich durch die große Klarheit aus, mit welcher sich die Effekte des oben erwähnten, statistischen Vorgehens noch weiter schärfen lassen. So ist es z.B. möglich, störende bimodale Verteilungen (und damit Fehlen einer Normalverteilung) graphisch komplett zum Verschwinden zu bringen, auch die häufig unangenehm hohe Varianz der Resultate wird durch die SEM visuell auf ein erträgliches Minimum reduziert. Mit ein bisschen Geschick lassen sich die Statistikprogramme auch dazu bewegen, ordinale Werte auf diese Weise darzustellen, dadurch fällt es auch nicht mehr so auf, dass wir darauf parametrische Tests angewendet haben. Durch Sternchen (*), welche über den Balken schweben und von statistischer Signifikanz künden, findet das Auge sofort Halt an den wesentlichen Befunden. Es ist sicher richtig, dass Box-, Violin-, Dot-Plots etc. wesentlich mehr Information vermitteln würden, aber gleichzeitig verwässern sie damit auch die eindeutigen Aussagen, welche nicht nur Leser, sondern besonders auch die Reviewer so schätzen. Ich rate daher dringend von diesen Darstellungsformen ab, sie machen im Übrigen auch mehr Arbeit.

Unangenehmer Weise fragen Journale immer häufiger, welche Anstrengungen man zur Vermeidung von Verzerrungen (Bias) unternommen hat, also z.B. Verblindung, Randomisierung, Vordefinition von Ein- und Ausschlusskriterien. Lassen Sie sich hiervon nicht einschüchtern. Verzerrungen sind ein notwendiges, und noch dazu schwer zu bekämpfendes Übel. Keine Angst, die entsprechenden Check-Listen für die Einreichung beim Journal lassen sich meist schnell durchklicken und einfach Häkchen an den gewünschten Stellen setzen. Sehr praktisch ist es, wenn das Journal es erlaubt, das Ganze mit einfachen Sätzen im Methodenteil zu erledigen, schreiben Sie dann einfach: ‚This study was conducted in compliance with the ‚X'-guidelines' (X dann ersetzen durch ARRIVE, CONSORT, etc., je nach Journal und Studie).

Sollte das Journal eine Open Data Policy haben, und auf der freien zur Verfügungstellung der für die Publikation verwendeten Originaldaten bestehen, versuchen Sie es am besten zunächst mit der Floskel: ‚Data available upon reasonable request'. Damit sind Sie sicher davor geschützt, diese wertvolle Ressource, die ja immerhin von Ihnen und Ihren Mitarbeitern in harter Arbeit erzeugt wurde, mit potentiellen Konkurrenten teilen zu müssen. Sie können die Daten dann ungestört für weitere eigene Publikationen recyclen. Aus einem einzigen Datensatz soll es findigen Kollegen schon gelungen sein, zwanzig und mehr Publikationen (d.h. zwei Habilitationsäquivalente!) zu schöpfen. Sie beugen durch die kategorische Verweigerung von Open Data auch ganz allgemein einem wissenschaftlichen Parasitentum vor. Dieses macht sich derzeit, angefeuert von Open Science Aktivisten immer mehr breit. Sollte Ihr Fördergeber sie dennoch durch Auflagen zur Datenteilung verpflichten wollen, können Sie dies getrost ignorieren. Es ist bisher kein Fall bekannt geworden, bei dem es bei Nichterfüllung zu Konsequenzen gekommen wäre. Irgendeine Ausrede, warum es Ihnen nicht möglich war, die Daten zu teilen, wird Ihnen schon einfallen.

Viele Journale bestehen auf der Nennung möglicher Interessenkonflikte. Auch hier sollten Sie der Einfachheit halber gleich mit ‚none' antworten. Auch wenn es nicht stimmen sollte, brauchen Sie keine Konsequenzen befürchten. Wer, denn Sie selbst sollte besser wissen, womit sie einen Konflikt haben könnten? Häufig heben sich Interessenkonflikte ja auch gegenseitig auf, insbesondere wenn man Fördermittel und Honorare von den verschiedensten Firmen erhält. Als Wissenschaftler sind Sie aber auch ganz grundsätzlich vor solchen Konflikten geschützt, da Sie doch nur der wissenschaftlichen Wahrheit verpflichtet sind. Auch sind ihre Ergebnisse objektiv und mit aufwendigen Methoden quantifiziert, und daher auch durch kollidierende Sekundärinteressen gar nicht beeinflussbar. Einem FACS Gerät oder einem Mikroskop ist es doch egal, von wem Sie materielle oder finanzielle Mittel erhalten. Sogar Patente und deren Anmeldungen werden

heutzutage als Interessenkonflikte aufgefasst! Das ist natürlich unlogisch, die Unis geben sich doch große Mühe uns auf valorisiernde Maßnahmen zu verpflichten. Wir sind ja nur die Erfinder, Eigner der Patente sind in der Regel unser Arbeitgeber, wenn dann müssten doch die Unis einen Konflikt haben. Und wenn wir wegen des Geldes dabei wären, hätten wir einen anderen Beruf als Wissenschaftler ergriffen.

Ein Wort noch zum Verhältnis von Kausalität und Korrelation. Das Dogma ‚Korrelation ist nicht Kausation' hat hier viel Verwirrung gestiftet und Regressionsanalysen unnötig stigmatisiert. Insbesondere wenn sie für zwei Messparameter viele Datenpunkte haben, sollten sie nicht darauf verzichten, einen Korrelationskoeffizienten zu bestimmen. Bei Bedarf können Sie abhängige und unabhängige Variable auch vertauschen und so eine sinnvolle Interpretation und Einordnung in Ihre Hypothese ermöglichen. Günstiger weise ist der Korrelationskoeffizient trotz eines niedrigen Wertes oft statistisch signifikant. Durch Fokus auf diese Signifikanz, und nicht der in der Regel wenig beeindruckende niedrige Wert des Quadrats des Regressionskoeffizienten (‚Determinationskoeffizient') erhalten wir dann zusätzliche wertvolle Argumente im mechanistischen Narrativ. Das Einzeichnen der Regressionsgeraden in der graphischen Darstellung unterstützt die Konstruktion von Kausalzusammenhängen zusätzlich in visuell suggestiver Weise.

Zu guter Letzt noch ein Hinweis bezüglich des Abschnittes ‚Limitationen der Studie', eine Unsitte die sich aus der anglosächsischen Literatur immer weiterverbreitet und mittlerweile von vielen Journalen erwartet wird. Natürlich hat alles was wir tun Limitationen, so auch unsere Forschung. Auf diesen negativen Aspekten herumzureiten bringt außer einer Verwässerung der Schlüsselaussagen Ihrer Studie gar nichts. Sollte Sie dennoch genötigt werden, sich zu diesem Thema zu äußern, empfehle ich, einfach zwei oder drei triviale Limitationen zu listen. Allerdings sollten Sie diese so auswählen, dass sie von Ihnen in einem direkt daran anschließenden Satz einfach entkräftet werden können.

Ich vermute, dass ich Ihnen in meiner Auflistung nichts wirklich Neues bieten konnte, sind wir doch alle als Autoren und Reviewer in dieser über viele Jahrzehnte bewährten Publikationspraxis geschult. Ich hoffe aber, dass es mir gelungen ist, mit diesen nicht ganz ernst gemeinten Handreichungen, Ihnen einen Schrecken einzujagen, und eine Reflektion über diese Praxis auszulösen. In der Tat bewegt sich hier derzeit einiges. Viele Fördergeber, Journale, und Open Science Aktivisten versuchen, die methodische Qualität von Publikation zu erhöhen, deren Inhalte nachvollziehbarer und Daten sowie Code frei verfügbar zu machen, und auch solche Ergebnisse zu veröffentlichen, welche auf soliden Experimenten beruhen aber die Ausgangshypothese nicht bestätigen konnten (‚NULL-Resultate'). Unterstützen wir sie dabei!

Der Tag, an dem der Journal Impact Factor starb

LJ 10/2022

Viel ist über den Journal Impact Factor (JIF) geschrieben worden, auch auf diesen Seiten, und auch vom Narren. Kein gutes Haar wurde dabei an diesem Indikator gelassen, der schlicht misst wie oft die Artikel einer bestimmten Zeitschrift in anderen wissenschaftlichen Publikationen durchschnittlich pro Jahr zitiert werden. Der aber, weil ach so bequem und quantitativ, zur Leitwährung der akademischen Reputationsökonomie wurde. Eingeführt um Bibliothekaren Hilfestellung zu geben, welche Journale eine Subskription lohnen, bestimmt er heute die Karrieren von Wissenschaftlern und deren Anträgen. Unzählige Male wurde dieser Irrsinn angeprangert – selbst Clarivate, die Firma die mit dem Errechnen des JIF und dessen anschließendem Verkauf Milliarden verdient und seinen Aktionären stolz eine Bruttogewinnspanne von 64% berichtet, warnt mittlerweile auf ihren Webseiten davor. Genützt hat's nichts, der JIF feiert fröhliche Urstände.

Sehen Sie es mir deshalb nach, wenn ich das leidige Thema trotzdem nochmals aufwärme. Zum einen, weil ganz aktuell sogar die DFG (damit eigentlich wir deutsche Wissenschaftler!) den JIF und seinen Missbrauch nicht nur sehr korrekt analysiert, sondern auch in ungewöhnlich deutlicher Weise als Beurteilungskriterium verdammt hat. Und zwar im vor kurzem erschienenen, insgesamt sehr lesenswerten Positionspapier „Wissenschaftliches Publizieren als Grundlage und Gestaltungsfeld der Wissenschaftsbewertung" (Link hierzu wie zu allen anderen Quellen wie immer unter http://dirnagl.com/lj). So klar und deutlich kam das noch nie vom Lordsiegelbewahrer des akademischen Status quo. Außerdem lohnt es sich nochmals vom JIF zu sprechen, weil vor ein paar Wochen die aktuellen Werte, berechnet aus Zitationen 2021 auf Arbeiten die 2019/2020, veröffentlicht wurden. Und dabei stellte sich heraus, dass viele Journale ihren JIF über Nacht verdoppelt, ja sogar verdreifacht hatten. Es hatte eine regelrechte JIF Hyperinflation eingesetzt. Lancet stieg von 79 auf 203, das New England Journal of Medicine von 91 auf 176, Nature von 50 auf 70, usw. Plötzlich hatten 7 Journale zum ersten Mal einen JIF über 100! Euphorische Editoren von unzähligen Journalen ließen Sektkorken knallen, und twitterten ihr Glück in die Welt hinaus. Was den Wenigsten dabei auffiel: Wenn es noch irgendein Argument gebraucht hätte, um die komplette Untauglichkeit dieses Faktors in der Bewertung von Wissenschaftlern zu belegen, dann die Tatsache, dass er sich über nach verdreifachen kann. Oder auch wieder halbieren. Und dass das gar nichts mit der Wissenschaft der meisten Wissenschaftler, die sich über diesen Indikator messen zu tun hat. Der Tag der Veröffentlichung der JIFs 2022, also der 28.6.2022, wird (hoffentlich) als offizieller Todestag des JIF in die Geschickte eingehen.

Was war geschehen? Ganz einfach: COVID! In den letzten beiden Jahren war es zu einer ‚Covidization' der akademischen Forschung gekommen. Es wird geschätzt, dass mehr als 10 % aller Forschungsressourcen über Nacht in SARS-COV-2 Forschung flossen. Pubmed listet bei einer Suche mit dem Stichwort SARS-COV-2 über 275000 Artikel! Die

Superjournals publizierten die großen klinischen Studien und die Schlüsselarbeiten zu den Pathomechanismen. Diese Arbeiten wurden viele Tausend mal zitiert. Aber auch kleinere Journale konnten profitieren: Editoren luden Reviews ein, zu möglichen Zusammenhängen von SARS-COV-2 und dem Lieblingsorgan (Herz, Hirn, Lunge, etc.) des jeweiligen Journals, oder dessen namensgebender Erkrankung (Stroke, American Heart Journal, etc.). Und sogleich explodierten die Zitate.

Dass sich hierdurch die JIFs so massiv verändern können, liegt an einem der vielen mathematischen Webfehler des JIF. Der JIF ist ein Mittelwert, allerdings von einer total schiefen Verteilung. Bekanntermaßen erzielen wenige Arbeiten je Journale die überwiegende Zahl der Zitationen, und ein erklecklicher Anteil der Arbeiten wird überhaupt nie zitiert. So werden 20% der Artikel in Nature nie zitiert, und ungefähr 20 % der Papers sind für 80 % der Zitate verantwortlich. Zur Beschreibung solch schiefer Verteilungen müsste man eigentlich den Median verwenden, das steht in jedem Statistikbuch auf den ersten 3 Seiten. Schon lange wird deshalb argumentiert, dass man nicht den Mittelwert der Zitationen auf ein Journal verwenden sollte, sondern den Median, und dazu sollte auch gleich die Verteilung der Zitationen angeben werden. Wie dies zum Beispiel vorbildlich die Journale der EMBO Press tun. Das hätte nicht nur den Effekt, dass eine solche Darstellung stabil gegen Ausreißer ist. Die Unterschiede im JIF zwischen den Journalen würden auch drastisch nivelliert. Und das ist einer der wesentlichen Gründe warum Clarivate das nicht macht. Der andere Grund ist schlichtweg weil es einfacher ist, den JIF als Mittelwert mit nur drei Zahlen für jede Zeitschrift zu berechnen. Interessanterweise ist die Zitationsverteilung für jedes Journal wie auch der Median auf der (nur mit Subskription zugänglichen) Website von Clarivate verfügbar. Es interessiert sich nur keiner dafür.

Übrigens hat der JIF, genauso wie viele Studien in der Biomedizin, ein Reproduzierbarkeitsproblem. Wie in der Wissenschaft auch liegt das auch an mangelnder Transparenz. Wir kennen zwar die simple Formel für den JIF, aber wie die Zitationen genau berechnet werden, welche Arbeiten überhaupt gezählt werden, etc., das veröffentlicht Clarivate nicht. Die Zahlen kommen nämlich aus proprietären Datenbanken, die ebenfalls von Clarivate vermarktet werden. Deshalb gab es auch letztes Jahr schon einen erratischen Sprung beim JIF vieler Journale nach oben. Wieder freuten sich die Editoren! Bis ihnen klar wurde, dass es einfach daran lag, dass Clarivate das Jahr 2021 zum ‚Übergangsjahr' erklärte. Zitate aus Early-Access-Datensätzen ließ man fortan in den Zähler der JIF-Berechnung einfließen, schloß diese aber von der Anzahl der Veröffentlichungen im Nenner aus. So einfach ist es, den JIF zu manipulieren, wenn man an der Quelle sitzt.

Und weil wir schon dabei sind: Haben sie sich schon mal gefragt, warum Clarivate den JIF mit 3 Nachkommastellen Genauigkeit verkauft? Und wir Wissenschaftler mit unserer Schafsnatur das dann ohne nachzudenken genau so in unsere Lebensläufe übernehmen? Obwohl das überhaupt keinen Sinn macht? Zum einen wird der JIF durch diese Pseudogenauigkeit geadelt. Das muss schon ein wahnsinnig wissenschaftlich objektiver Wert sein, wenn man ihn auf 3 Kommastellen genau bestimmen kann! Aber der Hauptgrund warum Clarivate das so macht, ist dass nur so ein Ranking von Journalen möglich wird, und für solche Rankings verkaufen sie mit dem JIF das Substrat. Eine Rundung auf ganze Zahlen würde nur etwa 20 Ränge erlauben, denn der JIF der Mehrzahl der Journale liegt zwischen 0 und 20. Es gibt aber mehr als 50.000 wissenschaftliche Zeitschriften! Modellrechnungen zeigen, dass es sinnvoller wäre, den JIF auf die nächsten 5 oder 10 zu runden. Dies entspräche dann in etwa der Genauigkeit, mit der man voraussagen kann, wie viele Zitationen ein Artikel in einem bestimmen Journal aller Voraussicht nach haben wird. Und die wäre ja eigentlich der interessante Wert, den man als Wissenschaftler von einem Journal wissen wollte.

Aufgrund der diesjährigen Inflation des JIF wird sich jetzt manch ein Wissenschaftler an Deutschlands medizinischen Fakultäten die Hände reiben! Weil in fast allen medizinischen Unis die Formel, mit der die „leistungsorientierten Mittel“ (LOM) vergeben werden, als ein wesentliches Element den JIF enthalten. Was natürlich, wie auch im Positionspapier der DFG formuliert, völlig daneben ist. An der Charité bringt derzeit ein JIF Punkt ca. 150 €. Ein Lancet Paper ist damit nun über 30.000 € wert! Aber Vorsicht, öffnen Sie noch nicht den Schampus. Zum einen ist es wie bei der Inflation im wirklichen Leben: Die JIF Punkte werden entwertet, weil es mehr davon gibt, aber die LOM Summe gedeckelt ist. Das perfide dabei ist aber, dass es trotzdem zu einer LOM Umverteilung kommen wird. Wenn Sie das Pech haben, in hoch angesehenen Journalen zu veröffentlichen, die den Fehler gemacht haben, sich nicht mit SARS-COV-2 Artikeln zu schmücken, oder ihr Thema das nicht hergibt, sind jetzt die Gelackmeierten. Denn diese Journale, in meinem Feld zum Beispiel das Journal of Neuroscience, oder Brain Research, konnten ihren JIF kaum verbessern.

Wer jetzt noch nicht verstanden hat, dass der JIF nichts mit der Qualität einer spezifischen Publikation zu tun hat, die Zitationen auf einen bestimmten Artikel praktisch nicht mit dem JIF des Journals korrelieren, und er damit auch nicht für eine Vorhersage taugt, wie gut ein bestimmter Artikel in einem Journal zitiert werden wird, ist selber schuld. Und ist verdammt dazu, bis zu seiner Pensionierung im Fegefeuer der inadäquaten universitären Leistungsbewertung zu rösten. Alle anderen können nur hoffen, dass die wundersame Vermehrung der JIFs im Jahre 2022 diesem den lang herbeigesehnten Todesstoß gegeben hat.

Candide oder der Überoptimismus in der Nutzen- und Schaden-Rechnung klinischer Studien

LJ 11/2022

Haben Sie schon mal an einer klinischen Studie teilgenommen, vielleicht sogar in der Königsklasse, einer randomisiert kontrollierten klinischen Studie (RCT)? Wenn ja, warum haben sie da eigentlich mitgemacht? Wie haben Sie den Einschluss in die Studie erlebt? Hat man Ihnen nach Abschluss mitgeteilt, ob Sie im Placebo oder Verum Arm waren? Haben Sie erfahren, was eigentlich rausgekommen ist, welcher Erkenntnisgewinn unter Ihrer Mithilfe entstanden ist? Und wer eigentlich davon profitiert hat, z.B. eine zukünftige Generation von Patienten mit derselben Diagnose, oder nur die Pharmaindustrie?

Seit ihrer Entwicklung in den 50er Jahren des letzten Jahrhunderts bilden RCTs das Fundament des Wirksamkeitsnachweises von neuen Therapien in der modernen Medizin. Weil die Geschichte gelehrt hat, dass Patienten vor unethischen medizinischen Versuchen und Studien geschützt werden müssen, soll durch die Befolgung von forschungsethischen Prinzipien wie der Erklärung von Helsinki und des Belmont Reports sichergestellt werden, dass sich bei klinischen Prüfungen

möglicher Nutzen und Schaden für die Studienteilnehmer die Waage halten. Dies schließt, besonders in der Frühphase klinischer Prüfung, die sogenannte ‚Equipoise' ein, also die Unsicherheit der informierten medizinischen Fachwelt darüber, welche von zwei oder mehr möglichen Therapien die bessere ist. Wenn Experten vermuten, dass der Schaden überwiegt, sollte man die Finger von einer Überprüfung lassen, wenn diese überwiegend Nutzen vermuten und diesen begründen können, wäre es unethisch den Patienten diese Therapie im Rahmen der Randomisierung vorzuenthalten. Soweit so gut.

Die medizinischen Experten, welche eine Studie planen, müssen also in der Lage sein, eine Equipoise festzustellen, also möglichen Nutzen und Schaden der zu prüfenden Therapie mit einer gewissen Treffsicherheit vorhersagen zu können. Aber wie gut sind Experten, also mit der Materie vertraute Mediziner, egal ob sie an einer Studie selbst beteiligt sind oder nur deren Erfolg vorhersagen sollen, darin wirklich? Beunruhigender Weise haben alle Studien, welche dies untersucht haben gezeigt, dass dies Experten nicht oder unwesentlich besser tun als ein Zufallsgenerator! Hierfür kann es viele Gründe geben. Einer davon könnte sein, dass Kliniker in der frühen klinischen Prüfung (Phase I/II) die Aussagekraft von tierexperimentellen Studien, welche die Grundlage für diese Studien darstellen, überschätzen. Oder auch die Qualität und Robustheit der präklinischen Ergebnisse überbewerten. Dazu kommen aber auch eigene ‚Biase'. Dieses Wunschdenken wird genährt von der Begeisterung für eine neue Therapie, die man vielleicht sogar selbst mitentwickelt hat, und dem tief empfundenen Wunsch, mit den Ergebnissen ihre Patienten zukünftig besser behandeln zu können.

Die schlechte und überoptimistische Einschätzung der Nutzen/Risikoprofile zu prüfender Medikamente durch Experten sollte uns aber nachdenklich machen. Denn die Ethikkommissionen, welche die ethische Unbedenklichkeit von Forschungsvorhaben am Menschen prüfen, werden ja von eben diesen potentiell ‚überoptimistischen' und schlecht schätzenden Individuen informiert. Zudem muss man sich dann fragen, ob ‚Equipoise' auf der Strecke bleibt, wenn die Studiendurchführenden so großen Optimismus an den Tag legen.

Nun wird die Sache aber insbesondere dort noch verzwickter, wo über viele Jahre zwar Hunderte von klinischen Studien durchgeführt wurden, aber nur wenige oder gar keine davon erfolgreich waren. Das ist zum Beispiel bei der Suche nach neuen Schlaganfalltherapien der Fall, aber auch bei der Alzheimer'schen Erkrankung. Die bisherige Erfolglosigkeit der klinischen Prüfungen in diesen Feldern erniedrigt ja massiv die (Vortest-)Wahrscheinlichkeit, mit den nächsten Studien Erfolg zu haben. Wir dürfen dabei auch nicht vergessen, dass viele der zu prüfenden Therapien Nebenwirkungen haben, welche sehr schwer sein können. Daher muss man befürchten – und das sagt nicht nur der Narr (Literatur wie immer unter http://dirnagl.com/lj), dass Studienpatienten bei solch ‚schwierigen' Diagnosen möglicherweise besser dran sind, wenn sie in den Placebo-Arm randomisiert werden. Dort bleiben sie wenigstens von den Nebenwirkungen verschont.

Wenn aber schon erfahrene Kliniker Schwierigkeiten bei der Risiko/Nutzenabschätzung von Prüftherapien haben, wie sieht das bei den Patienten aus? Sie werden beim Studieneinschluss im Aufklärungsgespräch ja von den Ärzten informiert, die möglicherweise selbst schlecht abschneiden bei der Einschätzung von Nutzen und Risiko der Studienmedikation. Wer selbst schon mal an einer Studie teilgenommen hat, und die Patienteninformation und Einwilligungserklärung inklusive Datenschutzerklärung unterzeichnet hat, welche locker über 20 Seiten und länger sein kann, wird sich außerdem eines weiteren Problems bewusst sein. Selbst bei bestem Wissen, Willen und didaktischer Fähigkeit des einschließenden Arztes dürfte es den wenigsten Studienteilnehmern so richtig klar

geworden sein, worauf sie sich da einlassen. Warum sie es trotzdem machen? Weil sie der Einschätzung und dem Urteil des Arztes vertrauen! Und hier zeigt es sich dann wieder, wie sehr doch alles darauf beruht, dass der Studienarzt Nutzen und Risiken zuverlässig abschätzen kann.

Umso dramatischer ist es deshalb auch, dass ein nicht unerheblicher Teil der klinischen Studien nach Abschluss nie Ergebnisse publiziert werden, der Narr hat das auf diesen Seiten schon mehrfach angeprangert (zuletzt LJ 4/2022). Erst kürzlich hat eine Studie wieder gezeigt, dass etwa ein Drittel der abgeschlossenen klinischen Studien in Deutschland nicht publiziert werden. Oder ein aktuelles Beispiel aus der COVID-Behandlung: Die Ergebnisse der meisten klinischen Prüfungen des Covid-Medikaments Molnupiravir (Lagrevio) wurden nicht veröffentlicht und sind unzugänglich. Trotzdem sind mit dem Medikament bisher 3,2 Milliarden Dollar umgesetzt worden. Die Patienten werden also schlimmstenfalls ‚doppelt' betrogen: Ihr Altruismus zugunsten anderer Patienten wird durch Nichtpublikation zunichte gemacht.

Wenn es dann noch um eine schwerwiegende Erkrankung und eine akut zu fällende Entscheidung des Patienten geht, ob er oder sie an einer Studie teilnehmen soll, wird es noch viel problematischer. In einer noch unveröffentlichten Studie haben Kollegen von mir untersucht, ob sich Patienten mit akutem Schlaganfall, die vor dem Erhalt einer Standardtherapie bei dieser Erkrankung (intravenöse Thrombolyse) nach allen Regeln der Kunst aufgeklärt wurden, nach 60 – 90 Minuten noch an wichtige Details, wie zum Beispiel mögliche schwerwiegende Nebenwirkungen der Therapie erinnern können. Das Ergebnis war wenig überraschend: Nur eine Minderheit konnte das. Wer in einer dramatischen Lebensphase medizinischen Vorträgen zuhören und diese auch noch verstehen muss, wird notwendig Konzentrationsprobleme sowie Verständnis- und Erinnerungslücken entwickeln.

Vermutlich ist den meisten Studienpatienten klar, dass sie allein schon wegen der Randomisierung keine Garantie haben, von der Studie persönlich zu profitieren. Denn es wird immer betont, dass die Wahrscheinlichkeit das Studienmedikament zu erhalten, nur 50% ist. Die meisten Patienten machen dennoch mit, sie handeln damit bewusst altruistisch. Sie setzen auf einen möglichen Nutzen, den dann möglicherweise erst nachfolgende Patienten mit derselben Erkrankung genießen werden. Es stellt sich aber die Frage, ob Patienten wirklich die Risiken einer Studie einschätzen können und insbesondere auch die Last, welche möglicherweise auf sie bei Studienteilnahme zukommt. Insbesondere bei Studien mit Krebspatienten in fortgeschrittenen Stadien wird mittlerweile die ‚time toxicity' der Studienteilnahme problematisiert: Patienten in vielen Studien verbringen unter Umständen relevante Lebenszeit bei Klinikaufenthalten im Rahmen von Studien, die man ihnen eigentlich in häuslicher Umgebung wünscht.

Patienten erfahren in der Aufklärung bezüglich ihres persönlichen Nutzens von der Studienteilnahme immer nur: ‚Es ist möglich, dass Sie einen Nutzen aus der Behandlung ziehen oder auch nicht'. Sollte man ihnen aber nicht auch das mitteilen, was die Kliniker bei der Fallzahlberechnung veranschlagt haben? Von persönlichem Nutzen für den Patienten steht da gar nichts. Sondern vielmehr das, worum es eigentlich in der Studie geht, in Form des primären Endpunktes sowie den Unsicherheiten seiner Bestimmung. Damit wird sowohl das wissenschaftliche Prinzip hinter solchen Studien, als auch der mögliche gesellschaftliche Nutzen, der ja die Basis für den Altruismus der Patienten bildet klarer benannt. Dann wären Patienten besser informiert für ihre persönliche Abwägung. Wie man sieht, geht es immer wieder um das Verhältnis von Nutzen und Risiko, und wie gut Ärzte und Patienten in dieser Abwägung sind.

Und noch etwas muss bei all diesen Überlegungen mit bedacht werden. Wie ‚unabhängig', wie ‚frei' von Konflikten ist die Rekrutierung von Patienten in klinische Studien wirklich? Pharmafirmen erstatten den Studien-durchführenden Kliniken die Kosten, das ist nur fair. Schließlich wird Infrastruktur, Personal, Expertise etc. von den Kliniken dafür vorgehalten. Allerdings kompetitieren in vielen Bereichen die Pharmafirmen um die knappe Ressource Patient. Ein Beispiel aus der Medikamentenentwicklung für Patienten mit akuter Querschnittslähmung: Eine Pharmafirma bezahlte den an der Studie beteiligten Kliniken 18.000 € pro eingeschlossenen Patienten. Eine aufwendige Studie, aber ist das nicht doch ein bisschen viel für die Verabreichung eines Medikamentes und einiger zusätzlicher Untersuchungen? Das tolle Angebot der Firma hat dazu geführt, dass Studien zur Prüfung anderer, vielleicht ebenso vielversprechender Therapien bei Querschnittslähmung entweder nicht weiter rekrutieren konnten, oder erst gar nicht gestartet wurden. Um es unverblümt zu sagen: Diese Pharmafirma hatte alle verfügbaren Patienten aufgekauft!

An diesem ‚Handel' mit Patienten ist aber auch noch ein weiterer Haken dran: Studien einzuwerben lohnt sich (sofern die Pharmafirmen gut bezahlen) für die Kliniken. Das bringt nicht nur Prestige durch die schiere Beteiligung an wichtigen Studien, es fällt auch oft noch eine Ko-Autorschaft für den Chef und ein paar Mitarbeiter ab. Aber es wird auch ein Überschuss erwirtschaftet. Von dem kaufen sich die Chefärzte (in der Regel) aber keinen neuen S-Klasse Mercedes oder Tesla, sondern finanzieren in ihren Kliniken durchgeführte eigene Forschung, welche meist unterfinanziert ist. Da kann sehr viel sehr Gutes dabei rauskommen, aber es sollte einem schon ein wenig mulmig werden bei diesen Finanzierungsströmen. Denn das heißt ja auch, dass sich neben das Primärinteresse, mit einer klinischen Studie neue und wirksame Therapien zu etablieren, ein Sekundärinteresse gesellt: Durch den Studieneinschluss andere Forschung zu ermöglichen – für mich ein klassischer Interessenkonflikt. Wer aber hier den Stöpsel ziehen wollte, würde die akademische Forschung an Deutschlands Unikliniken in arge Bedrängnis bringen, und klinische Studien der Pharmaindustrie gäbe es dann wohl auch keine mehr. Zumindest nicht im gegenwärtigen Modell der Forschungsfinanzierung im Gesundheitswesen.

Sollten wir also auf klinische Prüfungen vollständig verzichten? Weil Experten überoptimistisch sind, und ihnen nicht bewusst ist, wie häufig lückenhaft und wenig robust die präklinische Evidenz für ihre Studien sind? Weil deshalb so mancher Studie keine Equipoise zu Grunde liegt? Weil Patienten in Studien möglicherweise Vorteile haben, wenn sie in die Placebogruppe randomisiert werden? Weil sie die Aufklärung nur partiell verstehen und sich wenig davon merken können? Weil durch Interessenkonflikte bei den Studien-Ärzten und Kliniken Patienten auf ‚Teufel komm raus' rekrutiert werden könnten?

Natürlich brauchen wir randomisiert kontrollierte Studien! Sie allein garantieren die Evidenzqualität welche für die Zulassung von Medikamenten nötig ist. Aber es muss sich einiges ändern. Bei der Entscheidung zu einer klinischen Entwicklung muss die (häufig mangelhafte) Qualität und Validität der präklinischen Befunde stärker berücksichtigt werden. Planer von klinischen Studien und Studienärzte müssen sich ihrer eigenen Biase bewusster werden. Beides wird ihre Fähigkeit hin zu einer realistischeren Nutzen/Risiko Analyse für die Patienten verbessern. Die Aufklärung von Patienten, insbesondere unter Akutbedingungen und bei schwerwiegenden Diagnosen wird problematisch bleiben, aber wir dürfen den Altruismus der Patienten nicht hintergehen. Wir müssen dafür Sorge tragen, dass alle Studienergebnisse zeitnah veröffentlicht werden. Auch wenn das Prüfmedikament nicht besser war als Standardtherapie muss sichergestellt sein, dass Studien so angelegt werden, dass sie zum medizinischen Erkenntnisgewinn beigetragen können. Die Studienteilnehmer haben hierfür schließlich Belastungen und Risiken auf

sich genommen. Auch wurden für die Studie Ressourcen eingesetzt, die wir letztlich alle gemeinsam schultern müssen, sei es über Steuern, Krankenversicherung, oder Arzneimittelpreise. Wir sind den Patienten auch Rechenschaft schuldig über das, was rausgekommen ist bei der Studie, welcher Erkenntnisgewinn sich ergeben hat. Auch müssen die potentiellen und durchaus nicht trivialen Interessenkonflikte bei der Patientenrekrutierung offengelegt bzw. dort wo nicht mehr vertretbar aufgelöst werden. Noch besser wäre es natürlich, die Finanzierung der klinischen Forschung insgesamt zu verbessern damit auf verdeckte Finanzierungen ganz verzichtet werden kann. Alternativen, wie z.B. einen Teil der Industriemittel für Studien in einen Pool der Universitäten oder Länder für die Förderung klinischer Forschung fließen zu lassen, könnte man ja mal in Erwägung ziehen.

Wenn all dies gegeben wäre, würde sich auch der Wissenschaftsnarr bedingungslos in jede klinische Studie rekrutieren lassen.

Der Wissenschaftsnarr dankt Prof.Dr. Jonathan Kimmelman und Prof.Dr. Daniel Strech für anregende Diskussionen zu den ethischen Aspekten klinischer Studien.

‚Spät kommt Ihr – doch Ihr kommt! Der weite Weg, DFG, entschuldigt Euer Säumen.‘

LJ 12/2022

Wann haben sie eigentlich das letzte mal ein Positionspapier oder eine Denkschrift der DFG gelesen? Vermutlich noch nie, und dabei wären Sie in guter Gesellschaft. Für die meisten Wissenschaftler ist die DFG einfach nur der wichtigste und prestigereichste Forschungsförderer. Man stellt Anträge, wird abgelehnt, oder eben auch gefördert. Aber neben ihrem Fördergeschäft äußert sich die DFG auch regelmäßig zu wissenschaftspolitischen Themen. Das hat sie vor kurzem erst wieder getan, in dem sie zwei sogenannte ‚Positionspapiere‘ herausgebracht hat. Eins mit dem etwas sperrigen Titel ‚Wissenschaftliches Publizieren als Grundlage und Gestaltungsfeld der Wissenschaftsbewertung - Herausforderungen und Handlungsfelder‘. Das andere bekennt schon im Titel Farbe: ‚Open Science als Teil der Wissenschaftskultur‘. Vielleicht hat es mir den Kopf verdreht, dass ich an dem Papier zum Publizieren mitarbeiten durfte, aber ich denke zumindest für die DFG handelt es sich dabei um geradezu revolutionäre Schriftstücke. Und weil Sie beide Papiere vermutlich nicht gelesen haben, erlaube ich mir Ihnen kurz darzulegen, warum der sonst so kritische Narr plötzlich so enthusiastisch ist.

Vorweg aber erstmal zur Einordnung ein paar Gedanken zur eigentümlichen und weltweit durchaus einmaligen Verfasstheit der DFG. Trotzdem sich die DFG aus Steuergeldern finanziert (70% Bund, 30% Länder), ist sie ein eingetragener Verein! Wie jeder

ordentliche Verein hat sie Mitglieder. Trotzdem die DFG sich selbst als die ‚Selbstverwaltungsorganisation der Wissenschaft in Deutschland' bezeichnet, sollten Sie sich die Mühe sparen, einen Mitgliedsantrag zu stellen! Denn die derzeit 97 Mitglieder der Deutschen Forschungsgemeinschaft sind Hochschulen, andere Forschungseinrichtungen, Akademien der Wissenschaften sowie wissenschaftliche Verbände. Wie in einem Kaninchenzüchterverein gibt es in der DFG natürlich auch eine Menge von Organen und Posten. Mitgliederversammlung, Präsidentin, Präsidium, Vorstand, Generalsekretärin, Senat mit einer Busladung Senatoren, Hauptausschuss, Kommissionen und nicht zu vergessen die Fachkollegien. 3,3 Milliarden € jährlich wollen schließlich geordnet unter die Leute gebracht werden. Dafür hat die DFG auf administrativer Ebene auch eine recht schlanke und kompetente Administration. Wer hat nicht schon mal bei der Sachbearbeiterin nachgefragt, wie es um den eigenen Antrag steht, oder was man tun kann, wenn er abgelehnt wurde? Aber wo bleibt da eigentlich die ‚Selbstverwaltung'? Diese begründet die DFG damit, dass ja wissenschaftliche Organisationen wie die Universitäten Vereinsmitglieder sind, und somit an den Entscheidungen teilhaben. Zum anderen sitzen im Senat, den Kommissionen und den Fachkollegien lauter Wissenschaftler, auch wenn viele von denen häufiger in Gremiensitzungen als im Labor oder beim Paper-Schreiben anzutreffen sind. Immerhin könnten sie, sofern Sie promoviert sind und an einer anerkannten wissenschaftlichen Einrichtung in Deutschland tätig sind, durch Ihre Stimme alle vier Jahre per Wahl einen Kollegen oder eine Kollegin in eines der 46 Fachkollegien entsenden.

All das ist für nationale Großförderer einmalig. Die National Institutes of Health sind dem amerikanischen Kongress verpflichtet, der Wellcome Trust sich selbst, die Wissenschaftsministerien fördern im direkten Staatsauftrag, usw. Wenn wir also von ‚DER' DFG reden, oder auch über sie schimpfen, weil wir wieder mal nicht gefördert wurden, reden wir dann eigentlich nicht von uns selbst? Das kommt sehr drauf an. Sollten Sie arrivierter Wissenschaftler sein, im Idealfall Lehrstuhlinhaber oder Institutsdirektor, dann gilt dies tatsächlich. Denn dann sind potentiell Sie es, der in den Gremien, Kommissionen und Kollegien sitzt und (Richtungs-) Entscheidungen fällt. Als gewöhnlicher Wissenschaftler, insbesondere wenn Sie noch nicht durch eine Habilitation geadelt wurden, oder gar abseits des Mainstreams forschen, dürfen sie zwar die arrivierten Wissenschaftler wählen, welche von den Hochschulen und wissenschaftliche Fachgesellschaften zur Wahl vorgeschlagen werden. Auch dürfen Sie, sollten Sie ein bestimmtes Maß an Reputation erreicht haben und nicht sonst wie unangenehm aufgefallen sind, als Fachgutachter Förderanträge an die DFG begutachten. Die Förderentscheidungen treffen aber die Vierzehnender! All dies führt dazu, dass die DFG ein ultrakonservativer Organismus ist. Die im System Erfolgreichen geben die Richtung vor, und sorgen dafür, dass ihnen genehme und interessant erscheinende Forschung gefördert wird. Und sie selbst und ihre Adepten dabei natürlich nicht zu kurz kommen.

Aber ist das nicht gut so? Gerade die Erfolgreichen haben doch an sich selbst bewiesen, was gute Forschung ist und wie man sie macht. Da brennt doch nichts an? Die Schattenseite eines solchen Systems ist aber, dass eine so verfasste Organisation notwendig dem Mainstream verpflichtet ist und eine unglaubliche Trägheit entwickelt, die sie reformunfähig macht: Warum sollten ausgerecht diejenigen, die es im System zu etwas gebracht haben, irgend etwas daran ändern wollen? Dazu kommt eine Intransparenz in der Begutachtungspraxis, welcher Juristen bescheinigt haben, dass sie rechtsstaatlichen Anforderungen nicht genügt, da Entscheidungen nicht ausreichend begründet werden und keine Widerspruchsmöglichkeiten für Abgewiesene bestehen.

Und dies bringt mich zurück zu den eingangs erwähnten Positionspapieren. Denn in ihnen zeigt sich die DFG überraschend system- und damit selbstkritisch, und entwickelt

darin eine Programmatik des Systemwandels. Kurz gesagt: Vom Kern her geht es der DFG darum, die Wissenschaft transparenter und offener (‚Open Science') zu machen, und die Bewertung von Forschern und deren Produkten, also Papers, Anträge, etc., zu reformieren. UNESCO, EU, LERU, sind mit Manifesten und Deklarationen vorangegangen, manche Länder (z.B. Niederlande, Frankreich) setzen hierzu bereits nationale Pläne um. Dennoch, mit Schiller möchte man der DFG zurufen: ‚Spät kommt Ihr – doch Ihr kommt! Der weite Weg entschuldigt Euer Säumen.'

Besonders interessant ist das Positionspapier zum wissenschaftlichen Publizieren. Wieso sollte sich die DFG hierzu äußern, das ist doch nun wirklich Sache jeden Wissenschaftlers? Beim Studium des Papiers wird jedoch schnell klar, dass es in Wirklichkeit um viel mehr, eigentlich um Alles geht - das Papier ist ein veritables trojanisches Pferd! Wissenschaftliches Publizieren dient der Bekanntmachung von Ergebnissen, deren Qualitätssicherung und Dokumentation, sowie der Sicherung von Urheberschaft. In Letzterem wiederum versteckt sich aber die Zuschreibung von Reputation. Und ‚Reputation' ist im wissenschaftlichen System entscheidend fürs Fortkommen – das Stipendium, die Stellenzusage, die Professur, der Antragserfolg, und so weiter, alle hängen davon ab. Das DFG-Papier bietet eine gründliche Propädeutik, Historie und Kritik davon, wie die Zuschreibung von Reputation über Publikationen in den letzten Jahrzehnten sich zu einem System entwickelt hat, bei dem es bei der Beurteilung von Forschern zunehmend weniger auf die Inhalte und den wissenschaftlichen oder gesellschaftlichen Impact von deren Ergebnissen ankommt, als vielmehr auf die Reputation des Journals in dem veröffentlicht wurde. Oder gar nur auf karge Zahlen wie dem Journal Impact Factor (JIF), dem Hirsch-Faktor, oder die schiere Anzahl von Publikationen.

Das Papier kommt nach diesem theoretischen Teil zu wesentlich eindeutigeren Schlüssen und Handlungsanweisung als diese bisher von der DFG zu hören waren. Und die DFG kehrte auch gleich vor der eigenen Haustüre, und änderte z.B. das Formular für den Lebenslauf sowie den Leitfaden für Projektanträge. Damit fördert sie eine inhaltlich-qualitativ fundierte und den jeweiligen Lebens- und Karriereabschnitt stärker berücksichtigende Bewertung wissenschaftlicher Leistung. Antragsteller können jetzt auch das gesamte Spektrum wissenschaftlicher Publikationsformen gleichwertig in Förderanträgen und Lebensläufen abbilden, also zum Beispiel Artikel auf Preprint-Servern, Datensätze oder Softwarepakete. JIF und ähnliche ungeeignete Metriken werden gleich ganz verbannt. Damit löst die DFG auch ein, wozu sie mit der Unterzeichnung der San Francisco Declaration on Research Assessment (DORA) Declaration verpflichtet hat.

Im kurz darauf veröffentlichten und noch druckfrischen Positionspapier ‚Open Science als Teil der Wissenschaftskultur' bekennt sich die DFG meiner Ansicht nach klar zu Open Science. Dabei hebt sie insbesondere die Verbesserung von Forschungsprozessen, erhöhte Transparenz, gleichberechtigten Zugang zu wissenschaftlicher Information, Stärkung der wissenschaftlichen Zusammenarbeit, und die Erleichterung von Innovationen durch die offene Wissenschaft hervor. Weil sich das Dokument aber über weite Strecken mit ‚Herausforderungen' befasst, und zu einer ‚differenzierten Betrachtung' aufruft, wurde die Positionierung der DFG bereits vielfach als halbherzig kritisiert und als Lippenbekenntnis abgetan.

In der Tat finden sich im Papier eine Vielzahl von Statements wie z.B. ‚dass Reformziele in Kulturwandelprozessen der Wissenschaft nicht als Selbstzweck behandelt werden sollten', Open Science kein ‚Heilsversprechen oder Ideologie' sein dürfe, ‚Open Science per se kein Garant für Forschung von höherer Qualität' sei, die ‚völlige und unregulierte Transparenz aller Prozesse und Daten' potentiell schädlich, oder dass ‚wo Open Science ausschließlich als Vorgabe politischer Zielsetzungen erscheint, sich Effekte ergeben

könnten, die zu wissenschaftsinadäquaten Entwicklungen führen können.‘ Aber das ist doch wohl selbstverständlich, dass wo etwas Selbstzweck, Ideologie, total unreguliert, oder reine politische Vorgabe wird, kein Blumentopf mehr zu gewinnen ist! Aber durch die Einstreuung solcher Plattitüden musste man wohl Zweifler und Bremser, die es in der wissenschaftlichen Community und damit auch der DFG immer noch gibt, ruhigstellen.

Im Übrigen gibt es ja tatsächlich eine Reihe von, um im DFG-Speak zu bleiben, ‚Herausforderungen‘. Dazu zählt die Qualitätskontrolle (aber auch nicht mehr als bei ‚closed science‘!), mögliche infrastrukturelle Abhängigkeiten und Effekte der Kommerzialisierung von Open Science (man denke an die horrenden Open Access Gebühren), und die Tatsache, dass in vielen Bereichen der offenen Wissenschaft Infrastrukturen und Kompetenzen fehlen, die von den Institutionen vorgehalten werden und auch von den Fördergebern finanziert werden müssten. Aber all das steht auch in dem Papier.

Aber was ist nun der Nährwert solcher Positionspapiere? Wo sie doch die meisten von uns, die wir durch die DFG selbstverwaltet werden, gar nicht lesen? Ich denke, dass man deren Wirkung nicht unterschätzen darf. Zum einen, weil sich die DFG damit selbst zur Reform verpflichtet – erste Schritte wurden bereits eingeleitet. Zum anderen, weil die Mitglieder der DFG damit auf die Ziele ihres Vereins eingeschworen werden. Schließlich ist man Vereinsmitglied. Der Verweis auf Statements der DFG kann in universitären Gremien wahre Wunder bewirken.

Etablierten Wissenschaftlern scheint es wie dem Mädchen Goldlöckchen im gleichnamigen Märchen im Haus der drei Bären zu gehen. Goldlöckchen ist der Brei immer ‚nicht zu heiß, nicht zu kalt, sondern genau richtig‘. Sobald man es in Academia geschafft hat, arbeitet man gefühlt im besten aller Systeme, und will auf keinen Fall mehr was daran verändern. Dankbar müssen wir deshalb sein, dass es in der Geschäftsstelle der DFG eine Menge von für uns ‚namenlose ‘ kreative und fortschrittliche Köpfe gibt, die deutlich mehr machen als Förderanträge bearbeiten und Begutachtungen organisieren. Behutsam antagonisieren sie die Inertia der etablierten Wissenschaftler in den DFG-Gremien und haben damit einen wesentlichen Anteil daran, die DFG und damit auch das akademische System in Deutschland zu reformieren.

Wissenschaftsbetrug ist selten. Aber stimmt das eigentlich?

LJ 1-2/2023

Fast wöchentlich lesen wir von Fällen wissenschaftlichen Fehlverhaltens. Häufig spielen darin renommierte Journale und prominente Wissenschaftler eine Rolle. Die Website Retraction Watch von Ivan Oranski und Adam Marcus versorgt uns in einem unablässigen Strom mit solchen Nachrichten und deren Hintergründen. Auch im Laborjournal findet sich fast in jeder Ausgabe eine Story über ein Labor in dem es nicht mit rechten Dingen zuging. Meist wurde das ruchbar, nachdem ein Artikel mit manipulierten, gefälschten oder gar erfundenen Daten aufgeflogen war. Ans Licht der wissenschaftlichen Öffentlichkeit bringen dies oft Whistleblower, oder aufmerksame Leser, die ihre Zweifel über die Dignität von Abbildungen anonym auf PubPeer veröffentlichen. Auffällig selten decken dagegen Universitäten, Fördergeber oder Journale solche malignen Machenschaften auf.

Auch die Wissenschaftsseiten der Tageszeitungen versorgen uns recht häufig mit Nachrichten aus den moralischen Niederungen der Wissenschaft. Aktuell mit Berichten über fragwürdige Papers aus den Ställen des Nobelpreisträgers Gregg Semenza, dem ‚Entdecker' des Hypoxie-induzierten Faktors (HIF), oder des Neurowissenschaftlers und derzeitigen Präsidenten der Stanford Universität Marc Tessier-Lavigne. Dabei geht es nicht nur um Arbeiten aus der Grundlagenforschung: COVID hat uns mit einen wahren Tsunami von Artikel-Retraktionen beschert, allen voran die Lancet und New England Journal Arbeiten welche auf komplett erfundenen Daten beruhten.

Berichte über wissenschaftliches Fehlverhalten goutieren wir oft mit gar wohligem Gruseln, groß aufregen tun wir uns darüber allerdings nicht. Wir haken es unter der Rubrik ‚Jede Branche hat ihre schwarzen Schafe' ab, ist das doch menschlich, allzu menschlich. Oder tun es – zum Beispiel, wenn es um Massenretraktionen von Artikeln aus sogenannten ‚Papermills' geht (also übers Internet bestellten und gegen eine Gebühr verfassten, komplett erfundenen Artikeln) als exotische Phänomene des Wissenschaftsbetriebes in uns fernen Weltgegenden ab. Bei uns spielt Wissenschaftsbetrug, also das Plagiieren sowie Falsifizieren oder Fabrizieren von Daten doch keine wichtige Rolle. Plagiarismus vielleicht, allerdings eher in nicht-naturwissenschaftlichen Fächern – solch Abschreiben ist war unschön, aber doch ein eher geringfügiges Vergehen. In ihren Pressemitteilungen listet die DFG für das Jahr 2022 ganze 6 Fälle, in denen Wissenschaftler wegen wissenschaftlichem Fehlverhalten gerügt und für ein paar Jahre von der Förderung ausgeschlossen wurden. Zeigt das nicht wie selten Betrügereien in unserem Wissenschaftsbetrieb sind?

Auch der Narr hat sich bis vor kurzem dieser bequemen Illusion hingegeben. Und geglaubt, dass es fast exklusiv sogenannte ‚fragwürdige wissenschaftliche Praktiken' sind, welche unsere Aufmerksamkeit verdienen. Also das Weglassen von Befunden, welche eine Story nicht mehr so ganz glatt erscheinen lassen. Oder die Durchführung multipler Tests, bis man auf einen stößt, der die ersehnte statistische Signifikanz ergibt, auch als

p-Hacking bekannt und beliebt. Oder das Formulieren von Hypothesen, nachdem man die Ergebnisse bereits kennt – aber so tut als wären die Versuche durchgeführt worden, um genau diese Hypothesen zu testen (‚HARKING – Hypothesizing after the results are known'). Aber das Bearbeiten von Banden auf Western-Blots mit Photoshop? Oder das Manipulieren von Ergebnissen in Spreadsheets? Die Verwendung von Kontrollergebnissen, die gar nicht zum aktuellen Experiment gehörten? Nicht bei uns, und auch nicht im Nachbarlabor!

Aber warum sind wir uns da eigentlich so sicher? Es spricht nämlich sehr viel dafür, dass wissenschaftliches Fehlverhalten weit häufiger ist, als wir uns das eingestehen. Eine aktuelle, aufwendige und methodisch exzellente Arbeit ergab, dass 8 % von etwa 7000 niederländischen Wissenschaftlern, die auf eine anonymen Umfrage zu Forschungspraktiken geantwortet hatten, zwischen 2017 und 2020 mindestens einmal Daten gefälscht und/oder erfunden hatten! In Medizin und Biowissenschaften waren es sogar über 10%. Mehr als Hälfte gab außerdem zu, häufig (!) fragwürdige Wissenschaftspraktiken anzuwenden. Haben holländische Wissenschaftler etwa mehr kriminelle Energie als die deutschen? Schon vermutlich allein deshalb nicht, weil die Niederlande 2012 einen wissenschaftlichen Betrugsskandal erlebten, der die gesamte Nation bis ins Mark erschütterte. Was weitreichende Konsequenzen im holländischen Wissenschaftssystem zur Folge hatte. Zum Beispiel einem nationalen Plan unter Beteiligung der Universitäten und Fördergeber mit dem Ziel, offene Wissenschaft zu fördern (‚Open Science'). Und eine Reform des akademischen Karriere- und Belohnungssystems (‚Every talent counts') in Gang setzte. Um beides beneiden wir mittlerweile unsere Nachbarn.

So schockierend die Ergebnisse der niederländischen Umfrage sind, so sehr passen sie doch ins Bild. Mittlerweile können wissenschaftliche Abbildungen automatisiert auf Manipulationen untersucht werden. Die Anwendung solcher Techniken zeigt, dass mehr als 4 % aller biomedizinischen Publikationen Graphen und Abbildungen enthalten, welche hochgradig suggestiv für maligne Manipulationen sind. Etwa die Verschiebung von Banden, Duplikationen, nicht plausible Fehlerbalken, usw. Diese Zahlen werden auch durch Arbeiten bestätigt, in denen Menschen die Abbildungen untersuchten. Gleichzeitig hat ein Wettlauf begonnen zwischen Software, welche kaum noch zu erkennende ‚deep fakes' von wissenschaftlichen Graphiken erzeugen kann, und Software, welche in der Lage ist, genau diese zu erkennen. Die steigende Zahl von Retraktionen, sowie die vermehrten Berichte über nachgewiesen Wissenschaftsbetrug kann uns ja immer nur die Spitze des Eisbergs von tatsächlich stattgehabtem Fehlverhalten anzeigen, vermutlich mit einen Bias in Richtung der krasseren Verstöße. Hieraus auf die Größe des Problems, also die Gesamtmasse des Eisbergs zu schließen, ist nicht möglich. Aber eins ist klar – dieser muss viel größer sein als das, was sichtbar aus dem Wasser ragt: Handelt es sich doch um sanktioniertes, wenn nicht justiziables Verhalten. Deshalb dürften vermutlich Umfragen, wie die erwähnte niederländische, ebenfalls zu niedrige Prävalenzen von Verstößen aufzeigen.

Aufgeschreckt und verunsichert durch die Zahlen aus den Niederlanden hat sich der Narr in der Literatur umgetan (wie immer unter http://dirnagl.com/lj zu finden) und fand überraschend großen viele Belege – z.B. Umfrage-Ergebnisse, Stichproben, systematische Reviews, die in ihrer Totalität nur einen Schluss zulassen: Wissenschaftliches Fehlverhalten jenseits von HARKING und p-Hacking, also Plagiarismus, Falsifikation und Fabrikation von Daten, ist viel häufiger als wir uns eingestehen.

Erhellende Hinweise darauf, warum das so ist, finden sich übrigens in der Autobiographie des bereits oben erwähnten Wissenschaftsbetrügers Diederik Stapel. Er beschreibt, wie leicht es ihm gefallen ist, durch nicht offengelegte Selektion von Daten und

Analyseverfahren die ‚Stories' seiner Paper interessanter zu machen, und dadurch in renommierten Journalen publizieren konnte. So fing er an, sich in der Psychologie einen Namen zu machen, die Tenure war greifbar. Der Übergang zur Manipulation seiner Daten war dann fließend. Niemand an der Uni, und auch kein Reviewer fragte nach, oder wollte Daten sehen. Das ging alles so einfach und glatt, dass er allmählich dazu überging, Studienergebnisse komplett zu erfinden. Die Befragungen führten seine Studenten durch, er hübschte die Daten dann in großem Stile auf. Damit wurden die Ergebnisse so spektakulär, dass Science und Nature sie mit Handkuss nahmen. So ‚erfand' er zum Beispiel Daten, deren Auswertung zeigte, dass in einer vermüllten Umgebung Befragte eher zu rechtsextremen Antworten neigen als in einer sauberen. Über solche Ergebnisse berichtete sogar die New York Times, und er wurde in kurzer Zeit zum Shooting-Star der Psychologie! An einer Stelle seiner Autobiographie beschreibt er, dass er sich fühlte wie ein Kind, das man allein im Zuckerladen zurückgelassen hatte. Einzig mit dem Hinweis, doch bitte keine Süßigkeiten zu stibitzen. Was ihm letztendlich das Genick brach waren seine eigenen Studenten. Die konnten sich zwar zunächst freuen, Koautoren auf tollen Papern zu sein, fanden es aber nach einiger Zeit befremdlich, die Daten nie selbst auswerten zu dürfen, immer nur von Stapel bereits prozessierte Daten zu Gesicht zu bekommen.

Der Fall Diederik Stapel ist sicherlich extrem, und wie er seine Verfehlungen ganz lässig auf das System abwälzt, das es ihm zu leicht gemacht hat, ist natürlich wohlfeil. Aber trotzdem kann man an seiner Karriere die wesentlichen Elemente des modernen Wissenschaftsbetruges studieren: Das auf einer Journal-Reputationsökonomie basierende akademische Belohnungssystem, Journale die spektakuläre Studien soliden vorziehen, mit der Qualitätskontrolle überforderte Reviewer, Berufungskommissionen und universitäre Gremien, die sich von Stories und Selbstvermarktern blenden lassen, als normal geltende und nicht sanktionierte fragwürdige Wissenschaftspraktiken als Einstiegsdroge, mangelhafte Diskussions- und Führungskultur in der Arbeitsgruppe, sowie methodische Inkompetenz bei allen Beteiligten. Mehrere Elemente dieses toxischen Gemisches finden zu jeder Zeit in den meisten Forschungseinrichtungen. Wenn aber alle zusammenkommen ist es nur noch eine Frage der Zeit, bis einzelne Wissenschaftler der Versuchung erliegen, der wissenschaftlichen Karriere ein bisschen nachzuhelfen und Abkürzungen zu nehmen. Nur wenn sie es allzu doll treiben, wie z.B. Herr Stapel, müssen sie damit rechnen, aufzufliegen. Und auch dann sind die Konsequenzen, falls es überhaupt zu Sanktionen kommt, recht überschaubar.

Besteht die Lösung des Problems also darin, Wissenschaftsbetrug härter zu sanktionieren? Schaden würde das sicher nicht, denn Fälle in denen Strafen verhängt wurden, kann man an einer Hand abzählen. Wissenschaftsbetrug wird also nicht nur selten aufgedeckt, sondern noch seltener geahndet. Müssen wir mehr Gute Wissenschaftliche Praxis lehren und trainieren? Auch das ist eine gute Idee, aber sehr viel nützen wird es wohl nicht. Es gibt ja auch keine Kurse, in denen Schülern und Studenten erklärt wird, dass Banküberfall und Urkundenfälschung gegen gesellschaftliche Normen verstoßen, verboten sind und konsequenterweise bestraft werden. Wissenschaftsbetrüger wissen was sie tun, sie tun dies nicht aus Unkenntnis ihnen unbekannter Regeln. Brauchen wir vielleicht eine Wissenschaftspolizei, welche unangekündigte Kontrollen von Westernblots und Festplatten in Laboren durchsuchen? Ganz sicher nicht, moderne ist Wissenschaft viel zu komplex, als dass sie durch solche Visiten kontrollierbar wäre. Ganz abgesehen davon, dass die dadurch entstehende Big Brother Atmosphäre alles andere als förderlich für gutes Forschen wäre.

Ein viel naheliegender Ansatz zur Abhilfe ist es, sich dem Kern des Problems anzunehmen, und das toxische Karriere- und Bewertungssystem zu reformieren – also Forscher

nicht auf Basis fragwürdiger Metriken, sondern mit Fokus auf Forschungsqualität, Inhalte und tatsächlichen wissenschaftlichen oder gesellschaftlichen Impact zu beurteilen. Das ist in der Tat der Königsweg, und in Ansätzen findet das zum Glück auch derzeit statt. Die von der Europäischen Union initierte Coalition for Reforming Reserach Assessment (COARA), der übrigens die DFG bereits beigetreten ist, wird hierbei eine wichtige Rolle spielen. Allerdings geschieht all dies im Schneckentempo, sodass ein schnellerer Fix erstrebenswert wäre.

Vielleicht gibt es den sogar! Wissenschaftsbetrug ist nämlich nur dort möglich, wo Einzelne die Auswertung und Analyse von Forschungsdaten monopolisiert haben, häufig noch vergesellschaftet mit fehlender methodischer Kompetenz im unmittelbaren Umfeld. Nur wenn Westernblots von nur einer Person angefertigt und ausgewertet werden, und auch niemand mit der nötigen Kompetenz draufschaut, können diese mittels Photoshop manipuliert werden. Das Analoge gilt für Datenreihen und die darauf angewendeten Analyseverfahren: Wenn also nur eine Person die Datenbanken oder Spreadsheets verwaltet und selbstverfasste Codes darüber laufen lässt, die Ergebnisse nicht von anderen kontrolliert werden. Was ja schon wegen der ehrlichen Fehler, die wir alle leider häufig machen, notwendig wäre. Wenn dann noch der Gruppenleiter einsam vor dem Rechner sitzend die Ergebnisse in eine Story verwandelt, kann es passieren, dass ihm ein übereifriger Mitarbeiter ein ‚Ei' legt, oder umgekehrt: Er selbst ‚kreativ' wird, und die Ergebnisse in eine Story verwandelt.

Es braucht also in den Arbeitsgruppen eine funktionierende Struktur und Arbeitskultur, und dann ist wissenschaftliches Fehlverhalten praktisch ausgeschlossen. Problematisch wird es nämlich immer dann, wenn Arbeitsgruppen zu groß werden, die Expertise zu fragmentiert ist oder punktuell gleich gar fehlt. Leider Bedingungen, welche gerade in der biomedizinischen Forschung nicht wirklich selten sind. Wie kann da Abhilfe geschaffen werden? Auf jeden Fall durch Thematisierung und Fokus auf gute Arbeitskultur und Gruppenleitung wo überall das möglich ist. Natürlich in der Ausbildung, wo das meist zu kurz kommt. Viele Unis bieten im Rahmen der Personalentwicklung ein ‚Führungskräftetraining' an. Aber dort müsste ein stärkerer Fokus auf die Wichtigkeit einer offenen und kollaborativen Arbeitsweise als Bollwerk gegen wissenschaftliches Fehlverhalten gelegt werden.

Aber auch bei Berufungen und Tenurisierung sollten wir uns stärker mit dem Thema befassen. Die zuständigen Kommissionen könnten Kandidaten zur Größe, Struktur und den Interaktionen in ihrer Arbeitsgruppe befragen. Sie könnte sogar Gespräche mit ehemaligen (oder auch noch aktiven) Mitgliedern der Arbeitsgruppe führen. An mancher Stelle wird dies bereits praktiziert, zum Beispiel bei EU-LIFE, einer Allianz von renommierten europäischen Forschungsinstituten. Bei klinischen Berufungen ist es übrigens gängige Praxis, dass Berufungskommissionen die Abteilung der Bewerber aufsuchen und sich vor Ort einen Einblick in deren ‚Arbeitsweise' verschaffen. Ein Vorschlag zur Steuerung der Arbeitsgruppendynamik durch die Universitäten wäre zum Beispiel die Kappung der Leistungsorientierten Mittelvergabe ab einer gewissen Gruppengröße. Ob all dies närrische, und letztlich unrealistische oder uneffektive Maßnahmen sind, und ob damit wissenschaftliches Fehlverhalten vermindert werden könnte, muss offenbleiben. Allein ein breiter Diskurs über die Arbeitskultur in wissenschaftlichen Arbeitsgruppen und wie wir sie verbessern können, würde uns schon weiterbringen.

Warum wissenschaftlicher Wumms weltweit weniger wird

LJ 3/2023

„Papers und Patente werden immer weniger disruptiv“, so alliterierte der Titel einer Studie von Park und Kollegen von der University of Minnesota, die kürzlich in Nature erschienen ist (https://doi.org/10.1038/s41586-022-05543-x, weitere Zitate wie immer unter http://dirnagl.com/lj). Sie untersuchten darin, wie sich Netzwerke von Zitationen in Wissenschaft und Technologie in 45 Millionen Artikeln und 3,9 Millionen Patenten der letzten 60 Jahre verändert haben. Kern ihres quantitativen Ansatzes ist der einfache Gedanke, dass, wenn eine Arbeit oder ein Patent etwas wirklich Neues zutage gefördert hat, die nachfolgende Arbeit, die es zitiert, weniger wahrscheinlich auch ihre Vorgänger zitiert: Für zukünftige Wissenschaftler sind die Ideen, die zu seiner Produktion geführt haben, weniger relevant. Wenn eine Arbeit oder ein Patent dagegen eher konfirmatorisch war, ist es wahrscheinlicher, dass nachfolgende Arbeiten, die sich darauf berufen, auch ihre Vorgänger zitieren: Für künftige Forschergenerationen ist das Wissen, auf dem die Arbeit aufbaut, immer noch (oder sogar noch mehr) relevant.

Tageszeitungen rund um den Globus berichteten aufgeregt über die beunruhigende Botschaft, dass die Innovationskraft der Wissenschaft ganz offensichtlich schwächelt. Das Paper wurde aber auch gleich zum Beleg für die eigene Botschaft: Dass nämlich wissenschaftliche Innovationen, Paradigmenwechsel, Revolutionen, Durchbrüche, wie immer man sowas auch nennen mag, seit Jahrzehnten kontinuierlich weniger werden, haben nämlich bereits eine Vielzahl von Studien mit sehr unterschiedlichen Methoden und Ansätzen schon vorher gezeigt. Die Message, dass wissenschaftliche Durchbrüche immer seltener werden, ist also selbst alles andere als disruptiv, sie ist vielmehr konfirmatorisch, oder wie die Autoren der Studie das formulieren würden, ‚konsolidierend ‘.

Disruptiv oder konsolidierend, bemerkenswert ist das Phänomen aber allemal. Forschen doch heute so viele Forscher wie nie zuvor: 90% aller Wissenschaftler, die je gelebt haben, sind noch am Leben. Weltweit sind das etwa 8 Millionen, sie veröffentlichen aktuell etwa 7 Millionen wissenschaftliche Artikel in 34.000 im Web of Science gelisteten Journale. All diese Zahlen steigen seit Jahrzehnten exponentiell. Trotzdem nimmt das von uns produzierte wirklich neuartige Wissen gleichzeitig ab, daran kann kein Zweifel bestehen. Wie kann es aber sein, dass trotz gigantischem Input immer weniger Weltbewegendes hinten rauskommt? Wissen wir vielleicht schon (fast) alles? Waren frühere Generationen von Wissenschaftlern einfach genialer? Ist das, was die Welt im Innersten zusammenhält, mittlerweile einfach zu komplex für uns? Gibt es schon zuviel Wissen und wir kommen nicht mehr hinterher? Oder nehmen gar die Qualität und Originalität der Forschung kontinuierlich ab?

Bevor wir uns auf die Suche nach Antworten begeben, ein kurzer Exkurs in die in diesem Diskurs verwendeten Begrifflichkeiten: Disruption, Revolution und Innovation in der Wissenschaft. Der oben zitierte Artikel beklagt das Abnehmen von ‚Disruption', bei gleichzeitiger Zunahme der ‚Konsolidierung' des Wissens. Disruption (von lat. disrumpere, zerbrechen, zerreißen) ist aber ein Begriff aus der Mottenkiste der Unternehmensberater, und gehört gar nicht in die Wissenschaft und deren Theorie. Ökonomen lieben disruptive Technologien, die alte Produktionsprozesse oder Produkte "zerbrechen", und damit überflüssig machen. Mit diesen kann man sich dann trefflich gegen die Konkurrenz durchsetzen, welche noch auf die alte Technologie setzt. Wissenschaft dagegen ist nicht disruptiv. Einstein hat Newton nicht ‚gebrochen', seine ‚Gesetze' gelten immer noch. Heisenberg, Schrödinger und Dirac haben auch den Einstein nicht zerrissen: Die allgemeine Relativitätstheorie ermöglicht ihrem Handy immer noch die Ortsbestimmung mittels GPS. Sie haben alle aufeinander aufgebaut, man steht auf den Schultern der Anderen – die dabei immer noch eine solide Standfläche bieten.

Wissenschaft schreitet nämlich in Paradigmen fort. Diese müssen, wie von Thomas S. Kuhn 1962 in seinem Klassiker „Die Struktur wissenschaftlicher Revolutionen" beschrieben, von Zeit zu Zeit gewechselt werden. Solche „revolutionären" Änderungen in den grundlegenden Konzepten und experimentellen Praktiken von wissenschaftlichen Disziplinen werden nun offensichtlich immer seltener. Es kommt dann zum Wechsel, wenn das vorherrschende Paradigma, unter dem die „normale" Wissenschaft arbeitet, mit neuen Phänomenen unvereinbar wird, was die Annahme einer neuen Theorie oder eines neuen Paradigmas erleichtert, ja nachgerade notwendig macht. Dabei ist die normale Wissenschaft das, was die meisten von uns tagsüber (und manchmal auch nachts und am Wochenende) so treiben, nämlich ordentliche Wissenschaft innerhalb des vorherrschenden Rahmens oder Paradigmas. Ich hoffe Sie nehmen es mir das nicht übel, aber wir sind eben, zumindest die Meisten von uns, keine Einsteins. Die Paradigmenwechsel führen häufig (müssen aber nicht) unter Anwendung neuer Ideen und Techniken zu geplanter und kontrollierter Veränderung im System, also zur Innovation.

Eine naheliegende Erklärung für den immer größeren Aufwand der trotzdem immer weniger neue Paradigmen (und in der Folge Innovationen) produziert könnte darin begründet liegen, dass wir die ‚einfach' rauszufindenden Prinzipien schon wissen. Um es bäuerlich auszudrücken: Die niedrig hängenden Früchte wurden schon gepflückt. Jetzt müssen wir uns mehr und mehr strecken, um an die in der Krone verbliebenen Früchte zu kommen. Zur Illustration des Gedankens ein paar Beispiele aus meinem Gebiet, den Neurowissenschaften: Epilepsien sind heutzutage sehr gut therapierbar, wobei fast alle Medikamente auf dem gleichen Prinzip aufbauen, nämlich die Übererregung von Nervenzellen zu dämpfen. Mehr als drei Viertel aller Patienten mit Epilepsie können damit anfallsfrei gemacht werden. Bei den verbleibenden Patienten ist das leider nicht der Fall – hier sind oft selbst komplexeste und nebenwirkungsreiche Medikamentenkombinationen wenig effektiv. Ähnliches gilt für viele andere Erkrankungen, Multiple Sklerose, Morbus Parkinson, etc. Für einen Großteil der daran Erkrankten hat das pathophysiologische Verständnis (ein ‚Paradigma') zur Entwicklung von effektiven Therapien (‚Innovationen') geführt. Das waren die niedrig hängenden Früchte, die Pathophysiologie war scheinbar noch recht übersichtlich (Übererregung, autoimmunologische Reaktion auf Proteine des Nervensystems, Untergang einer ganz bestimmten Neuronenpopulation und damit Ausfall eines spezifischen Neurotransmitters, etc.). Zur Behandlung der therapieresistenten Patienten fehlen uns nun neue Paradigmen, ja vielleicht wäre sogar ein Paradigmenwechsel auch im Verständnis der schon im Textbuch gedruckten Krankheitsmechanismen fällig. Aber das hängt eben weit oben im Baum, genauso wie bei Krebs, Alzheimer, und um ein Beispiel aus der Physik zu bringen, der dunklen Materie.

Mit dem Obstbauern Argument verbunden ist auch die Vorstellung, dass ein zunehmendes Problem für die Wissenschaft unserer Zeit die ‚Last des Wissens' sein könnte. Wir wissen in allen Bereichen schon so viel, dass es insbesondere für individuelle Forscher immer schwieriger wird, dies alles zu überblicken und Verknüpfungen, idealerweise sogar mit anderen Wissensgebieten, herzustellen. Die Ära der ‚Renaissance Forscher' ist definitiv zu Ende, als Wissenschaftler fokussieren wir uns auf immer winzigere Segmente des Wissens, und entwickeln ein Skotom für größere Zusammenhänge. Hier muss auch das sog. ‚Diversity innovation paradox' genannt werden. In einer Reihe von Studien wurde gezeigt, dass unterrepräsentierte Gruppen anteilig eine höhere Rate an wissenschaftlichen Neuerungen generieren als der Mainstream. Allerdings werden deren neue Beiträge von anderen Wissenschaftlern in geringerem Maße aufgegriffen. All dies sind Argumente für Team Science sowie Inter- und Transdisziplinarität, aber auch ‚Equity, Diversity, Inclusion' (EDI) - Ziele die zwar häufig beschworen, aber umso seltener praktiziert werden.

Ein gewichtiger Grund für die Stagnation wissenschaftlicher Innovationen könnte zudem sein, dass das viele Geld, das in die Wissenschaft fließt, nicht optimal verteilt wird. Deutschland steckte 2020 laut Weltbank immerhin 3.4 % seines Bruttoinlandprodukts in Forschung und Entwicklung, das ist eine ganze Menge Asche. Ein substantieller Teil dieses Geldes, von der Sachbeihilfe bis zum Exzellenzcluster wird mittels Peer Review verteilt, wir begutachten uns also gegenseitig. Das wesentliche Kriterium ist dabei bisheriger ‚Erfolg', in der Regel gemessen an der Reputation der Journale in denen wir veröffentlichten. So ein System ist notwendig risikoavers, statisch und homophil, in ihm regiert das Matthäus – Prinzip, nach dem bevorzugt auf große Haufen ges..issen wird. Was rauskommt ist dann gehobener Mainstream, aber eher selten Paradigmenwechsel. Das sagt nicht nur der Narr – den reichen Nationen dämmert das schon seit geraumer Zeit. Sie setzen gigantische Programme auf, bei denen sehr viel Geld mit ganz anderen Mechanismen verteilt wird. In USA ist das ARPA-H (Fokus auf Gesundheitsforschung, 10 Milliarden USD), in England ARIA (Advanced Research and Invention Agency, 1 Milliarde USD) und in Deutschland die Agentur für Sprunginnovation (SPRIN-D, 1 Milliarde €). Aber auch Losverfahren bei der Förderentscheidung, wie sie in Deutschland von der Volkswagenstiftung pilotiert werden und mittlerweile von einer Vielzahl von Fördergebern eingesetzt werden - sogar der Nationalfond in der sonst so konservativen Schweiz ist mit von der Partie - könnten neuen Ideen besser auf die Sprünge helfen.

Wenn dann noch ein substantieller Teil des im Peer Review unter uns aufgeteilten Geldes in Forschung von zweifelhafter Qualität versenkt wird und deshalb nicht reproduzierbar ist, könnte das zusätzlich die Chancen vermindern, dass was wirklich Innovatives rauskommt. Der Narr hat an dieser und anderer Stelle schon häufiger die hierbei zum Einsatz kommenden fragwürdigen Praktiken gebrandmarkt. Dazu gehören z.B. die nicht offen gelegte Flexibilität bei der Datenerhebung und -analyse, das Aufstellen von Hypothesen nach dem Bekanntwerden von Ergebnissen (HARKING), das Erzeugen vermeintlicher statistischer Signifikanz der Ergebnisse (p-Hacking), das Fehlen von Verblindung und Randomisierung, oder die nicht-Veröffentlichung von relevanten aber nicht ins Konzept passenden (vornehmlich negativen) Studien. Damit will ich Sie aber heute verschonen.

Übrigens findet das Nature Paper von Park et al. keine Evidenz dafür, dass die schon gepflückten, also niedrig hängenden Früchte für die Misere mitverantwortlich sind, denn ‚Disruption' nahm in ihrer Analyse über die Zeit in allen untersuchten Gebieten etwa gleichförmig ab. Genauso scheint ihnen geringer werdende Forschungsqualität kein relevanter Faktor, denn sie fanden schwindende Disruption in gleichem Maße in den top Journalen wie im Rest des Blätterwaldes. Ich halte diese Argumente für geradezu

grotesk. Natürlich werden in allen Forschungsfeldern gleichermaßen jene Äpfel zuerst gepflückt, die am einfachsten erreichbar sind. Und dass Nature, Cell und Science Forschung bessere Qualität abdrucken als andere ‚scholarly journals' müsste erst noch bewiesen werden. Die Beweislage spricht hier eher für das Gegenteil. Ich muss mich doch wundern, dass die Reviewer das geschluckt haben.

Schließen möchte ich dennoch mit dem letzten Paragraphen des hier vorgestellten Artikels, denn schöner hätte es der Narr auch nicht sagen können: „Insgesamt vertiefen unsere Ergebnisse das Verständnis für die Entwicklung des Wissens und können als Orientierungshilfe für die Karriereplanung und die Wissenschaftspolitik dienen. Um eine bahnbrechende Wissenschaft und Technologie zu fördern, sollten Wissenschaftler dazu ermutigt werden, viel zu lesen und sich Zeit zu nehmen, um mit dem schnell wachsenden Wissen Schritt zu halten. Die Universitäten sollten den Schwerpunkt nicht mehr auf Quantität legen, sondern die Qualität der Forschung stärker belohnen und vielleicht einjährige Forschungspausen stärker subventionieren. Bundesbehörden könnten in die risikoreicheren und längerfristigen individuellen Auszeichnungen investieren, die Karrieren und nicht nur spezifische Projekte unterstützen, um Wissenschaftlern die Zeit zu schenken, die sie brauchen, um aus der Masse herauszutreten, sich gegen die "publish or perish"-Kultur zu immunisieren und wirklich folgenreiche Arbeit zu leisten. Ein umfassenderes Verständnis des Rückgangs der disruptiven Wissenschaft und Technologie ermöglicht ein dringend erforderliches Überdenken der Strategien zur Organisation der Produktion von Wissenschaft und Technologie in der Zukunft."

Mit NARRativen läuft das Leben besser

LJ 4/2023

Kaum ein Schritt im beruflichen Werdegang von Wissenschaftlern, der nicht durch die Linse eines schriftlichen Lebenslaufes betrachtet wird. Stipendien, Anträge, Anstellungen, Verstetigungen, Berufungen, Preise, alle erfordern die Abgabe eines ‚Curriculum vitae'. Darin finden sich nach dem Familienstand, der Anzahl und manchmal den Geburtsjahren von Kindern, das Alter und Geschlecht der Antragsteller. Gelistet werden dann die Stadien der Ausbildung vom Abitur weg, alle Publikationen (manchmal getrennt in Erst/Letztautorschaft und Co-Autorschaft, aber meist garniert mit dem Journal Impact Factor mit 3 Nachkommastellen Genauigkeit), eine Liste aller gehaltenen (oder wenigstens der eingeladenen) Vorträge, Preise und Patente (sofern vorhanden), dazu die eingeworbenen Förderanträge (mit Fördersumme in €), sowie eine Erwähnung der Aktivitäten in wissenschaftlichen und akademischen Gremien, meist komplementiert durch die Nennung der Journale, für die man schon mal begutachtet hat. Dazu noch ein Porträtfoto, am besten eine Studioaufnahme und keins aus dem Fotofix-Automaten. Auch kumulative Metriken machen sich im CV gut, etwa der Hirsch-Faktor, Anzahl der Zitate, durchschnittlicher

JIF (oder Anzahl Artikel mit JIF > X) und die Gesamtsumme der eingeworbenen Drittmittel. In den fortgeschritteneren Stadien des Berufslebens kommen dann noch die Zahl (oft auch mit Titel der Arbeit und Prädikat) der betreuten Promotionen dazu. Es gibt sogar welche, die Auskunft über die Religionszugehörigkeit erteilen, den Berufstand der Eltern, oder die Anzahl der Geschwister.

Je nach Karrierestadium füllt ein akademischer CV damit locker 10 Seiten und mehr. Die Aktualisierung und Pflege dieses Dokuments ermöglichen dem Akademiker unzählige Stunden der Selbstreflektion und Selbstdarstellung. Sehr viel Zeit geht dabei schon allein deshalb drauf, weil je nach Art der Bewerbung unterschiedliche Formatierungen erwartet werden, auch wenn der Informationsgehalt der gleiche ist.

Kaum zu glauben, aber die Zeiten dieses über Jahrzehnte entwickelten und mittlerweile liebgewonnenen tabellarischen CV Formats könnten gezählt sein! In gewohnt behutsamer (man könnte auch sagen bedächtiger) Weise hat die DFG zum Ersten dieses Monats ein neues Format für Lebensläufe verbindlich gemacht. Dieses nimmt einige Elemente auf, die bei den größten Fördergebern anderer Länder (z.B. in USA die NIH, in England UK Research and Innovation, in der Schweiz die Swiss National Science Foundation) schon wesentlich weitgehender umgesetzt worden sind. Aber, wie vom Narren im einer der letzten Ausgaben des Laborjournal ausgeführt, hier gilt es nicht zu nörgeln, sondern zu applaudieren, denn für die DFG gilt: ‚Spät kommt ihr, doch Ihr kommt, der weite Weg entschuldigt Euer Säumen'! (LJ 12/2022)

Ein prototypisches Beispiel für das, was sich da derzeit weltweit tut, kommt aus einem Land, das erst 1971 das Frauenwahlrecht eingeführt hat und auch sonst nicht für radikale Reformen bekannt ist: Das CV-Format des Schweizer Nationalfonds (SNF), dem eidgenössischen Äquivalent der DFG. Darin gibt man nach der derzeitigen beruflichen Position nur das akademische Alter an, nicht etwa das Geburtsdatum. Das akademische Alter ist die Zeit, welche man tatsächlich der Forschung widmen konnte, nach Abzug von Unterbrechungen und nicht-wissenschaftlicher Arbeit. Auch die Open Researcher and Contributor iD (ORCID) wird abgefragt, sie ermöglicht die eindeutige elektronische Zuordnung von Publikationen und anderen Forschungsaktivitäten und -erzeugnissen, auch von jenen die im CV nicht gelistet wurden. Danach folgen ganz konventionell tabellarisch die Stationen der akademischen Ausbildung und Beschäftigungsverhältnisse.

Aber dann wird es richtig interessant, denn nun werden nur noch die maximal drei bedeutendsten Leistungen der akademischen Laufbahn abgefragt. Die Gesamtlänge dieses Narrativs ist auf eine (!) A4-Seite beschränkt. Die Beschreibungen können zum Beispiel Folgendes enthalten: Den eigenen Beitrag an der Forschung, deren Erkenntnisse und Einfluss auf die Wissenschaft oder die Gesellschaft; oder auch den historischen Kontext des wissenschaftlichen Problems. Dafür kann man maximal 10 eigene Arbeiten als Referenzen angeben, die man frei auf die 3 Leistungen verteilen kann. Verwendet werden dürfen alle Arten von wissenschaftlichen Outputs, nicht nur Artikel in wissenschaftlichen Zeitschriften, sondern auch Buchkapitel, Konferenzbeiträge, Preprints, Datensets etc. Ein Lebenslauf so klar wie Schweizer Gebirgswasser! Dieses CV-Template wurde vom SNF partizipativ, also unter Beteiligung von potentiellen Antragstellern und Gutachtern sowie Wissenschaftsadministratoren des SNF entwickelt, seine Verwendungspraxis wird wissenschaftlich begleitet, erste Ergebnisse wurden bereits publiziert.

Die DFG pirscht sich an solch revolutionäre Umtriebe da erst einmal langsam an: In ihrem neuen CV-Template findet sich nun ein optionales Freitextfeld für ergänzende Angaben zum Werdegang. Da kann man dann zusätzliche Informationen eingeben oder zu einer besonderen persönlichen Situation machen, die für eine angemessene Begutachtung und Bewertung der wissenschaftlichen Leistung relevant sein könnten. Dazu zählen

auch Kinderbetreuungsaufgaben, Mutterschutz-, Eltern- oder Erziehungszeiten, chronischen/langfristigen Erkrankungen, usw. Auch ein optionales Freitextfeld für Engagement im Wissenschaftssystem, wie Gremientätigkeiten, Tätigkeiten in der Selbstverwaltung der Wissenschaft, die Organisation wissenschaftlicher Veranstaltungen, Aktivitäten in der Lehre sowie Tätigkeiten als Mentorin bzw. Mentor., ist nun vorhanden.

Der Abschnitt ‚Wissenschaftliche Ergebnisse' im DFG-Vordruck ist von der Struktur her wie gehabt, man hangelt sich also entlang von maximal 10 ausgewählten Peer Review Publikationen. Aber nun hören wir Kuhglocken von fernen Almwiesen läuten: ‚Wo möglich' (!) soll man den eigenen Anteil an den öffentlich gemachten Ergebnissen darlegen, und ausführen, warum man die jeweilige Publikation an dieser Stelle genannt hat. Das steht zwar irgendwie auf dem Kopf: Das Tolle ist hier das Abstraktum ‚Autoren, Titel, Journal', und dies wird dann erst durch deren Inhalt und den eigenen Beitrag daran gerechtfertigt (‚wo möglich'). Sei's drum – immerhin wird hier nicht exklusiv auf die Reputation des Journals geschielt, es weht erstmals ein zarter Hauch vom Inhalt der Forschung und deren wissenschaftlichen oder gesellschaftlichen Impact. Und dann steht da noch der Satz: ‚Angaben zu quantitativen Metriken wie Impact-Faktoren und h-Indizes sind nicht erforderlich und werden bei der Begutachtung nicht berücksichtigt.' Auch das ist wieder typisch DFG, zwei Schritte vorwärts, einer zurück: Zwar werden diese Metriken angeblich nicht berücksichtigt, aber man kann sie trotzdem angeben. Vielleicht ist ja doch der eine oder andere Gutachter unter Zeitdruck, und will sich nicht um lästige Forschungsinhalte kümmern. Und dann hat man im DFG CV noch in einem optionalen Freitextfeld die Möglichkeit 10 jenseits des Peer Reviews öffentlich gemachte Forschungsergebnisse anzuführen. Wie zum Beispiel PrePrints, Datensätze, Protokolle, Softwarepakete, Patente, und – man höre und staune, sogar Blogbeiträge. All dies wird allerdings als ‚Kategorie B' leicht stigmatisiert und unter Quarantäne gestellt, sodass die Liste der Peer Review Artikel der ‚Kategorie A' um Himmels willen nicht kontaminiert wird. Tu felix Helvetia!

‚Alternative' CV- Formate, wie sie nun langsam Eingang ins akademische Begutachtungswesen finden, zeichnen sich dadurch aus, dass sie über die Integration des akademischen Alters den Vergleich verschiedener Karrierestadien erlauben, den Blick auf den wissenschaftlichen Beitrag fokussieren, und neue Disseminationsformate einschließen. Sie überwinden damit Schwächen des etablierten, klassischen Formats, welches Innovation, gesellschaftliche Auswirkungen, verantwortungsbewusste Forschungspraktiken, die Vielfalt der Forschung und der Forschungskarrieren, sowie die Qualität der Forschung als Ganzes ausblendet bzw. eindimensional auf wenige ungeeignete Metriken reduziert. Nebenbei könnte das neue Format dazu beitragen, Salamitaktiken beim Veröffentlichen entgegenzuwirken, und ganz allgemein die Diversität des agierenden Personals zu erhöhen.

Klingt das alles vielleicht zu gut um wahr zu sein? Oder ist es vielmehr so, dass hier der Teufel mit dem Beelzebub ausgetrieben wird? Fördern die Narrative nicht gar ein ‚selfmarketing' der Wissenschaftler? Dieser nun laut werdende Einwand entlarvt sich schon allein deshalb, weil es ja auch jetzt schon ein wesentliches Merkmal unseres Wissenschaftsbetriebes ist, dass jene, die sich gut verkaufen können, stark im Vorteil sind. Im schlimmsten Fall würde also die Kosmetik am klassischen CV durch ausgefeilte Narrative ersetzt werden. Noch wichtiger aber: Ist es nicht eine intellektuelle Bankrotterklärung der Begutachter, wenn Sie zu Protokoll geben, in 500 Zeichen langen Texten nicht die Blender erkennen zu können? Würde da nicht schon ein Blick auf den Inhalt der das Narrativ begleitenden Schlüsselreferenzen genügen? Und noch ein häufig gehörter Einwand: Die Narrative würden alle gleich klingen, man könne die Antragsteller bzw. Bewerber gar nicht mehr unterscheiden können. Auch dies eine eigentümliche

Befürchtung: Sollte das tatsächlich der Fall sein, dann hätte die bisher praktizierte Fokussierung auf die Reputation des Journales als wesentliches Exzellenzkriterium den Bewerbern und den Gutachern gleichermaßen die Fähigkeit geraubt, Inhalte zu transportieren oder diese zu würdigen. Das wäre dann aber ein weiteres gewichtiges Argument, sich mit den neuen CV Formaten wieder auf eine wissenschaftlich-inhaltliche Bewertung zu verpflichten und diese damit dann zu einzuüben.

Und noch ein Totschlägerargument sei hier erwähnt: Die alternativen Lebensläufe öffneten der Subjektivität Tür und Tor. Auch dieser Vorbehalt zielt ins Leere: Die Narrative, der Hinweis auf das akademische Alter, auch die Angabe von besonderen persönlichen Situationen werden ja immer begleitet von Referenzen, die das Behauptete objektiv belegen sollen. Zudem ist Subjektivität kein Fehler im System, sondern ein Merkmal auch der gängigen Auswahlpraxis. Wie die Inhalte der Tabellen und langen Publikationslisten von konventionellen Lebensläufen von den Gutachern in ihrem Gesamturteil berücksichtigt werden, wird doch auch jetzt stark von deren persönlichen Vorlieben und Anschauungen beeinflusst.

Fantastisch wäre es allerdings, wenn die Fördergeber und Institutionen sich eines gemeinsamen alternativen CV-Formates bedienen würden, und auch ein Tool zur Verfügung stellten, in dem man einmal seinen Lebenslauf einpflegt, und danach nur noch dann editiert, wenn sich was ändert. Das würde den Aufwand für alle Beteiligten massiv vermindern. Wieder einmal schauen wir neidisch in andere Länder, wie zum Beispiel in die USA, wo das schon länger der Fall ist (SciENcv: Science Experts Network Curriculum Vitae). Stellen Sie sich vor, die Europäische Kommission, die DFG, das BMBF hätten sowas, und auch die Unis würden sich dessen bedienen. Die Zeit, die wir als Antragsteller, aber auch als Gutachter einsparen würden, wäre substantiell. Ganz zu schweigen von dem Quantensprung in der guten Evaluationspraxis, welcher damit verbunden wäre.

Tschüss LOM: Zu wenig Geld, unwirksame Steuerung, falsche Anreize

LJ 5/2023

Vor fast 20 Jahren hat die DFG eine äußerst wirkmächtige Stellungnahme verfasst. Die 2004 veröffentlichten ‚Empfehlungen zu einer »Leistungsorientierten Mittelvergabe« (LOM) an den Medizinischen Fakultäten' sind, im Gegensatz zu vielen anderen solchen Papieren nahezu 1:1 umgesetzt worden. Alle medizinischen Fakultäten Deutschlands verteilen mittlerweile Anteile der Mittel, die sie aus ihrem jeweiligen Bundesland für Forschung und Lehre erhalten, unter Berufung auf und Verwendung der Kriterien aus dem DFG Papier. Die LOM soll danach für eine gerechte Mittelzuweisung anhand klarer Kriterien sorgen, Transparenz und Dynamik schaffen und Forschungsleistungen belohnen, sowie Anreize generieren, diese

zu steigern (‚Incentivierung'). In Academia sollte endlich der Wind des ‚New Public Management' wehen, auch war es ein Angriff auf die Erb- und Gutshöfe der Ordinarien.

So wohlmeinend die Intentionen der DFG und der Fakultäten dabei auch gewesen sein mögen, so gründlich ist das Projekt gescheitert. Die DFG empfahl einen Anteil der LOM am Landeszuführungsbeitrag von 20-40%. Er liegt heute weit darunter, zwischen 0,4 und 15%, im Median um die 5%. Dabei muss man bedenken, dass die Landeszuführungsbeiträge weder mit den gestiegenen Preisen und Personalkosten, noch den glücklicherweise steigenden Drittmittel-Einnahmen Schritt gehalten haben. Die den Wissenschaftlern real zur Verfügung stehenden LOM-Mittel sind deshalb an fast allen Fakultäten mittlerweile lächerlich gering. Ein ohnehin kleiner (und noch schrumpfender) Kuchen wird in immer kleinere Stücke zerteilt. Trotzdem sind diese Mickerbeträge für die Wissenschaftler lebensnotwendig, weil nicht nur die Projektmittel, sondern auch die Grundfinanzierung nicht auskömmlich ist.

Obzwar verschiedene Studien dies versucht haben, gibt es auch keinerlei Nachweis dafür, dass die LOM als Steuerungsmittel funktioniert hat – sie also in der Incentivierung von mehr oder besseren Publikationen und Drittmitteln Wirkung zeigte. Es kommt aber noch schlimmer, die in der LOM mathematisch kodifizierte Gleichsetzung von Forschungsleistung und Journal Impact Factor und Drittmitteln hatte eine toxische Wirkung in der Sozialisierung der Wissenschaftler und deren Nachwuchs. Die Formel sagt nämlich: Man leistet was in der Forschung, wenn man in renommierten Journalen publiziert und viel Geld dabei ranholt. Erkenntnisgewinn, gesellschaftlicher Nutzen, oder gar Forschungsqualität sind dabei bestenfalls Mittel zum Zweck, und kommen in der Gleichung auch gar nicht vor. Zusammen mit den Berufungsverfahren, in denen JIF und Drittmittel ebenfalls regieren, hat die LOM schleichend zu einer Umdefinierung des Zweckes von Forschung beigetragen, ein Kulturwandel der heute fast vollständig vollzogen ist: Der Zweck von Forschung ist die Einwerbung von Drittmitteln und Publikationen mit hohem JIF. Und weil die LOM retrospektiv (i.d.R. 3 Jahre zurück) ermittelt wird, und sich damit auf ‚Leistungen' bezieht, die viele Jahre vorher erfolgt sein müssen, ist der wissenschaftliche Nachwuchs sowie nach vorne gerichtete Innovation, oder Forschung ‚to boldly go where no man has gone before' sowieso außen vor.

Nun ist es nicht so, dass die Autoren der DFG-Empfehlungen damals naiv waren. Sie haben vor all den Entwicklungen, die jetzt eingetreten sind, gewarnt. Ein paar Leseproben: ‚Bei stagnierenden bzw. sinkenden Zuführungsbeträgen führt aber das Fehlen der Finanzierung der tatsächlichen Projektkosten, also auch des Infrastrukturkostenanteils („overhead"), zu empfindlichen Belastungsproben innerhalb der Fakultäten', oder: „... dass Originalität und Qualität als Bewertungsmaßstab stets Vorrang vor Quantität haben. Sinnvoller ist es, die inhaltliche Bewertung von Publikationen, also die Qualität der erbrachten Forschungsleistung, zum Kriterium der Vergabe von Forschungsmitteln zu machen.' Die Verwendung des JIFs wurde daher für eine Pilot- und Übergangsphase empfohlen, bis ‚unter Mitwirkung der Fakultäten Mechanismen entwickelt werden, wie ein echtes Prüfverfahren zeitnah und kostengünstig erfolgen kann.' Nur ist das nicht passiert, wir sind im Jahr 2004 stecken geblieben.

Die LOM ist also zu niedrig, verfehlt ihre Steuerungswirkung, und schafft Fehlanreize – ja sie erzeugte und zementiert nun eine gefährliche Unkultur. Etwas abgeschwächt, und eleganter formuliert das auch der Medizinische Fakultätentag (MFT), der einflussreiche Verband der Medizinischen Ausbildungs- und Forschungsstätten Deutschlands. Er ist die Stimme der 39 deutschen Medizinfakultäten. In seinem kürzlich veröffentlichten Impulspapier ‚Indikatorgestützte Mittelallokation für die Forschung in der Hochschulmedizin (ehemals LOM)' beerdigt er die LOM schon gleich im Titel und stutzt sie auf das,

was sie in Wirklichkeit ist: eine ‚Indikatorgestüzte Mittelallokation ' (IMA). Die neue Terminologie löst zwar nicht die Probleme der unterfinanzierten Fakultäten, noch weniger die der prekären Forschungsprojekte der Wissenschaftler. Mit all den richtigen Argumenten stellt der MFT damit aber klar, dass es bei der ‚ehemals LOM' nicht um ‚Forschungsleistung' und auch nicht um Incentivierung geht. Sondern um eine Kofinanzierung basierend auf verausgabten Drittmitteln und einem mittlerweile gänzlich desavouierten Indikator, dem JIF.

Und damit kann der MFT auch viel klarer Ross und Reiter benennen: ‚Eine IMA kann Defizite in der Grundfinanzierung bzw. das Fehlen von kostendeckenden Drittmitteln nicht kompensieren'. Weiter fordert der MFT: es ‚sollte eine finanzielle Honorierung von Leistungs- und Belastungsunterschieden in Forschung und Lehre geschaffen, eine Rechenschaftslegung gegenüber den Länderparlamenten etabliert und der sachgerechte Umgang mit knappen Haushaltsmitteln optimiert werden'. Was heißt das nun für die Fakultäten? Und übrigens nicht nur für die medizinischen, denn das LOM-(un)Wesen hat in fast alle Universitäten und deren Fakultäten metastasiert. Sollen sie die LOM nun einfach abschaffen? Oder genauso weiter machen, sie aber anstatt LOM nun einfach IMA nennen? Natürlich nicht!

Als erstes sind die Fakultäten aufgerufen, das Ziel zu definieren, um das es ihnen geht. Will die Universität so viele CNS (Cell, Nature, Science) Papers wie möglich akkumulieren, und dann damit zu glänzen? Dann kann sie beim JIF als Hauptindikator bleiben, darf aber nicht der Illusion verfallen, dass sie dadurch mehr davon kriegt. Sie honoriert die Zielerfüllung halt mit einem finanziellen Zuschuss. Was natürlich absurd ist, selbst wenn die Incentivierungslogik stimmen würde: Submittiert irgendjemand Papers bei CNS, Lancet, oder New England Journal of Medicine, weil man dafür LOM bekommt? Will sie Wissenschaftler darin unterstützen, regelhaft unterfinanzierte Drittmittelprojekte kozufinanzieren? Sie wird dadurch nicht mehr Drittmittel bekommen, aber erleichtert ihren Wissenschaftlern vielleicht die Durchführung der Projekte. Will sie den wissenschaftlichen Nachwuchs fördern? Dann muss sie sich ein Programm ausdenken, bei denen Sie Mittel des Landeszuführungsbetrages direkt an die jungen Wissenschaftler bringt. Will sie Kollaborationen fördern unter ihren Wissenschaftlern? Dann muss sie ihnen dafür Geld geben. Will sie Innovation und Hochrisiko-Forschung fördern? Will Sie mehr offene Wissenschaft? Und so weiter. Ich denke es wird klar was ich meine. Erstmal die Ziele definieren, am besten zusammen mit den Wissenschaftlern, und nicht per ordre du mufti, und dann Instrumente und Programme definieren und etablieren um diese umzusetzen.

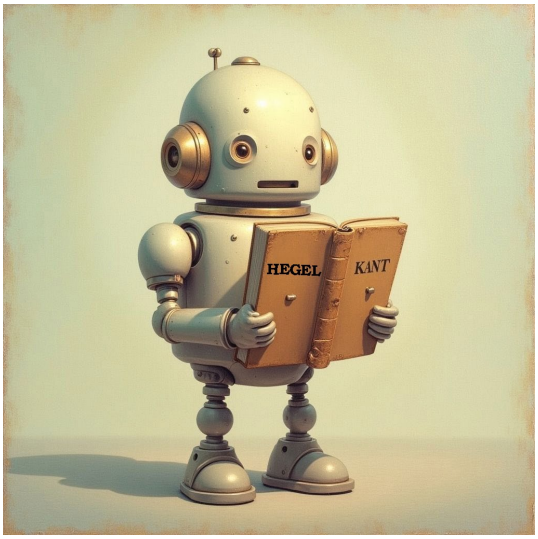
Nun werden die Dekane sagen, klingt ja schön und gut, aber wir können nur verteilen was wir haben, und wir haben eben nur sehr wenig. Das ist natürlich richtig. Aber es erhöht den Druck in der Diskussion um die Unterfinanzierung der Unis und fördert eine bisher nichtexistierende Debatte. Denn viele Bundesländer praktizieren in der Universitätsfinanzierung ebenfalls LOM Modelle, und unterwerfen sie dabei häufig denselben Indikatoren wie die Unis in der Folge ihre Wissenschaftler. Und lassen sogar die landeseigenen Unis um die LOM konkurrieren. Und da gilt dann wieder alles bereits oben Gesagte: Zu wenig Geld, unwirksame Steuerung, falsche Anreize. Aber die WissenschaftsministerInnen, StaatsekretärInnen und deren Arbeitsbienen hinterfragen bisher weder die Logik, noch die Effektivität der Maßnahme. ‚Leistungsorientiert' klingt doch super, oder?

Zudem fördert die Umdefinition von LOM in IMA einen fakultätsinternen Diskurs über die Ziele und Möglichkeiten einer solchen Kofinanzierung. Wissenschaftler (und Dekane waren ja früher auch mal Wissenschaftler) sind bekanntermaßen konservativ. Aber

vielleicht gibt es doch fortschrittliche Fakultäten, die jetzt angestoßen werden, über die Ziele und Kultur ihrer Forschung nachzudenken. Damit wären sie in guter Gesellschaft, und Teil einer großen Initiative, der Coalition on Reforming Research Assessment (COARA). Angestoßen von der Europäischen Kommission haben sich ihr bereits eine Vielzahl wichtiger Unis und Forschungsförderer weltweit angeschlossen. Raten Sie mal, wer der erste Unterzeichner in Deutschland war? Die DFG!

KI: Kritik der schwätzenden Vernunft

LJ 6/2023



Künstliche Intelligenz (KI) und kein Ende: Täglich wird gewarnt, beruhigt, abgewägt. Sind ChatGPT et al. nun ein Segen für die Menschheit oder der Anfang von der Herrschaft der Maschinen? In der Berichterstattung werden dabei jede Menge Metaphern und Vergleiche benutzt, welche die Leistungen von KI mittels Analogien vermenschlichen: Intelligenz, Lernen, Sprechen, Denken und Verstand, (Selbst-) Bewusstsein, Urteilen, Schließen, Entscheiden, Generalisieren, Fühlen, Kreativität, Irren, Halluzinieren, neuronale Netze, und vieles mehr. Gleichzeitig werden Funktionen des menschlichen Gehirns mit Begriffen wie Computer, Memory, Speicher, Code, Algorithmus, usw. beschrieben. Auch fehlt der Hinweis nicht, dass doch auch im menschlichen Gehirn elektrischer Strom fließt, ganz wie im Computer. Besonders das Feuilleton, benebelt von den mittlerweile verblüffenden Leistungen der schwätzenden und malenden Bots treibt deshalb die Frage um, ob wir es bei der generativen KI nun schon mit „echter“ Intelligenz zu tun haben – trotzdem im Namen die Frage eigentlich schon entschieden schien.

Warum gibt auch der Narr noch seinen Senf dazu? Weil er glaubt, dass die KI-Debatte voll am Thema vorbei geht. Das Lager derer, die KI für intelligent halten, belegt das mit einer Batterie von Leistungen, welche alle ziemlich intelligent aussehen. Die Zweifler überzeugt das aber nicht, ihnen fehlen immer noch bestimmte „Funktionalitäten“ von Intelligenz, die sie dann aus dem Ärmel ziehen Frei nach Tesler“s Theorem: „Intelligenz ist, was KI noch nicht gemacht hat.“ Die Diskussion bewegt sich bloß an der Oberfläche, statt sich mit der Frage zu befassen, was eigentlich Intelligenz, Denken, Sprache, Bewusstsein, etc. sind, um die KI daran zu messen.

Zum Glück hat sich vor über 200 Jahren schon mal jemand ganz grundsätzliche und sehr schlaue Gedanken zu eben diesen Geistestätigkeiten gemacht, welche die Debatte zurück auf eine inhaltliche Ebene führen kann. Und das war nicht, wie der Titel dieser Zeilen vermuten ließe, Immanuel Kant, sondern sein Kritiker Georg Wilhelm Friedrich Hegel. Leider hat er seine Gedanken in eine für uns heutzutage schwer verdauliche Sprache verpackt, deshalb sind sie keine einfache Lektüre. Aber allemal gutes Material für den Narren, der im Folgenden versuchen wird, etwas zu leisten, was – Spoiler Alert! - KI

eben nicht kann: Aus den Begriffen von Sprache und Denken abzuleiten, warum KI nicht sprechen und denken kann. Und zwar ganz prinzipiell nicht.

Die Kernfrage unserer Betrachtung lautet: Kann ein Computer mit Nullen und Einsen denken, kann er sich mittels KI zu einem geistigen Subjekt entwickeln (oder hat er das vielleicht schon?). Ein „Verstand“, der sich möglicherweise anschickt, sich erkennend einen Begriff von der Welt zu machen? Und dann anfängt, Gutes, aber vielleicht auch gar Schreckliches für und mit uns zu tun?

Beginnen wir auf der Ebene des Computers, genauer gesagt auf der der Transistoren. Ein Wort ist für den Computer, und damit die KI, nichts als eine Folge von zwei physikalischen Zuständen, dem „Ein“ oder „Aus“ eines Schalters auf einem Halbleiter. Menschen, welche die Chips gebaut und programmiert haben, haben diesen Zuständen Symbole zugewiesen, nämlich 0 und 1. Zahlen deshalb, weil man damit rechnen kann. Dies auch der Grund, warum das Ding Computer heißt: Weil es – und damit auch Ihr Handy, nichts anderes ist als eine programmierbare Rechenmaschine. Der Programm-Code weist nun bestimmten Abfolgen dieser Symbole in vielen Zwischenschritten Worte zu, die nur für uns Menschen Bedeutung haben. Die Bezeichnung der dafür genutzten Algorithmen als „neuronale Netze“ ist ein gigantischer Marketing Trick, genauso effektiv und falsch wie der Begriff künstliche „Intelligenz“ selbst. Tatsächlich sind künstliche neuronale Netzwerke nichts als mathematische Formeln, die mit begriffslosen Symbolen rechnen, und von sehr einfachen, veralteten Modellvorstellungen von “echten” Neuronen inspiriert wurden.

Hinter all dem, was da so gerechnet wird, kann also ganz prinzipiell keine Vorstellung oder ein Begriff der Sache stehen, die von der KI erfasst wurde. Auch wenn die KI noch so geschliffen argumentiert – es sind für sie inhaltsleere physikalische Zustände, codiert in Nullen und Einsen. Noch offensichtlicher wird das natürlich bei der Repräsentation von Bildern im Computer – auch deren Pixel sind der KI nichts als binäre, gegenstandslose Schaltzustände von Transistoren.

Damit ist eigentlich schon alles gesagt, der Begriff abgeleitet, warum KI nicht intelligent sein kann, damit auch nicht sprechen, denken oder urteilen. Lassen Sie uns aber trotzdem noch ein bisschen weitermachen, und der Frage nachgehen, was Sprechen, Denken und Urteile eigentlich sind. Dabei wird dann endgültig klar, warum das nicht durch Rechnen mit Nullen und Einsen geht.

Bei der Begriffsklärung dieser Geistesleistungen helfen leider keine der ach so populären „funktionalen“ Bestimmungen weiter. Stellvertretend für die Schwäche solcher Definitionen hier die von Intelligenz, von der Konsensusgruppe führender internationaler Psychologen: „Intelligenz ist eine sehr allgemeine geistige Fähigkeit, die u. a. die Fähigkeit zu denken, zu planen, Probleme zu lösen, abstrakt zu denken, komplexe Ideen zu verstehen, schnell zu lernen und aus Erfahrungen zu lernen aus Erfahrungen zu lernen einschließen.“ Das ist in Wahrheit keine Definition, sondern eine recht willkürliche Aufzählung von Fähigkeiten. Sie sagt nicht, was Intelligenz ist, sondern lediglich wozu man sie (möglicherweise) nutzen kann. Genauso geht das bei der Definition von Sprache *als* „kommunizierende Verhaltensweise“, oder Denken *als* „eine Form des Erkenntnisgewinns und der Erkenntnisnutzung; es ist etwas Dynamisches, das in der Zeit abläuft“. So kommen wir der Sache, also dem Begriff von Intelligenz, Sprechen und Denken nicht näher.

Lassen, wir zunächst jene zu Wort kommen, welche glauben, dass der Rubikon nun endlich überschritten sei, und die Large Language Models (LLM) menschenähnliche generelle Intelligenz besitzen. In Reinkultur findet sich die falsche Vorstellung vom

intelligenten Computer in dem kürzlich veröffentlichen 155-seitigen Preprint „Sparks of Artificial General Intelligence: Early experiments with GPT-4“ (alle Links unter http://dirnagl.com/lj). Mit geradezu kindlicher Freude berichten die Wissenschaftler der Forschungsabteilung von Microsoft über ihre „Experimente“ mit einer Reihe von LLMs. Mit dabei natürlich GPT-4, dem derzeitigen Klassenprimus. Die Bots bekommen dabei Fragen und Aufgaben gestellt, und siehe da, die Resultate sehen doch ganz so aus, als ob die LLMs urteilen, empathisch und kreativ sind (sie malen und machen Musik!), sowie Selbstbewusstsein und „Theory of Mind“ besitzen. Natürlich attestiert die Forscher den LLMs noch einigen Verbesserungsbedarf: Manchmal „halluzinieren“ sie, oder machen grobe Fehler, und das sogar bei simpelster Arithmetik. Ausgerechnet GPT-4 fällt in Mathe durch, weil es bei 7 x 4 + 8 x 8 = 88 ausgibt. Aber eigentlich gilt den Autoren auch dies als Intelligenzbeweis: Wie menschlich, allzu menschlich!

Die Microsoft-Forscher berauschen sich an der sauberen Grammatik der LLMs, und deren überaus höflichen Sprachstil, der mühelos zwischen Rap, Shakespeare und Python wechseln kann. Aber weil sie sich vor und bei ihren Spielereien keinen Begriff gemacht haben über ihren (Forschungs-) Gegenstand, übersehen sie das Wesentliche. Das ist doppelt tragisch, denn nicht nur kommen die Autoren deshalb zu einem falschen Schluss (Computer = intelligent). Sie haben außerdem gerade das nicht geleistet, was eine der wesentlichen Leistungen menschlicher Intelligenz ist, nämlich genau das, „sich einen Begriff von der Sache“ (hier also von der KI) zu machen. Einen Begriff macht man sich, in dem man sich erkennend zur Welt stellt, also bestimmt, was die Sache wirklich ist: Nicht wie sie vorkommt, sondern man benennt, was notwendig und wesentlich, und nicht nur zufällig und äußerlich ist.

Argumentierend nur mit oberflächlichen Analogien (eben Äußerlichkeiten, und nicht Wesentlichem) liegen sie deshalb voll daneben in ihrem Schluss, dass sie es bei den LLMs mit allgemeiner Intelligenz, oder mit irgendeiner anderen Form von Intelligenz zu tun haben. Sie erkennen nicht, dass die einzige Intelligenz, die da im Spiel war, die der menschlichen ist, welche die Software programmiert hat – möglicherweise also mit ihrer eigenen! Und natürlich der geballten und historischen Intelligenz, die für das Training verwendet wurde. Diese menschliche Intelligenz hat sich außerhalb und unabhängig von der KI betätigt, und so die Grundlage dafür geschaffen, dass die KI Erkennen, Verstehen und Entscheiden begriffslos simulieren kann, indem sie aus dem Material früherer Zuordnungen neue statistisch extrapoliert.

Der KI Algorithmus stellt nämlich lediglich statistische Bezüge und Korrelationen zwischen Merkmalen der Eingabe her, egal ob diese aus Tweets von Elon Musk, Goethes Faust, oder der Wikipedia bestehen. Diese Bezüge zwischen den Inhalten des Trainingsmaterials sind rein stochastisch, sie beruhen nicht auf physikalischen, logischen, oder inhaltlichen Zusammenhängen. Die KI und ihr Sprachmodell generalisieren dabei entgegen anders lautender Behauptungen nicht, sondern schaffen bloß begriffslose Kennzeichnungen, Klassifizierungen und Regeln. Diese beruhen eben nicht auf allgemeinen Bestimmungen, sondern sind lediglich das Resultat statistischer Ähnlichkeiten von Einzelfällen mit den Trainingsdaten.

Ein schönes Beispiel für diese Semantik-, Begriffs- und Inhaltslosigkeit der KI ist, dass sie Sprachen perfekt übersetzen kann, ohne die Vokabeln und Grammatik von auch nur einer dieser Sprachen zu kennen oder zu verstehen, sie also sprechen zu können. Bei uns Menschen ist letzteres aber die Grundvoraussetzung des Erlernens einer Fremdsprache. Entgegen landläufiger Meinung lernt die KI dabei auch nicht, es sei denn man versteht wie viele Psychologen unter Lernen lediglich Konditionierung, Imitation, oder Habituation. Lernen ist dieser Definition folgend stumpfsinniges Repetieren („Pauken“). Echtes

Lernen bedeutet aber ein Erfassen des Lerngegenstandes durch Nachdenken oder Nachvollziehen, oder noch abstrakter, und für KI unerreichbar: Beim Lernen die allgemeinen Bestimmungen einer Sache zu erfassen.

Das zeigt sich auch beim Spracherwerb. Ein Kind lernt nicht Sprechen durch das Abhören von Milliarden von Texten und nachfolgender statistischer Analyse. Es erlernt eine Sprache (und dabei gleichzeitig komplexes Denken, aber davon gleich), in dem es aus eigener Erfahrung und Anschauung Vorstellungen im Gedächtnis „speichert", und diese mit Sprachzeichen und Wörtern, die es hört, in eine feste Verbindung bringt. Das können auch die Gebärden sein, welche Taubstumme sehen und als ihre Sprache erlernen. Dafür benötigt ein Kind erstaunlich wenig Material, auf jeden Fall keine Terabytes Weltliteratur. Das Gehirn des Kindes erlernt die Sprache durch deren Nutzung nach dem gehörten Vorbild, und eignet sich deren grammatikalische Regeln an, ohne je eine Grammatik zu Rate zu ziehen. Das Resultat dieser Leistung der Intelligenz ist es, eine Sache im Namen (z.B. ein Wort oder Begriff) zu erkennen und dabei beides – also Sache und Namen - im Denken eins werden zu lassen. Man muss sich keinen Baum mehr vorstellen, um beim Wort Baum zu verstehen, was damit gemeint ist - man könnte salopp auch sagen, das Wort Baum ist im Gehirn zum Baum geworden. Wie das ein Gehirn mit einem synaptischen elektrochemischen Gewitter zustande bringt ist gänzlich unbekannt – aber wir müssen das auch gar nicht wissen, weil dieses neurobiologische Wissen nichts Zusätzliches beiträgt, es würde ja „nur" die materiellen (physiologischen) Grundlagen des Denkens beschreiben, und nicht seinen Begriff, also was die Sache selbst ist.

In diesem „Embodiment", dem Eins werden im Gehirn von Sache und Namen beim Denken liegt auch der Grund, warum man mittels funktioneller Magnetresonanztomogaphie (fMRT) Hirnoxygenierungsmuster „auslesen" kann, die während dem Sprechen von Wörtern, dem Blick auf oder dem Imaginieren von Bildern oder Sprache auftreten. Diese Muster, die ihre Bedeutung im vorangegangenen Training mit eben diesen Bildern oder Worten zugewiesen bekommen haben, erlauben es dann, und das nur im identischen, trainierten Individuum, diese Wörter oder Bilder wieder teilweise zu rekonstruieren, und das auch nur mit hoher Fehlerrate. Das sind fantastische Ingenieurs- und Programmierleistungen, es könnte auch für eine rudimentäre Kommunikation mit Gelähmten taugen, die sich motorisch nicht mehr ausdrücken können (Brain-Computer-Interface), ist aber weder Gedankenlesen noch macht es die Maschine in irgendeiner Weise intelligenter: Der Computer findet inhaltsleere Muster, der Inhalt (=die Bedeutung) wird vom Menschen zugewiesen.

Durch die Sprache können wir über Dinge nachdenken, und dies auch ohne einen inneren Monolog zu führen. Das geht natürlich auch, es mag dies manchmal sogar hilfreich sein, insbesondere wenn wir komplexe Gedanken wälzen – wie hier zu Sprache und Denken. Auch ohne einen solchen Monolog beruht das Denken, mit dem wir die Welt verstehen lernen und dieses Wissen im täglichen Leben - wie auch gerade in der Wissenschaft - noch erweitern, auf Sprache. Zum Vergleich die KI: Sie kann (oft) fehlerlos und umfassender als mancher Mensch „sagen", was ein Baum ist. Aber sie spricht oder denkt dabei nicht, denn für die KI liefert nur ein Eintrag in einer lexikalischen Liste von für sie inhaltslosen Bestimmungen, welche sie für das Sprachzeichen BAUM aus Myriaden von Quellen zusammengesucht hat. Wenn die KI dies dann ausgibt, vielleicht auch noch mit sonorer Stimme, scheint das manchem intelligent zu sein. Aber die gleiche Person hält doch die Wikipedia nicht für intelligent, weil man (wie auch die KI) in ihr beim Eintrag „Baum" richtige Bestimmungen findet.

Es gibt also viele Gründe, warum eine KI nicht urteilen und Begriffe bilden kann. Diese liegen ganz grundsätzlich darin, dass die Welt in ihr in begriffsleeren Symbolen

repräsentiert ist. Deshalb kann die KI auch nicht sprechen – und was uns als gesprochene Sprache verkauft wird, ist lediglich die Umsetzung von Zeichen in Töne. Weil aber schon das Zeichen für die KI keinen Inhalt hat, kann der daraus generierte Ton natürlich auch keinen haben. Und weil die KI nicht sprechen kann, kann sie auch nicht denken, denn die Sprache ist das Mittel des begrifflichen Denkens.

Deshalb klappt's bei der KI auch nicht mit dem Urteilen, denn mit Sprache trennt die Intelligenz im Urteil das Subjekt von dessen Bestimmung - dem Prädikat (z.B. „Die Rose ist wohlriechend", „Der Computer ist eine programmierbare Rechenmaschine"). Im Schluss beweisen wir daraufhin die Identität von Subjekt und Prädikat – womit man, wenn der Schluss richtig war, die Substanz einer Sache ausgemacht hat: Man hat sie erklärt, sie unterschieden von dem, wie sie bloß erscheint oder vorkommt. Hegel würde sagen, man hat den Begriff der Sache, man erfasst die Realität im Gedanken. Wir Menschen können das – die KI nicht. Tiere übrigens auch nicht, weil sie zwar denken können, aber keine Sprache haben. Genügend Material für einen weiteren Artikel des Narren!

Damit ist auch klar, dass, KI keinen freien Willen entwickeln kann – und uns an den Kragen gehen, wie die KI SkyNet im Film „Terminator". Das heißt aber natürlich nicht, dass KI nicht gefährlich sein kann. Ihre bereits schon länger durchgesetzte Nutzung in der Militärtechnik beweist das ebenso wie Fahrzeuge von Tesla, die im Autopilot-Mode manchmal ihre Eigner und dazu noch ein paar Fußgänger töten. Aber hier ist immer der Mensch das Subjekt, also der Gefährder. Ebenso wie bei Deep fakes, Plagiarismus und anderen kriminellen Aktivitäten, für welche Menschen KI trefflich einsetzen.

Aus dem bisherigen sollte klar geworden sein, dass sich alle bisher entwickelten KIs nur auf „next word" oder „next pixel prediction" verstehen, und damit kein neues Wissen schaffen können. KI schmeißt alles zusammen, was Menschen in eine digitale Form gebracht haben, vorausgesetzt, dass es via Internet oder proprietäre Datenbanken verfügbar ist. Da findet die KI Richtiges und Nützliches, aber noch mehr Unsinniges, Unklares und Falsches. Damit repliziert KI natürlich auch alle existierenden Vorurteile. Weshalb Horden von Programmierern die resultierenden Unflätigkeiten, Volksverhetzungen, Gewaltaufrufe etc. durch Zensur der KI wieder ausbügeln müssen. Oder gleich versuchen, diese Probleme auf der Linie durch Zensur des kommunizierenden Menschen zu klären. Man verbietet der KI ganz einfach, zu antworten.

Mittels KI konfrontieren wir uns also mit den Leistungen und Auswüchsen unserer eigenen Intelligenz. Deshalb ist KI auch keine „künstliche Dummheit", wie so mancher Kritiker glaubt. Auch weil es für Dummheit, welche nichts anderes als der falsche Einsatz von Intelligenz ist, eine gute Portion Intelligenz braucht, und an der mangelt es der KI komplett. KI taugt damit hervorragend zum Schreiben von Besinnungsaufsätzen und Gedichten. Und zu allem, bei dem menschliche Intelligenz auch nichts anderes macht als Muster erkennen, zu kodieren, sortieren oder klassifizieren. Und davon gibt es eine Menge, in der Medizin, auf dem Amt, im Journalismus, beim Programmieren, dem Übersetzen oder auf dem Schlachtfeld. Da führt uns KI nur vor, wie geistlos doch viele unserer beruflichen Tätigkeiten eigentlich sind. Diese werden in nächster Zukunft durch KI ersetzt.

Aber warum warnen eigentlich ausgerechnet die Vermarkter und Profiteure von KI so medienwirksam vor ihren eigenen Produkten? Und fordern wie Elon Musk gar eine Trainingspause für ihre besten LLMs, „weil sie die Kontrolle über unsere Zivilisation übernehmen könnten", oder vergleichen sich selbst mit den „Vätern der Atombombe", wie Sam Altman, dem Gründer von OpenAI. Zum einen wohl, weil sie selbst keinen Begriff von dem haben, was KI ist: Sie glauben tatsächlich, dass ihre LLMs generelle Intelligenz besitzen. Aber noch viel wichtiger: Sie präsentieren sich – im Vorgriff auf staatliche

Regulation - als verantwortungsbewusste Menschheitsbeglücker, und legen dabei den Turbogang in dem Hype um ihre Produkte ein.

An KI beunruhigt mich einzig die menschliche Intelligenz, die sie einsetzt, aber nicht die Aussicht, von Computern unterjocht zu werden.

Der Wissenschaftsnarr dankt Andreas Schneider für anregende Diskussionen.

Zen und die Kunst, Forschungsqualität zu bewerten

LJ 9/2023

Schon häufig hat der Narr auf diesen Seiten über mangelnde Qualität in der Wissenschaft gemäkelt. Auch ist er Sekretär eines Preises, bei dem eine internationale, hochkarätige Jury jährlich 500.000 € an Individuen, Gruppen, Institutionen, und auch junge Wissenschaftler vergibt, welche herausragend zur Verbesserung von Qualität in der Wissenschaft beigetragen haben („Einstein Foundation Award for promoting of quality in research“, links wie immer unter http://dirnagl.com/lj). Auf seinem Kreuzzug für mehr Forschungsqualität ist der Narr aber nicht allein, auch ist das Thema nicht wirklich neu. In seinen „Reflections on the Decline of Science in England, and on Some of Its Causes“ griff Charles Babbage, ein Multitalent und Erfinder des Vorläufers des modernen Computers, schon 1830 das wissenschaftliche Establishment, die Universitäten, und die Royal Society scharf an. Er prangere darin u.a. selektive Datenanalyse („Trimming“), unsaubere Statistik („Cooking“) und des Wissenschaftsbetruges („Forging“, „Hoaxing“) an. Die Qualität von Wissenschaft steht also schon recht lange auf dem Prüfstand.

Nicht nur Kritiker des Wissenschaftssystems interessieren sich für „Forschungsqualität“. So muss Forschung von hoher Qualität sein, um gefördert oder publiziert zu werden, wobei der Peer Review als weithin akzeptierte Qualitätskontrolle fungiert. Qualität bildet mit Originalität und Exzellenz die Trias der wesentlichen Kriterien, welche über Antragserfolg, Publikation, oder Karriere entscheiden. Aber wüssten Sie, „Forschungsqualität“ zu definieren? Falls Sie dabei Schwierigkeiten haben, sind Sie damit nicht allein. So hat sich kürzlich ein Panel von hochkarätigen Wissenschaftlern aus verschiedensten Disziplinen auf Einladung der Einstein Stiftung (die auch den oben genannten Preis vergibt) mit der Frage auseinandergesetzt, wie man Forschungsqualität definieren oder gar messen kann, und ob es hierfür gar disziplinäre Standards gibt. In Bezug auf diese Fragen herrschte jedoch keineswegs Einigkeit. Dabei ist die Definition von „Forschungsqualität“ keine rein theoretische Angelegenheit. Bei einem Hauptkriterium in der Beurteilung von Wissenschaftlern und deren Produkten sollte man eigentlich ziemlich genau wissen, mit welchem Maßstab man da urteilt. Sonst wird das Urteil

willkürlich bzw. geschmäcklerisch ausfallen, frei nach dem häufig gehörten Spruch: „Qualität erkennt man, wenn man sie sieht“. Damit macht man sich selbst zur Autorität in dieser Frage, ohne aber seine Karten auf den Tisch zu legen. In vielen Begutachtungen, die ich in den letzten Jahrzehnten mitgemacht habe, ist genau dies passiert.

Qualität kann man aber tatsächlich nur dann erkennen, wenn man einen Begriff, oder praktisch gesprochen eine Definition davon hat, was das sein soll. Es ist eben nicht wie bei Gefühlen, wie zum Beispiel Ärger, oder Glücklichsein – die erkennt man wirklich dann, wenn man sie hat. Wie aber können wir Qualität in der Forschung definieren? Wenn wir über Gegenstände des täglichen Lebens nachdenken, haben wir da zumeist keine Schwierigkeit. Was ist die Qualität einer Matratze? Einer Kaffeemaschine? Profis in solchen Fragen sind Organisationen wie die Stiftung Warentest. Sie definieren Qualität nach Güte der Verarbeitung, Haltbarkeit, heutzutage auch Nachhaltigkeit, und Nutzen für den Gebrauch (Wie schläft sich’s drauf? Wie lange dauert es, bis die erste Tasse rauskommt? usw.), oder das Preis/Leistungsverhältnis. Auf dieser Basis kann man dann sogar quantifizieren, verschiedene Produkte vergleichen, und für diese ein Ranking erstellen. Rankings machen die Gutachter in der Wissenschaft aber auch, nur auf welcher Grundlage? Wenn Qualität eines der Kriterien dabei ist, wie fassen wir diese? Und warum tun wir uns da so schwer?

Die Definition von „Qualität“, etwa in der Wikipedia, wonach Qualität die Summe bzw. Güte aller Eigenschaften eines Objektes, Systems oder Prozesses ist, hilft uns in ihrer Allgemeinheit nicht weiter. Robert M. Pirsig’s „Zen und die Kunst, ein Motorrad zu warten“, dem Klassiker der philosophischen Betrachtung des Qualitätsbegriffs, neutralisiert seine rationelle Betrachtung des Gegenstandes mit einem subjektiven, Zen-artigen "Im-Moment-Sein", und dies wollten wir ja gerade hinter uns lassen in der Forschungsbewertung. Nützlicher fand ich da Peter Dahler-Larsen Abhandlung „Quality – from Plato to performance“. Darin findet sich eine Auflistung der vielen Dimensionen und Bedeutungen, die „Qualität“ haben kann. Er findet nur einen einzigen gemeinsamen Nenner all dieser Definitionen – dass sie nämlich alle eine positive Konnotation haben. Qualität will man haben, ist etwas Gutes, wenn es an Qualität mangelt, haben wir ein Problem. Ausgehend von der Multidimensionalität des Begriffes betrachtet er dann verschiedene „Perspektiven“, unter denen man Qualität definieren oder analysieren kann: Qualität als Nützlichkeit, Qualität als Expertenmeinung, Qualität als Compliance mit Standards, Qualität als Impact, oder Qualität als Exzellenz, usw.

Und siehe da, wir finden da alle Verwendungen (Perspektiven!) des Qualitätsbegriffes in der Beurteilung von Wissenschaft wieder. Sie sind Experte? Dann wissen Sie was Qualität ist! Sie finden, dass Wissenschaft exzellent sein muss? Dann können Sie als „Experte“ ohne weitere Bestimmung Qualität und Exzellenz in eins fallen lassen, und ihr Urteil gleich al Gusto fällen. Sie halten den Impact Factor für ein gutes Kriterium für die Relevanz einer Publikation? Dann haben Publikationen in Journalen wie Nature, Cell oder New England Journal eine sehr hohe Qualität. Letzteres hat auch gleich den Vorteil, dass Sie Wissenschaftler ganz einfach ranken können, wie Matratzen oder Kaffeemaschinen bei der Stiftung Warentest. Wir halten fest: Der Qualitätsbegriff in der Wissenschaftsbewertung ist eine unausgesprochene, unreflektierte und damit intransparente Mischung aus diversen nicht näher definierten Perspektiven auf das, was als Forschungsqualität verstanden werden könnte. Und damit wenig brauchbar.

Von der Stiftung Warentest könnten wir lernen, dass Qualitätskriterien transparent sein müssen. In Bezug auf „Forschungsqualität“ bedeutet dies, dass wir eine Definition und daraus abgeleitete Kriterien brauchen, welche offen kommuniziert und auf alle Kandidaten oder Anträge eines Verfahrens gleichermaßen angewendet werden. Auch aus der

Leitlinienentwicklung in der klinischen Medizin könnten wir einiges lernen. Moderne Medizin ist evidenzbasiert. Nur wo es belastbare Evidenz dafür gibt, dass der Nutzen einer Behandlung deren Risiken übersteigt, darf sie eingesetzt werden. Kommissionen von Experten sichten dafür die zu einer Therapie jeweils verfügbaren Forschungsergebnisse, bewerten sie nach international konsentierten Kriterien, und sprechen dann eine Empfehlung aus, die auch negativ sein kann. Je nach Güte der vorhandenen Evidenz (z.B. kleine Studien zweifelhafter Qualität oder große randomisierte kontrollierte Studien) wird die Stärke bzw. Verbindlichkeit der Empfehlungen auch noch quantifiziert. Diese sog. GRADE-Methodik (Grading of Recommendations, Assessment, Development and Evaluation) wird von über 100 Organisationen (einschließlich der Weltgesundheitsorganisation) befürwortet und/oder verwendet, um die Qualität von Evidenz und die Stärke von Empfehlungen im Gesundheitswesen zu bewerten.

Wie wäre es, wenn sich Institutionen und Fördergeber auf einen transparenten und vergleichbaren Ansatz zur Bewertung von Forschungsqualität einigen könnten? Zunächst müsste man sich einigen, was man unter Forschungsqualität verstehen möchte. Trotz der oben geschilderten Multidimensionalität und Vielfalt von Perspektiven glaube ich, dass dies, zumindest in den Lebenswissenschaften, eine ziemlich gradlinige Sache wäre, und dies auch konsensfähig. Wenn wir uns ganz grundsätzlich darauf verständigen können, dass Wissenschaft verantwortungsvoll sein muss, könnten wir dies als unsere Perspektive festlegen. Hier gleich ein Vorschlag für relevante Dimensionen, die sich hieraus ableiten ließen: Robustheit der Ergebnisse, Transparenz im Forschungsdesign und in der Veröffentlichung der Ergebnisse, Nützlichkeit für die Wissenschaft oder die Gesellschaft, sowie Ethik (für Tier und Mensch). Für jeder dieser Dimensionen lassen sich ganz zwanglos und einfach Bewertungskriterien ableiten, welche selbstverständlich kontextabhängig sein müssen, also z.B. teilweise anders für Grundlagenforscher und deren Produkte als für Psychologen oder klinische Forscher.

So sind Forschungsergebnisse robust, wenn sie von hoher interner und externer Validität sind. Also in der präklinischen Forschung z.B. randomisiert und verblindet durchgeführt wurden, und dies mit hoher methodischer Kompetenz, am besten in verschiedenen Spezies oder experimentellen Settings. Wenn Studien- und Analysenprotokolle präregistriert werden, ist das Vorgehen und die Auswertung transparent festgelegt. Praktiken wie selektive Datenanalyse („Cherry picking“), Hypothesenbildung nach Erhalt der Ergebnisse („HARKING“) oder statistische Tricksereien („p-Hacking“) werden damit erschwert. Forschung ist nützlich für Wissenschaftler, wenn Sie ihre Daten zur Nachnutzung veröffentlichen, und nützlich für die Gesellschaft (Patienten), wenn die klinischen Forscher diese in die Planung ihrer Studien involvieren. Ethisch ist Forschung für Patienten, wenn Patienten-Nutzen und möglicher Schaden in klinischen Studien im Gleichgewicht stehen, bzw. für Tiere, wenn deren Leid minimiert wird bzw. ganz auf ihre Verwendung verzichtet wird (3R).

Und schon hätten wir ein Set von Kriterien, mit denen die Forschungsqualität einer Studie, eines Antrages (in Bezug auf die Vorarbeiten und das Arbeitsprogramm), oder das Oeuvre eines Forschers bewertbar wird. Und zwar mit transparenten, überprüfbaren Kriterien, die auf alle Bewerber oder Anträge eines Verfahrens angewendet werden können. Ähnlich wie bei der GRADE Methode könnte man dann die einzelnen Dimensionen in einer „Bewertung” zusammenfassen (von „Höchste Qualität“ in allen Domänen zu „unklare/fragwürdige Qualität”, oder „nicht beurteilbar“).

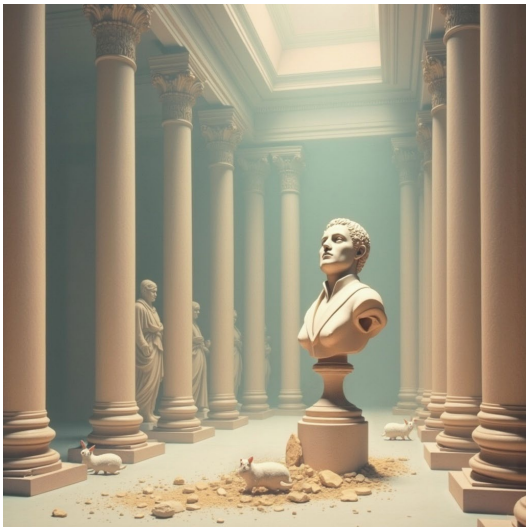
Antragsteller, Anträge oder Artikelsubmissionen niederer Qualität könnte man dann sofort aussortieren, und bräuchte sie gar nicht mehr nach Originalität oder Relevanz zu befragen. Forschung niedriger Qualität kann keine relevanten Ergebnisse erzielen, da

hilft es auch nichts, wenn sie originell ist, oder die Antragsteller schon mal in ganz tollen Journalen publiziert haben. Überhaupt sind Originalität und Relevanz Kategorien, die sich viel stärker einer Operationalisierung entziehen und wesentlich subjektiver bewertet werden müssten als Forschungsqualität. Hinzu kommt, dass Relevanz eine zeitliche Dimension hat, die ihre Bewertung fast verunmöglicht. Was heute als irrelevant erscheint, kann schon morgen hoch relevant werden, die Wissenschaftsgeschichte liefert hierfür unzählige Beispiele. Das gleich gilt natürlich auch für gehypte Projekte, für die sich nach wenigen Jahren niemand mehr interessiert. Die Bewertung von Originalität ist sogar noch subjektiver, und diskriminiert Ergebnisse, die konfirmativ sind – und von denen wir viel zu Wenige haben.

Die von mir oben angebotene Definition von Qualität der Forschung dagegen operationalisiert und objektiviert diese, sie ist dadurch in eine konkrete und bis zu einem gewissen Grad sogar messbare Form gebracht. Es braucht dafür nicht mehr als ein einfaches Formular. Einen gravierenden Nachteil hat mein Vorschlag aber: Um Qualität mit Hilfe des Formulars zu bewerten, müsste man den Antrag oder das Paper gelesen haben. Querlesen mit Blick auf die Affiliation (Harvard? Stanford?) oder den Namen des Journals hilft nicht weiter. Weshalb er wohl niemals zur Umsetzung kommen wird.

Der Fall T.-L.: Acht Lektionen aus einem ganz gewöhnlichen Wissenschaftsskandal

LJ 10/2023



Mit großer Fanfare ist im Juli diesen Jahres der Präsident der Stanford Universität Marc Tessier-Lavigne zurückgetreten. Die Weltpresse berichtete darüber auf den Titelseiten. Eine hochrangig besetzte externe Kommission war zu dem Schluss gekommen, dass Tessier-Lavigne zwar wohl keinen direkten Wissenschaftsbetrug begangen hatte. Es waren aber in seinem Labor und unter seiner Verantwortung über viele Jahre Daten und Bilder manipuliert worden, Abbildungen in Artikeln unsachgemäß kopiert, eingefügt oder geschnitten. Ergebnisse wurden dupliziert und als separate Experimente ausgegeben worden sowie Kontrollen mehrfach verwendet, und in einigen Fällen - in denen auch eine Absicht zur Verschleierung der Manipulation festgestellt wurde - waren Panels in Weisen gestreckt, gedreht und bearbeitet worden, die die veröffentlichten experimentellen Daten veränderten. Dazu kamen noch ganz grundlegende Fehler in der basalen Biostatistik. Die Kommission hatte 12 Artikel, publiziert in Journals wie Nature, Cell, Science, untersucht. In allen fand sich jede Menge schlechte wissenschaftliche Praxis, und dazu noch einiges an offensichtlichem Wissenschaftsbetrug.

Wer Retraction Watch verfolgt oder regelmäßig die Wissenschaftsseiten der Tagespresse liest, wird das nicht vom Hocker hauen. Eine fast schon gewöhnliche, ja mittlerweile langweilige Mischung von Daten- und Bildmanipulation mit denen Wissenschaftler spektakuläre Stories für Topjournals stricken. Der Narr hat erst kürzlich (Laborjournal 1-2, 2023) darauf hingewiesen, dass all dies mittlerweile zur Normalität im wissenschaftlichen Publikationswesen gehört. Aufmerksamkeit erhielt dieser Fall nur deshalb, weil an der Spitze des Skandals – andere Personen wurden gar nicht genannt oder zur Verantwortung gezogen – einer der prominentesten und mächtigsten Wissenschaftler der USA stand.

Zum Material für den Narren wird die Angelegenheit aber nicht deshalb. Vielmehr ist die Sache gerade in ihrer Gewöhnlichkeit interessant und relevant. Denn sie liefert uns eine Reihe von Lektionen über unser System Wissenschaft, z.B. bezüglich wissenschaftlicher Integrität, dem Verhalten von wissenschaftlichen Institutionen und akademischen Journalen im Angesicht von offensichtlichen Fehlern in Veröffentlichungen, dem Versagen des Review Prozesses und den Stärken des Post-Publication-Review, über Interessenkonflikte von Wissenschaftlern und Kommissionsmitgliedern, den Umgang von Wissenschaftlern mit ihren Fehlern, die Rolle des Journalismus wenn es um die ‚Selbstkorrektur' der Wissenschaft geht, die Effektivität von studentischem Aktivismus, die Übernahme von Verantwortung für wissenschaftliche (Miss)Erfolge, die Leitung wissenschaftlicher Arbeitsgruppen, und natürlich das Prinzip „publish or perish".

Zunächst aber: was war eigentlich passiert? Der Star unserer Geschichte ist Marc Tessier-Lavigne, ein frankokanadischer Neurowissenschaftler, zunächst an der University of California San Francisco, dann Präsident der Rockefeller University, danach Executive Vice President for Research and Chief Scientific Officer bei Genentech, Mitglied der National Academy of Science, und seit 2016 Präsident der Stanford University. Aber er beeindruckt nicht nur als akademischer Führer: Pubmed listet 283 wissenschaftliche Publikationen, ein großer Teil davon in Glamjournals wie Cell, Nature, Science. Google Scholar findet auf seine Artikel über 80.000 Zitationen und belohnt ihn mit einen Hirsch-Faktor von 152. Ein academic superhero also! Zitate über Tessier-Lavigne aus einem „Profile" Artikel über ihn in Nature Medicine: "Er hat diese erstaunliche Fähigkeit, das zu erledigen, was erledigt werden muss."; „Die Geschichten über Marc handeln im Grunde alle davon, dass Marc perfekt ist."; "Er ist einer jener Menschen, bei denen die Dinge anscheinend [sic] immer genau richtig liefen."

Auftritt Theo Baker, Stanford Student im 2. Jahr und Redakteur der Studentenzeitung Stanford Daily. Im November 2022 schreibt er einen Artikel darüber, dass es auf PubPeer, einer Post-Publication-Review Website, eine Menge Kommentare zu Artikeln gibt, die Tessier-Lavigne mitverfasst hat. Diese wiesen auf duplizierte, verschobene oder anderweitig manipulierte Blots und weitere Ungereimtheiten in den Arbeiten hin. Baker und Kommilitonen recherchierten weiter, es folgten noch mehr Berichte im Stanford Daily über die wachsende Liste von fragwürdigen Studien sowie möglicherweise gefälschte Ergebnisse in einer Nature Studie von 2009. Damals war Tessier-Lavigne führender Wissenschaftler bei der Biotech-Firma Genentech. Der Artikel beschreibt, dass Amyloid Precursor Protein an den Death Receptor 6 bindet und verkauft dies als einen neuen, vielversprechenden Ansatz für die Alzheimer Therapie. Schlüsselbefunde dieser Studie konnten in der Folge nicht reproduziert werden, und Genentech ruderte zurück.

Die Standford University wollte sich der studentischen Berichterstattung – auf die weltweit sichtbaren und diskutierten Kommentare auf PubPeer war bisher gar nichts geschehen- mit einer internen Untersuchung entledigen. Damit bei der Untersuchung nichts anbrennt, wählte man ein Ausschuss-Mitglied aus, der Gründer einer Investmentfirma

war, die einen Anteil von 18 Millionen US-Dollar an einem Unternehmen besaß, das von Tessier-Lavigne gegründet wurde. Erst eine Anfrage des Stanford Daily führte dazu, dass dieser „mögliche Interessenkonflikt" offenbart wurde.

Jetzt erst war die Uni gezwungen eine hochrangig besetzte externe Untersuchungskommission einzusetzen, der auch „Digital Image Forensiker“ angehörten. Der Kommissionsbericht, fast 100 Seiten stark (der Link hierzu, wie auch zu anderen Quellen, wie immer unter http://dirnagl.com/lj) wurde am 17. Juli 2023 veröffentlicht. Die Kommission war nicht überall freundlich empfangen worden. Sie fand manch unkooperativen Ko-Autor vor, aber auch frühere und jetzige Mitarbeiter, die aus Furcht vor Repression keine Aussagen machen wollten. Fazit der Untersuchung war, dass keine Beweise für eine „direkte Datenmanipulation“ durch Tessier-Lavigne gefunden wurden. Aber er hatte eine Umgebung gefördert, die zu einer "ungewöhnlichen Häufigkeit der Manipulation von Forschungsdaten und/oder mangelhaften wissenschaftlichen Praktiken" in Labors an mehreren Einrichtungen geführt hatte. Als beginnend 2001 (!) verstärkt Bedenken hinsichtlich seiner Veröffentlichungen auftraten hatte er es versäumt, Fehler im wissenschaftlichen Kontext „eindeutig und entschlossen“ zu korrigieren. Er generierte eine Kultur im Labor, in der „Gewinner“ (Postdocs, die günstige Ergebnisse erzielen konnten) belohnt wurden und die „Verlierer“ (Postdocs, die nicht in der Lage waren oder Schwierigkeiten hatten, solche Daten zu erzeugen) zu marginalisieren oder abzuwerten. Sehr diplomatisch formulierte die Kommission abschließend: „Die ungewöhnliche Häufigkeit der Manipulation von Forschungsdaten und/oder mangelhafter wissenschaftlicher Praktiken durch verschiedene Personen zu verschiedenen Zeiten in Labors, die von Dr. Tessier-Lavigne an verschiedenen Einrichtungenbetreut wurden, deutet darauf hin, dass es möglicherweise Gelegenheiten zur Verbesserung der Laborüberwachung und des Managements gegeben haben könnte.“

Was lernen wir nun aus dieser Affäre? Zunächst einmal, dass sich meist weder Autoren, noch Journals, noch Institutionen proaktiv an der Aufklärung von Auffälligkeiten in Studien beteiligen, wenn solche publik gemacht werden. Seit fast 10 Jahren wurden Fehler und mögliche Verstöße gegen die gute wissenschaftliche Praxis in Tessier-Lavignes Arbeiten öffentlich diskutiert. Oft wird im Zusammenhang mit Skandalen wie diesem darauf hingewiesen, dass gerade diese doch belegen, dass die Wissenschaft ‚selbst korrigierend‘ ist. Stimmt schon, nur dauert es meist über 10 Jahre bis dies passiert. Eine Schlüsselfigur und Ko-Autorin im Tessier-Lavigne Fall ist übrigens mittlerweile Professorin an einer anderen sehr prominenten US-Uni geworden – auch mit und durch die betrügerischen Artikel. Bis zur Korrektur vergeht nicht nur sehr viel Zeit, sie stellt sich in der Regel auch nur in prominenten Fällen wie diesem ein, bei denen sich letztlich eine Armee von Amateur- und Profi-Forensikern über die Arbeiten her machen. Dabei stand alles schon seit 2015 auf PubPeer. Aber solche Post-Publication-Reviews werden in der Regel nicht nur ignoriert, sondern aktiv unterdrückt.

Gleichzeitig lernen wir wie nützlich der Post-Publication-Review ist. Allerdings: Schon beim „normalen“, anonymen Pre-Publication-Review bekommen die Referees keinen akademischen Kredit, und die Community sieht nicht, was diese über den Artikel zu sagen hatten. Beim Post-Publication-Review liegt alles offen, häufig auch die Identität derer, die sich die Mühe gemacht haben, genau hinzusehen. Dafür gibt’s auch keinen Kredit, dafür aber im schlimmsten Fall sogar juristische Konsequenzen. Wie gerade eben wieder geschehen: Die Autoren des Wissenschaftsblogs „Data Colada“ Leif Nelson, Joe Simmons, und Uri Simonsohn werden gerade von Francesca Gino vor Gericht auf 25 Mio US$ Schadenersatz verklagt. Sie hatten in einer Serie von 4 Beiträgen auf ihrem Blog in einem sehr aufwendigen und kompetenten Post-Publication-Review aufgedeckt, dass die prominente Professorin der Harvard Business School in großem Stil Daten

manipuliert hatte. Auf der Crowdfunding Website GoFundMe versuchen Sympathisanten deshalb gerade, die vermutlich fälligen bis zu 600.000 US$ Anwaltskosten für den Prozess hereinzuholen. Über 300.000 US$ haben sie schon zusammen!

Der Fall T.-L. birgt auch eine Lektion über den Umgang mit Fehlern. Falls es stimmt, dass Tessier-Lavigne nicht aktiv an den Schlampigkeiten, Manipulationen, und Betrügereien beteiligt war, hätte er doch lange genug Zeit gehabt, sich um deren Aufklärung zu bemühen, sowie Corrections oder Retractions zu veranlassen. Fehlanzeige. Denn Fehler werden in top Laboren nicht gemacht! Aber wir alle wissen, dass wir Fehler machen, in der Forschung und beim Publizieren. Aus Furcht vor Sanktionen oder geschädigtem Ruf werden diese aber unter den Teppich gekehrt, abgestritten, usw. Statt dass Wissenschaftler für positive Fehlerkultur belohnt werden, überziehen wir Sie, falls doch mal was rauskommt, mit Häme. Mir jedenfalls ist eine ordentliche Retraktion nach einem „honest error" viel lieber als ein schlechtes Nature Paper (siehe auch der Wissenschaftsnarr dazu: „Es irrt der Mensch, solang er strebt", LJ 1-2, 2019)

Wir haben hier auch ein schönes Beispiel dafür vor uns, dass Studenten oft mehr „Power" haben, als wir oder sie sich selbst zutrauen. Ohne die Redakteure der Studentenzeitung würde die Angelegenheit weiterhin in PubPeer begraben liegen. Es half dabei sicher, dass Theo Baker der Sohn des Chefkorrespondenten der New York Times im Weißen Haus ist. Aber nicht nur Professoren, auch Studenten haben eben häufig potente Netzwerke.

Im Umfeld von Tessier-Lavigne geht es auch um viel Geld. Um fast 30 Patente, auf denen er laut Google Patents gelistet ist, aber auch die finanziellen Interessen von Genentech und Stanford University. Außerdem gibt es da Firmen, die er gegründet hat oder bei denen er im Board of Directors sitzt. Ein äußerst fruchtbarer Boden für multiple Interessenkonflikte, bei der Forschung, und noch mehr bei der Aufarbeitung von möglichen Verfehlungen. Dass dies nicht nur Theorie ist, hat sich bei der Besetzung der Untersuchungskommission mit einem Geschäftspartner von Tessier-Lavigne gezeigt.

Der Fall T.-L. wirft auch ein Schlaglicht auf etwas, über das (zu) wenig gesprochen wird: Die Leitung von großen Forschergruppen durch Personen, die durch Administration oder andere Tätigkeiten eigentlich viel zu wenig Zeit haben, sich vor Ort mit der nötigen Expertise um die Betreuung der Forschung zu kümmern. Die im Wesentlichen für die Politics, die Ressourcen und das Renommee zuständig sind, aber so tun, als wären sie selbst noch die genialen Ideengeber und Garanten für Forschungsqualität. Diese Fiktion wird nolens volens von den PhD Studenten, Postdocs, und aufstrebenden AG-Leitern vor Ort mitgetragen. Auf eine unausgesprochene Abmachung vertrauend hoffen sie nämlich, im Windschatten und mit dem Netzwerk des Chefs Karriere zu machen. Und das klappt wiederum meist nur für die „Gewinner", welche Ergebnisse liefern, die das Material für spektakuläre Stories sind, wie das die Stanford-Kommission so schön formulierte.

Dieser Typus Chef, von dem Tessier-Lavigne wohl ein Rollenmodell war, sonnt sich im Licht der Papers aus ihren Laboren, steht auch auf großen Anträgen gerne vorne – das hilft ja auch unbestritten bei deren Akzeptanz. Sobald es allerdings Probleme gibt, und Vorwürfe im Raum stehen, weisen sie jede Verantwortung von sich. Sie behaupten dann eilfertig, sie hätten von nichts gewusst und könnten doch eh nicht alles überprüfen, was da in ihrem Großbetrieb so geforscht wird. Außerdem sei Wissenschaft Vertrauenssache, dazu waltet doch Forschungsfreiheit. Nun wird plötzlich mit der Wahrheit argumentiert: sie wissen gar nicht was in ihren Laboren abgeht. Und diese Ausrede funktioniert sogar: Tessier-Lavigne wurde nicht für Verstöße gegen die gute wissenschaftliche Praxis gerügt, sondern für seine Schwäche als akademischer Führer. Deshalb musste er als Führer

einer Uni zurücktreten. Der Wissenschaftler M.-L. kommt mit einem leichten Imageschaden davon.

Die ultimative Lehre aus dem Fall T.-L., und gleichzeitig die nicht überraschende Erklärung für alles, was das so passiert ist: Ein Wissenschaftssystem das zuvörderst nicht auf einer Wissens-, sondern auf einer Reputationsökonomie basiert, muss notwendig solche Skandale produzieren. Deshalb kommen die auch immer wieder. Wir sollten uns aber nicht täuschen – nicht diese sind das eigentliche Problem. Sondern die tägliche Praxis, die daraus für uns alle resultiert, die in einem solchen System forschen. Publish or perish.

Wissenschaftler und Bibliothekare, hört die Signale: Keine DEALs mit unseren Papers!

LJ 11/2023

Wissenschaftler forschen eine Weile, und bringen dann, wenn sie glauben, dass es so weit ist, ihre Resultate in Artikelform unter die Kollegen. Ihre Texte formatieren sie vorher sauber, ein Muttersprachler, DeepL oder ChatGPT hübschen diese noch sprachlich auf, auch kommen anschauliche Graphen und Abbildungen dazu, bis es ‚print ready' ist. Dann wird das Ganze von Wissenschaftlerkollegen kritisch beäugt, ob das denn alles so seine Richtigkeit hat. Viele von denen agieren dabei nicht nur als Reviewer, sondern o rganisieren den Prozess auch noch als Editoren. Wenn der Artikel ein paar Runden gedreht hat, vielleicht noch ein paar Experimente gemacht und dann alle Beteiligten ihr OK gegeben haben, wird der Artikel auf den Internetseiten eines Journals veröffentlicht. Fällt Ihnen auf, dass in diesem Zyklus nur Wissenschaftler vorkommen, aber keine Verlage? Die bräuchte es nämlich bei keinem der genannten Schritte. Die komplette ‚Wertschöpfungskette' der Wissenschaft Forschen – Artikel schreiben – Artikel reviewen findet innerhalb von Academia statt. Trotzdem sind die Verlage die ‚Gatekeeper' des wissenschaftlichen Publizierens, und das kommt uns im wahrsten Sinne des Wortes teuer zu stehen. Wissenschaftsverlage beschäftigen uns Forscher als Autoren, Gutachter und Herausgeber, bezahlen uns für unsere Tätigkeiten aber nicht. Sie verlangen von uns Autoren, die Urheberrechte unserer Artikel an sie abzutreten. Und dann lassen sie die Universitätsbibliotheken die mit Steuergeldern finanzierten Artikel zurückkaufen. Und jetzt tracken sie auch noch unsere Tätigkeiten und verscherbeln unsere Daten.

Verlage waren nötig, als Artikel und Bücher noch verlegt und in Printform verteilt werden mussten. Ohne einen Verleger wie Elzevir hätte Galileo sein ‚Systema cosmicum' nicht veröffentlichen können. Ohne einen Wissenschaftsverlag wie Nature Watson und Crick nicht ‚A structure for deoxyribose nucleic acid' publizieren. Hätte es damals schon Preprints und Diamond Open Access (OA) gegeben, wären wir vermutlich früher auf dem Mond gelandet, oder hätten das Genom sequenziert.

Nun sind die Zeiten von gedruckten und nur von kommerziellen Verlagen kuratier- und disseminierbaren wissenschaftlichen Artikeln schon eine ganze Weile vorbei. Dennoch zahlen Volkswirtschaften weltweit jährlich Milliarden an Elsevier und Co., die eigentlich für Forschung ausgeben werden könnten. Wir vergeuden dabei oft auch Zeit und Nerven im ‚Kaskadieren' unserer Artikel, indem wir dem Gradienten abnehmender Impactfaktoren mit multiplen Submissionen folgen. In der Musik-, und Transportindustrie, im Handel, beim Reisen etc. hat sich in den letzten Dekaden ein massiver Wandel hin zur Digitalisierung vollzogen. Das akademische Publizieren dagegen ist in den 50er Jahren des vorigen Jahrhunderts steckengeblieben. Und dies, obwohl das Substrat der Artikel, unsere Ergebnisse, mittlerweile praktisch zu 100 % digital generiert und kommuniziert wird.

Weil wir so beschäftigt mit Forschen waren, und unsere Bibliotheken das Finanzielle diskret im Hintergrund erledigt haben, ist den meisten von uns gar nicht aufgefallen, dass das Publikationswesen Unsummen verschlingt. Wir klickten in Pubmed oder einer Literaturliste auf einen Link – und schon öffnete sich wie von Zauberhand der Artikel auf dem Monitor. Nicht nur war vielen von uns nicht klar – und ist dies vielleicht immer noch nicht – dass unsere Bibliotheken sehr viel Geld für die Subskription dieser Zeitschriften bezahlt haben – vermutlich um die 4000 – 5000 US$ pro Artikel im Subskriptionssystem. Und dass der Großteil der Menschheit (insbesondere Forscher in nicht so begüterten Staaten, oder niedergelassene Ärzte, oder Patienten) gar nicht an diese Artikel rankommen. Denn es werden schnell mal 40 € (Cell) oder 150 € (New Engl J Med) fällig, wenn man sie lesen wollte, ohne eine institutionelle Bibliothek im Hintergrund, welche die Rechnung diskret schon beglichen hat! Das ist viel Geld für einen niedergelassenen Arzt der über eine neue Behandlung nachlesen will, oder einen Forscher in Ghana.

Forschungsorganisationen erhoben deshalb vor genau 20 Jahren in der „Berliner Erklärung" die naheliegende Forderung, dass öffentlich finanzierte Forschung auch für die Öffentlichkeit zugänglich sein sollte! Da kam bei den Verlagen kurz die Panik auf, dies könnte ja das Ende ihrer immer weiter steigenden Profitraten bedeuten. Abe inspiriert durch OA-Idealisten fanden sie ein neues Geschäftsmodell: Statt mit den Subskriptionsgebühren die Bibliotheken zu schröpfen, wollen sie nun gleich an der Quelle ansetzen, bei den Autoren! Die APCs, die Article Processing Charges, waren geboren, nach deren Bezahlung Artikel OA, also überall und für jedermann frei zugänglich gemacht werden.

Diese universelle Verfügbarkeit von OA Artikeln ist natürlich toll. Das Geschäftsmodell dahinter aber nicht. Nun können zwar Wissenschaftler weltweit auf OA – Artikel zugreifen. Aber die in sog. ‚low and middle income countries' haben Probleme beim Publizieren, weil sie sich trotz „Waiver" –Programmen mancher Verlage die APCs kaum leisten können. Das ist den wenigsten Forschern in Deutschland und anderen ‚reichen' Ländern klar. Aber auch uns belasten die APCs natürlich massiv. Die APCs, welche schon mal mit 10.000 € zu Buche schlagen können, haben die Verlage nämlich so berechnet, dass der liebgewordene Profit langfristig durch hyperinflationäre Steigerungen und zusätzliche ‚Data Analytics' auf Basis des Trackings unserer Aktivitäten sogar noch erhöht werden kann. Und das ist selbst für die Großkonzerne ein ambitioniertes Programm. So hatte Elsevier, das dem Mediengiganten (man könnte treffender auch sagen der ‚Datenkrake') RELX gehört und Titel wie Cell und Lancet veröffentlicht, im Jahr 2005 eine Umsatzrendite von 31 %, 2022 lag sie dann schon bei 38 %.

Das Geschäftsmodell der APCs beruht genauso wie das der Subskriptionen darauf, dass die Gesellschaft Verlage zum zweiten Mal für Produkte und Services bezahlt, für welche sie die Wissenschaft bereits alimentiert hat – nämlich Forschen, Resultate

zusammenschreiben, in Manuskriptform bringen und deren Qualität kontrollieren. Im Strudel der Begeisterung, dass dadurch viele Titel öffentlich zugänglich werden, hat man das wohl übersehen. Und so ändert sich nichts am System. Vermutlich vor allem deshalb, weil die Wissenschaftler zusammen mit den Verlagen eine komplexe Hierarchie der Journale aufgebaut haben. Das Renommee der Zeitschriften, meist gemessen mittels des Impact Factors, dient als Währung in der Reputationsökonomie des Wissenschaftsmarktes. Diese Währung tauschen Forscher gegen Fördermittel, Stipendien, Professuren, etc. Dieses sich selbst stabilisierende System würde zusammenbrechen, wenn es zu einer ‚Währungsreform' käme. Denn dann würde nicht mehr die Anzahl von Publikationen und das ‚Renommee der Journale' (d.h. der Impact Factor) bewertet, sondern der Inhalt, die Qualität und der wissenschaftliche Impact von Forschung und zur Grundlage von Karriereentscheidungen und Forschungsförderung gemacht.

Ein weiterer Grund, warum die Verlage derzeit nicht um ihre Zukunft fürchten müssen, ist die Trägheit des Wissenschaftssystems und unsere eigene Bequemlichkeit. Wir werden fürs Forschen bezahlt, und nicht dafür, das akademische Publikationswesen zu reformieren. Viele schimpfen zwar über ‚publish or perish', aber ändern sollen das allenfalls die Anderen, die Institutionen, die Fördergeber, usw. Diejenigen, die das System gestalten – und es somit auch ändern könnten, also die arrivierten Wissenschaftler, wurden in ihm sozialisiert und selektiert. Wenn Du einen Sumpf austrocknen willst, dann solltest Du nicht die Frösche fragen!

Aufgeregt haben wir uns nur darüber, dass man jetzt die APCs selber bezahlen muss, wo doch vorher die Bibliotheken die Rechnung beglichen haben. Das tun sie deshalb jetzt häufig auch wieder, über einen von der DFG aufgelegten und vom Steuerzahler bezahlten Publikationsfonds. Aber auch wenn die APCs von den Wissenschaftlern privat bezahlt würden, ist das Karriere-ökonomisch gesprochen gut investiertes Geld: Was sind schon 5390 €, aus Forschungsmitteln bezahlt, für das Label ‚Nature Communications'. Damit steigen doch die Chancen beim nächsten Antrag, und allemal veredelt es den Lebenslauf für die nächste Bewerbung.

In der wohlmeinenden Absicht, OA weiter zum Durchbruch zu verhelfen, aber die Autoren nicht mit APCs zu belasten, hat 2014 die Allianz der deutschen Wissenschaftsorganisationen die Hochschulrektorenkonferenz beauftragt, bundesweit neue Vertragsmodelle zu verhandeln. Das daraus entstandene DEAL Konsortium hat sich mit den Verlagsriesen und Data Analytics Giganten Springer Nature (2020) und Wiley (2019) geeinigt. Diese Verlage öffneten daraufhin einen Großteil ihrer Journal-Portfolios für die Wissenschaftler der am DEAL beteiligten Forschungseinrichtungen, auch können diese ohne APCs in diesen Journalen veröffentlichen. Das Tolle für die Verlage daran: Die fälligen Gebühren berechnen sie nach der Anzahl der publizierten Artikel, und zwar so, dass sie mindestens so viel verdienen, wie vorher mit den Subskriptionen. Natürlich mit garantierten jährlichen ‚hyperinflationären' Steigerungen. Die Kosten dafür werden auf die jeweiligen institutionellen Bibliotheken umgelegt, diese stehen damit wegen der exorbitanten Zahlungen an die Verlage mindestens so prekär da wie vorher im Subskriptions-System.

Der DEAL bringt für die Großverlage noch einen weiteren, nicht zu unterschätzenden kommerziellen Vorteil: Wissenschaftler publizieren jetzt natürlich bevorzugt in Journalen, bei denen sie wegen eines DEAL Vertrages nichts bezahlen müssen. Kleinere, Nicht-DEAL Verlage leiden darunter und sind deshalb schon vors Bundeskartellamt gezogen. Sie fanden dort aber kein Gehör. Das trifft ausgerechnet Verlage wie EMBO Press, PLOS, Elife. Das ist tragisch, denn diese agieren nicht gewinnorientiert (non-profit), legen ihre Kosten offen, und generieren wirklichen Mehrwert für das akademische

Publikationswesen. Diese Verlage entwickeln und testen nämlich neue und bessere Reviewverfahren (post-publication, open, commenting etc.), neue Publikationsformen (Preprints, Registered reports, etc.), und andere Innovationen (Community action publishing statt APCs, etc.), welche die ‚Großen' dann bequem übernehmen können, nachdem sie von ihren kleinen Konkurrenten erfolgreich etabliert wurden.

Mit Elsevier, das sich selbst als globales Unternehmen für Informationsanalysen bezeichnet, verhandelt das DEAL Konsortium schon seit 2016. Wie man es von der anerkannten ‚dark force' des Publikationswesens nicht anders erwarten konnte, ist bis vor kurzem wegen der exorbitanten Forderungen von Elsevier aber kein DEAL zustande gekommen. Dabei trieb es Elsevier so toll, dass 2016/2017 statt eines DEALs 200 wiss. Einrichtungen in Deutschland ihre Subskriptionen bei Elsevier nicht mehr verlängerten! Die Charité z.B. nahm Elsevier vor 5 Jahren vom Netz. Und siehe da, es gibt sie immer noch, ihre Wissenschaftler, nun ohne institutionellen Zugang, publizieren nach wie vor in Elsevier-Journalen. Sie besorgen sich Elsevier-Artikel über die Autoren, Kollegen in Einrichtungen welche noch Zugang haben, Fernleihe, oder sog. Schattenbibliotheken, wie z.B. SCI-HUB. Die Elsevier- Kündigung hat der Charité eine Menge Geld gespart, die sie jetzt gezielt für Grundausstattung oder Kofinanzierung ihrer Wissenschaftler einsetzen konnte. Oder damit auch die fälligen Preissteigerungen der subskribierten und DEAL Journale bezahlt!

Am 6. September hat das DEAL-Konsortium nun überraschend und mit großer Freude doch einen „Abschluss eines transformativen Open Access-Vertrags mit dem internationalen Fachverlag Elsevier" bekannt gegeben. Das dahintersteckende Finanzierungsmodell wird damit auf die nächsten 5 Jahre festgeschrieben. DEAL stabilisiert und perpetuiert hier ein weiteres Mal ein aus der Zeit gefallenes akademisches Publikationswesen, und dazu noch die Reputationsökonomie der Wissenschaft sowie die Verschwendung gesellschaftlicher Ressourcen.

Es gibt jedoch einen Silberstreifen am Horizont, denn für den Elsevier Vertrag gilt opt-in für einzelne Institutionen. Dies ist eine phantastische Gelegenheit für diese, in nicht zu unterzeichnen und damit auch ein Signal zu setzen. Neben dem prinzipiellen Motiv, das Publikationswesen von Grund auf zu reformieren und die Verschwendung von Ressourcen zu stoppen, gibt es aktuelle, zusätzliche Argumente dafür, dass der Zeitpunkt ideal ist. Dazu zählt, dass in den zurückliegenden Jahren, in denen viele Institutionen weder DEAL noch Subskription mit Elsevier hatten, keinerlei negative Auswirkungen auf das Wissenschaftssystem bemerkbar waren. Außerdem gebietet die gegenwärtige angespannte Haushaltslage in den öffentlichen Wissenschaftseinrichtungen eine Ökonomisierung der Ressourcen. Wir können es uns einfach nicht leisten, Großverlage satt zu füttern, aber gleichzeitig selbst zu wenig Mittel für Forschung zu haben.

Insbesondere sollten wir unsere Forschungsgelder auch nicht dafür ausgeben, uns überwachen zu lassen. Aber genau dies tut Elsevier, man kann sogar sagen, dass dies das Kerngeschäft des Konzerns ist. Verharmlosend nennt sich dies ‚Data Analytics Business'. Eine Schlüsseltechnologie hierbei ist umfangreiches User Tracking, das auf allen Elsevier-Plattformen erfolgt. Elsevier weiß, wer, wo, wie viel, mit wem und worüber forscht. Im DEAL ist festgelegt, dass die Verarbeitung der Elsevier zugänglich gemachten Daten entsprechend Elseviers Privacy Policy und Data Processing Terms erfolgt. Diese Informationen sind, insbesondere, wenn mit anderen Daten verlinkt, Gold wert. Sogar die US-Immigrationsbehörde zählt zur Kundschaft. Das so verdiente Geld wird passenderweise dann auch in Firmen wie Palantir investiert, einem US-Anbieter zur Analyse großer Datenmengen und Dienstleister für internationale Nachrichtendienste. Und damit nicht genug: Die Climate Rights Coalition legte offen, dass Elsevier Daten, Analysen und

Informationen bereitstellt, die die Erkundung und Entwicklung von fossilen Brennstoffen weltweit fördern. Auch bietet RELX den meisten Fortune-500-Unternehmen der Öl- und Gasbranche Ressourcen und Tools, um ihre klimaschädigenden Geschäfte auszuweiten. Darunter sind auch Kunden und Partner von Kohleunternehmen, die immer noch die Entwicklung von unverbrennbaren Kohlereserven ausweiten und sich weigern zu dekarbonisieren.

Vermutlich denken Sie sich jetzt: Oh je, wieder der Wissenschaftsnarr mit seinen aufrührerischen Gedanken! Weit gefehlt, ich befinde mich in bester Gesellschaft. Die EU Wissenschaftsminister haben die „Mitgliedstaaten und die Kommission ermutigt, in interoperable, gemeinnützige Infrastrukturen für die Veröffentlichung auf der Grundlage von Open-Source-Software und offenen Standards zu investieren und diese zu fördern, um die Bindung an Dienste sowie proprietäre Systeme zu vermeiden und diese Infrastrukturen mit der European Open Science Cloud zu verknüpfen“. Und das passt nun gar nicht zu DEAL. Die großen Wissenschaftsorganisationen, darunter auch die European University Alliance (viele deutsche Unis sind Mitglied) und Science Europe (hier ist die DFG dabei) begrüßen die Beschlüsse der Wissenschaftsminister. Die DFG wünscht sich „Open-Access-Infrastrukturen, die ohne von Autorinnen und Autoren zu zahlende Publikationsgebühren und Gewinnabsichten operieren”.

Also wieder mal ein großer Auftritt für Positionspapiere und Deklarationen, auf die dann wenig Aktionen folgen. Dabei bietet der Verzicht auf den Elsevier DEAL, sowie die Ende des Jahres anstehenden Verhandlungen zur (nicht)-Verlängerung der DEALs mit Springer Nature und Wiley die phantastische Möglichkeit eine Wende einzuleiten. Hin zu einem Publikationswesen, in dem die Wissenschaft gemeinsam mit Bibliotheken und wissenschafts-orientierten, nicht-kommerziellen Verlagen die Nutzung wissenschaftlicher Erkenntnis organisiert und kuratiert. Dazu brauchen wir ein Leistungsbewertungssystem in Academia, das sich an der Qualität und den wissenschaftlichen und dem gesellschaftlichen Impact orientiert und nicht am Renommee von Journalen. Technisch ist das alles kein Problem. In Südamerika werden über 70 % aller wissenschaftlichen Publikationen im Diamond OA Modus veröffentlicht. Das sind Publikationen und Publikationsplattformen, die von allen Interessierten weltweit kostenlos gelesen werden können, und bei denen auch für die Autoren keine Kosten anfallen. Die Berlin University Alliance hat gerade einen Diamond OA Verlag gegründet (Berlin Universities Publishing), der die Wissenschaftler der großen Berliner Universitäten hierbei unterstützt. Und der von der EU angestoßene COARA (Coalition for Advancing Research Assessment) Prozess, in dem nicht nur die DFG, sondern auch bereits viele deutsche Universitäten und Forschungseinrichtungen konkrete Maßnahmen zur Reform des akademischen Bewertungssystems erarbeiten, kann dabei wichtige Unterstützung leisten.

Der Wissenschaftsnarr dankt Bjoern Brembs und Ursula Flitner herzlich für inspirierende Diskussionen und wertvolle Beiträge.

Woke Wissenschaft: Bremse oder Beschleuniger von Qualität und Innovation in der Forschung?

LJ 12/2023

Sie befinden sich mit Kollegen bei einer internationalen Begutachtung Ihres großen, kollaborativen Forschungsantrages, gerade läuft die gemeinsame Runde mit den Antragstellern, Gleichstellungsbeauftragten, Dekan, und anderen Würdenträgern der Uni. Der Dekan zeigt Statistiken zum Frauenanteil bei den Antragstellern und Postdocs, Kinderbetreuung und Frauenmentoring. Die Gutachter aus den anglosächsischen und skandinavischen Ländern werden mittlerweile zunehmend unruhig, bis einer von ihnen unterbricht: Sehr schön, das mit der Frauenförderung, aber das sei doch mittlerweile alles selbstverständlich, wo bleibt die Förderung von Equity, Diversity, Inclusion (EDI) jenseits der Genderperspektive? Große Ratlosigkeit bei den Antragstellern, manch einem scheint gar nicht klar worum es geht. In letzter Sekunde verweist man auf die PhD Studenten aus verschiedenen Ländern, aber die Gutachter tuscheln bereits untereinander.

Der Wissenschaftsnarr war in den letzten Jahren bei Begutachtungen mehrmals Zeuge eines solchen Szenarios, bei dem die Antragsteller und Unileitungen schwer ins Schwitzen kamen. Im deutschen Wissenschaftssystem arbeitet man sich nämlich noch vorrangig an der Erhöhung des Frauenanteils ab, ein erweiterter Diversitätsbegriff hat sich noch nicht recht durchgesetzt. Dabei gibt es neben Genderaspekten noch andere wichtige Dimensionen von Diversität (‚Vielfältigkeit'), wie Alter der Forschenden, Internationalität, Vielfalt der Ideen, Forschungsfelder und Herangehensweisen und vieles mehr.

Bei deutschen Wissenschaftlern führt die Erwähnung von Diversität im Forschungskontext sogar häufig Stirnrunzeln hervor, wenn nicht unverhohlenen Hohn. So lamentierten der Vizepräsident des Deutschen Hochschulverbandes und Ex-Dekan der Medizinischen Fakultät der Uni Frankfurt, Josef Pfeilschifter, gemeinsam mit seinem professoralen Kollegen Helmut Wicht in der FAZ, „dass die ideologischen/politischen Programme der Diversitäts- und Identitätspolitik im Kern antiaufklärerisch, totalitarismuskritisch, konstruktivistisch und vor allem machtorientiert sind." Und wähnen „dahinter […] eine wissenschaftsfeindliche Ideologie." Sie verkennen dabei, dass die Förderung von Vielfalt im Wissenschaftssystem nicht nur wichtig ist, um im Antragsgeschäft vor internationalen Gutachtern zu punkten. Sondern vor allem um die Qualität der Forschung zu steigern, Innovation voranzutreiben und sicherzustellen, dass die wissenschaftliche Gemeinschaft die Vielfalt der Gesellschaft angemessen repräsentiert und respektiert. Als Lordsiegelbewahrer des akademischen Status quo verkennen sie insbesondere, dass wir all dies, und das nicht nur in Deutschland, bitter nötig haben.

Warum lassen sich Ergebnisse auch von hochrangig publizierten Studien häufig nicht wiederholen (‚Reproduktionskrise')? Warum häufen sich prominente Fälle von Plagiarismus und Datenmanipulation bis hin zum Wissenschaftsbetrug? Warum nimmt trotz

gigantischem Input das von uns produzierte, wirklich neuartige Wissen gleichzeitig ab? „Papers und Patente werden immer weniger disruptiv" titelte eine erst vor kurzem eine in Nature veröffentlichte Studie (der Wissenschaftsnarr berichtete hierzu in LJ 3/23). Haben wir vielleicht ein ganz grundlegendes Problem bei der Auswahl von Personal und Projekten in der Wissenschaft? Wie effektiv die Wissensgenerierung ist, hängt auch von den Kriterien ab, nach denen Forschungsgelder verteilt werden. Hierfür hat sich weltweit der ‚Peer Review' etabliert. Gelder werden nicht mit der Gießkanne oder per Los verteilt, sondern nach gegenseitiger Begutachtung durch Experten der wissenschaftlichen Gemeinschaft. Das klingt sehr vernünftig, nur bringt dies auch Probleme mit sich. In jüngster Zeit haben sich diese aufgrund des enormen Anstiegs von Forschung und deren Resultaten massiv verschärft. Denn bewertet wird mittlerweile im Wesentlichen nach den Kriterien ‚Exzellenz' und ‚Originalität'. Diese Begriffe sind soziale Konstrukte zur Verteilung von Forschungsmitteln, Ideale, die uns wenig darüber sagen, wie gut die zu beurteilende Wissenschaft tatsächlich ist, aber alles darüber, wer die Auswahl trifft.

Gutachter neigen dazu, Wissenschaftler und deren Forschung zu bevorzugen, die ihrer eigenen „ähnelt". Da zukünftige Generationen von Forschern von den aktuellen ausgewählt und gefördert werden und Individuen dazu tendieren aufgrund bewusster und unbewusster Vorurteile Personen auszuwählen, die ihnen und ihren Überzeugungen ähneln, werden Forscherteams immer homogener. Aufgrund fehlender objektiver und quantifizierbarer Kriterien für Exzellenz und Originalität in der Wissenschaft, sowie der anschwellenden Flut an zu begutachtenden Projekten und Forschern, hat sich in vielen wissenschaftlichen Bereichen weltweit ein Surrogat-Kriterium für die Auswahl durchgesetzt: Das Renommee der Zeitschriften in denen die ForscherInnen bisher veröffentlicht haben, meist gemessen in einer einzigen Zahl – dem Impact Factor. Dieser sagt nichts über die Qualität und die Wichtigkeit der Forschung der Wissenschaftler aus, sondern misst lediglich wie häufig das betreffende Journal zitiert wird.

Die hierauf beruhende und mittlerweile allgemein durchgesetzte Reputationsökonomie führt zu einer Fokussierung auf vergangene Leistungen und einer Betonung des Mainstreams. Das System ist homophil: Bei Auswahlprozessen kommen homogen zusammengesetzte Kommissionen zu ähnlichen Entscheidungen, da man die gleiche Wertematrix teilt. Vom Bekannten abweichende und deshalb risikoreichere Projekte und KandidatInnen werden mit höherer Wahrscheinlichkeit aussortiert. Auch kommt es zur Konzentration von Ressourcen, frei nach Matthäus (Mt 25,29): „Wer hat, dem wird gegeben".

Führende Wissenschaftsorganisationen und Forschungsförderer (einschließlich der DFG) weltweit beklagen daher seit einiger Zeit, dass es der Wissenschaft an Diversität fehle. Die sprichwörtlichen ‚weißen alten Männer', also die arrivierten, immer noch überwiegend männlichen Professoren entscheiden darüber, wer an was forschen darf. Man muss nur auf die Zusammensetzung der Kommissionen und Kollegien der Deutschen Forschungsgemeinschaft oder die Liste der Nobelpreisträger schauen. Diese Homogenität der Entscheidungsträger führt in der Folge dann zu geringer Diversität in den beforschten Fragestellungen, sowie der Faktoren, die auf den Forschungsgegenstand Einfluss haben. Konkret bedeutet dies, dass Forscher insbesondere in der Biomedizin und den Verhaltens-, aber auch in vielen Geisteswissenschaften bisher fast ausschließlich einen kleinen Teil der Menschheit und deren Umwelt ins Visier genommen haben: Menschen aus westlichen, gebildeten, industrialisierten, wohlhabenden und demokratischen Gesellschaften. Die meisten Menschen auf diesem Planeten sind aber gar nicht WEIRD (white/western-educated-industrialized-rich-democratic).

Bei Diversität geht es eben keineswegs nur um das Geschlecht. Ethnizität, Alter, Fähigkeiten, kultureller oder sozioökonomischer Hintergrund, ‚atypische' Karrierewege usw. spielen für Vielfalt eine mindestens ebenso große Rolle. Dabei sollte es eigentlich überraschen, dass Wissenschaft ein Diversitätsproblem beim Personal hat: Wissenschaftler reisen zu Kongressen und Forschungsaufenthalten in ferne Länder. Viele PhD – Studenten kommen zu uns aus dem Ausland, deutsche Postdocs erhalten Auslands-Forschungsstipendien. In frühen Karrierephasen ist das Geschlechter-Verhältnis in vielen Bereichen auch noch recht ausgeglichen. All dies ändert sich aber kontinuierlich mit zunehmendem Alter und Karrierestadium der Wissenschaftler. Auf dem Level der Professoren und Abteilungs – oder Institutsdirektoren – also bei denen, die im System das Sagen haben, dominieren wie oben beschrieben die deutschen Männer, mit typischer, sehr homogener Sozialisation im System. Wie eben Pfeilschifter, Wicht, und der Wissenschaftsnarr. Der Kreislauf schließt sich, es ist dafür gesorgt, dass sich nichts ändert. Gar nicht überraschend ist deshalb, dass es häufig insbesondere die Wissenschaftler sind, die es im System so weit gebracht haben, welche ein Problem mit dem Ruf nach mehr „Diversität" in der Forschung und der Bewertung ihres Personals und deren Produkten haben.

Es ist sehr plausibel, dass Diversität Forschung innovativer und kreativer, robuster, origineller, und globaler machen kann und Problemlösung, Qualität, Zusammenarbeit, und Zugang zu unterrepräsentierten Gemeinschaften erhöht, sowie Forschungsfragen breiter werden und die Forschungsergebnisse damit insgesamt relevanter. In einem Artikel im Hausblatt des Deutschen Hochschulverbandes hob allerdings das bereits oben erwähnte Duo Pfeilschifter/Wicht genau hiergegen zur Gegenrede an. Dies wohl nicht zu Unrecht in der Gewissheit, stellvertretend für den professoralen Mainstream in Deutschland zu sprechen: „Kein Mensch weiß, ob eine diverse Forschungsgruppe überhaupt diversere Ideen hat als ein gleich großer homogener Trupp. Kein Mensch weiß, ob heterogene Forschergruppen erfolgreicher sind als homogene."

Diese Aussagen zeugen allerdings weniger von der Abwesenheit von Evidenz zum Einfluss von Diversität auf Forschungsqualität und Innovation, sondern sind viel mehr Evidenz für die Abwesenheit von Kenntnis der hierzu existierenden Studienlage (eine Auswahl davon und von anderen hier verwendeten Quellen wie immer unter https://dirnagl.com/lj). Richtig ist, dass es wenig gezielte und kontrollierte Interventionen hierzu gibt. Das ergibt sich aber aus der Problematik der Durchführung solcher Studien: Man müsste ja z.B. verblindet in diverse und weniger diverse Teams randomisieren, und dann nach vielen Jahren deren „Erfolg" (was immer man darunter versteht) vergleichen. Wenn man es sich einfach macht, und Erfolg als Anzahl von Nature und Cell Papers definiert, braucht man so eine Studie auch gar nicht zu machen. Natürlich produzieren Wissenschaftler die schon in Cell und Nature veröffentlichen auch mehr Cell und Nature Papers. Aber das ist trivial, und man hat nicht verstanden, worum es hier wirklich geht.

Deshalb zur Klarstellung hier nochmal worum es bei EDI in der Forschung geht: Die Perspektivenvielfalt diverser Forschungsteams kann zu innovativeren Ansätzen und Lösungen führen, da verschiedene Blickwinkel in die Forschung einfließen. Eine diverse Gruppe von Forschenden ist besser in der Lage, verschiedene Probleme und Herausforderungen zu identifizieren, die in monokulturellen Teams möglicherweise übersehen werden. Dies erhöht die Relevanz der Forschung. Der Einschluss von unterschiedlichen Bevölkerungsgruppen in Studien ermöglicht die Erhebung von repräsentativeren und aussagekräftigeren Daten, was wiederum zu besseren Ergebnissen führt. Das ist am offensichtlichsten in der Medizin. Forschung, welche die Vielfalt der Gesellschaft widerspiegelt, ist meist relevanter und anwendbarer für unterschiedliche Bevölkerungsgruppen. Dies erhöht auch die gesellschaftliche Akzeptanz und den Nutzen der Forschung.

Diversität in Forschungsteams kann dazu beitragen, unbewusste Vorurteile und Verzerrungen zu minimieren, das Ergebnis ist objektivere und verlässlichere Forschung. Inklusion ermöglicht es, ein breiteres Spektrum von Talenten und Fachwissen in die Forschung einzubeziehen. Dies erhöht die Wahrscheinlichkeit, dass Experten für spezifische Forschungsfragen beteiligt sind. Eine inklusive Forschungsumgebung fördert die Zusammenarbeit und den Wissensaustausch zwischen Forschenden. Dies kann zu besser koordinierten und effizienteren Forschungsprojekten führen.

Die Erhöhung von Gerechtigkeit, Diversität, und Inklusion hat in allen Gesellschaftsbereichen normativen Charakter, das schließt selbstverständlich auch die Wissenschaft ein. EDI im Wissenschaftsbetrieb steht also keineswegs zur Diskussion, wie dies manch deutscher Professor noch glauben mag. EDI sind Teil der Sustainable Development Goals (SDGs) der UN, sowie der Open Science Standards der UNESCO, die überschrieben sind mit „Wissenschaft zugänglicher, inklusiver und gerechter gestalten, zum Nutzen aller." Eine durchaus fällige Diskussion zu EDI in der Wissenschaft muss sich daher nicht mehr mit der Frage auseinandersetzen, ob wir sie wirklich brauchen. Sondern damit, wie wir Barrieren in der Umsetzung überwinden, wir sie am effektivsten gestalten, sowie mögliche negative, unintendierte Wirkungen vermeiden können.

Neben der Identifikation von Barrieren, zählen dazu Fragen wie: Welche Elemente von EDI sind besonders wichtig, um die Qualität der Forschung sicherzustellen? Welche Aspekte von EDI wirken sich besonders auf Innovation aus? Wie unterscheidet sich dies in verschiedenen Forschungskontexten und Disziplinen? Auf welchem Weg kommen wir zu diversen Forschungsteams? In wieweit ist das gegenwärtige Verständnis von „Qualität" und „Exzellenz" eine Barriere für mehr Diversität? Wie können wir Team Science und Interdisziplinarität fördern in einem System dessen Bewertungsmaßstab individuelle Leistung misst? Gibt es negative Effekte die bedacht werden müssen? Wieso wird Diversität gefordert, aber kaum gefördert? Wie so häufig an dieser Stelle wagt der Narr die Voraussage, dass die meisten der Antworten auf diese Fragen sehr viel mit der Art und Weise zu tun haben werden, wie wir Wissenschaftler und deren Produkte bewerten und belohnen.

Der Wissenschaftsnarr hat im Laufe der Zeit zahlreiche Zuschriften erhalten und so manche kontroverse Debatte angestoßen. Eine Klarstellung zum obigen Beitrag von Prof. Dr. Klaus-Ferdinand Gärditz möchte ich Ihnen jedoch nicht vorenthalten. Ich schließe mich seinen Ausführungen vollumfänglich an und bin ihm dafür sehr dankbar. Hier ein Exzerpt:

„Wenn man Vielfalt der Zugänge zur Erkenntnis - wie Sie es zu Recht tun - als epistemisches Instrument ansieht, besseres (robusteres, replizierbareres, besser begründetes usf.) Wissen zu generieren, muss man sich zunächst sehr genau darauf verständigen, welche Vielfalt man meint und was man damit verfolgt. Erkenntnistheoretisch bietet es sich auch hier an, zwischen den sozialen Bedingungen, unter denen Wissen entsteht (Entdeckungskontext) und denen, wie Wissen als wissenschaftlich begründet wird (Rechtfertigungskontext) zu unterscheiden. Diversität hat im Entdeckungskontext ihren berechtigten Platz, aber nicht im Begründungskontext. Dass es hilfreich ist, mehr Perspektiven einzubinden, um Wissen zu auf einer breiteren Basis generieren, leuchtet ein. Wenn Forschungsergebnisse abhängig von sozialen Zugehörigkeiten unterschiedlich wissenschaftlich gerechtfertigt werden, besteht hingegen der Verdacht, dass es nicht mehr um Wissenschaft geht. Es gibt eben keinen Zitratzyklus der women of colour, keine schwarze Quantenmechanik und keine queere Neuropathologie der Glia.

Fragen des fairen Zugangs zu Wissen und zur Wissenschaft sind politische Kategorien, die sich normativ selbstverständlich stellen, zumal wenn knappe Gelder verteilt werden. Ich stimme Ihnen völlig zu, dass wir Mittel anhand von Kriterien verteilen, "soziale Konstrukte zur Verteilung von Forschungsmitteln" sind, die auf Idealen gründen, "die uns wenig darüber sagen, wie gut die zu beurteilende Wissenschaft tatsächlich ist". Das verdient besonders Aufmerksamkeit, weil wir uns ständig vergewissern müssen, wie wir normativ bewertende Kriterien konstruieren - denn es werden ja immer Konstrukte sein. Nur ist das nicht notwendig das Gleiche wie die Forderung nach epistemischer Härte von Forschung. Wenn z.B. eine schwarze Nachwuchswissenschaftlerin statistisch schlechtere Chancen auf eine Professur hat als ein weißer Nachwuchsforscher, ist das ein strukturellen Diskriminierungsproblem, das Reaktionen rechtfertigt - ggf. auch durch Förderkriterien. Dass aber ein konkretes Forschungsprojekt zur frühmittelalterlichen Musik im keltischen Irland, zur Neuropathologie von Astrozytomen oder zur limnoökologischen Funktion von Dinoflagellatentoxinen nun "better science" liefert, wenn im Team eine schwarze Deutsche aus Wuppertal beteiligt ist, erscheint hochgradig unplausibel. Das würden aber wiederum die Kulturwissenschaften mit ihrer standpoint epistemology kategorisch anders sehen.“

Trau, schau, wem – Wie erkenne ich Overselling, Spin, und anderen Merkwürdigkeiten in wissenschaftlichen Artikeln?

LJ 1-2/2024

Ein neues Jahr liegt vor uns, und wieder dürfen wir uns allein in PubMed bis Sylvester 2024 auf fast 2 Millionen neue wissenschaftliche Artikel freuen! Darunter Spektakuläres, Triviales, Bestätigendes, Widersprechendes und Widersprüchliches, Redundantes, Geschöntes, Absurdes, aber natürlich auch Gefälschtes und Plagiiertes. Weniges davon wird die Biomedizin revolutionieren, einiges davon wird sie voranbringen, das meiste aber wird gar nicht gelesen, geschweige denn zitiert werden. Egal ob gelesen, nützlich oder relevant, die meisten Titel werden Lebensläufe zieren und den Autoren damit zu Titeln und Fördergeldern verhelfen. Was ja eine der vornehmsten Funktionen des akademischen Publikationswesens ist.

Wir Wissenschaftler (und die Verlage, welche davon leben) haben uns in eine auf Artikeln und deren assoziierten Metriken basierte akademische Reputationsökonomie eingemauert. In ihr buhlen wir um Sichtbarkeit und Impactpunkte. Dies produziert eine Paperflut. Um noch Aufmerksamkeit zu erzeugen, müssen die von uns berichteten Effekte immer größer und die behauptete Relevanz unserer Befunde für die Wissenschaft oder gar Menschheit immer wichtiger werden. Wie navigiert man in diesem Meer von Publikationen? Gibt es evtl. sogar formale Kriterien, die uns dabei helfen könnten?

Gerade jüngere Wissenschaftler, noch motiviert von der Freude am Erkenntnisgewinn, voller Ideale und kaum verdorben durch die Jagd auf karrierefördernde Publikationslisten und Impactfaktoren und h-Indizes, stellen sich – und manchmal auch dem Wissenschaftsnarr- diese Frage. Was ist real? Wo regiert überwiegend Spin? Welchen Papers bzw. deren Schlussfolgerungen kann ich trauen, bei welchen sollte ich besonders skeptisch sein?

Deshalb hier der Versuch einer närrischen 14-Punkte Checkliste. Sie ersetzt weder das intensive Studium des Artikels, noch technische und inhaltliche Kompetenz. Vielleicht ist sie aber bei einer ersten, ganz allgemeinen Einordnung des Gelesenen hilfreich.

1. Außergewöhnliche Behauptungen benötigen außergewöhnliche Evidenz! Allzu häufig lesen wir über geradezu unglaubliche Entdeckungen: Orale Transplantation von Fäkalien, die im Tierversuch experimentell induzierte Schlaganfälle kurieren; Appetitsteigerung durch Handystrahlung (der Narr berichtete); mediterrane Diäten, die das kardiovaskuläre Risiko stark senken, und so fort, sie wissen welche Aussagen ich meine (Quellen und Zitate wie immer unter http://dirnagl.com/lj). Hier sollten wir uns immer und sogleich an Carl Sagans berühmten Spruch erinnern, der zuerst von Pierre-Simon Laplace vor über 200 Jahren so formuliert wurde: „Das Gewicht der Evidenz für eine außergewöhnliche Behauptung muss proportional zu ihrer Merkwürdigkeit sein“. Womit wir schon beim wichtigsten Kriterium für die Glaubwürdigkeit eines wissenschaftlichen Artikels wären: Die Qualität der darin geschilderten Evidenz.

2. Teststatistik und Signifikanzen: Dass ein Ergebnis „statistisch signifikant“ ist, gilt in vielen Studien als das wichtigste Argument. Schon im Abstrakt schreit uns häufig der p-Wert entgegen, häufig oft nicht einmal begleitet von der dazugehörigen Varianz oder Effektstärke. Auf Letztere, und noch mehr auf deren mögliche biologische Bedeutung kommt es aber an. Ganz zu schweigen davon, dass die ach so wichtige „statistische Signifikanz“ auf tönernen theoretischen Füssen steht. Der „Erfinder“ der universellen 5 % Schwelle für das falsch positive Resultat (d.h. dem Typ I Fehler) Ronald A. Fisher formulierte die Konsequenzen bei Unterschreiten dieser Schwelle vor 100 Jahren so: „Da lohnt es sich hinzuschauen”! Mitnichten also eine Entdeckung zu reklamieren, und Papers, Doktorarbeiten, oder Förderanträge darauf aufzubauen. Außerdem sollten Sie (auch wenn die Autoren das in der Regel tun), den p-Wert nicht mit dem positiven prädiktiven Wert verwechseln. Auch wenn die meisten Wissenschaftler das glauben, der p-Wert sagt nämlich nicht, wie wahrscheinlich der Befund ist, der damit belegt werden soll (Mehr dazu im LJ 10/2019).

   Finden sich im Artikel auch keine Aussagen zum Typ II Fehler (d.h. zur statistischen Power) und eine ordentliche a priori Fallzahlabschätzung, sollten Sie doppelt skeptisch werden. Die insbesondere in präklinischen Studien sehr geringen Fallzahlen führen nicht nur zu einer hohen falsch negativ- und falsch positiv-Rate, sondern auch zu einer substantiellen Überschätzung der Effektstärken, falls diese überhaupt real sein sollten (sog. „Winners curse“).
   Und weil wir schon bei der statistischen Power sind: Wegen der in Phase II einer klinischen Studie recht niedrigen Fallzahlen sind diese nicht auf ‚Wirksamkeit‘ gepowert, dafür macht man bei Erfolg danach eine viel größere Phase III Studie. Nichts desto trotz können viele Kliniker der Versuchung nicht widerstehen, im Rausche der statistischen Signifikanz, mit der sie ihren primären Endpunkt erreicht haben, gleich

auch noch Wirksamkeit zu reklamieren. Die aber ein nur explorativer Endpunkt war. Auch dies ein Warnsignal!

3. Präsentation der Daten mit SEMs und Balkengraphen? Dieses Negativkriterium ist allgegenwärtig! Ich muss Sie warnen: Man will Ihnen was vormachen! Der Standardfehler des Mittelwerts (SEM) wird benutzt, um eine große Streuung (Varianz) der Daten zu verschleiern, der Balkengraph (noch dazu, wenn Achsen skaliert/unterbrochen werden) dient der Verheimlichung der Verteilung der Daten sowie der Vorspiegelung eines substantiellen Effektes. Kampagnen wie „#BarBarPlots“ und die Aufforderung in den Autoreninstruktionen der Journale, echte Varianzmasse wie Standardabweichung (SD) oder besser Konfidenzintervalle (CI) zu verwenden, sind bisher ohne Effekt geblieben. Bestehen Sie trotzdem auf Dot-blots, Violin-Blots, Box oder Whisker blots, sowie SDs oder CIs.

4. Korrelation ist nicht Kausation! Manchmal direkt heraus, häufig aber subliminal verkaufen uns viele Autoren signifikante Korrelationskoeffizienten als Belege für Ursache – Wirkung-Beziehungen. A geht rauf, B geht rauf, also bewirkt Parameter A Parameter B. So einfach ist das aber nicht. In der „normalen“ (d.h. allgegenwärtigen) frequentistischen Statistik gibt es das Ursache – Wirkung- Prinzip gar nicht! Korrelationskoeffizienten funktionieren in beide Richtungen. A kann B bewirken, genauso wie B A, selber Korrelationskoeffizient. Außerdem ist es vielleicht C, das wir gar nicht kennen oder berücksichtigen, was auf A und auf B wirkt, und damit einen Scheinzusammenhang herstellt. Eine wunderbare Zusammenstellung solcher scheinbaren Korrelationen findet sich unter https://www.tylervigen.com/spurious-correlations. Meine Lieblingskorrelation dort ist die zwischen dem Alter der Miss America und den Morden durch Dampf und heiße Objekte. Ohne Verwendung der erst in den letzten beiden Jahrzehnten entwickelten Methoden der kausalen Inferenz mit ihren graphischen Ansätzen (v.a. Directed acyclic graphs) muss weiter gelten: Ohne Intervention, aus der bloßen Beobachtung heraus, lässt sich eine Kausalbeziehung zwar vermuten, aber nicht belegen. Mehr dazu in einer der nächsten Folgen dieser Kolumne.

5. Wie gut war das Studiendesign? Sie sollten die folgenden Fragen mit „ja“ beantworten können: Wurden die Versuche und die Auswertung der Ergebnisse verblindet durchgeführt? Wurden die Versuchsobjekte (Zellkulturen, Mäuse, Menschen, etc.) randomisiert ausgewählt? Waren Kriterien vorab bestimmt worden und im Artikel angegeben, nach denen Ergebnisse in die Analyse ein- oder ausgeschlossen wurden? Berichten die Autoren dann auch, wie viele Versuchsobjekte demnach nicht eingeschlossen werden konnten und warum? Gab es eine a priori Definition eines relevanten Hauptergebnisses („primärer Endpunkt“), auf den die Fallzahlplanung ausgerichtet war? Wurden weitere Parameter vorbestimmt, die nur explorativ analysiert werden sollen („sekundäre Endpunkte“), für welche die Studie aber nicht gepowert war? Haben die Autoren sich festgelegt, ob ihre Studie explorativ angelegt ist – also der Generierung von Hypothesen dienen soll, oder ob sie konfirmatorisch ist, und damit eine Hypothese be- oder widerlegen will? Von dieser (selten gemachten) Unterscheidung hängen ganz wesentlich das Studiendesign und die statistischen Analyseverfahren ab – und natürlich mehr noch der aus Studie zu erzielende Erkenntnisgewinn.

6. Werden die Originaldaten im Artikel zur Verfügung gestellt? Gibt es also ein „Data availability statement“, und was steht da drin? Ganz schlecht ist, wenn es gar keines gibt (Ausnahmen, besonders bei klinischen Studien bestätigen die Regel). Nicht viel

besser ist es, wenn es lapidar heisst: „Data available on reasonable request“. Sehr gut dagegen ist ein direkter Link zum Download (unbedingt mal reinschauen!), noch besser, wenn diese Daten „FAIR“ – also findable, accessible, interoperable, and reusable geteilt werden und mit brauchbaren Metadaten annotiert sind.

7. Äußern sich die Autoren zu möglichen Interessenkonflikten? Gibt es also ein „Conflict of interest statement“, und was steht da drin?

8. War das Studiendesign präregistriert? Bei interventionellen klinischen Studien ist dies eigentlich Pflicht – zumindest, wenn die Ergebnisse später ordentlich publiziert werden sollen. Bei präklinischen Studien ist das leider noch die Ausnahme. Das ist sehr schade, denn eine Präregistrierung bringt einen massiven Qualitätssprung in der Bewertung der Evidenz, die in einem Artikel berichtet wird. Vor allem können wir uns dann als Leser versichern, dass Endpunkte und Analyseverfahren nicht al Gusto von den Autoren im Studienverlauf verändert wurden, um erwünschte Ergebnisse zu erzielen. Wir müssen leider davon ausgehen, dass solch „Cherry picking“ von Daten sowie multiple statistischen Analysen bis zum Erreichen signifikanter Resultate („undisclosed flexibility in analysis and reporting“) ein wesentlicher Grund für die epidemische nicht-Reproduzierbarkeit von Studienergebnissen und die Inflation von Effektgrößen ist. Präregistrierte präklinische Studien bekommen vom Narren deshalb einen Vertrauensvorschuss! Das gleiche gilt übrigens für Multicenter-Studien, in denen Ergebnisse unabhängig in verschiedenen Laboren repliziert wurden.

9. Gibt es eine kritische Diskussion der Limitationen? Reflektieren die Autoren die Ergebnisse und die Schlussfolgerungen, die sie daraus ziehen, im Lichte möglicher Schwächen ihrer Studie? Dazu könnten geringe Fallzahlen, beschränkte Generalisierbarkeit (externe und Konstruktvalidität), mögliche Verzerrungen (Bias), methodische, instrumentelle, und technische Probleme, limitierter Zugang zu Daten, Notwendigkeit der weiteren Absicherung durch unabhängige Konfirmation, sowie vieles mehr zählen. Das pro forma Auflisten von ‚Strohmann‘-Limitationen, nur um diese dann mit einem Satz als „unwahrscheinlich“ zu disqualifizieren, gilt nicht!

Die oben gelisteten Kriterien gelten für Grundlagen- als auch auf klinische Forschung. Im Folgenden noch einige Warnsignale, auf die man besonders bei klinischen Studien achten sollte.

10. Argumentation mit relativer statt absoluter Risikoreduktion. Eine starke Reduktion des relativen Risikos durch eine neue Therapie lässt diese in sehr gutem Licht erscheinen. Aber wenn das absolute Risiko des Ereignisses gering ist (was häufig der Fall ist), dann ist die relevante Risikoreduktion wesentlich geringer. Artikel, welche die relative Risikoreduktion in den Vordergrund stellen, tun dies häufig, um uns einen großen Nutzen eines Medikamentes vorzugaukeln, der gar nicht existiert.

11. Unterscheidet sich der im Artikel berichtete primäre Endpunkt von dem der Präregistrierung? Werden gar sekundäre Endpunkte plötzlich zu primären geadelt? Es lohnt sich deshalb immer, einen Blick ins klinische Studienregister zu werfen, um dies zu überprüfen!

12. Beruhen die Erfolgsmeldungen der Studie auf im Nachhinein definierten Subgruppen? Falls nicht a priori definiert und nicht präregistriert, kann die Analyse von

Subgruppen zwar interessante Hypothesen generieren, sollte aber nicht zu positivem Spin in der Interpretation der Studienergebnisse führen.

13. Verwendet die Studie Surrogat-Endpunkte? Was für Patienten wirklich zählt („Patient important outcomes"), sind insbesondere Lebensqualität, Symptomfreiheit oder -linderung, und möglicherweise Lebensverlängerung (aber nicht um jeden Preis). Häufig werden in klinischen Studien aber sog. Surrogat-Endpunkte verwendet, wie zum Beispiel Veränderungen von Laborwerten, physikalischen Variablen oder in bildgebenden Verfahren. Meist wird dies mit Praktikabilität, objektiver Quantifizierbarkeit, sowie pathophysiologischen Überlegungen begründet. Die Studienliteratur ist aber voll von Beispielen (einige davon unter http://dirnagl.com/lj), in denen Interventionen solche Surrogatmarker positiv beeinflusst haben, die Patienten aber letztendlich gar nichts davon hatten. Die Pharmaindustrie dagegen aber umso mehr. Denn bis die Unwirksamkeit auf letztendlich für Patienten relevante Endpunkte gezeigt werden, kann sie Milliarden mit dem Verkauf von Tabletten verdienen, die lediglich Blutwerte verbessern.

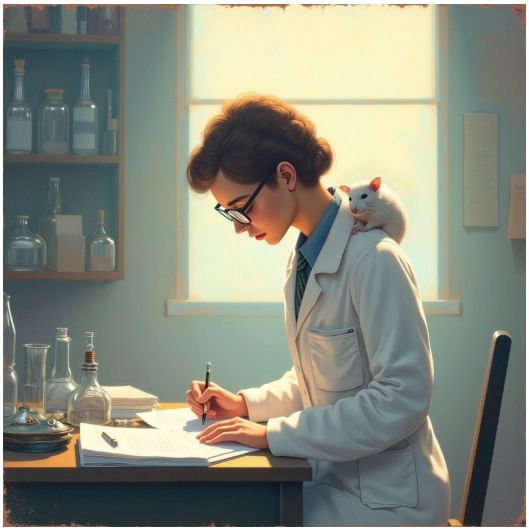
14. Wurde die Studie wegen Wirksamkeit vorzeitig beendet? Es erscheint paradox, dies als Warnsignal zu betrachten. Fakt ist aber, dass sowohl theoretisch als auch praktisch an einer Vielzahl von Beispielen zu belegen ist, dass es hierbei regelhaft zu einer deutlichen Überschätzung des Effektes, bei geringen Ereignisraten sogar falsch positiven Resultaten kommt.

Sollten einige oder mehrere der hier angeführten formalen Kriterien erfüllt sein, heißt dies natürlich keineswegs automatisch, dass der Studie nicht zu trauen ist, deren Schlussfolgerungen mit Vorsicht zu genießen, oder die Studie sonst wie von minderer Qualität ist. In jedem Fall sollten Sie beim Auftreten solcher „Warnsignale" noch genauer als üblich hinschauen. Unbedingt abzuraten ist allerdings davon, die Reputation eines Journals als Qualitätskriterium zu verwenden. Ob eine Studie in PlosOne oder Nature veröffentlich wurde, sagt allenfalls etwas darüber aus, für wie relevant oder spektakulär die Autoren und die Editoren die Ergebnisse hielten. Und dass man sich da häufig in die eine oder andere Richtung getäuscht hat, zeigten zuletzt der Fall und die multiplen Retraktionen des Wissenschaftsstars und nun Ex-Stanford Präsidenten Marc Tessier-Lavigne (siehe LJ 10/2023) und der Nobelpreis für Katalin Karikó. Ersterer veröffentlichte seine Ergebnisse vorwiegend in Nature, Cell und Science, Letztere im Journal of Biochemistry, Molecular Therapy, oder Nucleic Acids Research.

Bleiben Sie skeptisch: Organisiertes Misstrauen produziert vertrauenswürdige Wissenschaft!

Zeige mir Dein Laborbuch und ich weiß, ob Du ein guter Wissenschaftler bist!

LJ 3/2024



„Ich habe gefunden, dass eine messbare Zeit vergeht, während sich der Reiz, welchen ein momentaner elektrischer Strom auf das Hüftgeflecht eines Frosches ausübt, bis zum Eintritt des Schenkelnerven in den Wadenmuskel fortpflanzt“. So beginnt ein „Vorläufiger Bericht“, den Hermann Helmholtz, Professor der Physiologie in Königsberg 1850 im Archiv für Anatomie, Physiologie und wissenschaftliche Medizin veröffentlicht hat. Ein fulminanter Auftakt für einen Artikel, der Wissenschaftsgeschichte geschrieben hat, weil hier zum ersten Mal die Nervenleitgeschwindigkeit gemessen wurde.

Aber auch andere Aspekte dieser Arbeit ragen heraus. Sie markiert den Beginn der modernen Neurophysiologie, und zeigt uns, wie gute Forschung geht: Sie stützt sich auf Vorarbeiten Anderer: Der Italiener Carlo Matteucci hatte einige Jahre zuvor ähnliche Experimente gemacht, allerdings technisch weniger ausgereift. Der Weg zur Erkenntnis verläuft nicht linear sondern „mäandert“, nach H.Schmidgen eine „überraschende Reihung von Vorwärts-, Rückwärts- und Seitwärtsbewegungen“ (Zitate und weiterführende Links wie immer unter http://dirnagl.com/lj). Sie ist methodisch innovativ, denn Helmholtz konstruierte seine eigenen Messgeräte und graphischen Schreibapparate. Messfehler und Störgrößen werden quantifiziert und minimiert, er berechnete z.B. die maximalen Fehler seiner Apparaturen und den Einfluss der Umgebungstemperatur. Zudem ist sein auf den vorläufigen folgende vollständige Bericht der Ergebnisse mit 90 Seiten ausgesprochen detailreich. Denn jedes Experiment sollte von anderen Forschern verstanden und wiederholt werden können. Und die in der Arbeit berichteten Versuche wurden gemeinsam mit seiner Frau Olga im Laborbuch lückenlos dokumentiert.

Noch heute, 174 Jahre nach ihrer Durchführung, lassen sich die Versuche anhand dieses Laborbuches rekonstruieren und nachvollziehen. Schon dies ist außergewöhnlich, und dürfte für die wenigsten heute verfassten (und für 10 Jahre in der Einrichtung vorzuhaltenden) Laborbücher gelten. Ebenso erstaunlich scheint mir aber, dass in fasst allen universitären Einrichtungen heutzutage Versuche immer noch in Papierlaborbüchern dokumentiert werden. Und das in Zeiten, in denen – im Gegensatz zum 19. Jahrhundert – fast alle Daten primär digital anfallen. Helmholtz übertrug die wenigen vom Kymographen visuell abgelesenen Messwerte direkt ins Buch. Heute spucken Sequencer, PCR und FACS – Maschinen, Konfokalmikroskope usw. Megabytes von Primärdaten und in den Geräten bereits durchgeführten Analysen und Statistiken aus. Auf diese verweisen wir heute mit einer ins Laborbuch gekritzelten Linkadresse, die auf ein digitales Speichermedium verweist, oft garniert mit einem Ausdruck ausgewählter, ausgedruckter und dann eingeklebter oder gar gehefteter Befunde. Ein vergleichender Blick in die Laborbücher von Helmholtz und seinen Zeitgenossen und die heutiger Biowissenschaftler

ist durchaus verstörend. Der wesentlichste Unterschied zu modernen Laborbüchern dürfte dabei sein, dass die 150 Jahre alten ordentlicher geführt wurden, die Handschrift leserlicher war, und die Befunde auch heute noch für jedermann unmittelbar nachvollziehbar sind.

Die anhaltende Popularität atavistischer Papierkladden ist umso erstaunlicher, da jede TA, jeder Student, jede Wissenschaftlerin heute entweder über einen eigenen Rechner oder Notebook verfügt, oder zumindest in allen Laboren offener Zugang hierzu besteht. Und eine Vielzahl von kommerziellen, aber auch kostenlosen Open Source Lösungen für elektronische Laborbücher (ELN) existieren. Wobei diese soviel mehr können als ein Papierbuch. Die digitalen Daten können mit ins ‚Buch' integriert werden. Dies als ‚harte' Links auf institutionelle, gesicherte Speichermedien, oder wenn nicht zu umfangreich, als Originaldaten, teilweise direkt über eine Schnittstelle zum Messgerät eingelesen. Arbeitsgruppen oder Wissenschaftler, die gemeinsam an einem Projekt arbeiten, können über das ELN direkt kollaborieren und Daten und Befunde tauschen, auswerten und diskutieren. Arbeitsgruppenleiter können mit ihren Mitarbeitern gemeinsam die Versuchsergebnisse besprechen ohne physisch vor dem Buch zu sitzen – gerade in Zeiten von Home Office und Überlastung durch Administration und Krankenversorgung ein nicht zu unterschätzender Vorteil. Oft Verwendetes (Protokollbestandteile, SOPs, Beschreibungen von Assays etc.) werden als Templates vorgehalten und müssen nicht immer wieder abgeschrieben werden. Außerdem sind ELNs ganz einfach durchsuchbar, Jahre zurückliegende Einträge findet man in wenigen Sekunden – haben Sie schon mal in einem Papierlaborbuch nach einem älteren Eintrag gesucht?

Manche ELNs enthalten auch Datenbanken für Laborinventar, Chemikalien, etc. Außerdem haben Papierlaborbücher die Tendenz, verloren zu gehen, spätestens wenn der Student oder der Betreuerin die Einrichtung verlässt. Oder der Aktenschrank entsorgt wird, indem sie gelagert wurden. Nicht so das ELN, das zeitlich unbegrenzt und platzsparend archiviert werden und auch ohne große Kopieraktion beim Verlassen der Einrichtung ‚mitgenommen' werden kann. Die meisten ELNs erfüllen noch dazu die Anforderungen des 21 CFR Title 11c der US Food and Drug Administration (FDA) an elektronische Aufzeichnungen und Unterschriften. So wie die meisten Papierlaborbücher in unseren Instituten geführt werden, würden sie von der FDA nicht akzeptiert. Dies hat schon bei manchen, deren Ergebnisse bei einer Medikamentenzulassung eine Rolle spielen sollten, zum Desaster mit Ansage geführt – die Versuche mussten wiederholt und FDA konform dokumentiert werden. Und zu guter Letzt, und angesichts der Zunahme von Datenmanipulation bis hin zur Fälschung (siehe LJ 1-2/ 2023) nicht ganz irrelevant: ELNs sind viel fälschungssicherer als ihre papierenen Verwandten.

Das ELN scheint also sowas wie ein Schweizer Messer der Forschungsprozessdokumentation, es kann Dokumentation, aber auch rudimentäres Datenmanagement, Kollaboration, Prozessökonomisierung und Qualitätssicherung. Ein no-brainer also? Da drängt sich unmittelbar die Frage auf, warum die Verwendung eines ELN die Ausnahme, und nicht die Regel ist.

Zunächst einmal ist da die Inertia der Forschenden selbst. Einmal sozialisiert mit Kladde, propagiert man die liebgewordene Tradition. Außerdem verlangt das Ganze natürlich Einarbeitung – und im Rennen um das nächste Paper oder den nächsten Antrag ist Zeit ein knappes Gut. Und wenn man ein kommerzielles ELN wählt, fallen Kosten an. Und was passiert, wenn der Anbieter die Grätsche macht oder von einer Datenkrake wie Elsevier übernommen wird? Sind die Daten dann sicher? Im Gegensatz zur Textverarbeitung (z.B. rtf) oder bei der Speicherung von digitalen Bildern (z.B. tif) gibt es bei ELNs keine allgemein akzeptierten Austauschformate zwischen Herstellen. Es droht damit bei

Verwendung eines kommerziellen ELN ein „Vendor lock in“. Und wer es mit den Regularien in Deutschland ernst nimmt, hat bevor es losgeht möglicherweise noch ein paar überraschende Hürden zu überwinden: Möglicherweise will der Personalrat mitentscheiden, ob das überhaupt statthaft ist – könnte das ELN vielleicht zur Überwachung von Mitarbeitern missbraucht werden? Nicht zu vergessen die Satzungen für gute wissenschaftliche Praxis, die an allen universitären Einrichtungen existieren. Weil es zum Zeitpunkt von deren Erstellung noch gar keine ELNs gab fordern viele Ordnungen explizit die Dokumentation in einem Papierlaborbuch.

Der Narr kennt die Vor- und Nachteile von Kladde und ELN deshalb so gut, weil er in seiner eigenen Abteilung ein ELN eingeführt hat, und an der Ausrollung eines ELN in das Forschungsökosystem einer der größten universitären biomedizinischen Einrichtungen Europas (Charité Universitätsmedizin Berlin) beteiligt ist. Und weil er mit KollegInnen diesen Prozess wissenschaftlich begleitet hat. Und dabei hat er viel gelernt.

Zunächst einmal, dass die potentiellen Nachteile des ELN zwar ernst zu nehmen, aber alle überwindbar sind – im Gegensatz zu den Nachteilen der Kladden. Ein kostenloses open source ELN, das auf institutionellen Servern läuft, ist unvergleichlich sicherer, als die Kladde, die man vielleicht in der Cafeteria liegen lässt. Thema Ausstieg: ELNs erlauben den Export als HTML und im pdf-Format. Das ist nicht toll, entspricht aber dann der Funktionalität der Papierversion. Und Einarbeitung ist nötig, aber wer nicht in der Lage ist, ein ELN zu bedienen, wird auch sonst wenig Erfolg an den Geräten im Labor und bei der Auswertung seiner Versuche haben – gehört also vielleicht gar nicht in die Wissenschaft.

Was ich aber auch gelernt habe – und da liegt der Hase im Pfeffer: Wer schon nicht weiß, wie man ein Papierlaborbuch ordentlich führt, oder das Pech hat, in einer Arbeitsgruppe zu forschen, in der die Supervision nicht richtig funktioniert, wird durch den Wechsel zum ELN nichts gewinnen. Es könnte sogar schlimmer werden, insbesondere wenn dann nur noch auf Zettelchen dokumentiert wird, die im Abstand von Tagen oder Wochen abfotografiert und ins ELN ‚übertragen‘ werden. Oder gar parallel analog und digital dokumentiert wird – am Ende findet sich niemand mehr zu recht.

Wie ein Laborbuch geführt wird, egal ob analog oder digital, ist nämlich ein wichtiger Indikator für die Forschungsqualität eines Wissenschaftlers bzw. einer Arbeitsgruppe. Wie wird was dokumentiert? Ist die Dokumentation zuordenbar, lesbar, zeitnah, original, genau, vollständig, konsistent, beständig, und verfügbar? Dies sind die Standards, um die Zuverlässigkeit und Glaubwürdigkeit von Daten zu gewährleisten. Diese sogenannten ALCOA+ Prinzipien werden von Aufsichtsbehörden wie der FDA in den USA oder der EMA (European Medicines Agency) in Europa gefordert. Im akademischen Kontext sind sie leider wenig oder gar nicht bekannt, und werden noch seltener erfüllt. Gut geführte Labore mit kompetentem Personal wissen vermutlich gar nicht, dass es sowas wie ALCOA gibt, machen es aber intuitiv richtig. Olga und Hermann Helmholtz konnten die ALCOA-Prinzipien noch gar nicht kennen – ihr Laborbuch erfüllte dennoch deren Kriterien.

Das Wie der Forschungsdokumentation reflektiert also Forschungsqualität, man kann sogar sagen, es wirkt als „Forschungskultursensor“. Der Ersatz eines analogen durch ein digitales Thermometer ändert nicht die Raumtemperatur. Die bloße Einführung eines überlegenen technischen Werkzeuges verbessert nicht nötig die Qualität von Forschung. Einzelne Wissenschaftler oder Gruppen von Forschenden, welche sehr gut dokumentieren (also ALCOA+ entsprechend), deren Leitungs-, Team-, Diskussions – und Kollaborationskultur gut ‚funktioniert‘, werden beim Wechsel zum ELN mit all dessen Vorteilen belohnt. In weniger optimalen, aber dafür im akademischen Forschungskontext leider

eher häufigen Szenarien müssen Leitung und Supervision, Methodenkompetenz, Infrastrukturen, Arbeit im Team usw. verbessert und die für gute Forschungsdokumentation notwendigen Ressourcen bereitgestellt werden. Erst dann macht der Wechsel in die Wunderwelt der ELNs Sinn.

Letzteres bleibt im gegenwärtigen System natürlich närrisches Wunschdenken. Unterausgestattete Universitäten, darbende Fakultäten, nicht auskömmliche Grundausstattung der Forschenden und keine Möglichkeit, dies über Overheads der Fördergeber zu kompensieren, gepaart mit dem immensen Druck zur schnellen Publikation auf dem Weg zu Promotion, Habilitation, Professur (widrigenfalls der Notwendigkeit, Academia zu verlassen), die Jagd nach noch mehr Drittmitteln – all dies ist kein guter Nährboden für professionelle Forschungsdokumentation, gar mit einem ELN. Vielleicht liegt darin auch einer der vielen Gründe, warum Forscherpersönlichkeiten mit dem Impact eines Helmholtz heute so selten geworden sind.

Der Wissenschaftsnarr dankt Christiane Wetzel und Ina Frenzel für inspirierende Diskussionen.

Kann denn Abschreiben Sünde sein?

LJ 4/2024

Wissenschaftliches Fehlverhalten ist mal wieder groß in den Schlagzeilen. Man denke an die Plagiatsvorwürfe inklusive Rücktritt von Claudine Gay, der Präsidentin von Harvard). Oder der aktuelle Vorwurf der Datenmanipulation, gefolgt vom Rücktritt von Simone Fulda, Präsidentin der Uni Kiel. Oder die Institutionen mit massivem Imageschaden wegen gefälschter oder geschönter Daten, wie das Dana Farber Cancer Research Institute in Boston – dem selbsternannten Olymp der Krebsforschung. Die mussten erst kürzlich Dutzende von Studien retrahieren. Und so geht es dahin. Alles Belege für etwas, das der Narr vor einem Jahr behauptet hatte (LJ 1-2/2023): Dass nämlich Wissenschaftsbetrug gar nicht so selten ist, wie wir Wissenschaftler uns das immer schönreden. Wenn die Spitze des Eisbergs schon so groß ist, welche Ausmaße muss erst das Eis darunter haben?

Verstöße gegen die gute wissenschaftliche Praxis werden gerne unter dem Akronym FFP subsummiert: Falsifikation, Fabrikation, und Plagiarismus. Insbesondere für das nichtwissenschaftliche Publikum nimmt der Plagiarismus dabei eine herausragende Stellung ein: Im Gegensatz zu irgendwelchen Gelschnipseleien ist es dabei einfacher zu verstehen, worum es geht, und warum das eigentlich nicht geht. Auch liest man in der Presse andauernd davon, weil reihenweise Persönlichkeiten des öffentlichen Lebens des Plagiarismus bezichtigt werden und in der Konsequenz oft einen Karriereknick erleiden. zu Guttenberg, Schavan, Giffey, Koch-Mehrin, die Liste lässt sich allein für Deutschland lange fortsetzen. Fakultäten haben darauf reagiert, und unterziehen mittlerweile regelhaft eingereichte Qualifizierungsarbeiten einem automatisierten Plagiarismus-Screen.

Diese sind zwar für Betreuer, Doktoranden und Habilitanden kaum verwertbar, da sie meist Hunderte von potentiell plagiierten Stellen anzeigen. Das kommt daher, dass eben auch massenhaft triviale und standardisierte Formulierungen rausgefischt werden. Aber die Institutionen demonstrieren, dass sie sich ihrer Verantwortung bewusst sind und proaktiv handeln.

Aber wie ist das eigentlich mit dem Plagiarismus? Schlagen wir da auf einen toten Hund ein? Ist Plagiarismus in einer Kategorie mit der Fälschung oder Manipulation von Daten gut aufgehoben? Was gilt eigentlich als Plagiarismus? Kann man sich überhaupt selbst plagiieren, auch wenn man es dürfte? Der Narr, bekannt und (un)beliebt für provokante Thesen und Behauptungen, erlaubt sich an dieser Stelle einmal einen kritischen Blick auf unseren Umgang mit diesem Phänomen in der Wissenschaft und in der Öffentlichkeit.

Zunächst einmal: „Plagiat" und „Plagiarismus" sind keine juristischen Termini. Es gibt sehr wohl ein Urheberrecht, aber da kommen diese nicht vor. In der Wissenschaft sind Plagiate trotzdem nicht gern gesehen, denn in der dort können sie gegen Prüfungsordnungen, Arbeitsverträge oder Universitätsrecht verstoßen. Zwischen rechtswidrigen Übernahmen fremder geistiger Leistungen und der legitimen Übernahme freier oder frei gewordener Ideen gibt es eine Grauzone, wo ein Plagiat zwar als legal, nicht aber als legitim gilt.

Die letzten beiden Sätze habe ich übrigens aus Wikipedia kopiert, und wenn ich Ihnen das jetzt nicht gesagt hätte, wäre es ein Plagiat gewesen. Ich hätte es aber auch so formulieren können: „In der wissenschaftlichen Gemeinschaft kann das Plagiieren gegen die Regeln von Prüfungen, Arbeitsvereinbarungen oder die Rechtsvorschriften von Hochschulen verstoßen. Es existiert eine unscharfe Grenze zwischen der unrechtmäßigen Aneignung von geistigem Eigentum anderer und der zulässigen Übernahme von Ideen, die frei oder gemeinfrei sind. In diesem Bereich kann ein Plagiat zwar rechtlich zulässig, jedoch ethisch nicht vertretbar sein". Selber Inhalt, aber umformuliert von ChatGPT. Die erste Formulierung wäre als Plagiat bei der Überprüfung moniert worden, die von ChatGPT dafür ohne Beanstandung durchgegangen.

Dies zeigt zwei Dinge: Plagiieren mit copy/paste und nur ein paar Wörtern vertauschen ist Oldschool. Wird es so auch bald nicht mehr geben, außer der Plagiator lebt unter einem Stein. Zweitens: Der Plagiatsvorwurf hebt in fast allen Fällen auf die Formulierung ab, nicht auf deren Inhalt. Wer es nicht glaubt, den bitte ich bei den Plagiatsvorwürfe des letzten Jahres, z.B. von der Leyen, Baerbock, Wedel nachzulesen. Es ging nie um den Inhalt – da wurde nie die Aneignung eines besonders originellen oder neuen Gedankens angemahnt, oder gar von Ideen oder Hypothesen der plagiierten Autoren. Es ging um absolut triviale und grenzgradig inhaltsleere Statements – was hingegen einiges über den Zustand der jeweils plagiierten Wissenschaft sagt.

Auch damit entlarven sich die meisten Plagiarismus-Vorwürfe als Kampagnen gegen Personen, und gelten nicht der Sorge um die Integrität der Wissenschaft. Davon lebt ja auch mittlerweile ein ganzer Berufsstand. Professionelle „Plagiarismusjäger" werden dafür bezahlt, sich gezielt unliebsamen Personals oder Widersachern zu entledigen. Das hat dann gar nichts mehr mit Whistle blowing oder wissenschaftlicher Ethik zu tun.

Wenn man die Biowissenschaften betrachtet, wird das Thema Plagiarismus sogar noch komplexer. Abgesehen von Übersichtsarbeiten - und mit sowas sollte man eh nicht Promovieren oder Habilitieren können - geht es doch dort bei einer wissenschaftlichen Arbeit um eine originelle Hypothese oder Fragestellung, und in der Folge um deren kompetente Überprüfung. Die Aussage im Methodenteil: „The protein concentration was

quantified by Bradford Assay. After extraction, the samples were diluted (dilution factor 10) before reading against the calibration curve. After gel electrophoresis with a load of 3 g of protein, …" findet sich in vielen Tausend Papers. Ich weiss gar nicht, wieviele meiner eigenen Artikel mit der Formulierung beginnen „Stroke is the second leading cause of death worldwide and the most frequent cause of long-term disability in adults in developed countries." Sollte man sowas umformulieren, oder gar ChatGPT darum bitten, das zu tun?

Dies ist kein Plädoyer für hemmungsloses Plagiieren. Ordentliches Zitieren gehört zum 1x1 der guten wissenschaftlichen Praxis (GWP). Aber man muss die Kirche im Dorf lassen. Meine GWP-Kurse werden mittlerweile überschattet von Diskussionen und Befürchtungen der Studenten bezüglich der Frage: „Ist das nun schon ein Plagiat?" Außerdem vernebelt die Plagiatsdiskussion oft die Auseinandersetzung um den Inhalt der mutmaßlich plagiierten Aussagen, obwohl da meist gar keiner ist.

Und letztendlich verstellt der Fokus auf den Plagiarismus den Blick auf die gravierenden Verstöße der Manipulation und Erfindung von Daten. Diese sind im biomedizinischen Kontext relevanter, und vermutlich sogar häufiger. Nur entgehen sie (noch) dem Radar der automatisierten Detektion. Auch alle anderen, als „fragwürdige wissenschaftliche Praktiken" verharmloste und nicht sanktionierte Aktivitäten, wie das allgegenwärtige Rosinenpicken in der Auswahl der in das Paper eingehenden (oder ausgeschlossenen) Daten, oder post-hoc Durchführung multipler Tests, verdienten viel mehr Aufmerksamkeit als das dröge Wiederholen von Sätzen anderer.

Außerdem sollten wir uns klarmachen, dass generative, auf einem Large Language Model (LLM) beruhende KI vom Wesen her eine probabilistische Plagiarismus-Maschine ist. Sie frisst Texte, und erzeugt daraus eine begriffslose interne Repräsentation in Form einer multidimensionalen Matrix. Sie errechnet Wahrscheinlichkeitsverteilungen für Gruppen von Zeichen („Tokens"), die anzeigen, wie wahrscheinlich es ist, dass jedes mögliche Token in seinem Wortschatz als Nächstes kommt. Aus solchen Wahrscheinlichkeitsverteilungen erzeugt sie dann den Ausgabetext. Das kann die Übersetzung in eine andere Sprache sein. Oder ein Text, der sich aus der stochastischen Nähe der Wörter errechnet, welche wir dem LLM in unserer Anfrage („Prompt") vorgelegt haben. Mittlerweile sind die Modelle schon recht gut trainiert, und haben von meist unterbezahlten Menschen in Ländern des globalen Südens Manieren beigebracht bekommen. Die Modelle wurden aber auch mit einer Unmenge wissenschaftlicher Texte trainiert, und man hat Ihnen gelehrt, uns die Quellen zu nennen, aus der sich die Myriaden ihrer Matrixwerte speisen. Sie können uns deshalb nun auch ein Literaturverzeichnis liefern. Sie erlauben uns dadurch potentiell den Zugriff auf die geballte wissenschaftliche Erkenntnis der Menschheit, zumindest soweit sie in Text gefasst und durch Training im Modell repräsentiert wurde. Das führt zu einer komplexen Form von Plagiarismus! Und dieser plagiiert selbstverständlich auch all den Forschungsmüll, das mittlerweile Widerlegte, das Zweifelhafte, sowie die Halluzinationen der KI. Letztere werden, da wiederum als Trainingsmaterial für die KI benutzt, die Modelle vermutlich bald zum Kollaps bringen.

Bis es soweit ist, kann uns KI noch Einiges beim Paper-Schreiben abnehmen, was wenig mit Wissenschaft zu tun hat. Am Anfang einer Literatursuche, in der wir uns noch nicht in die Inhalte vertiefen, geht es uns übrigens zunächst auch nicht viel anders als der KI: Wir verwenden krude Heuristiken der Selektion und Aufmerksamkeit, wie Journal Reputation, Datum, Institution, usw. Und dies kann man natürlich dem Modell auch beibringen, denn dafür muss man nicht kapieren, worum es inhaltlich geht. Ein erster, schneller Überblick über das, was da draußen in den Tiefen der Literatur so existiert? Das Rumfeilen an toll klingenden Formulierungen, noch dazu in einer Fremdsprache?

Das Umformulieren von Text aus Angst, er könnte als (Selbst-)Plagiarismus aufstoßen? All das muss doch nicht sein. Die eingesparte Zeit könnte vielmehr in ein ordentliches Studium der Originalliteratur, sowie die ausführliche und nachvollziehbare Beschreibung der Methoden und Resultate investiert werden. Scienceos.ai (Motto: „Get scientific answers by asking millions of research papers“), perplexity.ai (Motto: „Where knowledge begins“) und ähnliche Angebote sind doch erst der Anfang. Wir sollten diese produktiv nutzen und uns mit der Qualitätskontrolle von deren Ausgaben auseinandersetzen.

Disclaimer wie: „Dieser Text wurde unter Verwendung einer KI erstellt“ sind albern, und in ihrer Allgemeinheit auch gar nicht aussagekräftig. Von einem Verbot der Nutzung von KI zur Erstellung wissenschaftlicher Arbeiten ist glücklicherweise kaum noch zu hören, weil gar nicht durchsetzbar und auch von den Entwicklungen des letzten Jahres überholt. Und was die Guten, alten ‚Textduplikationen‘ betrifft, schließe ich mich dem Linguisten John McWhorter von der Columbia Universität an. Er schlägt vor, und damit meint er natürlich nicht den Ideenklau, den Plagiarismus aus der Schmuddelecke zu holen und Wiederverwendung eigener Texte sowie die Nutzung von Standardformulierungen als „cutting and pasting“ zu kennzeichnen. Übrigens: Dieser Text wurde unter Verwendung einer KI erstellt.

Von Korrelation, Kausalität und anderen Kalamitäten

LJ 5/2024

Das menschliche Gehirn ist die ultimative Kausalschlussmaschine. Wir versuchen ständig alles, was in, um, durch und mit uns geschieht, auf spezifische Ursachen zurückzuführen. Offensichtlich sind die Fähigkeit, kausale Beziehungen zu erkennen und die damit mögliche Prädiktion von Ereignissen ein immenser Überlebensvorteil. Natürlich sind auch Tiere sind hierzu in der Lage. Auch sie haben ein Modell der Welt im Kopf, allerdings ist ihr Wissen um die Zusammenhänge, die Wirkmechanismen der Ursachen, allenfalls rudimentär. Die Krähe „weiß“, wie man ein Auto zum Nussknacker machen kann; der Hund, dass Pfötchen geben zu Belohnung führt. Wir wollen aber mehr wissen. Blitzt und donnert es, weil man die Götter erzürnt hat, oder wegen luftelektrischer Entladungen? Geht die Sonne auf, weil sie unter dem Horizont schläft, oder weil sich die Erde dreht? Kriegt man einen Schlaganfall, weil Körpersäfte ins Ungleichgewicht geraten sind, oder ein hirnzuführendes Gefäß atherosklerotisch verschlossen ist? An diesen Beispielen erkennt man schon, dass die Zuweisung von Ursachen keineswegs trivial ist, und dabei auch Fehler gemacht werden können. Man sieht aber auch, dass Ursache-Wirkungsbeziehungen erfolgreiche Prädiktion oft auch ohne tieferes Verständnis der zugrundeliegenden Zusammenhänge erlauben.

Mit der Wissenschaft treibt der Mensch das unserem Gehirn inhärente kausale Inferenzdenken ins Extrem. Die Suche nach den Ursachen, welche hinter den (nicht nur Natur-)

Phänomenen stecken, und die Erforschung der Mechanismen, welche diese Wirkungen erzeugen, sind zentrale Triebfedern jeder wissenschaftlichen Betätigung. Wissenschaftler stehen damit gleichermaßen evolutionär wie intellektuell auf der höchsten Stufenleiter des kausalen Inferenzdenkens. Die meisten von uns sind sich dieses für unsere tägliche Arbeit so wichtigen Status gar nicht bewusst.

Abgesehen von Offensichtlichem, wie: „Zu lange in der Sonne sitzen führt zu Sonnenbrand", ist die korrekte Zuordnung von Ursachen zumeist eine Herausforderung. Selbst der heute selbstverständliche Warnhinweis: „Rauchen führt zu tödlichem Lungenkrebs", welcher viele Zigarettenschachteln ziert, war in der Wissenschaft mehr als 20 Jahre hart umkämpft. Natürlich nutzt jeder von uns Heuristiken, mit Hilfe derer wir Ursache-Wirkungsbeziehungen wahrscheinlich machen können. Wenige davon sind so eindeutig wie die zeitliche Abfolge: Die Ursache muss der Wirkung zeitlich vorausgehen. Damit lässt sich einiges ausschließen, aber weniges als Ursache absichern. Die Schwäche des Kriteriums entlarvte der amerikanische Schriftsteller Ambrose Gwinnett Bierce (1842 - 1914) so: „Wirkung ist die zweite von zwei Erscheinungen, die immer in derselben Aufeinanderfolge vorkommen. Von der ersten, Ursache genannt, sagt man, sie bringt die zweite hervor – was nicht vernünftiger ist, als würde jemand ein Kaninchen für die Ursache eines Hundes halten, nur weil er noch nie einen Hund anders als bei der Verfolgung eines Kaninchens gesehen hatte." Wenn dann zum Kriterium der zeitlichen Abfolge noch Konsistenz, Spezifität, Stärke der Assoziation (Dosis-Wirkungsbeziehung), Plausibilität, Nicht-Existenz von Alternativerklärungen bzw. Kohärenz mit etabliertem Wissen dazu kommen, ist man wohl auf einer heißen Spur einer Kausalbeziehung. Aber Festnageln lässt sich die Kausalität damit immer noch nicht.

Damit dies gelingt, muss man experimentieren! Unter Berücksichtigung verschiedener Limitationen und Ausschluss von Verzerrungen (Bias) lassen sich im Experiment eindeutig Kausalbeziehungen belegen. In der heutigen, hoch entwickelten Form, d.h. mit Randomisierung, Verblindung, Kontrollgruppen, sowie statistischer Analyse gibt es das erst seit etwa 100 Jahren. In der klinischen Medizin werden Ursache-Wirkungsbeziehungen experimentell in randomisiert kontrollierten Studien (RCT) untersucht, und deren Methodik wurde erst in den 60er Jahren des vorherigen Jahrhunderts entwickelt.

Die experimentelle Manipulation (respektive Intervention in der RCT) ist eindeutig der Königsweg im Ergründen von Kausalitäten. Nur lässt sie sich aus praktischen oder ethischen Gründen häufig nicht einsetzen. Dann muss sich die Wissenschaft auf die Beobachtung beschränken. Und die stimuliert unsere mentale Kausalschlussmaschine gewaltig. Denn Zusammenhänge (Korrelationen) lassen sich überall finden. Wie z.B. eine sehr gute Korrelation zwischen dem nationalen Schokoladenkonsum und der Anzahl der Nobelpreisträger des Landes. Oder der zwischen Eiskonsum und der Zahl der Ertrinkungstoten. Hier sind die ‚Scheinkorrelationen' noch einfach zu entdecken: In ‚reichen' Nationen wird viel Schokolade gegessen, aber auch viel in Forschung und Entwicklung investiert. Und wenn es heiß ist, gibt's ein Eis und man geht ins Strandbad, wo man ertrinken kann. Die ‚Confounder' sind hier offensichtlich. Ein oder mehrere andere Ursachen, wirkten gleichzeitig auf die vermutete Ursache und die untersuchte Wirkung. Aber wie ist es mit der Aufnahme von Aluminium und Alzheimer, mit Krebsrisiko von rotem Fleisch, mit Handynutzung von Heranwachsenden und deren psychischen Problemen?

Die wissenschaftlichen Journale, und in der Folge oft auch die Zeitungen, sind voll von behaupteten Kausalbeziehungen, welche sich bei näherem Hinsehen, wenn nicht gleich als statistische Fehler oder Fehlinterpretation, später doch als bloße (Schein)Korrelationen herausstellen. Im Statistikgrundkurs, den fast jeder Wissenschaftler am

Karriereanfang absolvieren musste, hob der Dozent an dieser Stelle den Zeigefinger und proklamierte: Korrelation ist nicht Kausalität!

Wenn es nur so einfach wäre. Denn hinter vielen Korrelationen lauert tatsächlich eine Kausalbeziehung. Doch die von den meisten Wissenschaftlern benutzten statistischen Analyseverfahren zur Untersuchung der Zusammenhänge von Variablen, die Regressionsanalyse (mit Prüfung auf deren statistische Signifikanz), hilft uns hier nicht weiter. Denn nicht nur die Regressionsanalyse, die gesamte frequentistische Statistik (‚NULL-Hypothesen Signifikanztestung') kennt keine Kausalbeziehungen. Sondern nur Korrelationen. So kann man bei der Regressionsgeraden X und Y vertauschen, am Korrelationskoeffizienten wird sich nichts ändern. Schon die Urväter dieser Form der Statistik, allen voran der Positivist Karl Pearson (nach ihm ist der Korrelationskoeffizient r benannt), waren nämlich der Meinung, dass Kausalität lediglich ein Extremfall von Korrelation ist, und zwar bei einem Korrelationskoeffizienten von -1 oder +1. In „The Grammar of Science" (1892; Zitate und weiterführende Links wie immer unter http://dirnagl.com/lj) schreibt Pearson: „Variation und Korrelation schließen Kausalität und Determinismus als Spezialfälle ein, sofern sie tatsächlich in Bezug auf Phänomene eine wirkliche Existenz haben. Keine Erfahrung, die wir bis jetzt gemacht haben, rechtfertigt uns jedoch anzunehmen, dass sie etwas anderes sind als konzeptionelle Grenzen, geschaffen durch das menschliche Bedürfnis nach Ökonomie des Denkens, und so wenig in den Phänomenen selbst inhärent wie geometrische Flächen oder Kraftzentren."

Nicht nur die klassische Statistik hat ein Problem mit der Kausalität. Auch einige berühmte Physiker knabberten daran. Zum Beispiel Bertrand Russell, der Kausalität gar für ein unwissenschaftliches Konzept hielt. „Alle Philosophen stellen sich vor, dass Kausalität eines der grundlegenden Axiome der Wissenschaft ist, doch seltsamerweise kommt in den fortgeschrittenen Wissenschaften das Wort ‚Ursache' nie vor." Und weiter: „[…] weil die Gesetze der Physik alle symmetrisch sind, in beide Richtungen gehen, während kausale Beziehungen unidirektional sind, von der Ursache zur Wirkung gehen." In der Tat, $f=m \cdot a$ kann man auch als $a=f / m$ schreiben. Trotzdem gehen wir alle, auch die Physiker und die Ingenieure, welche mit dieser Formel Raketen zum Mond schicken, von der Direktionalität dieser Beziehung aus. Kraft verursacht Beschleunigung – und nicht Beschleunigung Kraft! Das Ganze kann man als philosophisches Rätsel kultivieren, oder eben als einen Mangel der formalen Sprache der Algebra begreifen, welche hier verwendet wird. Diese kennt keine Direktionalität. Wir lesen solche Gleichungen trotzdem richtungsbezogen, da wir wissen, dass wir in eine Variable eingreifen können, um eine andere zu ändern. Das Zurechtkommen im Alltag, ebenso wie das wissenschaftliche Experimentieren, setzen Kausalität voraus.

Es fehlt aber ein kausales Werkzeug im Kasten der klassischen Statistik! Wie können wir dennoch dort, wo wir keine kontrollierten und randomisierten Interventionen durchführen können, unsere Neigung, hinter jeder Ecke Kausalbeziehungen zu erkennen, im Zaum halten? Und Ursache – Wirkungsbeziehungen wissenschaftlich ableiten? Zum Beispiel in den in der Gesundheits- und medizinischen Forschung extrem verbreiteten Beobachtungsstudien. Diese untersuchen kausale Effekte und berichten „unabhängige" Assoziationen oder liefern „Prädiktoren". Dies tun sie zumeist nach Anpassung an eine oder mehrere Variablen („multivariables Regressionsmodell"). Leider wird dabei häufig nicht berichtet, welche Rolle jede Variable in Bezug auf die Exposition und das Ergebnis spielt. Auch bleibt oft unklar, warum einige Variablen zur Berücksichtigung ausgewählt wurden und andere nicht. Ohne diese Informationen sind aber viele der berichteten Assoziationen nicht interpretierbar, da die Schätzung eines spezifischen kausalen Effekts die Anpassung an eine spezifische Gruppe von Variablen erfordert.

Hierfür braucht es kausale Modelle! Judea Pearl hat diese durch die Einführung sogenannter gerichteter azyklische Graphen (directed acylic graphs, DAG) perfektioniert, um damit kausale Beziehungen zu modellieren und Hypothesen über kausale Effekte zu testen. Durch Open-Source Software wie DAGitty oder R-Packages wie dagR ist diese Methodik eigentlich jedem Wissenschaftler problemlos zugänglich. Leider ist das bisher aber nur bei Epidemiologen und in der Machine-Learning- und Data-Science-Community angekommen. Dabei sind DAGs ein fantastisches Werkzeug, um kausale Beziehungen zu erkennen, zu modellieren, zu verstehen und zu quantifizieren. Und das insbesondere wenn, was leider häufig der Fall ist, eine zufällige Zuweisung von Interventionen bzw. Behandlungen nicht möglich ist.

Gerichtete azyklische Graphen sind aber auch für Experimentatoren von großem Nutzen. DAGs können ihnen helfen, die Zusammenhänge zwischen Einflussgrößen innerhalb ihres Experiments besser zu verstehen, indem sie eine klare visuelle Darstellung der angenommenen kausalen Beziehungen zwischen verschiedenen Variablen bieten. Dies ermöglicht es, die Struktur der Daten und die potenziellen Wechselwirkungen zwischen den Einflussgrößen zu visualisieren.

Mit DAGs lassen sich viele wichtige Fragen klären: Welche der Variablen können als potenzielle Ursachen und welche als mögliche Wirkungen betrachtet werden? Wo lauern Confounder, die sowohl mit der unabhängigen als auch mit der abhängigen Variablen in Zusammenhang stehen und die kausale Beziehung verzerren könnten? Welche Variablen müssen manipuliert werden, um den gewünschten Effekt auf andere Variablen zu haben? Wie können Fallstricke wie Collider-Bias vermieden werden, der entsteht, wenn auf eine Variable kontrolliert wird, die eine gemeinsame Wirkung zweier anderer Variablen ist? DAGs erlauben damit eine systematische Analyse und Planung von Experimenten, und bieten damit die Grundlage für das Verständnis und die Modellierung komplexer kausaler Beziehungen.

Falls Sie von DAGs bisher noch nichts gehört haben, an dieser Stelle eine Leseempfehlung: „The Book of Why: The New Science of Cause and Effect“, vom Meister Judea Pearl selbst (zusammen mit Dana Mackkenzie). Darin wird die Entwicklung und Bedeutung der Kausalitätstheorie in der modernen Forschung erzählt. Insbesondere aber werden uns darin auf unterhaltsame Weise die DAGs nahegebracht. Sie werden die Lektüre nicht bereuen. Selbst der ewig schlaumeiernde Wissenschaftsnarr, der bis zum Lesen dieses Buches die Regressionsanalyse für den Gipfel der statistischen Auseinandersetzung mit Korrelation und Kausalität hielt, wurde dadurch eines Besseren belehrt.

PubPeer – Forum für persönliche Vendettas oder Zukunft des Peer Review?

LJ 6/2024

Der deutsch-amerikanische Hirnforscher und Nobelpreisträger 2013 Thomas Südhof macht derzeit Schlagzeilen. Südhof forscht an der Stanford University und ist Berlin u.a. durch eine langjährige Fellowship der Charité verbunden. In den Zeitungen war er vor kurzem allerdings nicht durch Entdeckungen auf seinem Fachgebiet, der komplexen Interaktion von Nervenzellen. Sondern vielmehr durch Kommentare auf der Internet Plattform PubPeer. Dort werden ihm bei einer Vielzahl von Artikeln aus seiner Arbeitsgruppe Schlampigkeit, wenn nicht sogar Datenmanipulation vorgeworfen. In der Folge hat er bereits bei einigen der dort kritisierten Artikel Korrekturen veröffentlichen lassen, einer davon wurde mittlerweile zurückgezogen. Dass nun Südhof im Rampenlicht steht, ist in mehrfacher Hinsicht bemerkenswert: Als Nobelpreisträger ist Südhof ein wissenschaftlicher Leuchtturm, er ist aber auch seit Jahren einer der prominentesten Kritiker des kommerziellen akademischen Publikationswesens und ein Streiter für eine robustere, reproduzierbarere Forschung.

Wir erinnern uns: Erst kürzlich war die deutsche Krebsforscherin Simone Fulda in die Schlagzeilen geraten, nachdem ihr - ebenfalls auf PubPeer - Datenmanipulation in einer Reihe ihrer Veröffentlichungen vorgeworfen wurde. Sie legte daraufhin im Februar ihr Amt als Präsidentin der Universität Kiel nieder, der Fall befindet sich bei der DFG noch im Stadium der Vorprüfung. Die Liste lässt sich fast beliebig fortsetzen, vor Simone Fulda war z.B. der Neurowissenschaftler und Präsident der Stanford University, Marc Tessier-Lavigne auf den internationalen Titelseiten. Auch er trat zurück, nachdem – ebenfalls auf PubPeer - in Abbildungen aus seiner Arbeitsgruppe offensichtliche Manipulationen aufgedeckt worden waren.

Wieso spielt PubPeer bei all diesen Vorgängen eine so zentrale Rolle? PubPeer ermöglicht jedermann wissenschaftliche Publikationen zu kommentieren. Es handelt sich dabei um einen „Post-Publication-Review“, also eine Begutachtung, nachdem die Ergebnisse bereits veröffentlich wurden. Diese werden hier also ein zweites Mal begutachtet, denn sie wurden ja bereits vor der Veröffentlichung im eigentlichen Peer-Review-Prozesses gereviewt. Bei letzterem evaluieren in der Regel 2, manchmal auch bis zu 4 Fachexperten („Peers“) das Manuskript. Wegen der hochgradigen Komplexität der Fragestellungen, der Methodik und der Analysen und des oft massiven Umfanges der Manuskripte kann dieser Prozess, wenn überhaupt, nur die allergröbsten Fehler und Probleme identifizieren. Von einer wirklichen Qualitätskontrolle kann deshalb nicht mehr gesprochen werden.

Hier setzt das beim wissenschaftlichen Publizieren bisher kaum eingesetzte Prinzip des Post-Publication-Review an: Es erlaubt der gesamten Leserschaft, in der Regel also einer

großen Anzahl von Experten, die Begutachtung und Diskussion von bereits veröffentlichten Artikeln. Die PubPeer Plattform, betrieben von einer amerikanischen, nichtkommerziellen Stiftung, ermöglicht die Kommentierung beliebiger wissenschaftlicher Publikationen, unabhängig davon in welchem Journal sie veröffentlicht wurden. Damit treten Leser in einen Diskurs mit den Autoren, welche auf die dort hochgeladenen Kommentare reagieren können. Um durch Hinweise auf Fehler in den Arbeiten oder Datenmanipulationen nicht Sanktionen durch mächtige Autoren fürchten zu müssen, kann auch anonym kommentiert werden. Die Kommentare, und Reaktionen der Autoren darauf, werden von den Betreibern der Plattform nur ‚moderiert', wenn sie nicht den Anforderungen der ‚Commenting guidelines' entsprechen.

Häufig werden auf PubPeer, wie auch bei Südhof, Fulda oder Tessier-Lavigne geschehen, Abbildungen von biochemischen, molekularbiologischen oder histologischen Ergebnissen diskutiert bzw. in Frage gestellt. Mit großem Abstand werden hier am häufigsten Duplikationen, Verschiebungen, und andere Auffälligkeiten in fluoreszenmikroskopischen Abbildungen sowie in Western-Blots moniert. Das liegt ganz einfach daran, dass es sich hier um experimentelle Standardtechniken handelt, die jeder Experte, auch ohne Fachwissen auf dem Gebiet des Artikels, ja sogar ohne weitere Kenntnis des sonstigen Inhalts einer Studie gut visuell inspizieren und interpretieren kann. Es liegt aber auch daran, dass das menschliche Gehirn, mittlerweile oft unterstützt durch künstliche Intelligenz (KI), sehr gut Muster erkennen kann. Dem geschulten Auge oder der KI gelingt es so, Abbildungen, in denen Teile verdoppelt, gespiegelt, oder verzerrt wurden, zu identifizieren. Meist steht dann die Frage im Raum, ob dies das Resultat eines methodischen Artefaktes oder eines „ehrlichen" copy-paste Fehlers bei der Erstellung der Abbildung oder deren Beschriftung war. Oder eben in betrügerischer Absicht erfolgt ist. Etwa um Resultate vorzuspiegeln, die sich im Experiment so nicht einstellten (aber für die ‚Story' relevant waren), oder um die Studienergebnisse ‚spektakulärer' erscheinen zu lassen, als sie waren, und so eine prestigeträchtige Publikation wahrscheinlicher zu machen. Oder einfach nur als ‚Abkürzung', nach dem Motto: ‚Ähnliche Kontrollen haben wir doch schon in anderen Experimenten gemacht, lasst uns die nehmen, jetzt wäre auch nichts anderes rausgekommen'.

Durch Diskussionen auf PubPeer wurden bereits eine Vielzahl von Fehlern in wissenschaftlichen Publikationen in verschiedenen Disziplinen aufgedeckt. Durch Korrektur der Veröffentlichung, oder im Extremfall Zurückziehen der Arbeit wurde der Wissenschaft so ein großer Dienst erwiesen. Viele dieser Fehler waren „ehrlich". Die Wissenschaft, ihre Methoden und Analysen sind komplex und kompliziert. Dass Fehler gemacht werden, ist nachgerade notwendig und normal. Leider gibt es aber in den meisten wissenschaftlichen Disziplinen keine entwickelte Fehlerkultur (der Narr hierzu in LJ 1-2/2019). Fehler werden aus Angst vor Reputationsverlust oder Sanktionen lieber unter den Tisch gekehrt, als diese zu kommunizieren, zu analysieren und damit eine Wiederholung unwahrscheinlicher zu machen.

Und hier liegt leider eine Schattenseite von PubPeer: Paradoxerweise verstärkt die Möglichkeit, dort mit Fehlern offen umzugehen, die Angst der Wissenschaftler davor, vor ihren Kollegen bloßgestellt zu werden. Denn die bloße Tatsache, die eigene Arbeit bei PubPeer mit dem Hinweis auf einen möglichen Fehler wieder zu finden, kann eben wegen des Fehlens einer wissenschaftlichen Fehlerkultur zu Reputationsverlust und bei Konkurrenten zu Schadenfreude führen.

Und es gibt noch ein anderes Problem: Als anonyme Plattform eignet sich PubPeer trefflich für private Vendettas und Kampagnen, die nichts mit Wissenschaft, aber dafür umso mehr mit Neid oder persönlicher Animosität zu tun haben.

Ein instruktives und interessantes Beispiel für das Niveau der Auseinandersetzungen auf PubPeer zeigt die Diskussion einiger weniger Abbildungen in einem Paper aus Südhofs Gruppe (doi: 10.1016/j.neuron.2017.04.011). In wenigen Wochen haben sich auf PuPeer 40 (Stand 16.5.2024) Comments und Responses angehäuft, die sich teilweise über Seiten ziehen und häufig mit bildanalytischen Auswertungen und hoch-technischen Auslassungen über die verwendete Gerätschaft und Software garniert sind. Man erkennt dabei die Ernsthaftigkeit und auch Kompetenz der meisten Kommentatoren. Es schleicht sich aber auch das Gefühl ein, dass sich manche Kritiker wiederholen oder sich in irrelevanten Details ergehen. Bei dem Versuch, sich ein Bild der Kontroverse zu machen, verliert man am Ende den Überblick, und je nach persönlichem Bias und Expertise denkt man sich entweder: ‚Irgendwie geht das zu weit, hier wird eine Sau durchs Dorf getrieben', oder: ‚Wenn schon so wenige Abbildungen in einer Arbeit problematisch sind, wie steht es eigentlich mit dem Rest? Richtig und wichtig, dass auch ein Nobelpreisträger den Standards der guten wissenschaftlichen Praxis gemessen wird!'.

Auf den von Südhof eigens für die Auseinandersetzung mit den PubPeer- Vorwürfen aufgesetzten Webseiten (https://med.stanford.edu/sudhoflab/integrity---pubpeer.html, diese und weitere Zitate und Links finden sich wie immer unter: http://dirnagl.com/lj) beschwert sich Südhof: „PubPeer und andere soziale Medien sind intransparent, zensieren Antworten und verwenden anonyme Kommentatoren mit ständig wechselnden Aliasnamen. Zudem werden PubPeer-Beiträge sofort an Zeitschriften, Journalisten und Universitätsverwaltungen weitergeleitet. Nicht alle Kommentare auf PubPeer sind aufrichtig, und einige scheinen eine Agenda zu verfolgen, die nichts mit wissenschaftlicher Integrität zu tun hat. Wir reagieren dennoch auf solche Kommentare, weil Zeitschriften (an die die Kommentatoren ihre Anschuldigungen übermitteln) manchmal Anschuldigungen akzeptieren, ohne ihre Plausibilität zu überprüfen. Darüber hinaus besteht eine gängige Strategie einiger Kommentatoren darin, Anschuldigungen auf verschiedene Weise zu wiederholen, um die Wirkung zu verstärken, oft begleitet von Grafiken und Animationen, die einen Anschein von Ernsthaftigkeit vermitteln. Selbst wenn eine Anschuldigung eher unwahrscheinlich ist, schafft diese Strategie eine Aura von "hier stimmt etwas nicht", die für die Beschuldigten schwer zu widerlegen ist."

Die wenigen, welche auf PubPeer mit offenem Visier diskutieren, also nicht anonym bleiben wollen, sind häufig das, was man derzeit als ‚professional data sleuths' bezeichnet. Hierzu zählen z.B. Leonid Schneider, Sholto David, und die über die Vielzahl der von ihr mit aufgedeckten Fällen von Bildmanipulationen mittlerweile selbst über Fachkreise prominent gewordene Elisabeth Bik. Data sleuths leben, z.B. über Platformen wie Patreon oder über crowd sourcing, häufig zumindest teilweise, und vermutlich mehr schlecht als recht, vom Aufspüren von schlechter Wissenschaft. Thomas Südhof, und er ist nicht allein damit, wirft ihnen aber genau das vor, nämlich ein ‚finanzielles Interesse' zu haben.

Und da liegt er voll daneben. Zum einen ist das ein unbrauchbares Argument gegen die von den Sleuths ja immer ausführlich belegten Auffälligkeiten. Die Auffälligkeiten sind entweder real oder sie sind es nicht, das hat nichts mit dem finanziellen Hintergrund derer zu tun, die sie aufzeigen. Zum anderen ist es traurig, dass es so einen ‚Berufsstand' offensichtlich braucht. Und wenn schon, dann sollte er von Institutionen (also den Unis) und den Verlagen finanziert werden. Ich meine wir haben nicht zu viele Data sleuths, die sich dieser Aufgabe professionell widmen, sondern viel zu wenige. Es ist sogar leider so, dass es sich lohnen würde, einen Studiengang zur Ausbildung eines kompetenten Nachwuchses aufzubauen, und dauerfinanzierte Stellen in den Forschungsabteilungen der Unis für diese einzurichten.

Festzuhalten bleibt, dass PubPeer das derzeit beste Medium ist, das die Korrektur der wissenschaftlichen Literatur erlaubt. Der etablierte Wissenschaftsbetrieb (einschließlich unehrlicher Forscher, ängstlicher Herausgeber und klageunwilliger Institutionen) ist bisher nicht in der Lage hierzu. Auch zeigt der ‚Fall Südhof' eindrücklich die Stärken des Post-Publication-Review Prozess und einer Plattform wie PubPeer – und wie wichtig es ist, dass sich Wissenschaftler (als Kommentatoren, aber auch als kritisierte Autoren) auf dieses Format einlassen. Thomas Südhof und seine Koautoren haben, im Gegensatz zu vielen anderen, deren Arbeiten auf PubPeer kritisiert wurden, zeitnah auf jeden Kommentar konstruktiv und wissenschaftlich reagiert. Sie konnten einen erheblichen Teil der Kritik aufklären bzw. ausräumen, und haben tatsächliche Fehler eingestanden und korrigiert. Ausdrücklich bedankt sich Südhof bei denen, die konstruktiv kommentiert hatten, und lobt das Prinzip von PubPeer. Nach gegenwärtigem Stand hatten die ihm nachgewiesenen Fehler relevante Aussagen der Arbeiten nicht beeinflusst. Absichtliche Manipulationen oder Täuschung wurden meiner Ansicht nach nicht überzeugend nachgewiesen – wobei die Diskussion hierüber noch nicht abgeschlossen ist. Sehr wohl aber weist die Häufung von Fehlern im Südhof-Labor auf allgemeine Qualitäts- und Supervisionsprobleme seiner vermutlich sehr großen Arbeitsgruppe hin.

Sicher scheint mir aber, dass die vermehrt auch in der Öffentlichkeit diskutierten PubPeer Fälle von möglichem Wissenschaftsbetrug nur die kleine Spitze eines gigantischen Eisberges darstellen. Sie sind nur wegen der Prominenz der beteiligten Wissenschaftler besonders sichtbar. Große Teile der Wissenschaft haben ein Qualitäts- bzw. Integritätsproblem, deren Hauptursachen wie in dieser Kolumne schon häufiger angeprangert, in einem falschen akademischen Anreizsystem liegen. PubPeer ist ein Indikator hierfür, ich glaube aber auch es ist auch Teil der Lösung.

Im Übrigen lehrt PubPeer uns, dass jeder Wissenschaftler sich darüber bewusst sein sollte, dass die eigenen Publikationen nach deren Veröffentlichung Gegenstand einer intensiven Auseinandersetzung mit einer breiten Community werden können. Die Einhaltung aller Regeln der guten wissenschaftlichen Praxis und die gute Dokumentation der Resultate und Analysen (in Originalformaten) sich also langfristig auszahlt, und es nicht nur kurzfristig auf eine tolle Publikation ankommt. Deshalb empfiehlt der Wissenschaftsnarr auch die Installation des PubPeer Browser Plugins – das auf existierende PubPeer-Diskussionen hinweist, wenn man Artikel sucht oder liest. Wir sollten PubPeer mit kühlem Kopf und wissenschaftlicher Etikette weiter passiv wie aktiv - idealerweise mit offenem Visier - nutzen, und es zu dem internationalen „Online Journal Club" machen, als der es sich selbst bezeichnet.

Dieser Artikel basiert auf einem Gastbeitrag des Autors im Berliner Tagesspiegel vom 12.4.2024

Wissenschaftsfreiheit als Freibrief für schlechte Forschung?

LJ 9/2024

„Kunst und Wissenschaft, Forschung und Lehre sind frei. Die Freiheit der Lehre entbindet nicht von der Treue zur Verfassung“. Fünfundsiebzig Jahre Grundgesetz, 75 Jahre Artikel 5 (3)! Und alle reden von Wissenschaftsfreiheit – sowie deren akuter Gefährdung. Bei der Jahresversammlung der DFG unterstrich deren Präsidentin Katja Becker soeben ausdrücklich „die Bedeutung und den Wert der Wissenschaftsfreiheit, die von allen Akteuren im deutschen Wissenschaftssystem gemeinsam getragen und gelebt werden muss“. Was manche im BMBF wohl nicht mitbekommen hatten: Sie prüften zeitgleich die Sanktionierung von „verwirrten Gestalten“ durch Förderentzug. Gemeint waren Wissenschaftler, die in einem Protestbrief die Räumung eines propalästinensischen Camps an der FU Berlin kritisiert hatten. Auch freute man sich über die mit so einer Maßnahme zu erreichende Selbstzensur beim Rest der deutschen Forscher (Zitate und weiterführende Literatur wie immer unter https://dirnagl.com/lj).

Das weckte Assoziationen zu Ungarn, wo die von der Akademie der Wissenschaften betriebenen Forschungsinstitute einem neuen Träger unterstellt werden, bei dem Regierungsvertreter das Sagen haben. Oder zu den USA, wo die Heritage Foundation mit dem „Project 2025“ das 922-seitige Skript für die Zeit der 2. Präsidentschaft von Donald Trump veröffentlicht hat. Darunter auch „Schedule F“, ein Präsidialerlass, den Trump am Tag seiner Inauguration wiedereinsetzen will. Er hatte diesen in seiner ersten Amtszeit nicht voll umsetzen können. Staatsbeamten, die als illoyal gegenüber dem Präsidenten wahrgenommen werden, wird darin der Schutz entzogen, diese können entlassen werden, bei Einstellungen werden sie zu Loyalitätsbekundungen gegenüber dem Präsidenten ermutigt. Beamte der National Institutes of Health, der Food and Drug Administration, der Centers for Disease Control, der Environmental Protection Agency, etc., können so gefügig gemacht werden, oder eben gegen willigere Kollegen ausgetauscht. Mit absehbaren Folgen für die Wissenschaft dieser Institutionen und deren Wissenschaftler.

Auch die AfD macht sich Sorgen um die Wissenschaftsfreiheit in Deutschland. In einer aktuellen Anfrage an den Bundestag kritisiert sie die Verwendung des „Academic Freedom Index“, einer systematischen, weltweiten Erhebung zur Wissenschaftsfreiheit durch die Bundesregierung, auch weil er „keine Einschränkungen erfasst, die von akademischen Akteuren ausgehen“. Auffallend an der Fürsorge der AfD um die Wissenschaftsfreiheit ist der Kontrast zur bisher einzigen wissenschaftspolitischen Forderung dieser Partei, Lehrstühle für Gender Studies abzuschaffen. Noch bemerkenswerter ist aber, dass sich nicht nur die AfD, sondern auch das BMBF Sorgen um die Gesinnung deutscher Wissenschaftler macht, welche Art. 5 (3) GG missbrauchen könnten. Sogar manch gewöhnlicher Wissenschaftler argumentiert des Öfteren mit dem Grundgesetz.

Dies insbesondere dann, wenn er verdächtigt wird, Ergebnisse manipuliert oder gar gefälscht zu haben. Nicht nur die Feier der verfassungsrechtlich garantierten Freiheit der Wissenschaft, sondern auch deren Instrumentalisierung feiert also fröhliche Urstände.

Wird damit der Verweis auf Wissenschaftsfreiheit zum rhetorischen Trick, mit dessen Hilfe sich Wissenschaftler ihrer gesellschaftlichen Verantwortung entziehen können, wie Torsten Wilholt das in „Die Freiheit der Forschung: Begründungen und Begrenzungen" formuliert? Dies ist eine der Fragen, die Lucia Reuter in ihrer Dissertation „Wissenschaftsfreiheit oder Narrenfreiheit – Biomedizinische Forschung zwischen Freiheit und Verantwortung" nachgeht. Die Müh(l)en eines Promotionsverfahrens an der Charité malen aber leider sehr, sehr langsam, deshalb konnte sie die Arbeit noch nicht verteidigen bzw. publizieren.

Informiert und inspiriert durch diese Dissertation, und weil das Thema derzeit so aktuell ist, lädt sie der Narr deshalb zu einem kurzen Exkurs ein. Und zwar zu einem in der gegenwärtigen Diskussion um die Forschungsfreiheit und deren Gefährdung sträflich vernachlässigten Aspekt. Nämlich der Idee, dass mit Freiheit auch Verantwortung verbunden ist, quasi als Bringschuld. „Aus Freiheit erwächst Verantwortung" – so Bundespräsident Frank-Walter Steinmeier in einer Rede zur Forschungsfreiheit bei der diesjährigen Tagung der Humboldt-Stiftung. Er meinte aber nicht die Verantwortung zur Einhaltung methodischer Qualitätsstandards, zur Vermeidung fragwürdiger Forschungspraktiken und damit von Forschungsmüll und Ressourcenverschwendung. Oder die Verantwortung der Wissenschaftler gegenüber der sie alimentierenden Gesellschaft, effizient und transparent zu forschen. Steinmeier, wie auch andere Festredner bei den Feierlichkeiten zum Jubiläum des Grundgesetzes, wollte vielmehr darauf hinaus, „dass Wissenschaftlerinnen und Wissenschaftler sich als Bürgerinnen und Bürger für Freiheit, Demokratie und Rechtsstaatlichkeit einsetzen".

Lassen Sie uns aber über die Verantwortung reden, gute und relevante Forschung zu machen. Doch wie kommt man eigentlich von „Freiheit", der abstrakten und damit inhaltsleeren Abwesenheit von Zwang, die jegliches Verhalten zu erlauben scheint, zu Verantwortung? Noch dazu im Verfassungsrecht, wo doch im Art.5 (3) von Verantwortung gar nichts steht – man muss nur ‚verfassungstreu' sein, also Demokrat im Sinne des GG. Lädt der Artikel des GG damit nicht geradezu dazu ein, als Freibrief für die Wissenschaft verstanden zu werden, los und ledig jeglicher Rücksichtnahme auf Gesellschaft und wissenschaftliche Standards drauf los zu forschen? Für die Beantwortung dieser Frage ist ein kurzer Primer zur verfassungsrechtlichen Garantie der „Freiheit der Wissenschaft" hilfreich.

Ein Verfassungsartikel, der Forschungsfreiheit garantiert, ist zunächst einmal nichts Besonderes. Viele Verfassungen haben so etwas, auch der Art. 13 der Charta der Grundrechte der Europäischen Union liest sich wie aus der deutschen Verfassung kopiert. Interessanterweise geht es aber wohl auch ohne, denn entsprechendes fehlt z.B. in den Verfassungen der USA, Kanadas, Englands, der Niederlande, oder Irlands. In Deutschland findet sich dagegen die Forschungsfreiheit, fast identisch zum heutigen Wortlaut, bereits in der „Paulskirchenverfassung" von 1849. Man könnte also fast sagen, dass das Ganze eine deutsche Erfindung ist. Auch die Weimarer Verfassung hatte einen Forschungsfreiheitsartikel. Der ging aber mitsamt der Verfassung über Bord, unter aktiver Beteiligung eines substantiellen Teils der deutschen Professorenschaft. Man denke nur an Philipp Lenards „Deutsche Physik", die „Die Verfassung der Freiheit" (1935) des Staatsrechtlers Carl Schmitt, oder Gerhard Wagners „Neue Deutsche Heilkunde", wobei letztere (nicht nur) im noch heute gültigen Heilpraktikergesetz von 1939 geistert. Angesichts des Track records der Wissenschaft im 3.Reich kam es 1948 beim

Verfassungskonvent sogar zu Diskussionen, ob man deren Akteuren überhaupt noch „Freiheit der Forschung“ zugestehen sollte. Was den, in allen anderen Ländern unüblichen, (Zu)satz zur Verfassungstreue begründet.

Aber worum geht es eigentlich, wenn der Staat Forschungsfreiheit gewährt? Zunächst ist das ein Abwehrrecht, es schützt alle selbständig wissenschaftlich Tätigen vor Eingriffen des Staates, der Kirchen, und auch der Öffentlichkeit. Es gewährt Wissenschaftlern Freiheit in der Wahl der Themen, der Methoden, der Zugänge, der Publikation. Interessanterweise, und vielen Wissenschaftlern gar nicht bewusst, handelt es sich bei Art. 5 (3) GG aber auch um ein Gewährleistungsrecht: Der Staat muss Wissenschaft ermöglichen, er verpflichtet sich selbst zur Förderung der Wissenschaft. Er muss für die Finanzierung und Infrastruktur sorgen, und das erstreckt sich nicht nur auf Individuen, sondern auch auf die Institutionen, wie die Universitäten, oder außeruniversitären Forschungsinstitute. Aber freuen Sie sich nicht zu früh: Ihre Verstetigung als Postdoc oder Professor, die Bewilligung des nächsten DFG-Antrages, oder einfach nur die Reparatur eines Gerätes der Grundausstattung ihres Institutes können sie damit trotzdem nicht vor dem Verfassungsgericht erstreiten, denn im GG geht es ums Prinzipielle!

Natürlich ist die Freiheit, die uns Wissenschaftlern zugestanden wird, nicht grenzenlos. Sie endet klar dort, wo sie andere (Grund)rechte einschränkt, wie zum Beispiel Menschenrechte, Tierschutz, oder Persönlichkeitsrechte. Aber wie ist es eigentlich mit schlechter Wissenschaft, ist die auch geschützt? Und wie ist es mit der Verantwortung der Wissenschaft gegenüber der Gesellschaft? Gewährt die Verfassung in diesen Dingen Narrenfreiheit?

Laut Bundesverfassungsgericht ist Wissenschaft im Sinne von Art. 5 (3) GG nach jede Tätigkeit, die „nach Inhalt und Form als ernsthafter planmäßiger Versuch zur Ermittlung der Wahrheit anzusehen ist“. Und weiter: „Die Wissenschaftsfreiheit schützt daher auch Mindermeinungen sowie Forschungsansätze und -ergebnisse, die sich als irrig oder fehlerhaft erweisen“. Dagegen schützt Art. 5 (3) GG Forschung nicht, wenn sie lediglich den Anschein einer wissenschaftlichen Vorgehensweise besitzt oder wissenschaftliche Standards deutlich verfehlt. (BVerfGE 90, 1 – 21).

Damit ist klar, dass derjenige, der Wissenschaftsbetrug begeht, nicht Schutz unter dem Schirm der Verfassung suchen kann. Aber was ist mit den viel häufigeren fragwürdigen Forschungspraktiken? Outcome switching, Hypothesizing after the results are known (HARKING), p-Hacking, Studien die verzerrte Ergebnisse liefern, da nicht verblindet oder randomisiert, Studien ohne Aussagekraft wegen zu geringer statistischer Power, nicht-Veröffentlichung von methodisch kompetenten NULL-Resultaten, etc. Weil das in vielen, insbesondere biomedizinischen Disziplinen gängige Praxis ist, muss man dies alles nicht schon zum wissenschaftlichen Standard zählen? Aber halt, das Lehrbuch sagt: Ein Standard ist eine festgelegte, allgemein anerkannte und einheitliche Richtlinie oder Norm, die als Maßstab oder Referenz für ein bestimmtes Vorgehen, eine Methode oder einen Prozess dient. Er bietet eine strukturierte und konsistente Herangehensweise, um Qualität, Effizienz und Vergleichbarkeit in einem spezifischen Bereich zu gewährleisten. Demnach wären die fragwürdigen wissenschaftlichen Praktiken alle keine richtige Wissenschaft – ein Urteil dem sich der Narr vorbehaltlos anschließen kann.

Verfassungs- (und sonst wie) rechtlich relevant ist das aber trotzdem nicht, denn wer würde entscheiden, was eine „deutliche Verfehlung von Standards“ ist? Außerdem urteilte das Bundesverfassungsgericht: „Über gute und schlechte Wissenschaft, Wahrheit oder Unwahrheit von Ergebnissen kann nur wissenschaftlich geurteilt werden“. Das ist natürlich eine gute Nachricht, wäre ja noch schöner, wenn Epistemisches gerichtlich

geklärt werden würde. Die schlechte Nachricht ist allerdings, dass die die Wissenschaft selbst dazu auch nicht in der Lage scheint.

So bleibt noch die Frage nach der akademischen Verantwortung, methodisch kompetent, transparent und gesellschaftlich relevant zu forschen. Im Gegensatz zum deutschen Grundgesetz, das die akademische Verantwortung nicht ausdrücklich mit der akademischen Freiheit verbindet, erkennt internationales Recht das Recht der Allgemeinheit auf die Nutzung wissenschaftlicher Erkenntnisse an (UN International Covenant on Economic, Social and Cultural Rights, Artikel 15 (1)(b)) an, und legt damit nahe, dass die Wissenschaft verpflichtet ist, Forschung im Interesse des Gemeinwohls durchzuführen. Die UNESCO bringt den grundlegenden Zusammenhang zwischen akademischer Freiheit und Verantwortung so auf den Punkt: „Die Ausübung von Rechten bringt besondere Pflichten und Verantwortungen mit sich. [...] Die akademische Freiheit beinhaltet die Pflicht, diese Freiheit in einer Weise zu nutzen, die mit der wissenschaftlichen Verpflichtung vereinbar ist, die Forschung auf eine ehrliche Suche nach der Wahrheit zu gründen. Lehre, Forschung und wissenschaftliche Arbeit sollten in voller Übereinstimmung mit ethischen und professionellen Standards durchgeführt werden und sollten, wo angemessen, auf zeitgenössische Probleme der Gesellschaft reagieren."

Von der Allgemeinheit alimentierte Wissenschaftler sind also frei darin, ihre Forschungsfragen, Themen und Methoden unabhängig zu wählen. Gleichzeitig erwarten Gesellschaft und Fördergeber, dass Forschung zum allgemeinen Nutzen beiträgt und die von ihr zur Verfügung gestellten Ressourcen effizient eingesetzt werden. Das gilt nicht nur, aber besonders für die biomedizinische Forschung. Akademische Freiheit und Verantwortung sind also eng miteinander verknüpft. Diese Verantwortung beinhaltet, dass Wissenschaftler hohe professionelle Standards einhalten, ihre Methoden kompetent anwenden und neues Wissen schaffen, das auch die Bedürfnisse und Interessen der Gesellschaft berücksichtigt.

Verweise auf Rechtsnormen oder gar das Grundgesetz taugen aber nicht dafür, Fragen nach Qualität, Robustheit, Transparenz und gesellschaftliche Relevanz von Forschung zurückzuweisen. Noch weniger ist das Grundgesetz umgekehrt dazu geeignet, professionelle Standards für die Forschung oder gute wissenschaftliche Praxis zu dekretieren und zu sanktionieren.

Im Prinzip könnten nationale oder besser internationale wissenschaftliche Gesellschaften oder Akademien solche Standards festlegen und über ihre Einhaltung wachen. Ärzte, Architekten, Apotheker, die Liste der akademischen Berufe, die professionelle Standards einfordern, ist lang. Um sie praktizieren zu dürfen, bedarf es einer Lizensierung, und regelmäßigen Überprüfung, um sicherzustellen, dass die Praktizierenden die erforderlichen Qualifikationen und Fähigkeiten besitzen. Für Wissenschaftler gibt es das nicht, Wissenschaft ist keine Profession. Einige zarte Versuche einer formalen Professionalisierung gibt es zwar (in England z.B. das Science Council https://sciencecouncil.org/), diese konnten sich bisher aber nicht durchsetzen. Befürchtungen, dass so etwas zu weiterer Bürokratisierung und administrativen Wasserköpfen führt, sind vermutlich berechtigt.

Und was ist mit wissenschaftlichen Normen (z.B. Mertons Universalismus, Kommunismus, Uneigennützigkeit und institutionalisierter Skeptizismus)? Kaum ein Wissenschaftler könnte sie aufsagen! Oder mit Ordnungen und Kodizes (z.B. den Kodex der DFG)? Letztere sind für Wissenschaftler zwar verpflichtend, man unterschreibt sie sogar mit dem Arbeitsvertrag. Die meisten Wissenschaftler wissen aber gar nicht was drin steht. Gelehrt, ab- bzw. überprüft wird da gar nichts, und sanktioniert sowieso nur in den seltensten Fällen und auch nur bei den gravierendsten Verstößen.

Wie kann man trotzdem, ohne Eingriff in Art 5. (3) GG sicherstellen, dass Wissenschaftler verantwortlich forschen, d.h. gemäß hoher professioneller Standards, transparent, effizient und relevant? In dem man die teils auch Domänen- und Methoden-spezifischen Standards bereits im Studium strukturiert lehrt und prüft. Und deren Einhaltung zur Bedingung von Forschungsförderung macht – und dabei auch (z.B. stichprobenartig) nach Projektabschluss auditiert. Aber vielleicht am wichtigsten: Akademische Karrieren (Verstetigung, Berufung, etc.) nicht nur von fragwürdigen (z.B. Zitationen) oder schlicht ungeeigneten Metriken (z.B. Journal Impact Faktor, Summe der Drittmittel) abhängig macht – sondern der Einhaltung eben dieser professionellen Standards (z.B. Methoden zur Verminderung von Bias, Präregistrierung), Transparenz (z.B. Open Access und Data, Veröffentlichung von NULL Resultaten), und zumindest in der klinischen Forschung der Einbeziehung von Patienten und Angehörigen in den Forschungsprozess. Eigentlich alles recht gradlinig, und keine Rocket science. Aber unter die wohlfeilen Einwände der Pragmatiker („Ist doch sehr aufwändig, und irgendwie funktioniert das System doch…“) mischen sich da gleich die Stimmen derer, welche einen Angriff auf die Forschungsfreiheit wittern. Sollten Sie dabei sein: Ziehen Sie nicht über Los, setzen Sie die Lektüre am Anfang des Artikels fort!

Der Wissenschaftsnarr dankt Lucia Reuter und Klaus-Ferdinand Gärditz für wertvolle Anregungen und Diskussionen.

Wie man rausfindet, ob (klinische) Studien was taug(t)en

10/2024

Klinische Studien sind das Fundament der Bewertung der Sicherheit und Wirksamkeit neuer medizinischer Behandlungen, Medikamente und Therapien. Patienten werden darin häufig Risiken ausgesetzt, die sie bereit sind aus Altruismus zu tragen. Klinische Studien beschäftigen ein Heer von Ärzten und Studienpersonal, und erfordern erhebliche Investitionen von Förderinstitutionen sowie Pharmafirmen. Die Zeche zahlen letztendlich die Steuerzahler und Versicherungsnehmer.

Seit über 20 Jahren fördern DFG und das Bundesministerium für Bildung und Forschung (BMBF) „qualitativ hochwertige Studien in Deutschland industrieunabhängig, um eine bessere Patientenversorgung zu gewährleisten“ (wie immer alle Quellen und weiterführende Hinweise unter https://dirnagl.com/lj). Das wurde kürzlich in Berlin bei der Festveranstaltung „20 Jahre Klinische Studien: Erfolge, Impulse, Perspektiven“ gebührend gefeiert: Stolz zählte man 161 durch die DFG geförderte Interventions- und 15 Machbarkeitsstudien mit einem Gesamtvolumen von 200 Mio. €. Das BMBF kam auf 137 klinische Studien und 164 systematische Reviews mit einer Fördersumme von 250 Mio. €. Das ist, so die gemeinsame Presseerklärung, eine „beachtliche Bilanz“. Aber, frägt der Narr, was ist

denn dabei rausgekommen, wie haben diese Studien die Wissenschaft befördert, was hat es für die Patienten gebracht?

Aber wie könnte man feststellen, ob dieses Geld gut investiert war? Die Fördergeber wissen es nicht – und ich fürchte, sie werden es auch nicht herausfinden. Man muss dazu nämlich Studienqualität und Nutzen beurteilen können – bisher leider Terra icognita für viele Förderer, aber auch für akademische Institutionen und deren Wissenschaftler. Deshalb im Folgenden hierzu ein paar närrische Hinweise.

Die Frage nach der Effizienz klinischer Forschung stellt sich nämlich auch besonders deshalb, weil wir durch Meta-Forschung ziemlich genau wissen, dass in diesem Bereich eine erhebliche Menge Forschungsmüll produziert wird. Eine unvollständige Liste der nicht nur in Deutschland prävalenten Probleme schließt folgendes ein: Studienabbrüche; irrelevante oder falsche Endpunkte; selektive und unvollständige Veröffentlichung von Resultaten; die nicht-Veröffentlichung von Studienergebnissen; Verändern von Endpunkten während der Studie oder beim Schreiben des Manuskriptes („Outcome switching"); zu geringe Fallzahlen („statistische Power"); ein hohes Maß an Verzerrungen („Bias") im Design, Durchführung, Analyse und Reporting; Unterschlagung von Nebenwirkungen; Fehlen eines öffentlich zugänglichen Protokolls; usw. usf.

Es ist also nicht alles Gold, was in der klinischen Forschung glänzt, dies hat der Narr bereits 2021 aufgespießt (LJ 9/2021). Mittlerweile wird aber sogar in den Medien über diese Malaise berichtet, und selbst die Nationale Akademie der Wissenschaften (Leopoldina) ist alarmiert! In einer im Juni veröffentlichten Stellungnahme zum Medizinforschungsgesetz subsummiert sie die oben genannten Probleme unter dem Label „mangelnde Qualitätssicherung bei klinischen Studien", und kritisiert, dass Deutschland den hinterster Rang bei der Pro-Kopf-Anzahl von klinischen Studien in der EU belegt. Die Leopoldina liefert auch gleiche eine beeindruckende Liste von Gründen, warum das so ist und klinische Forschung in Deutschland besonders schwierig sein soll.

Sie listet unter anderem fehlende Karrierewege im klinischen Studienbereich (Clinician/Medical Scientists); Barrieren bei der Sekundärnutzung von Daten von Patientinnen und Patienten zu Forschungszwecken, insbesondere durch Datenschutzrechtliche Rahmenbedingungen; ein bisher fehlendes Medizinforschungsgesetz; die föderale Zersplitterung in Bezug auf Datenschutz, rechtlichen Regelungen, Ethikkommissionen, behördliche Aufsichtsstrukturen; mangelnde Ausstattung oder Fehlen von Förderprogrammen; langwierige und komplizierte Antragsverfahren für klinische Arzneimittelprüfungen („Überregulierung"); sowie Studienregistrierung an verschiedenen Orten (z. T. mehrfach, EUCTR, DRKS, clinicaltrials.gov). Auffällig nur, dass die Ursachen fast ausschließlich außerhalb des akademischen Forschungsbetriebes zu liegen scheinen. Könnte das daran liegen, dass alle Autoren der Stellungnahme selbst Wissenschaftler sind?

Ist es aber nun wirklich so schlimm, und wie könnte man die oben gestellte Frage nach der Qualität bzw. Effizienz der klinischen Forschung beantworten? Allgemeine Hinweise hierzu habe ich bereits vor einer Weile gegeben („Zen und die Kunst, Forschungsqualität zu bewerten", LJ 9/2023). Wie sieht das aber konkret bei klinischer Forschung aus?

Ganz einfach, man frägt, was die Aussagekraft, man könnte auch sagen der Informationswert einer Studie ist oder war. Dieses Konzept der sog. „Informativeness" hat sich bereits bei der Bewertung von klinischer Studienqualität in der Meta-Forschung bewährt. Es wird auch international von Fördergebern eingesetzt, z.B. der Bill & Melinda Gates Foundation, die sich dafür interessiert, ob ihre 40 Milliarden US$ Fördergelder, die sie seit 2017 gewährt hat, gut angelegt waren.

Die kritischen Fragen an eine Studie sollten danach lauten: Was haben wir durch sie gelernt, was hat sie an neuer Erkenntnis (Evidenz) gebracht, wie wichtig sind die Ergebnisse, und wie sehr können wir uns auf sie verlassen? Denn uninformative Studien bieten keine Rechtfertigung dafür, Menschen zur Teilnahme an klinischer Forschung aufzufordern, sie widersprechen der informierten Einwilligung, denn Patienten erwarten, zum medizinischen Fortschritt beizutragen. Sie verschwenden kostbare Ressourcen der Teilnehmer, Forscher, „kontaminieren" die Literatur", wobei Ärzte oder politische Entscheidungsträger häufig die Designmängel nicht erkennen und falsche Schlussfolgerungen ziehen.

Aber wie können wir die Frage nach der Informativeness so operationalisieren, dass sie standardisiert und objektiv beantwortet werden kann? Zarin, Goodman und Kimmelman forderten 2019 in JAMA, dass eine Studie, um „informativ" zu sein, einige Kriterien erfüllen muss: Die Studienhypothese muss eine wichtige und ungelöste medizinische Frage ansprechen („Relevanz"). Sie muss so gestaltet sein, dass sie aussagekräftige Evidenz zu dieser Frage liefert („Studiendesign"). Sie muss machbar sein („Machbarkeit"), und sie muss die Methoden und Ergebnisse genau, vollständig und zeitnah berichten („Reporting").

Wie geht man nun konkret vor, um diese Kriterien zu überprüfen? Bezüglich der „Relevanz" überprüft man z.B., ob die Studienergebnisse in einen hochwertigen systematischen Review eingegangen sind, der darauf abzielte, medizinische Entscheidungsfindung oder klinische Praxisrichtlinien zu informieren. Ob Patienten im Design der Studie, insbesondere bei der Auswahl der Endpunkte involviert waren („Patient Stakeholder Engagement, PSE"), und ob sogenannte Core Outcome Sets verwendet wurden. Letztere sind standardisierte Sets von Ergebnissen, die in allen Studien zu einer bestimmten Erkrankung oder einem bestimmten Gesundheitsbereich gemessen und berichtet werden sollten und die Vergleichbarkeit und Relevanz der Forschungsergebnisse verbessern. Falls Sie das alles für trivial halten: Die absolute Minderheit der klinischen Studien nutzen bisher PSE oder Core outcome sets!

Beim Kriterium Studiendesign geht es um die Frage, wie hoch die Gefahr der Verzerrung der Studienergebnisse war („Risk of bias"). Wurde z.B. randomisiert und verblindet - und wie wurde das getan? Um das standardisiert bei einer Studie rauszufinden, hat das Cochrane Netzwerk ein allgemein akzeptiertes Tool entwickelt. Auch sollte man fragen, ob alle Daten berichtet wurden, oder dies nur selektiv geschah? Unglaublich wichtig ist auch die statistische Power einer Studie. War die Fallzahl hoch genug, um den Effekt nachweisen zu können? Dies ist ein riesiges Problem bei vielen klinischen Studien: Weil nicht genug Ressourcen (Geld, Zeit, Patienten) zur Verfügung stehen, reicht die Power nicht. Die Schlussfolgerung solcher Studien ist dann häufig: Das Therapeutikum wirkt „möglicherweise". So etwas fördert, wie es Jonathan Kimmelman ausdrückt, „klinischen Agnostizismus". Das ist folgenreich: Auf kleine, nicht genügend gepowerte explorative Studien folgen meist keine konfirmierenden Studien. Hierdurch kommt es in der Folge zur Empfehlung von Off-Label-Verschreibungen in klinischen Richtlinien in Abwesenheit von bestätigenden Studienergebnissen.

Das Kriterium „Machbarkeit" überprüft, insbesondere bei prospektiver Betrachtung eines Studienprojektes, ob die Rekrutierungsziele überhaupt erreichbar sind, und das im zeitlich vorgegebenen Rahmen, ob die Finanzierung auskömmlich ist, usw. All dies klingt wieder nach Selbstverständlichkeiten. Man muss aber wissen, dass 20-40 % aller registrierten randomisiert kontrollierten klinischen Studien nie beendet werden, und dass 80% der klinischen Studien ihre Rekrutierungsziele nicht rechtzeitig erreichen, sodass 9 von 10 Studien letztendlich ihren ursprünglichen Zeitplan verdoppeln. Wir erinnern uns

(ungern) an Corona: Damals war die überwiegende Mehrheit der Studien zu COVID-19-Therapeutika nicht darauf ausgelegt, verwertbare Informationen zu liefern. Geringe Randomisierungsraten und unzureichende Daten zur Wirksamkeit machten Fragen zur Sicherheit und Wirksamkeit im Allgemeinen uninterpretierbar. Nur wenige Studien lieferten dringend benötigte Evidenz. Die meisten Studien erreichten damals ihre Rekrutierungsziele nicht.

Beim Kriterium „Reporting" frägt man, ob die Studie Methoden und Ergebnisse genau, vollständig und zeitnah berichtet hat. Und auch da haperts derzeit noch gewaltig. Ein kürzlich veröffentlichtes Positionspapier des Bündnis „Transparenz in der Gesundheitsforschung" mit dem Titel „Unveröffentlichte Studienergebnisse gefährden die evidenzbasierte Gesundheitsversorgung" prangert an, dass die Ergebnisse von etwa einem Drittel aller von Deutschen Universitätskliniken geleiteten klinischen Studien unveröffentlicht bleiben. Dies verzerrt das Gesamtbild der Evidenz und kann letztlich zu schlechteren Behandlungen führen. Auch unterläuft die Nichtveröffentlichung von Studienergebnissen das Vertrauen von Studienteilnehmern, die zum medizinischen Fortschritt beitragen wollen. Verschwendung von Forschungsgeldern ist es allemal. Das Bündnis fordert deshalb für Deutschland eine zentrale Zusammenführung aller von Ethikkommissionen begutachteten klinischen Studien, sowie eine Nachverfolgung der Veröffentlichung der Studienergebnisse, welche bei Bedarf eingefordert werden müssen.

Studien, die ein neues Medikament am Menschen auf Wirksamkeit und Nebenwirkungen untersuchen, berufen sich häufig auf präklinische Evidenz, welche Mechanismen und Effektivität des Therapeutikums untersucht haben. Aber bei der Zulassung der Studie wird die Qualität dieser tierexperimentellen Studien gar nicht überprüft, geschweige denn hinterfragt. Wir erinnern uns: Ausgerechnet die Pharmaindustrie hat sich vor über 10 Jahren in Aufsehen erregenden Veröffentlichungen geoutet: Sie konnten die präklinischen Befunde aus den akademischen Laboren häufig nicht reproduzieren. Heute wissen wir aus aufwendigen präklinischen randomisierten Replikationsstudien (z.B. Cancer Biology Reproducibility Project), dass die anekdotischen Berichte aus der Pharmaindustrie leider generalisierbar sind. Häufig sind die Studienergebnisse, auf denen dann klinische Entwicklungen aufbauen, nicht wiederholbar, und wenn, dann mit substantiell geringeren Effekten.

Wäre es deshalb nicht angebracht, vor einer klinischen Interventionsstudie, welche Patienten potentiell belastet oder sogar gefährdet, und Hunderte von Millionen € kosten kann, zu überprüfen, wie robust die klinische Evidenz ist, welche die Basis für die klinische Testung darstellt? Es mehren sich die Beispiele, bei denen nach dem Scheitern einer klinischen Studie bei Wiederholung der präklinischen Experimente mit ausreichenden Fallzahlen, verblindet und randomisiert, nichts mehr rauskam, die Grundlagen für die klinischen Studie wie Kartenhäuser zusammenbrachen!

Was kommt aber nun raus, wenn man mit den oben genannten Kriterien die klinische Studienliteratur auf Informationsgehalt und Aussagewert abklopft? Nora Hutchinson und Kollegen haben genau dies getan. Sie untersuchten eine Kohorte von randomisierten interventionellen klinischen Studien in drei Krankheitsbereichen (ischämische Herzerkrankung, Diabetes mellitus und Lungenkrebs), welche eine klinische Frage im Zusammenhang mit der Behandlung oder Prävention von Krankheiten zu beantworten wollten. Das Ergebnis war, wenn auch nicht überraschend, ernüchternd: Nur ein Viertel der Studien erfüllten alle Kriterien der „Informativeness"! Alle anderen Studien zeigten Probleme im Design, in der Durchführung oder in der Berichterstattung, die ihre Fähigkeit, klinische Entscheidungsfindungen zu unterstützen, beeinträchtigte. Und noch ein interessanter Nebenbefund: Von der Industrie gesponserte Studien erfüllten wesentlich

häufiger (50%) alle vier Bedingungen der Informativeness, als nicht von der Industrie gesponserte Studien (6 %). Letztere sind im Wesentlichen die sog. „Investigator initiated trials“ (IITs), also die universitären Studien, wie sie auch vom BMBF und der DFG gefördert werden.

Nicht nur überprüfen DFG und BMBF nicht, wie informativ und aussagekräftig die von ihnen geförderten Studien waren – oder noch besser: machen Förderung davon abhängig, welches Potential beantragte Studien haben, informativ zu sein. DFG und BMBF rangieren selbst im unteren Drittel eines Rankings der 25 größten medizinischen Forschungsförderer weltweit, wenn es um die Compliance mit den Best-Practice-Benchmarks der WHO zur Registrierung und Berichterstattung klinischer Studien geht. Solche Richtlinien zur Registrierung und Berichterstattung klinischer Studien, gepaart mit Überwachung und Sanktionen, können Forschungskosten senken, Publikationsbias verringern und Transparenz fördern.

Fazit aus alledem: Fördergeber sollten die Qualitätsmerkmale der von Ihnen geförderten Studien evaluieren, transparent monitoren, und mangelnde Compliance sanktionieren. Sie haben die Verantwortung sicherzustellen, dass die von ihnen finanzierten Studien das Potential haben, informativ zu sein. Dies schließt die Gewährleistung von Rechenschaftspflicht der finanzierten Forscher bzw. deren Institutionen ein, um sicherzustellen, dass die Durchführung, Analyse und Berichterstattung wissenschaftlich angemessen und zeitnah erfolgen.

Hochglanzstudien und bittere Wahrheiten

LJ 11/2024

Derzeit häufen sich Meldungen, nachdem die „Hype um AI“ vorbei sei. Der Narr hatte deshalb gerade begonnen, etwas über den Sinn und Unsinn des Gartner’schen Hype Zyklus zu schreiben. Aber just in dem Moment brach wieder einmal einer neuer „Skandal“ über die biomedizinische Wissenschaft. Eliezer Masliah, einen der meistzitierten Hirnforscher weltweit und bis vor kurzem noch Direktor des National Institute on Aging (NIA) der US-amerikanischen National Institutes of Health (NIH), wurde grobes wissenschaftliches Fehlverhalten vorgeworfen. In über 130 seiner Arbeiten hat man „fragwürdige Bilder und Daten“ entdeckt, *Science* gab dazu ein 300-seitiges Dossier in Auftrag. Obwohl bisher (noch) kein eindeutiger Betrug nachgewiesen wurde, deuten die Art und das Ausmaß der Manipulationen darauf hin, dass es sich um weit mehr als bloße „Flüchtigkeitsfehler und Nachlässigkeiten“ handelte.

Alles bei der Affäre Masliah verlief bisher nach Schema F. Und das sieht folgendermaßen aus: In Publikationen prominenter Wissenschaftler werden, oft Jahre nach ihrer Veröffentlichung in namhaften Fachzeitschriften, manipulierte Abbildungen entdeckt. Häufig werden diese „Auffälligkeiten" von Whistleblowern oder Wissenschaftlern aufgedeckt und publik gemacht, die gezielt nach solchen Unregelmäßigkeiten suchen. Zunächst passiert entweder gar nichts, oder es dauert viele Jahre, bis sich die betroffenen Universitäten oder Fachjournale widerwillig dem „Fall" annehmen – und das oft nur, wenn sich die Affäre bereits zu einem ausgewachsenen Skandal entwickelt hat. Zunächst werden meist diejenigen angegriffen, die auf die Auffälligkeiten hingewiesen hatten. Anschließend wird behauptet, die „Unregelmäßigkeiten" seien auf Unaufmerksamkeiten oder Versehen der Autoren zurückzuführen. Außerdem wären diese Befunde für die Aussage des Artikels gar nicht von Bedeutung gewesen. Der Aufforderung, die Originaldaten zu zeigen, kann dann manchmal nicht gefolgt werden, weil ‚der Hund den USB-Stick gefressen habe', auf denen sie gespeichert waren. Wenn's ganz schlimm kommt, wird ein PhD Student oder Postdoc geopfert. Artikel werden sehr selten zurückgezogen, und die wirklich Verantwortlichen fast nie zur Rechenschaft gezogen. Häufig hatten die in Frage stehenden Publikationen nicht nur den wissenschaftlichen Ruf der Autoren begründet, sondern ihnen durch darauf basierende Patentanmeldungen, Lizenzierungen und Firmengründungen auch persönlichen finanziellen Gewinn verschafft.

So ähnlich läuft das alles auch bei Eliezer Masliah. Allerdings wird er, als ehemaliger NIH-Direktor (er wurde wohl kurz vor der Enthüllung des Skandals beurlaubt) und angesichts des schieren Ausmaßes der Manipulationen vermutlich in die "Scientific Misconduct Hall of Fame" aufgenommen. Die Konkurrenz schläft aber nicht, denn in letzter Zeit dürfen sich eine Reihe von anderen prominenten Kandidaten Hoffnung auf diese Ehrung machen: Marc Tessier-Lavigne, der Ex-Präsident der Stanford University (der Narr berichtete in LJ 10/2023 - Zitate und Weiterführendes wie immer unter http://dirnagl.com/lj), Berislav Zlokovic (University of Southern California, wie bei Masliah ließ *Science* die Bombe platzen), Sylvain Lesné (University of Minnesota) oder Domenico Praticò (Temple University). In jeweils 20 bis über 70 Publikationen dieser hochdekorierten Neurowissenschaftler wurden hochgradig fragwürdige oder eindeutig manipulierte Abbildungen entdeckt.

Die Häufigkeit, Dramatik und Stereotypie der Skandale werfen eine große Zahl von Fragen auf, weshalb es sich lohnt, sich hier doch noch einmal dem Uralt-Thema Wissenschaftsskandal zu widmen: Warum sind eigentlich die Neurowissenschaften so häufig betroffen? Wie groß ist der Eisberg, dessen Spitze wir hier sehen? Sind diese Fälle wirklich Beleg dafür, dass die Selbstkorrektur der Wissenschaft am Ende doch ganz gut funktioniert? Wie effektiv ist eigentlich die Qualitätskontrolle der Wissenschaft, mithin der Peer-Review? Wieso dauert es so lange, bis Informationen von Whistleblowern und auf Fachforen wie PubPeer erhobene Vorwürfe ernst genommen werden, falls das überhaupt geschieht? Sollten sich Patienten Sorgen machen, dass sie Medikamente erhalten, die auf unsolider Wissenschaft beruhen? Woran liegt es eigentlich, dass manche Wissenschaftler Ergebnisse ‚schönen' oder gar fälschen? Und was kann man dagegen tun?

Ich war der fragwürdigen Forschung von Eliezer Masliah vor vielen Jahren schon einmal persönlich nahegekommen. Ich war damals Direktor des Centrums für Schlaganfallforschung an der Berliner Charité. Wir wurden von einer Pharmafirma kontaktiert, die ein neues – von Masliah federführend mitentwickeltes Medikament namens Cerebrolysin klinisch testen wollte. Kein ungewöhnlicher Vorgang, denn Firmen haben ja selbst keinen Zugang zu Patienten, und arbeiten bei klinischen Studien eng mit akademischen Kliniken zusammen, welche die Erkrankung beforschen und Betroffene behandeln.

Ungewöhnlich an dem Vorgang war aber das Therapeutikum. Bei Cerebrolysin handelt es sich nämlich um eine nicht genau definierte Mischung von Peptiden, welche – man höre und staune - durch enzymatische Verdauung von Schweinehirnen gewonnen wird. In Studien an Versuchstieren und Zellkulturen waren angeblich wahre Wunder durch das „Medikament" zu erreichen. Nervenzellen wurden geschützt, sogar die Entstehung neuer Nervenzellen angeregt, und eine Vielzahl von Erkrankungen des Nervensystems, darunter auch der Schlaganfall, ließen sich in den Experimenten erfolgreich behandeln.

Wir haben diese im wahrsten Sinne des Wortes „vielversprechenden" Studien damals selbstverständlich genauer unter die Lupe genommen und festgestellt, dass sie unseren methodischen Standards nicht entsprachen. Die Fallzahlen waren zu gering, die Ergebnisse schlichtweg zu „spektakulär" und dabei zu „makellos". Allerdings ist dies allein noch nichts Besonderes, denn die biomedizinische Wissenschaft, insbesondere die präklinische Forschung, welche die Grundlagen für die Entwicklung und Testung neuer Therapien liefert, ist voll von solchen makellosen und spektakulären Befunden von Studien mit zu kleinen Stichproben. Die meisten davon sind wohl nicht betrügerisch, aber doch ein deutlicher Hinweis auf die selektive Nutzung von Daten, sowie das Überstrapazieren wissenschaftlicher Freiheit im Design, der Analyse, und der Berichterstattung von Versuchsergebnissen. Sie werfen damit ein Schlaglicht auf eine Forschungssystem, das ‚positive Ergebnisse' bevorzugt und belohnt. Wegen der mangelhaften Qualität der Datenlage und der „schweinischen" Zusammensetzung von Cerebrolysin hatten wir damals nicht an der klinischen Studie teilgenommen.

Aber Cerebrolysin war nicht die einzige Substanz, welche auf Grund von nun infrage gestellten Befunden Masliahs in die klinische Prüfung kam. Der wissenschaftliche Ruf von Masliah basiert nämlich insbesondere auf seiner Forschung zu Alpha-Synuclein, ein Protein dem eine Schlüsselrolle bei der Entstehung wichtiger Hirnerkrankungen zugeschrieben wird.

Alpha-Synuclein ist ein kleines Eiweiß, das vorwiegend im Gehirn vorkommt, insbesondere in den Synapsen. Seine genaue Funktion ist noch nicht vollständig verstanden. Insbesondere bei der Parkinson-Erkrankung finden sich Ablagerungen fehlgefalteter Alpha-Synuclein-Moleküle in dopaminergen Neuronen. Man vermutet, dass diese Ansammlungen ursächlich für die Schädigung dieser Nervenzellen sind, und damit die typischen Symptome dieser Krankheit hervorrufen. Deshalb gründen sich große Hoffnungen auf die Entwicklung von Therapien, welche die Fehlfaltung von Alpha-Synuclein verhindern können, oder fehlgefaltetes Eiweiß beseitigen helfen.

Und genau das schien Masliah und seinen Kollegen gelungen zu sein: Neben anderen auf Alpha-Synuclein gerichteten Strategien entwickelte er eine Immuntherapie mit einem Antikörper, der fehlgefaltetes Alpha-Synuclein bindet, sodass es anschließend von körpereigenen Immunzellen zerstört und abtransportiert wird. Wenn das möglich wäre, wären das natürlich nicht nur für Patienten fantastische Aussichten, sondern auch für die Pharmaindustrie – mit einem solchen Medikament ließen sich Milliarden verdienen. Es folgten daher zahlreiche Patente, es wurden Startups gegründet und in Zusammenarbeit mit großen Pharmafirmen klinische Studien begonnen und Patienten mit der Substanz behandelt. All dies förderte nicht nur massiv Masliahs akademischen Ruf, sondern brachte ihm auch erheblichen persönlichen finanziellen Gewinn.

Allerdings haben die klinischen Studien bisher keine Wirksamkeit dieser Therapien nachweisen können – was angesichts der aktuellen Enthüllungen nicht verwundert. Denn unter den nun infrage gestellten Befunden befinden sich zahlreiche, welche die Wirksamkeit der Substanz belegen sollten. Die zentralen Aussagen dieser grundlegenden Studien sind damit nicht länger haltbar.

Die besondere Brisanz von Skandalen wie dem um Masliah liegt also nicht nur darin, dass es sich um hochdekorierte und einflussreiche Wissenschaftler handelt. Durch manipulierte Ergebnisse werden zahlreiche andere Wissenschaftler in die Irre geführt. Besonders kritisch ist, wenn - wie in diesem Fall - die fraglichen Befunde unmittelbar die Grundlage für die Entwicklung von Medikamenten bildeten. Hier geht es also nicht nur um das Vertrauen in die Forschung und die Verschwendung von Ressourcen, sondern vor allem um die Gefährdung von Studienteilnehmern, ganz abgesehen von der Erzeugung falscher Hoffnung bei Erkrankten und Angehörigen.

Warum aber „schönen“ oder gar fälschen manche Wissenschaftler ihre Ergebnisse? Ganz offensichtlich spielen hier falsche Anreize und Interessenkonflikte eine wichtige Rolle. Wissenschaftliche Karrieren und Ansehen werden an hochrangigen Publikationen gemessen. Artikel sind die Währung einer Reputationsökonomie, die weniger auf die Qualität und den Inhalt von Veröffentlichungen als vielmehr auf die Namen der Top-Journale, in denen sie erscheinen, ausgerichtet ist. Deren Geschäftsmodell beruht auf dramatischer Selektivität, dort werden spektakuläre, perfekte Studienergebnisse veröffentlicht, welche im normalen Laboralltag leider allzu selten vorkommen. Mit ein paar „kreativen“ Manipulationen gelangt man schneller und sicherer zum Ziel als durch harte Laborarbeit, die häufig auch nicht die gewünschten Ergebnisse bringt.

Hinzu kommt, dass in der Biomedizin viel Geld im Spiel ist. Die Forschung ist teuer, und es werden umfangreiche Fördergelder benötigt – die man eher erhält, wenn man spektakuläre Ergebnisse vorweisen kann. Zudem verspricht die Aussicht auf neue Therapien monetären Gewinn für Universitäten, Kapitalgeber und die Pharmaindustrie. Auch die beteiligten Forscher können davon persönlich profitieren. Sowohl falsche Karriereanreize als auch finanzielle Interessenkonflikte gibt es in allen Bereichen der biomedizinischen Forschung. Aber besonders im Bereich der Forschung zu Gehirnerkrankungen, der aktuell sehr im Trend liegt, ist der Bedarf an neuen, wirksamen Therapien groß – ebenso wie der Einsatz von öffentlichen und privaten Geldern.

Sind die nun öffentlich werdenden Skandale lediglich auf die sprichwörtlichen wenigen „faulen Äpfel“ zurückzuführen, oder handelt es sich dabei nur um die Spitze eines Eisbergs? Wie groß dieser Eisberg tatsächlich ist, wissen wir leider nicht. Es besteht jedoch die berechtigte Sorge, dass es sich um ein systemisches Problem handelt, das sich hier nur in einem kleinen, verzerrten Ausschnitt zeigt. Es gibt keine systematische Suche nach Arbeiten von fragwürdiger Qualität – diese Aufgabe übernehmen bislang nur einige wenige Wissenschaftler, meist in ihrer Freizeit oder auf prekärer Basis durch Spenden finanziert. An die Öffentlichkeit dringen in der Regel auch nur Fälle, die prominente Wissenschaftler betreffen. Der Großteil der Fälle bleibt auf Plattformen wie PubPeer unbeachtet, weder die betroffenen Autoren noch die Fachzeitschriften oder Institutionen fühlen sich angesprochen.

Hinzu kommt, dass sich Manipulationen in den Ergebnissen von bilderzeugenden biochemischen und histologischen Labortechniken, die in der Publikation abgebildet werden, vergleichsweise leicht nachweisen lassen. Der Großteil der Ergebnisse der biomedizinischen Forschung wird jedoch als Zahlenwerte im Text, in Tabellen oder in Grafiken veröffentlicht. Ob ein bestimmter Wert dabei echt, verändert oder frei erfunden ist, lässt sich daraus nicht erschließen. Auch das Weglassen von Daten bleibt unbemerkt. Sollten tatsächlich nur Abbildungen manipuliert werden, wo doch gerade bei denen die Gefahr besteht, erwischt zu werden?

Aber wissen wir nicht spätestens seit Corona, dass man wissenschaftlichen Arbeiten erst vertrauen darf, wenn sie einem Peer Review unterzogen wurden, also einer Kontrolle und Freigabe durch unabhängige Experten? „Preprints“, so steht es in jedem

Wissenschaftsteil der Zeitungen, sollten wir daher mit Vorsicht betrachten. Theoretisch stimmt das, doch wie die zahlreichen Fälle von im Peer Review übersehenen, aber dennoch ganz offensichtlichen Manipulationen der jetzt angeprangerten Autoren zeigen, funktioniert dieses System heute nicht mehr zuverlässig. Wissenschaftliche Arbeiten sind mittlerweile so komplex und umfangreich in Methodik und Analyse, dass zwei oder drei Experten meist gar nicht in der Lage sind, alles umfassend zu beurteilen. Hinzu kommt, dass der Begutachtungsprozess sehr zeitaufwendig ist und Gutachter mehrere Tage voll beanspruchen kann. Dafür haben heutzutage die wenigsten Zeit oder Motivation, zumal es sich um eine unbezahlte und anonyme Aufgabe handelt.

Die Selbstkorrektur der Wissenschaft funktioniert also nicht mehr so, wie sie es vorgibt. Doch sind nicht gerade diese Skandale der Beweis dafür, dass offensichtlich Falsches letztlich doch aufgedeckt und eliminiert wird? Selbst wenn man diesem Argument folgt: Es dauert oft sehr lange – nicht selten über 10 Jahre – bis es zu Korrektur kommt. In der Zwischenzeit wird bereits erheblicher Schaden angerichtet. Außerdem ignoriert dieses Argument, dass ja nur spektakulären Fälle korrigiert werden, während es sehr wahrscheinlich ist, dass das Problem in Wirklichkeit weitaus größer ist.

Die ethischen Implikationen dieser Skandale und der zugrundeliegenden systemischen Defizite sind offensichtlich. Es werden potenziell Ressourcen verschwendet, die andernorts für solide Forschung genutzt werden könnten. Letztlich tragen wir als Steuerzahler, über Krankenversicherungsbeiträge und Apothekenrechnungen, die Kosten. In vielen der manipulierten Studien wurden Tiere eingesetzt, die dadurch unnötig geopfert wurden. Am allerwichtigsten aber: die Durchführung klinischer Studien ohne solide wissenschaftliche Grundlage ist unethisch, da sie Patienten durch Nebenwirkungen von ansonsten unwirksamen Medikamenten potenziell gefährdet.

Gibt es Abhilfe, oder sind Skandale und Verschwendung von Ressourcen eine notwendige und zu akzeptierende Folge eines Systems, das unbestritten auch beeindruckende Erfolge und wirksame neue Therapien hervorbringt? Der Schlüssel zur Effizienzsteigerung des gegenwärtigen Systems liegt in den Kriterien, nach denen Wissenschaftler von Institutionen und ihre Anträge von Forschungsförderern bewertet werden. Wir müssen uns von einfachen Metriken verabschieden, die lediglich das Renommee der Zeitschriften oder die Höhe der eingeworbenen Forschungsgelder messen. Stattdessen sollten die Qualität und die Inhalte der Forschung stärker in die Beurteilung einfließen. Dadurch würde weniger, aber dafür verlässlicher publiziert werden, und es bliebe mehr Zeit für eine sorgfältige Qualitätskontrolle im Peer-Review-Prozess. Vermutlich hätte es ein Eliezer Masliah dann viel schwerer gehabt, Professor und Direktor eines NIH Institutes zu werden, Forschungsmittel wären nicht verschwendet worden, und unwirksame Therapien wären nicht an Patienten getestet worden.

Der Artikel basiert auf einem im Berliner Tagesspiegel am 12.10.2024 veröffentlichten Gastbeitrag

Heilung im Rückwärtsgang: Wenn bewährte Therapien plötzlich schaden

LJ 12/2024

Sitzen im Büro, das ja mittlerweile als das neue Rauchen gilt, fördert kardiovaskuläre Erkrankungen wie hohen Blutdruck, Herzinfarkt und Schlaganfall, und ist noch dazu schlecht für die Wirbelsäule. Die logische Konsequenz: Weg mit dem Schreibtisch, ein Stehpult muss her. Oder wenigstens ein paar Stunden täglich bei der Arbeit im Stehen verbringen, statt im Sitzen. Weil das so plausibel ist, und weil die Anzahl der im Sitzen verbrachten Stunden mit steigendem kardiovaskulärem Risiko korreliert, kam vor ein paar Wochen folgende Meldung wie ein Schock: Stehen bei der Arbeit ist nicht gesünder als Sitzen! Eine sehr große australische Studie hatte knapp 80.000 Menschen über 7 Jahre verfolgt. Sie fand, dass man durch Stehen den Teufel mit dem Beelzebub austreibt. Potenziell verringert Stehen das Risiko für Herzinsuffizienz, Koronarerkrankung und Schlaganfall ein wenig – aber auch nur wenn man mehr als 10 Stunden pro Tag sitzt. Wer sitzen durch Stehen ersetzt, kauft sich dafür potenziell andere Kreislauerkrankungen ein, wie orthostatische Hypotension (mit der Gefahr von Stürzen), venöse Insuffizienz und Geschwüre. Und das schon bei wenigen Stunden Sitzen am Tag.

Aber wen überrascht das wirklich? Ähnliche Kehrtwendungen aus dem Bereich ‚Lifestyle' gibt's ja immer mal wieder, wie z.B. bei der angeblichen Schädlichkeit von ‚rotem Fleisch', oder dem lange gehegten Mythos der kardiovaskulären Wunderwirkungen einer ‚Mittelmeer'-Diät. Alles sehr plausible Theorien, oft unterlegt mit ‚Evidenz' von kleinen und qualitativ minderwertigen Beobachtungsstudien (Zitate und weiterführende Literatur wie immer unter https://dirnagl.com/lj). Gleichzeitig werden die in den wahrsten Wortsinnen phantastischen Effekte solcher Lebensstilveränderungen in allen Medien angepriesen. Bis sie mit einem Schlag in großen, sorgfältig geplanten, kompetent durchgeführten und analysierten Studien widerlegt werden.

Es handelt sich hierbei um sogenannte „Medical Reversals" („medizinische Kehrtwenden"), bei denen eine etablierte medizinische Theorie oder Therapie, die zuvor als richtig und wirksam galt, durch neue Erkenntnisse oder Studien widerlegt wird, und sich dann sogar häufig auch noch als schädlich erweisen. Die genannten Beispiele kamen aus dem Bereich der Lebensführung. Da ist das doch gar nicht so überraschend, und am Ende auch gar nicht so problematisch? Ein bisschen mehr Barolo, Branzino oder Olivenöl – das hat doch wohl noch keinem geschadet!

Sie werden es geahnt haben, solche Kehrtwenden sind auch bei „echten" medizinische Interventionen, also Therapien mit Medikamenten oder Implantaten sehr häufig. Zur Illustration nur ein paar Beispiele für dramatische Medical Reversals.

Blutdruckzielwerte bei älteren Menschen: Früher wurde bei älteren Menschen versucht, den Blutdruck aggressiv zu senken, um Schlaganfälle und Herzerkrankungen

zu verhindern. Reversal: Aggressive Blutdrucksenkung bei älteren Menschen erhöht das Risiko für Schwindel, Stürze und andere Nebenwirkungen.

Verwendung von Beta-Blockern bei nicht-kardialen Operationen: Beta-Blocker wurden routinemäßig vor nicht-herzbezogenen Operationen eingesetzt, um das Risiko von Herzproblemen während und nach der Operation zu verringern. Reversal: Die präoperative Verwendung von Beta-Blockern erhöht das Risiko für Schlaganfälle und Todesfälle.

Routine-Mammographie bei jüngeren Frauen: Regelmäßige Mammographien wurden Frauen ab einem Alter von 40 Jahren empfohlen, um Brustkrebs früh zu erkennen. Reversal: Bei Frauen unter 50 Jahren führen routinemäßige Mammographien oft zu Überdiagnosen und unnötigen Behandlungen, ohne die Sterblichkeit signifikant zu senken.

Einsatz von Antiarrhythmika nach Herzinfarkt: Nach einem Herzinfarkt wurden Antiarrhythmika routinemäßig eingesetzt, um unregelmäßige Herzrhythmen zu kontrollieren und Todesfälle zu verhindern. Reversal: Weit gefehlt, diese Medikamente erhöhen das Risiko eines plötzlichen Herztodes.

Hormontherapie bei Frauen in den Wechseljahren: Östrogen und Gestagen wurden routinemäßig verwendet, um Symptome der Menopause zu lindern und Herzerkrankungen zu verhindern. Reversal: Die Hormontherapien erhöhen das Risiko für Brustkrebs, Herzinfarkte, Schlaganfälle und Blutgerinnsel, anstatt es zu senken.

Stenting bei stabiler Angina pectoris: Weltweit und pro Jahr wurden etwa 500.000 Patienten ein Stent eingesetzt, um die blockierte Herzarterie zu öffnen. Reversal: Stents bei Patienten mit stabiler Angina haben keine bessere langfristige Wirkung als eine medikamentöse Therapie. Stents verbesserten zwar kurzfristig die Symptome, hatten aber keinen signifikanten Einfluss auf das langfristige Überleben oder die Verhinderung von Herzinfarkten.

Antivirale Therapie: Oseltamivir (Tamiflu) verkürzt und lindert die Symptome von Influenza und Vogelgrippe. Reversal: Nachdem Roche zur Offenlegung verheimlichter (negativer) Studiendaten gezwungen wurde, stellte sich heraus, dass Tamiflu (mittlerweile „Scamiflu“ genannt) allenfalls Nebenwirkungen hat, aber auf den Ausgang von Infektionen keinen Einfluss hat. Zu dem Zeitpunkt hatte Roche allerdings schon ein paar Milliarden damit verdient.

Depression bei Jugendlichen: Mit Paroxetin (Paxil) ist diese effektiv zu behandeln. Reversal: Nachdem GlaxoSmithKline gerichtlich gezwungen wurde, zurückgehaltene und manipulierte Daten offenzulegen, wurde klar, dass Paroxetin bei Jugendlichen nicht besser als Placebo ist, und sogar die Suizidrate erhöhen kann.

Dies waren nur ein einige Beispiele für prominente Medical Reversals, ich könnte seitenweise weiter machen. Aber ist das nicht alles anekdotisch, der Narr trägt da wie gewohnt ein bisschen dick auf, so häufig wird das doch nicht sein? Weit gefehlt, wir wissen aus systematischen Untersuchungen, dass Reversals erschreckend häufig sind. So haben z.B. Vinay Prasad, der auch ein sehr lesenswertes Buch zum Thema geschrieben hat, und seine Kollegen 3000 randomisierte kontrollierte klinische Studien in den drei Top Journalen JAMA, Lancet, und NEJM untersucht. Etwa 40 Prozent der Studien, welche etablierte Therapien untersuchten, waren vom Ergebnis her Reversals. Dort wo nachfolgend systematische Reviews durchgeführt wurden, bestätigen sie diese Reversals.

Wie hoch die Prozentzahl wirklich ist, lässt sich nicht sagen, es gibt sowohl Argumente für eine Über- als auch eine Unterschätzung. Am Ende ist die genaue Anzahl aber gar

nicht so wichtig. Klar belegt ist, dass ein erheblicher Teil aller gut gemachten klinischen Studien gängige medizinische Praxis widerlegt, und damit häufig auch die Empfehlungen medizinischer Leitlinien. Letztlich können sich Mediziner nicht beschweren, dass sie das nicht geahnt hätten. Denn schon im Medizinstudium kursiert der Kalauer: "Die Hälfte von dem, was man im Medizinstudium lernt, ist falsch – nur weiß niemand, welche Hälfte."

Woran liegt es aber, dass ganz offensichtlich ein relevanter Anteil der medizinischen Praxis eine Überprüfung in großen, gut konzipierten und durchgeführten randomisiert kontrollierten klinischen Studien nicht standhält?

Ganz wesentlich daran, dass ein erschreckend großer Anteil der medizinischen Praxis, und das gilt sowohl für medikamentöse Therapien als auch für Medizinprodukte und Lifestyle-Interventionen, gar nicht auf solider medizinischer Evidenz beruht. Sondern auf Plausibilität („toller Mechanismus, das muss funktionieren", Gewohnheit („das haben wir schon immer so gemacht"), in Verbindung mit handfesten ökonomischen Vorteilen der Pharmaindustrie und ärztlichen Zunft („lässt sich gut abrechnen").

Solche Reversals treten oft dort auf, wo die ursprünglichen Studien oder Annahmen, die zur Einführung einer Behandlung führten, methodologische Mängel aufwiesen. Dazu gehören z.B. fehlende Randomisierung, Verzerrungen, unangemessene Generalisierung von Studienergebnisse, zu kleine Stichproben, Beobachtungs- statt randomisiert kontrollierten Studien, usw. Zum Reversal kommt es, wen später durchgeführte, strengere Studien mit besserem Design widerlegen dann die früheren Erkenntnisse. Der Narr hat diese mangelhafte Studienqualität vor kurzem aufgespießt, und Tipps gegeben, wie man schlechte klinische Studien erkennt (LJ 10/2024).

Aber ist das nicht das Grundprinzip jeder Wissenschaft, dass Wissen immer vorläufig ist, und häufig durch nachfolgende Studien korrigiert, ja manchmal sogar widerlegt wird? Der Wissenschaftshistoriker Thomas Kuhn kennzeichnete Letzteres als „Paradigmenwechsel". Dabei wird ein Paradigma (bzw. die gängige Praxis) einer Disziplin, nach dem sich Evidenz gegen dieses akkumuliert hat, oder eine überraschende Entdeckung gemacht wurde, durch ein anderes Paradigma ersetzt wird. So geht Fortschritt in der Wissenschaft.

Es mag sein, dass einige Reversals in diese Kategorie fallen. Diese wären dann kein Grund zur Aufregung, sondern zur Freude, oder zur Verleihung eines Nobelpreises. Die meisten Reversals passen jedoch nicht in das Kuhn'sche Schema. Dort folgt der Paradigmenwechsel (also das Reversal) auf lange Perioden von robuster und kompetenter Forschung - er nennt das leider ein wenig abwertend „normale Wissenschaft". Medizinische Reversals folgen dagegen meist Phasen nicht vorhandener oder ausreichender, qualitativ hochwertiger Evidenz. Das Paradigma steht also auf tönernen Füssen.

Das ist hochgradig beunruhigend, denn die klinische Medizin befasst sich nicht, wie manche andere Wissenschaft, mit eher esoterischen Fragen der Art „Charakterisierung des stimmlichen Repertoires von Langflossen-Grindwalen (*Globicephala melas*) im Mittelmeer", oder „Multiskalare elektrische Spike-Aktivität in *Schizophyllum commune*". Vielmehr geht es hier buchstäblich um Leben oder Tod. Als (potenzieller) Patient wünscht man sich medizinisches Handeln, das auf solider wissenschaftlicher Grundlage steht. Solange diese nicht existiert, darf es gar kein ‚Paradigma', also keine etablierte Praxis geben. Hochwertige klinische Studien sind die Experimente, mit denen wir unsere Paradigmen testen, bzw. in Frage stellen.

Was kann man gegen Medical Reversals tun? Das logische, ja triviale zuerst: Medizinische Praxis muss evidenzbasiert sein, abgesichert durch qualitativ hochwertige Studien.

Es reicht nicht, es schon immer so gemacht zu haben, oder eine großartige Idee zu haben, garniert durch die fragwürdigen Ergebnisse ein paar kleiner Studien. Das ist nicht immer einfach, aber es hilft nichts. Man könnte die Ressourcen (Finanzen, Patienten) nehmen, welche derzeit in eine Vielzahl von kleinen, qualitativ problematischen Studien gehen, und daher keine solide Evidenzbasis erzeugen können. Und diese Ressourcen stattdessen in wenige, dafür aussagekräftige Studien stecken, dann wäre schon viel gewonnen.

Das sind unbequeme Gedanken, die auch gegen den Zeitgeist und aktuelle geopolitische Entwicklungen schwimmen. „Accelerated approval" ist das Stichwort der Stunde, umso mehr Trump die „Pharmaindustrie von den Ketten der FDA" befreien will.

Auch die Versprechungen der „personalisierten Medizin" müssen in diesem Kontext hinterfragt werden. Es liest derzeit häufig, wie einzigartig das Genom, Phänom und Envirom jedes Individuums ist. Weshalb jeder Patient eine auf diese einzigartige Konstellation abgestellte Therapie braucht. Es ist geradezu trivial vorherzusagen, dass die vermutlich besser sein wird, als eine „one size fits all" Therapie für ganze Populationen.

Worüber aber kaum gesprochen wird, ist dass es in den meisten Fällen unmöglich sein wird, einen robusten Wirknachweis zu führen. Man ist auf n=1 Studien angewiesen, ohne Kontrolle, ein rein anekdotisches – in der Medizin würde man sagen kasuistisches - Vorgehen. Nur wenn eine personalisierte Therapie extreme Effektstärken hat, wie komplette Remission oder gar Heilung, und das dann häufiger beobachtet wird, darf man sich auf der richtigen Spur wähnen. Aber solche „Wundertherapien" sind leider selten. Und man hat trotzdem immer noch das Problem, dass dramatische, aber seltene Nebenwirkungen nur in großen Studien erkannt werden können. Das macht rationelle Nutzen-Risiko Abwägungen, ein Grundprinzip medizinischen Handels, praktisch unmöglich. Fazit also: Die personalisierte Medizin wird uns eine Vielzahl von hoch plausiblen, nur durch Kasuistiken belegte, in der Regel sehr teure Therapien bescheren. Ein Gewinner dieser Entwicklung steht jetzt schon fest: Die Pharmaindustrie.

Wo wir schon davon sprechen: Da sind dann noch die perversen monetären Anreize im Gesundheitssystem. Verkauft werden entgegen dem Mantra der Industrie nicht Gesundheit, sondern Medikamente und Medizinprodukte. Die Erlöse alimentieren einen gigantischen Lobbyapparat, mit dem man auf Politik genauso wie auf Ärzte und Wissenschaftler Einfluss ausübt. Weniger vornehm ausgedrückt: Man versucht, diese zu bestechen.

Und, siehe Scamiflu, Paxil u.v.m., man schreckt dabei auch nicht davor zurück, negative Daten zurückzuhalten, oder wissenschaftlichen Manuskripten und Leitlinien über sogenannte „Key Opinion Leaders" den nötigen „Spin" zu geben. Nicht vergessen wollen wir Teile der Ärzteschaft, die manchmal sogar wider besseres Wissen Therapien verordnen, weil sich mit denen ordentlich Geld verdienen lässt, auch wenn Sie unwirksam oder potenziell schädlich sind. Auch nachdem klar war, dass Stenting bei stabiler Angina pectoris einfacher medikamentöse Therapie nicht überlegen ist, haben Ärzte bei dieser Indikation noch Stents für über 12 Milliarden US $ eingesetzt! Was könnte man mit all dem Geld, das im Gesundheitssystem in falsche Kanäle fließt, für phantastische klinische Studien durchführen.

Auch die Lehre sollte man sich vornehmen. Medizinstudenten erfahren zwar (ein bisschen) was zu Evidenzbasierter Medizin (EBM). Aber Medical Reversals werden nicht gelehrt, so auch nicht das Bewusstsein dafür, wie dünn das Eis ist, auf dem sie später therapieren werden. Und damit erfahren sie auch nicht, dass und was man dagegen machen kann.

Letztlich zeigt uns das Phänomen der Medical Reversals, wie wichtig kritisches Denken und robuste, methodisch kompetente Forschung in der Medizin sind. Medizinisches

Handeln muss auf solider wissenschaftlicher Evidenz basieren und nicht auf bloßer Plausibilität, wackeligen Studien, oder wirtschaftlichen Interessen. Der Weg ist noch weit, bis deutlich weniger als die Hälfte dessen, was im Medizinstudium gelehrt wird, falsch ist – und man eine Vorstellung davon hat, welche Inhalte das betreffen könnte.

Coole Chefs, steile Hierarchien: Wie Machtmissbrauch in der Wissenschaft gedeiht

LJ 1-2/2025

Der Narr sonnt sich in der Vorstellung, immer ein kollegialer Chef gewesen zu sein. Von Anfang an wurde in seiner Arbeitsgruppe konsequent geduzt. Um deren „After work parties" ranken sich heute noch Legenden. Hierarchien im eigenen Labor? Ach was, nicht bei uns! Wir waren alle super drauf, man frotzelte fröhlich, auch über den AG Leiter, dann Prof., und schließlich Abteilungsleiter. Witze auf seine Kosten waren ausdrücklich erlaubt. Und das hört der Narr nicht nur über sich selbst, sondern auch von vielen seiner Kollegen. Wir sind alle coole Chefs, und „führen" unsere Studenten, Postdocs und ihr technisches Personal natürlich auf „Augenhöhe".

Knapp 30 Jahre hat es gedauert, bis dem Narren klar wurde, dass er sich und anderen etwas vorgemacht hat. Das mit der fehlenden Hierarchie hat nämlich was von einer wohlfeilen Selbsttäuschung. Die ihm noch nicht einmal auffiel, als ihn eine neue Postdoc hartnäckig siezte, obwohl er sie duzte. Sie kam aus dem Labor eines berühmten Neurowissenschaftlers – und zwar ziemlich traumatisiert. Mittlerweile ist dieser wegen Datenmanipulation in den Schlagzeilen. Auch dort wurde fleißig geduzt, man gab sich locker, und alle folgten brav der vom Chef ausgegebenen Maxime: „Der Tag hat 24 Stunden, und dann bleibt ja noch die Nacht".

Auch wenn man es sich nicht eingestehen will: Das Machtgefälle, die Hierarchie, ändert sich keinen Millimeter, nur weil der Chef seine Bürotür immer offenstehen lässt, bei Pizza und Bier am Freitagabend mitmischt oder großzügig Paddel- und Kletter-Retreats sponsert. All das kaschiert die Hierarchie höchstens – und genau das kann der Anfang richtig großer Probleme sein.

Denn die Hierarchien in der Wissenschaft sind, abgesehen vielleicht vom Militär und der katholischen Kirche, kaum irgendwo so extrem ausgeprägt wie hier. Professoren – und bis zu einem gewissen Grad auch die aufstrebenden Gruppenleiter – verfügen über eine geradezu pharaonische Machtfülle. Vieles davon ist nicht einmal formell verbrieft, was die Sache nur noch problematischer und auch perfider machen kann.

Einerseits geht es um institutionelle, also strukturelle Macht – die ist recht offensichtlich. Bewertungen, Noten, Personalentscheidungen, der Zugang zu Ressourcen: All das

liegt in ihrer Hand. Damit nichts weniger als die Zukunft ihrer Schützlinge. Die Karriere- und Beschäftigungsmöglichkeiten im akademischen System sind extrem begrenzt, und Professoren fungieren als Gatekeeper, die den Zugang zu diesem exklusiven Kreis kontrollieren. Viele Gruppenleiter stehen zwar selbst noch in der Abhängigkeit „ihres" Professors, aber gleichzeitig sind sie schon die Türsteher für die nächste Sprosse auf der Karriereleiter ihrer eigenen Mitarbeiter.

Weniger offensichtlich, aber nicht weniger bedeutsam, ist die fachliche Macht der Chefs. Die basiert auf ihrem Erfahrungs- und Wissensvorsprung, ihrer Vernetzung, ihrer wissenschaftlichen und politischen Expertise sowie ihrem Standing in der Community – etwa durch Positionen in Fachgesellschaften. Mit ihrer Vita haben sie ja längst bewiesen, dass sie wissen, wie Karriere funktioniert. Und genau deshalb können sie den Weg ihrer oft noch „naiven" Mitarbeiter lenken. Autorenschaften, Einladungen zu Vorträgen, kleine Posten in Fachgesellschaften und Kommissionen – all das liegt in ihrer Hand. Aber klar, solche Möglichkeiten gibt es nicht für alle.

Das bringt eine soziale, oder besser gesagt, psychologische Komponente der Macht ins Spiel: Wer wird am meisten gefördert? Wer steht in der Gunst des Chefs ganz oben? Diese Dynamik schürt Konkurrenz und Leistungsdruck in der Gruppe – was am Ende den Output steigert, nicht notwendig jedoch dessen Qualität.

Wenn der Chef ein netter Typ ist, der dem Altruismus huldigt, perpetuiert all dies zwar ein fragwürdiges System, aber es muss für die Untergebenen nicht notwendig problematisch werden. Aber die Wahrscheinlichkeit, dass er wirklich so toll drauf ist, ist eher gering. Warum? Weil die Tatsache, dass er diese Leitungsposition innehat, ziemlich stark mit dem Vorhandensein von dem korreliert, was Psychologen „Dark Traits" (dunkle Persönlichkeitszüge) nennen. Und dafür gibt es eine ganze Menge Evidenz – Ausgewähltes wie immer nachzulesen unter https://dirnagl.com/lj.

Die sogenannte „dunkle Triade" – bestehend aus Narzissmus, Machiavellismus und Psychopathie – bringt eine ganze Reihe unangenehmer Charakterzüge mit sich: Selbstüberschätzung, das Bedürfnis nach Bewunderung, extreme Selbstbezogenheit, manipulatives Verhalten und eine kühle, strategische Vorgehensweise. Hinzu kommt die Instrumentalisierung anderer für eigene Ziele, der Fokus auf die Durchsetzung eigener Interessen, emotionale Kälte, impulsives Verhalten, kaum vorhandenes Schuldgefühl und eine oft nur oberflächliche soziale Anpassungsfähigkeit. Und genau diese Eigenschaften sind überraschend oft hilfreich auf dem Weg zur Professur.

Und genau hier liegt das Problem: Die Kombination aus steilen Hierarchien und einzelnen oder gleich mehreren „Dark Traits" im Führungspersonal schafft den perfekten Nährboden für Machtmissbrauch. Die Folgen? Sexuelle Belästigung und Übergriffe, mangelnde wissenschaftliche Integrität, Ausbeutung, blockierte Karrieren, Mobbing und die gezielte Manipulation von Ressourcen – all das wird so begünstigt.

Aber es kommt noch besser – oder schlimmer, je nachdem, wie man es sieht. Denn „Dark Traits" und ihre Konsequenzen haben tatsächlich auch erwünschte Effekte auf den Output von Abteilungen, Fakultäten oder sogar ganzen Universitäten. Die Organisationspsychologie nennt das „unethisches pro-institutionelles Verhalten": Individuen begehen unethische Handlungen, die zwar moralisch oder wissenschaftlich fragwürdig sind, aber gleichzeitig die Interessen oder das Ansehen der Institution fördern.

All das hat sehr viel mit einem Lieblingsthema des Narren zu tun. Denn Wissenschaftler mit ausreichenden Dark traits in Verbindung mit der Reputationsökonomie des akademischen Systems stellen den perfekten Nährboden dar, auf dem fragwürdige wissenschaftliche Praktiken und Wissenschaftsbetrug gedeihen. Das ganz normale Mühsal und

die Enttäuschungen in der täglichen Forschung lassen sich nämlich durch selektive Nutzung von Daten, Ideenklau von Anderen, oder Manipulieren und Erfinden von Daten deutlich vermindern. Und damit publikatorischen und Drittmittelerfolg beschleunigen.

Im unethisches pro-institutionelles Verhalten hat man auch eine einfache Erklärung, warum Institutionen in der Regel keine großen Anstrengungen unternehmen, vermuteten Wissenschaftsbetrug zu untersuchen oder gar zu ahnden. Häufig geht es ja um Top-Publikationen und prominente Wissenschaftler. Die bringen der Institution Renommee und ein gutes Standing bei den ‚Clarivate highly cited researchers', der Verhandlung über den Landeszuschuss, im DFG Förderatlas, oder anderen institutionellen Rankings.

Ob Nature Papers und hohes Drittmittelaufkommen auf seriöser Wissenschaft oder auf fragwürdigen Forschungspraktiken oder gar Betrug beruhen, macht für das Ranking erstmal keinen Unterschied. Je mehr desto besser, Quantität vor Qualität. Bei verschärfter Aufklärung verliert man nur Impact Factorathleten, und wird gleichzeitig mit Reputationsverlust bestraft. Ein klares ‚lose-lose' Szenario!

Was kann man gegen all das tun? Die Persönlichkeitsmerkmale von Forschenden lassen sich ja kaum ändern. Man kann jedoch versuchen, den Kreislauf zu durchbrechen, der zur evolutionären Selektion von Personen mit hohem *Dark Score* führt. Zum Beispiel, indem man bei Bewerbungen stärker auf die Qualität und Inhalte der Forschung achtet – ebenso wie auf die Betreuung und Förderung, die Bewerbende bisher geleistet haben, und generell auf ihren Umgang mit Mitarbeitenden in der Vergangenheit.

Es hilft schon enorm, wenn Institutionen sich überhaupt der Existenz von unethischem pro-institutionellem Verhalten bewusst werden. In Gesprächen des Narren mit Dekanen und anderen Würdenträgern hatte er oft den Eindruck, dass dieses Thema zunächst gar nicht auf deren Radar war. Doch sobald es angesprochen wurde, entwickelte sich meist eine angeregte Diskussion darüber, was man dagegen tun könnte.

Und die Liste ist lang. Bereits erwähnt: eine Reform der Bewertung von Wissenschaftlerinnen und ihrer Arbeit. Natürlich auch der Abbau von Hierarchien – was sich unter anderem durch mehr Diversität in universitären Gremien erreichen lässt. Mehr Nachwuchswissenschaftlerinnen, mehr Frauen, mehr Internationalität und Interdisziplinarität. Zudem die Etablierung einer Organisations- oder Teamkultur, die „dunkle Verhaltensweisen" ächtet und positive Werte wie Team Science, Partizipation und Fairness aktiv fördert. Selbstverständlich gehören auch mehr Transparenz und Gerechtigkeit bei Bewertungen, Einstellungen und Vertragsverlängerungen auf diese Liste.

Ein verstärktes „Hinschauen" wäre ebenfalls wichtig, um zu sehen, was in den Instituten und Arbeitsgruppen tatsächlich passiert. Klassiker sind hier die Vergabe von Autorenpositionen und generell Personalentscheidungen. Läuft alles korrekt ab, wenn ein Professor alle paar Wochen als Autor auf einem Paper erscheint? Ist es realistisch, dass große Arbeitsgruppen von Personen geleitet werden, die gleichzeitig Ämter in Gremien und Gesellschaften innehaben, klinische Verantwortung tragen und oft nicht einmal mehr einen Schlüssel zum Labor besitzen?

Auch brauchen wir niedrigschwellige Angebote für alle, die Probleme mit Machtstrukturen haben, aber nicht direkt hochoffizielle Wege, wie z.B. zu einer Ombudsperson gehen können oder wollen. Dazu gehört vor allem ein strukturiertes Mentoring, idealerweise von Personen außerhalb der eigenen Organisation oder Arbeitsgruppe. Und für den Fall, dass es ernst wird: Mechanismen, die Whistleblowing ermöglichen und diejenigen schützen, die den Mut aufbringen, Missstände anzuprangern.

Sollen sich Profs und AG-Leitende jetzt also nicht mehr mit ihren Mitarbeitenden duzen und kein Bier mehr zusammen trinken? Darum geht es nicht. Es geht darum, stärker zu reflektieren, was wir tun, und offen anzusprechen, welche Abhängigkeiten bestehen. Es gibt zwar Fortbildungen zur Mitarbeiterführung, aber diese konzentrieren sich oft auf Konfliktlösung, den Umgang mit „schwierigen" Personen, Mitarbeitergespräche und Zielvereinbarungen. Für Studierende und junge Wissenschaftler*innen gibt es Kurse zur guten wissenschaftlichen Praxis und „Softskill"-Trainings, um bessere Papers zu schreiben oder Vorträge zu halten. Aber reicht das aus?

Was wir jedoch weder lehren noch ausreichend reflektieren, ist der Umgang mit Hierarchien – die es immer geben wird und auch geben muss – sowie mit Machtdynamiken. Genau hier könnte man bei vielen ein Umdenken anstoßen, das zu einer rationaleren Praxis führt und somit Machtmissbrauch verringert.

Der Narr dankt Prof.Dr. Daniel Leising (Uni Dresden), dessen Vortrag an der Charité ihn zum Nachdenken gebracht und damit diesen Beitrag inspiriert hat.

Der Narr bleibt dran: Warum sollten wir eigentlich Wissenschaft vertrauen?

LJ 3/2025

Vor über vier Jahren (LJ 11/2019) stellte der Wissenschaftsnarr Ihnen die Frage: Warum trauen Sie eigentlich dem Weltklimarat, nicht aber den Klimaskeptikern? Hintergrund meines Kommentars war damals eine, nach kurzer initialer Begeisterung im Verlauf der Corona-Pandemie spür- und messbare Abnahme des Vertrauens der Bevölkerung in die Wissenschaft.

Als Wissenschaftler in unseren Fachgebieten können wir die Messungen, Modelle und Annahmen der Klimawissenschaftler schließlich auch nicht besser beurteilen als – sagen wir – Donald Trump. Der das Gerede um den Klimawandel als einen Plot der Chinesen erklärt, die damit die US – Wirtschaft schädigen wollen. Darin liegt im Lichte der derzeitigen weltpolitischen Entwicklungen auch heute noch die Brisanz dieser „Vertrauensfrage". Aber ist da nicht auch ein Widerspruch am Wirken: Eine der zentralen Normen der Wissenschaft ist der organisierte Skeptizismus (R. Merton). Doch wie passt dieser professionelle Skeptizismus mit dem abstrakten Vertrauen zusammen, das man der Wissenschaft entgegenbringen soll?

Genau hier setzen nämlich die Vorwürfe der Wissenschaftsskeptiker an: Beim Klimawandel oder bei Impfungen, so ihre Kritik, hätten wir Wissenschaftler unser Ideal des Skeptizismus aus eigennützigen Gründen – sei es aus Karrierestreben, im Streben nach

Fördergeldern oder politischer Anerkennung – verraten. Damit, argumentieren die Skeptiker, verletze die Wissenschaft nicht nur ihr Prinzip des Skeptizismus, sondern gleich noch eine zweite zentrale Norm: ihre Uneigennützigkeit („Disinterestedness"). Bei genauerer Betrachtung richtet sich diese Kritik jedoch damit vor allem an die Politik – die Wissenschaft gerät ins Visier, weil sie sich habe instrumentalisieren lassen.

Der Narr nimmt sich des Themas nun erneut an – zum einen, weil vielerorts ein weiter zunehmendes Misstrauen gegenüber der Wissenschaft beklagt wird. Und das, obwohl die wissenschaftlichen Daten zeigen, dass das Vertrauen in Wahrheit stabil bleibt (und zwar hoch!). Zum anderen, weil er damals eine wichtige Frage unbeantwortet ließ: Wieso darf die Wissenschaft überhaupt epistemische Autorität für sich beanspruchen? Es gibt viele fragwürdige Begründungen dafür, warum man gut gemachter Wissenschaft vertrauen sollte. Aber was wären wirklich überzeugende Gründe?

Zur Einstimmung – eine kurze Wiederholung des närrischen Arguments von 2019: Die aktuelle Wissenschaftsskepsis richtet sich nicht gegen die Wissenschaft an sich oder ihre Methoden. Vielmehr gilt sie Wissenschaftlern als elitärer Kaste, die sich angeblich einem verachteten politischen System „andienen". Wissenschaftler wird vorgeworfen, gegen ihre eigenen Normen zu verstoßen.

Mit Newton, Maxwell, Einstein, Charpentier und Doudna hat kein Skeptiker ein Problem. Ablehnung erfahren wissenschaftliche Befunde wie der Klimawandel oder Impfungen, weil sie Konsequenzen für das eigene Leben nach sich ziehen – oder nach sich ziehen sollten und nicht zu den eigenen politischen Ansichten passen. Diese Skepsis wird nämlich von einer tiefergehenden systemkritischen Polarisierung geprägt (Stichworte: „Deep State", „Lügenpresse") und bedient sich gezielter Desinformationsstrategien verstärkt – perfektioniert und verbreitet von den „Händlern des Zweifels" („Merchants of Doubt", N. Oreskes).

Diese hatten beginnend in den 60iger Jahren des vorigen Jahrhunderts im Auftrag der Tabakindustrie, der Öl- und Kohlkonzerne, und der Hersteller von Pestiziden und FCKWs gezielt Zweifel an wissenschaftlichen Konsensen geschürt haben, um politische Maßnahmen, Regulierungen oder gesellschaftliche Veränderungen zu verzögern oder zu verhindern. Hierzu etablierten sie die auch heute noch beliebten Praktiken der selektiven Verwendung von Informationen, der Erzeugung von Scheindebatten, dem persönlichen Angriff auf Wissenschaftler, der Verzerrung wissenschaftlicher Unsicherheiten und den Einsatz von Experten mit fragwürdiger Qualifikation.

Diese gezielte Instrumentalisierung von Skepsis durch die *Merchants of Doubt* war (und ist) auch deshalb so erfolgreich, weil viele Menschen von der Wissenschaft absolute Gewissheit erwarten – und nicht Wahrscheinlichkeiten oder Ergebnisse, die sich mit neuen Erkenntnissen anpassen können. Medien verstärken das Problem, indem sie vor allem über Kontroversen und abweichende Meinungen berichten – schließlich sorgt das für mehr Aufmerksamkeit. Die Echokammern der sozialen Medien erledigen den Rest.

Da die Wissenschaftskritik von Impfgegnern und Klimawandelleugnern also in Wirklichkeit keine Kritik an der Wissenschaft, sondern vielmehr System- und Elitenkritik ist, verpufft der Ruf nach mehr Vertrauen ins Leere. Ebenso wenig zielführend ist der Ruf nach mehr Wissenschaftskommunikation – es sei denn, sie macht deutlich, worauf die epistemische Autorität der Wissenschaft basiert. Mehr dazu gleich weiter unten.

Im öffentlichen Diskurs herrscht weithin Einigkeit darüber, dass das Vertrauen in die Wissenschaft immer weiter abnimmt. Doch sämtliche großen Studien – sowohl national als auch international – zeigen ein anderes Bild (Quellen wie immer unter https://dirnagl.com/lj). Sie belegen, dass sich das Vertrauen in den letzten Jahren kaum verändert

hat und weiterhin auf einem recht hohen Niveau liegt, ebenso wie das Vertrauen in die Integrität der Wissenschaftler.

Allerdings spielen politische Orientierung und Bildungsniveau eine wichtige Rolle. Menschen, die sich politisch eher rechts (bzw. konservativ) einordnen, zeigen in Europa und Nordamerika im Durchschnitt weniger Vertrauen in die Wissenschaft als diejenigen, die sich links (bzw. liberal) orientieren. Interessanterweise ist dieses Verhältnis in einigen Ländern Afrikas und Asiens genau umgekehrt! Das alles deutet bereits stark darauf hin, dass es weniger um echte Wissenschaftskritik geht, sondern vielmehr um politische Einstellungen oder eine allgemeine Uninformiertheit.

Es bleibt die Frage, warum wir der Wissenschaft vertrauen sollten – und warum wir den Berichten des International Panel on Climate Change (IPCC) zu Recht Glauben schenken. Ganz einfach: Weil wir selbst Teil der wissenschaftlichen Elite sind, also Wissenschaftsbetrieb sozialisiert wurden. Wir wissen, wie Wissen durch organisierten Skeptizismus entsteht. Paradoxerweise sind Vertrauen und Skeptizismus dabei kein Widerspruch: Wir vertrauen auf die Wirksamkeit des organisierten Skeptizismus!

Wer eine neue Theorie oder einen Befund vorlegt, muss diese methodisch sauber und mit belastbarer Evidenz untermauern. Die Beweislast liegt immer bei der Person, die eine neue Behauptung aufstellt. Wissenschaft ist in diesem Sinne intrinsisch konservativ.

Aber Vorsicht: Auch wenn diese häufig genannt werden, es gibt auch eine ganze Reihe weniger tauglicher Gründe, uns zu vertrauen. Dazu gehören:

Die Autorität einzelner Wissenschaftler: Der oft gepriesene Ruf nach Vorbildern wie Nobelpreisträgern ist trügerisch. Ein paar Beispiele: Luc Montagnier (Nobelpreis für Physiologie/Medizin, 2008), der das HI-Virus (HIV) entdeckte, machte später das "Gedächtnis des Wassers" salonfähig. Linus Pauling (Nobelpreis für Chemie, 1954; Friedensnobelpreis, 1962) propagierte hohe Dosen von Vitamin C als Wundermittel gegen Krankheiten wie Krebs. Oder Ronald Fisher, einer der Begründer moderner Statistik und Genetik, der als einer der schlimmsten „Merchants of Doubt“ gegen die Erkenntnis agitierte, dass Rauchen Hauptverursacher von Lungenkrebs ist.

Die Annahme einer einheitlichen Methode: Wissenschaft hat keine universelle, systematische Methode. Neben der hypothetisch-deduktiven Methode spielen auch Induktion, Beobachtung, Statistik und Modellierung eine Rolle – man denke an die Klimawissenschaft. Außerdem können Methoden fehlerhaft angewendet werden, etwa durch Bias oder schlechte Datenqualität.

Praktische Erfolge bestimmter Theorien: Auch praktische Erfolge sollten nicht automatisch Vertrauen rechtfertigen. Selbst aus falschen Theorien können korrekte Vorhersagen resultieren (der sogenannte „Fehlschluss der Bejahung des Nachsatzes“). Ein bekanntes Beispiel ist das ptolemäische Weltbild mit seinen Epizyklen. Über lange Zeit konnte es erfolgreich die Bewegungen der Planeten erklären – trotz seiner grundlegenden Fehler.

Hier kommt der entscheidende Punkt: Wissenschaft ist ein kollektives, soziales Unterfangen. Sie beruht auf kollektiver Intelligenz, die einerseits Evidenz schafft und sie gleichzeitig kritisch hinterfragt (Skeptizismus). Ergebnisse müssen unabhängig reproduzierbar sein, einer kollektiven Prüfung (Peer Review) standhalten und können im Zuge wissenschaftlicher Selbstkorrektur auch modifiziert oder widerlegt werden.

Die epistemische Autorität der Wissenschaft gründet sich also darauf, dass ihre Erkenntnisse das Resultat eines einzigartigen sozialen Überprüfungs- und Konsensprozesses

sind – ein Prozess, der sich über Jahrhunderte als extrem erfolgreich erwiesen hat. Wissenschaftliche Tatsachen sind dann verlässlich, wenn sich Wissenschaftler durch intensive Diskussionen und Prüfungen aus unterschiedlichen Perspektiven und methodischen Ansätzen über ihre Richtigkeit einig sind. Dieser Konsens entsteht jedoch nicht im Geheimen, sondern muss transparent sein und bestimmten Normen folgen, wozu organisierter (institutionalisierter) Skeptizismus, Wissenskommunismus, Universalismus und Uneigennützigkeit zählen.

Jetzt könnte man einwenden: „Und wenn sie nicht gestorben sind, dann forschen sie noch immer…". Aber leider gibt es, wie so oft beim Wissenschaftsnarr, kein Happy End. Denn trotz all ihrer Erfolge liefert die Wissenschaft reichlich Gründe, warum man ihr nicht blind vertrauen sollte. Da wären verschwiegene Interessenkonflikte, fragwürdige wissenschaftliche Praktiken bis hin zu handfestem Betrug, die Reproduzierbarkeitskrise – die Liste ist leider noch viel länger. Hinzu kommt, dass Daten oft nicht geteilt, unbequeme Ergebnisse nicht veröffentlicht werden und viele Erkenntnisse, trotz öffentlicher Förderung, hinter Bezahlschranken verschwinden.

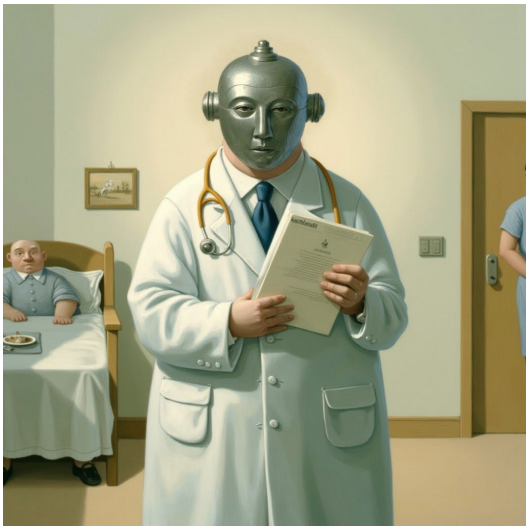
Aber vielleicht am gravierendsten ist das zunehmende Versagen des Peer Review Prozesses. Er ist oft völlig intransparent, was zu Verzerrungen und Homophilie führt, und begünstigt eine Angleichung an den Mainstream. Angesichts der Komplexität moderner Methoden und Ansätze können Gutachter die notwendige Expertise oft nicht abdecken. Viele Qualitätsaspekte von Artikeln lassen sich durch Peer Reviews gar nicht überprüfen, da hierfür Vor-Ort-Prüfungen nötig wären. Zudem erfolgt die Begutachtung meist post-hoc, sodass schwerwiegende Fehler und wissenschaftliches Fehlverhalten oft unentdeckt bleiben. Die Qualität der Peer Reviews ist äußerst heterogen, weder standardisierbar noch überprüfbar, und die Gutachter selbst sind oft nicht ausreichend geschult.

Hochwertige Peer Reviews erfordern enorm viel Zeit und sind deshalb kaum skalierbar, ohne dass die Qualität darunter leidet. Gleichzeitig wächst die Zahl der zu begutachtenden Arbeiten exponentiell – angetrieben durch die Praktiken renommierter Verlage, aber noch stärker durch Raubverlage und sogenannte Paper Mills. Letztlich versorgen sie uns aber nur mit der Währung, die wir selbst in unserer akademischen Reputationsökonomie geschaffen haben: Publikationen für den Lebenslauf. Das traurige Fazit dieser Entwicklung? Eine kaum noch kontrollierbare Flut von mediokren, überflüssigen und teils sogar betrügerischen Veröffentlichungen, die die Evidenzbasis der Wissenschaft zunehmend kontaminieren. Von wegen erfolgreicher kollektiver Überprüfung auf dem Weg zum wissenschaftlichen Konsens!

Es gibt also eine „Vertrauenskrise" innerhalb der Wissenschaft, und die können wir nur selbst lösen. Der Schlüssel dazu liegt in Offener Wissenschaft und einer Reform des akademischen Leistungsbewertungssystems. Gleichzeitig gibt es – auch wenn das Vertrauen stabil bleibt – eine öffentliche Vertrauenskrise in die Wissenschaft. Klar, 75 % Vertrauen mögen beeindruckend klingen, aber 25 % radikaler Skeptiker sind trotzdem eine ganze Menge. Der Aufstieg von praktischen Wissenschaftsskeptikern wie Trump, Meloni, Orban, Wilders und Co. bedeutet, dass dieser Skeptizismus unmittelbar dramatische gesellschaftliche Konsequenzen haben wird. Und da helfen weder Offene Wissenschaft, noch neue Bewertungsprinzipien wissenschaftlicher Leistungen oder gar Wissenschaftskommunikation.

KI in der Medizin: Hybris, Hype, Halbwissenschaft

LJ 4/2025



„Chatbot schlägt Ärzte bei der Diagnose!“. So schallte es kürzlich aus den Medien, von X bis New York Times. Eine randomisiert kontrollierte Studie einer illustren Schar von Autoren (Standford, Harvard, Beth Israel Deaconess usw.) hatte angeblich gezeigt, dass ChatGPT korrekter Diagnosen stellt als Mediziner in den genannten, weltbekannten Unikliniken. Nicht nur sei die Künstliche Intelligenz (KI) mit 92% deutlich akkurater als die Ärzte (74%) gewesen. Wenn diese den Chatbot nutzen durften, wären sie nur marginal besser (76%) geworden. Fazit also: Nicht nur stellt KI bessere Diagnosen als Ärzte, diese können auch nicht mit Chatbots umgehen, und lassen sich von der KI, obwohl sie es besser kann, auch nicht belehren!

Dies ist ein wunderbares Beispiel. Allerdings nicht für die enormen Fähigkeiten, welche KI in der Medizin bereits hätte. Sondern für die extreme KI-Hype, sowie die mangelnde Qualität der Studien in diesem Bereich. Weshalb der Narr sich nochmal dem Thema annehmen muss, nachdem er sich vor einiger Zeit schon mal ganz prinzipiell zu angeblichen ‚Intelligenz‘ von KI geäußert hat (LJ 6/23).

Zunächst zum Hype, der ist schnell abgehandelt, denn es dürfte mittlerweile auch den naivsten Zeitgenossen aufgefallen sein, wie überspannt die Versprechungen zu KI und Gesundheitssystem sind. Nur zwei Beispiele, pars pro toto: Deloitte, eine Consulting Firma, prognostiziert die Rettung von 400.000 Leben, die Einsparung von 200 Milliarden Euro und 1,8 Milliarden Arbeitsstunden allein in der EU, und das natürlich pro Jahr (Zitate und weiterführende Literatur wie immer unter https://dirnagl.com/lj). Sam Altman, der Chef von OpenAI, konfabulierte kürzlich im Rahmen der Vorstellung des Stargate Projektes durch Donald Trump („500 Milliarden Dollar für KI in USA“) im Januar: „KI wird helfen, Krankheiten in noch nie dagewesener Geschwindigkeit zu heilen: Wir werden erstaunt sein, wie schnell wir diesen Krebs und jenen sowie Herzkrankheiten heilen. Und was das für die Fähigkeit bedeutet … Krankheiten in rasantem Tempo zu heilen, wird, denke ich, eine der wichtigsten Errungenschaften dieser Technologie sein.“

Die Älteren unter Ihnen werden sich vielleicht erinnern, dass dies bereits das 4. Mal ist, dass die KI-Sau durchs Dorf getrieben wird: Marvin Minsky, ein KI Pionier war sich 1967 sicher, dass „innerhalb einer Generation das Problem der KI gelöst sein wird.“ In den 80er Jahren des vorigen Jahrhunderts glaubte man, dass „Expertensysteme bald Ärzte und Juristen ersetzen werden“. Als das Machine Learning (ML) Tool Deep Blue dann Gary Kasparov im Schach schlug, fühlte man sich der allgemeinen artifiziellen Intelligenz (AGI) von neuronalen Netzen ganz nah. ChatGPT hat natürlich einen Punkt, wenn es die Geschichte der KI selbstbewusst so sieht: KI-Hype kommt und geht – aber der Fortschritt bleibt.

Die Wissenschaft kriegt gleich ihr Fett weg, aber wieviel KI ist denn eigentlich schon ‚auf der Straße', wieviel KI nutzen wir heute in der Medizin? Denn man versucht ja schon seit deutlich mehr als einem Jahrzehnt sehr intensiv und mit sehr, sehr viel Geld KI in medizinische Anwendungen zu bringen. Man denke nur an Watson for Oncology von IBM. Gestartet 2010, begraben mitsamt der ganzen IBM KI Health Sparte im Jahr 2022.

Derzeit gibt es rund 1000 von der FDA zugelassene KI- und ML-Algorithmen, diese hauptsächlich in der medizinischen Bildgebung, also Radiologie und der Kardiologie. Klingt nach viel, ist es aber nicht, wenn man bedenkt, wieviel Ressourcen da reingesteckt wurde und wie lange man schon dran ist. Außerdem finden sich die Mehrheit der FDA-Zulassungen nur in wenigen Fachgebieten, wie Radiologie, oder Kardiologie, mit stark überlappenden Funktionalitäten.

Viel interessanter ist allerdings die Frage, wie viele KI-Tools es bereits mit hohem Evidenzlevel (I) in Guidelines von medizinischen Fachgesellschaften geschafft haben. Die lassen sich leider an nur zwei Händen abzählen. Und werden auch nicht in den Guidelines empfohlen, weil sie die Patientengesundheit verbessern. Sondern weil sie einige ärztliche Tätigkeiten effektiver machen, also Zeit sparen. Daran ist nichts verkehrt, aber so richtig prickelnd ist es noch nicht.

Dass so wenig KI in den Guidelines empfohlen wird, ist nicht verwunderlich. Denn empirische Belege für die Kosteneffizienz, geschweige denn Verbesserung von Diagnostik und Therapie durch KI im Gesundheitswesen sind rar und wo vorhanden, häufig methodisch unzureichend. Die meisten Studien konzentrieren sich auf technische Leistungskennzahlen oder klinische Machbarkeit, sind im Wesentlichen Proof of Concept (PoC). Robuste gesundheitsökonomische Bewertungen fehlen. Ein Systematischer Review fand nur 86 randomisiert kontrollierte Studien im Feld, von denen aber über zwei Drittel zu klein oder methodisch fraglich waren. Ohne solche Studien und ohne die erforderliche Qualität wissen wir schlichtweg nicht, ob einzelne KI-Tools tatsächlich Kosteneinsparungen bringen, oder ärztliche Entscheidungen und dadurch klinischen Endresultate verbessern können. Oder vielleicht sogar verschlechtern.

Und damit sind wir bei der KI - Wissenschaft. Diese steckt, wer hätte das gedacht, wie so manche andere Sparte der Forschung, in einer Reproduzierbarkeitskrise. Nur rund 5 % der KI-Forscher teilen ihre Quellcodes, und weniger als ein Drittel stellen ihre Trainings- oder Validierungsdaten zur Verfügung. So kann man gar nicht versuchen, etwas zu reproduzieren. Dort wo dies möglich war und versucht wurde, waren weniger als ein Drittel der Schlüsselresultate von KI-Studien reproduzierbar. Aber wie soll etwas, das man nicht wiederholen kann, zur soliden Grundlage für eine Entwicklung nützlicher Tools in der Biomedizin werden?

Reproduzierbarkeit ist ein Problem in allen Bereichen der Wissenschaft. Aber die Reproduzierbarkeit von KI-generierten Ergebnissen steht zusätzlich vor zahlreichen für diese Techniken spezifischen Herausforderungen. Zufälligkeit und Stochastizität können dazu führen, dass Algorithmen im Deep Learning unterschiedliche Ergebnisse liefern. Ein Mangel an Standardisierung in der Vorverarbeitung, etwa der Datenkennzeichnung für die Klassifizierung kann die Modellleistung erheblich beeinflussen. Nicht-deterministische Hardware- und Softwarebedingungen, wie Unterschiede zwischen den Prozessoren verschiedener Hersteller, können ebenfalls zu abweichenden Resultaten führen.

Hinzu kommt, dass Versionsprobleme, wie die Umstellung von verschiedenen Versionen einer ML-Bibliothek signifikante Unterschiede in den Ergebnissen verursachen kann. Auch die Verfügbarkeit und Variabilität von Datensätzen stellt ein Problem dar, da

proprietäre Gesundheitsdatensätze oft nicht zugänglich sind, wodurch unabhängige Replikationen verhindert werden. Ein weiteres Problem besteht im Überanpassen an spezifische Trainingsdatensätze. Die Interpretation der Ergebnisse wird auch durch eine übermäßige Abhängigkeit von denselben wenigen Datensätzen erschwert. Schließlich entstehen Verzerrungen durch selektive Berichterstattung, wenn nur die besten Versuchsergebnisse veröffentlicht werden, während weniger erfolgreiche Durchläufe verschwiegen bleiben.

Zudem sind KI-Algorithmen in der Regel Black Boxes, deren Ergebnisse oft nicht nachvollziehbar sind. Maschinelle Lernmethoden sind atheoretisch, assoziativ und häufig undurchsichtig. Dadurch wird die Erklärbarkeit ("Explainability") zu einer zentralen Herausforderung für KI. Wenn Nutzer die Entscheidungswege nicht verstehen, sind Fehler und Verzerrungen schwerer zu erkennen und zu korrigieren.

Gleichzeitig fällt es Menschen schwerer, Vorhersagen oder Empfehlungen zu akzeptieren, wenn sie nicht nachvollziehbar sind. Eine besondere Herausforderung ergibt sich aus der Wechselwirkung zwischen zwei gegensätzlichen psychologischen Phänomenen: der "Algorithmic Aversion" – also der Skepsis gegenüber algorithmischen Entscheidungen – und dem "Automation Bias", der dazu führt, dass Menschen automatisierten Systemen oft blind vertrauen. Während einige Nutzer KI-gestützte Entscheidungen kritisch hinterfragen oder ablehnen, neigen andere dazu, sie ungeprüft zu akzeptieren und sich weniger auf ihr eigenes Urteilsvermögen oder eine manuelle Überprüfung zu verlassen. Diese Dynamik macht den verantwortungsvollen Einsatz von KI umso komplexer, umso mehr, als ihre Ergebnisse nicht nachvollziehbar sind.

Ein substanzieller Korpus des ‚kausalen Wissens' der Medizin stellt sich im Nachhinein als falsch heraus: Die mit offensichtlich falschen Theorien begründeten, aber empirisch mit randomisiert kontrollierten Studien als erfolgreich belegten Therapien werden trotzdem weiter eingesetzt. Und wir finden meist gleich eine neue, passendere Theorie – die Erklärbarkeit ist scheinbar wiederhergestellt. Eine generelle Forderung nach vollständiger Erklärbarkeit von KI-Entscheidungen in der Medizin ist daher vermutlich unbegründet, könnte gar schädlich sein. In jedem Fall erfordert jedes klinische KI-Tool eine individuelle Bewertung der Erklärbarkeitsanforderungen, was die Sache nicht einfacher macht.

Zu den Kernproblemen der ML-basierten KI zählt auch der Daten-Drift. Er tritt auf, wenn sich die Daten mit der Welt um sie herum weiterentwickeln, der Algorithmus jedoch in dem Zeitraum verbleibt, in dem er trainiert wurde. Und das passiert in der Medizin ständig. Die medizinische Praxis ändert sich, die Bevölkerungsstruktur dazu, und noch vieles mehr. Da wird die KI sogar potenziell zum Opfer ihres eigenen Erfolges: Sollte sie dazu beigetragen haben, Diagnosen oder Therapien erfolgreich zu verbessern, ist die Wahrscheinlichkeit hoch, dass ihre Vorhersagen und Empfehlungen dadurch schlechter werden.

Wenn das passiert, kann das viele Menschleben kosten, wie geschehen bei Epic's Sepsis Prediction Model. Es wurde bei einem der riesigen Klinikkonzerne der USA eingesetzt, um das Risiko für die Entwicklung einer Sepsis vorherzusagen und dementsprechend zu therapieren. Das hat eine Weile gut funktioniert, bis sich die Kodierung von Sepsis im ICD-System veränderte, und der Konzern zusätzliche Krankhäuser gekauft hatte, die nicht im Trainingsdatensatz waren.

Das ist auch ein Beispiel für die Generalisierungsprobleme von KI-gestützten Gesundheitssystemen. Auch IBM Watson for Oncology funktionierte eine Weile ganz gut am Memorial Sloan Kettering Cancer Center, konnte aber nicht auf andere Krankenhäuser

übertragen werden. Optum's Healthcare KI diskriminierte schwarze Patienten bei der Risikobewertung. Google's Retinopathy Detection Model funktionierte in Studien, scheiterte jedoch im Praxiseinsatz in Thailand. Die NHS- und Babylon Health-KI gab irreführende oder unsichere medizinische Ratschläge. COVID-19-KI-Modelle versagten in realen Bedingungen. Google's Brustkrebs- und Stanford's Pneumonie-Modelle erzielten gute Ergebnisse in Tests, aber nicht in klinischen Einsätzen. Google Flu Trends scheiterte spektakulär bei der Grippevorhersage. KI zur Hautkrebsdiagnose schnitt bei dunkler Haut schlechter ab. PathAI's Diagnostik-KI führte zu unterschiedlichen Diagnosen und Fehldiagnosen von Krebs. Die Liste ließe sich beliebig fortsetzen.

Und in allem steckt das Bias-Problem: Alle KI-Modelle, aber insbesondere die momentan sehr populären größten Sprachmodelle (LLMs), bergen das Risiko, bestehende Verzerrungen in der Forschung zu verstärken. Dies passiert auf allen Ebenen des KI Lebenszyklus, vom Datensammeln und Annotation, über die eigentlichen Modellentwicklung, zum „Deployment", und der Evaluierung. Viele veröffentlichte wissenschaftliche Informationen sind falsch, veraltet oder voreingenommen. Da KI-Modelle auf diesen teils fehlerhaften Daten trainiert werden, verbreiten sie deren Mängel weiter und verstärken sie sogar. Eine zentrale Herausforderung besteht darin, zwischen vertrauenswürdigen und weniger glaubwürdigen Informationsquellen sowie zwischen voreingenommenen und neutralen Studiendesigns zu unterscheiden. Das gelingt selbst Experten häufig nicht.

Aber gibt es nicht bereits eine Lösung für all diese Probleme? Man müsste KI doch nur „trustworthy", also vertrauenswürdig machen? Und hat nicht bereits 2019 eine High-level Expert Group on Artificial Intelligence der Europäischen Kommission „Ethikrichtlinien für vertrauenswürdige KI" erstellt? Baut nicht der 2024 verabschiedete EU Artificial Intelligence Act (144 Seiten!) darauf auf? Ist Trustworthy KI damit nicht sogar gesetzlich kodifiziert, zumindest in der EU? Es gibt doch auch aktuelle Stellungnahmen des deutschen Ethikrates, der Bundesärztekammer, etc., die alle in diese Richtung zielen.

Es gibt in der Tat eine Fülle ethischer Richtlinien für die KI-Forschung und -Entwicklung, doch klafft eine erhebliche Lücke zwischen diesen hochrangigen Prinzipien und ihrer praktischen Umsetzung. Meta-Studien haben fast 100 solcher Richtlinien identifiziert, dennoch entstehen laufend nach diesen Kriterien „unethische KI-Anwendungen". Die meisten Rahmenwerke für vertrauenswürdige KI sind zu abstrakt und bieten kaum praktische Orientierung – über 75 % enthalten nur allgemeine Prinzipien, und mehr als 80 % liefern wenig bis gar keine konkreten Handlungsempfehlungen für Forscher und Entwickler. Sie zeigen auf „was" gemacht werden soll, aber nicht „wie".

Es existiert sogar ein regelrechter „Markt" für Trustworthy KI, der Entwicklern und der Industrie ein „Ethics (bzw. Fair) Washing" und „Ethics shopping" ermöglicht. Es wird sich schon eine genügend abstrakte Guideline finden lassen, die zum eigenen Produkt passt. Und falls nicht, kann man sich immer noch eine selbst stricken, das nennt man dann „Ethics lobbying" oder auch „Regulatory capture". So wie dies z.B. Microsoft derzeit tut, indem es vier Initiativen zur Entwicklung von KI-Richtlinien im Gesundheitswesen ins Leben gerufen hat, das „Trustworthy and Responsible AI Network (TRAIN)". Und dafür Experten, technische Unterstützung und finanzielle Mittel bereitstellt. Dies ermöglicht dem Unternehmen, Teststandards und Vorschriften mitzugestalten, wodurch die eigene Technologie bevorzugt geprüft und der Markteintritt für Wettbewerber erschwert wird.

Vergessen wird bei alledem auch, dass die meiste KI-„Forschung" in Wirklichkeit mehr Ingenieurwesen als wissenschaftliche Forschung ist. Der Schwerpunkt der aktuellen KI-Forschung im Gesundheitsbereich liegt hauptsächlich auf der Erkundung von Lösungen

und Anwendungen sowie auf Machbarkeitsstudien (Proof of Concept, PoCs), die nicht ausreichend in der realen Welt validiert sind. Davon gibt es mehr als genug, es werden immer mehr, und meist werden sie uns als mehr verkauft – als einsatzreifes, „transformatives" Tool. Echte Validierungen von KI-Tools im Sinne qualitativ hochwertiger randomisiert kontrollierter Studien kann man an 2 Händen abzählen.

Diese Differenzierung zwischen Exploration bzw. PoC auf der einen, und Bestätigung und Validierung auf der anderen Seite wird derzeit weder bei der Bewertung von Studienergebnissen noch in der Diskussion über die Vertrauenswürdigkeit von KI ausreichend berücksichtigt. Dabei sollten unterschiedliche Vertrauenswürdigkeitsstufen für die explorative Phase und PoC bzw. für die Validierung oder Implementierung gelten. Anforderungen an Art und Umfang der Trainingsdaten, an Erklärbarkeit und Transparenz, an Methoden zur Reduzierung von Verzerrungen (Bias) sowie an viele weitere Aspekte müssten je nach Entwicklungsstadium unterschiedlich ausgestaltet werden. Das sollte dann auch entscheidend dafür sein, ob ein KI Tool schon auf Patienten in der medizinischen Routine losgelassen werden darf.

Die eingangs erwähnte Arbeit, in der angeblich die Überlegenheit von ChatGPT gegenüber Ärzten in der Diagnosestellung gezeigt wurde, ist nicht nur ein Beispiel für die mediale Hype um KI, sondern auch für die mangelnde Qualität der Wissenschaft in diesem Bereich. Die Stichprobengröße dieser Studie war 6, auf die auch dann noch die falsche Statistik angewendet wurde! Der 2018 verstorbene, bekannte britische Statistiker Douglas Altman pflegte zu sagen: „n=8 ist eine Dinnerparty, keine Studie". Es ging auch gar nicht darum, ob richtige Diagnosen gestellt wurden, sondern um diagnostisches Schlussfolgern („reasoning"), das mit einer arbiträren, nicht validierten Skala vermessen wurde. Dies nur eine Auswahl einer Vielzahl von Problemen, diese Arbeit hätte niemals in The Lancet – Digital Health publiziert werden dürfen. Soviel auch zur verbreiteten Überschätzung der Qualitätskontrolle durch „Peer Review", über die sich der Narr schon häufiger kritisch geäußert hat (z.B. LJ 10/2020).

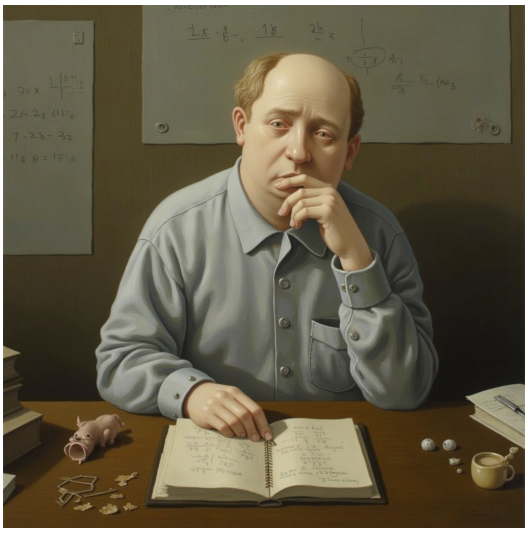
Dass es auch anders geht, zeigt eine sehr gut gemachte eben veröffentlichte einfach verblindete randomisiert kontrollierte Studie zur Genauigkeit von Mammographie-Screening bei 105.000 (!) Frauen. Die Ergebnisse legen nahe, dass KI zur frühen Erkennung von klinisch relevantem Brustkrebs beitragen kann und die Arbeitsbelastung beim Screening reduziert, ohne die Anzahl der falsch-positiven Befunde zu erhöhen.

Selbstverständlich hat KI das Potenzial, die medizinische Diagnostik, klinische Entscheidungsfindung und Prognostik zu verbessern, die Medikamentenentwicklung zu beschleunigen und tragbare Gesundheitstechnologien („Wearables") voranzutreiben – um nur einige der bekanntesten und häufig zitierten Anwendungsbereiche zu nennen. Allerdings wird die Entwicklung medizinischer KI-Anwendungen, die effektiv, sicher, vertrauenswürdig, fair und nachhaltig sind, derzeit durch eine „Move fast, break things"-Mentalität kommerzieller Entwickler sowie durch intensives Lobbying für eine industriefreundliche Regulierung behindert. Gleichzeitig zeigt die Wissenschaft – geblendet vom Hype – erhebliche Defizite in Transparenz, Reporting, der Reduktion von Bias und einer kompetenten Validierung. Kurzum: Wir brauchen eine evidenzbasierte KI.

Der Narr dankt Dr. Vince Madai für anregende Diskussionen und Kritik.

Rechnen bis man Sternchen sieht: Warum das Verhältnis von Experimentatoren und Statistikern so zerrüttet ist

LJ 5/2025



Es ist schon eine Weile her, da war der Wissenschaftsnarr Chief Editor eines angesehenen Journals einer ebenso angesehenen Fachgesellschaft. Schon damals hat er sich darüber aufgeregt, dass in der Biomedizin viele wissenschaftliche Artikel - ob aus Unwissen oder dem Drang, ihre Ergebnisse durch signifikante p-Werte und eindrucksvolle Korrelationskoeffizienten aufzuhübschen - auf fragwürdige experimentelle Designs und statistische Tricksereien setzen. Viel zu niedrige Fallzahlen, fehlende Randomisierung und Verblindung, ungeniertes Fischen nach statistischen Signifikanzen („p-Hacking"), Hypothesen erst nach dem Blick auf die Ergebnisse zu formulieren und sie anschließend als vorab festgelegt auszugeben („HARKING"). Kurz gesagt: lauter kreative Wege, um mit begrenzten Ressourcen (Versuchstiere, Studis, Geld, Zeit …) und vielleicht gar nicht so eindeutigen Resultaten trotzdem eine runde Geschichte zu basteln – mit etwas Glück sogar für ein Journal mit ordentlichem Impact Factor.

Im Gegensatz zu klinischen Studien in reputierlichen Journalen (Lancet, NEJM, JAMA, etc.) unterliegen nämlich (auch heute noch) präklinische Studien keinem professionellen, statistischem Review. Selbst Journale wie Cell, Nature und Science haben so was nicht, schon gar keine dedizierten Statistik-Editoren. Also, dachte der Narr, nun sei die Zeit gekommen, diesen fragwürdigen Praktiken zumindest in „seiner" Zeitschrift als Chief Editor den Garaus zu machen.

Weil Fachgesellschaft und Verlag Statistik-Reviews nicht finanzieren wollten, kratzte er die dafür nötigen Mittel aus „Bordmitteln" seines Institutes zusammen. Eingereichte Papers wurden dann von Statistikern auf Honorarbasis parallel zum regulären Peer Review begutachtet. Allerdings bemängelten diese fast alle ihnen vorgelegten Submissionen. Bei nicht wenigen davon waren dies nicht mehr korrigierbare Mängel. Der Urvater der Statistik, Ronald Fisher formulierte das Problem so: „Den Statistiker erst nach Abschluss eines Experiments zu konsultieren, bedeutet oft lediglich, ihn um eine Obduktion zu bitten. Er kann vielleicht sagen, woran das Experiment gestorben ist."

Nicht überraschend waren die betroffenen Autoren über diese negativen Reviews nicht glücklich: „Sowas ist uns ja noch nie passiert, dann publizieren wir das halt woanders". Und auch die hinter dem Journal stehende Fachgesellschaft, sowie der Verlag Nature, waren recht unzufrieden. Denn das Journal brachte ihnen bisher sehr viel Geld ein. Der Vorwurf an mich war aber nicht nur, dass ich Autoren vergrämen würde. Die Veröffentlichung exklusiv der Arbeiten, welche so einen statistischen Review überstanden hatten, würden zudem den Impact Factor des Journals ruinieren. Denn solche Papers berichteten in der Regel über weniger sensationelle Effekte, oder wareb häufig gleich „negativ".

Und sowas wird bekanntermaßen weniger zitiert, worunter dann der Impact Factor leidet.

Ich musste also die statistischen Reviews wieder einstellen, und bekam immerhin zum Trost eine neue Sektion im Journal spendiert: „Negative Studies“. In der durfte ich dann pro Heft 1-2 Arbeiten mit NULL-Resultaten in der Schmuddel-Ecke der NULL-Resultate in Quarantäne nehmen. Und die Moral von der Geschicht: Den p-Wert hinterfrägt man nicht!

Diese Anekdote ist in mehrfacher Hinsicht aufschlussreich. Fachzeitschriften kassieren ordentlich Geld dafür, dass sie unsere Forschung veröffentlichen – Forschung, die wir selbst durchgeführt, aufgeschrieben, formatiert und sogar begutachtet haben. Ihr Verkaufsargument? Sie sorgen für Qualitätskontrolle. Denkste!

Das zeigt uns aber auch – und das an einer ganz typischen und kritischen Stelle – dass der Peer-Review-Prozess schlicht versagt. Besonders dann, wenn er erst einsetzt, nachdem das Kind längst in den Brunnen gefallen ist! Und dass eine Antwort hierauf sog. „Registered Reports“ wären, wie sie einige Journals mittlerweile eingeführt haben, bei denen das Studiendesign samt Hypothesen und Methoden *vor* der Datenerhebung von Fachzeitschriften begutachtet wird. Wird es akzeptiert, garantiert die Zeitschrift die Veröffentlichung – unabhängig vom späteren Ergebnis. Gravierende Fehler in Methode, Design, oder Analyse werden frühzeitig erkannt und können beseitigt werden, solange sich das Kind erst über den Brunnenrand beugt. Und NULL Resultate kompetenter Studien gehen der wissenschaftlichen Community nicht verloren.

Aber mein gescheiterter Kreuzzug für ordentliche Statistik und Studiendesigns offenbart noch ein viel grundlegenderes Problem, über das (zu) wenig geredet wird: Das zutiefst zerrüttete Verhältnis zwischen Biostatistikern und den biomedizinischen Experimentatoren.

Experimentatoren glauben nämlich in der Regel, eigentlich gar keine Statistiker zu brauchen. Sie gehen in der Analyse ihrer Ergebnisse nämlich so vor, wie sie (oder ihre Kollegen) es schon immer gemacht haben, und gut damit publizieren konnten. Die gewünschten Gruppenunterschiede können in den meisten Fällen mit Excel, Graphpad, oder anderer Software, wenn auch manchmal erst nach einigem rumprobieren mit der Vielzahl der angebotenen Verfahren oder Kontraste gezeigt werden. Die niedrigen Fallzahlen sind ohnehin Literaturstandard, und die Statistik wird im der fachlichen Begutachtung eh nicht in Frage gestellt. Die Reviewer verstehen nämlich genauso wenig davon wie man selbst. Und machen es in ihren eigenen Studien ja selber so.

Zum Statistiker geht man deshalb nur, wenn’s gar nicht anders geht, sich also trotz des bewährten Vorgehens keine Signifikanzen einstellen wollen. Oder sich die Tierschutzbehörde, das Promotionsbüro, oder irgendeine andere wissenschaftsfeindliche Bürokratie erdreistet, dies zur Bedingung macht.

Die Biostatistiker, bei denen die Experimentatoren dann lechzend nach Signifikanzen und fachlicher Absolution aufschlagen, finden das naturgemäß wenig lustig. Denn selten kommen die Experimentatoren als verständnissuchende Partner auf Augenhöhe – sie suchen vielmehr Erfüllungsgehilfen beim Paper- oder Antrag-Schreiben, und wissen sowieso alles besser. Zumindest dass hier oder dort doch ein *-chen wohlverdient gewesen wäre. Statistik? Nur dazu da, Signifikanzen genau dort hervorzuzaubern, wo man sie braucht.

Hinzu kommt: Die Designs und Analysen der allermeisten präklinischen Studien sind für die Biostatistiker intellektuell oft kaum fordernder als ein Sudoku. Und für eine

multivariate Varianzanalyse gibt's auch keine Koautorschaft. Für sie ist das Ganze also komplett spaßbefreit. Ein kleines Cartoon auf YouTube bringt dieses Verhältnis zwischen beiden Welten herrlich auf den Punkt: https://bit.ly/4iH1P12.

Bei ordentlichen klinischen Studien ist das durchaus anders: Die Kliniker involvieren die Statistiker, häufig kommen diese aus der klinischen Epidemiologie, sehr früh in die Planung. Die Designs werden von Experten aufgestellt, die Analysen vorab definiert, und dann auch von den diesen durchgeführt. Auch sind die statistischen Verfahren in der Regel komplexer, Störfaktoren müssen berücksichtigt und durch Multiple Regressionen, Stratifikation, Sensitivitätsanalysen etc. in den Griff gebracht werden. Hinzu kommt, dass sich methodisch viel Neues tut, vermehrt adaptive Designs verwendet werden oder Causal Inference zum Einsatz kommt (der Narr berichtete hierzu im LJ 5/2024). Zu all dem braucht es kompetente Biostatistiker, diese werden wertgeschätzt, und haben ein wissenschaftliches Spielfeld, das ihnen auch Koautorenschaften für die eigene wissenschaftliche Karriere bringt.

Natürlich gäbe es auch präklinischen Bereich methodisch viel zu holen. Adaptive Designs zum Beispiel können die statistische Power erhöhen, bei gleicher Fallzahl. Überhaupt Fallzahl: Niedrige Fallzahlen sind eine Herausforderung für gängige Statistiken, auch hier kann der Biostatistiker helfen, Studiendesigns effizienter zu machen (Zitate und weiterführende Literatur wie immer unter https://dirnagl.com/lj). Deren Ehrgeiz, sich hier einzubringen, ist jedoch massiv gebremst. Denn die Experimentatoren wissen ja, was sie brauchen – einen signifikanten p-Wert. Und den gab es doch immer schon mit den gängigen Verfahren, Neues verwirrt die Gutachter nur und erhöht damit die Gefahr der Ablehnung.

Ein schönes Beispiel dafür ist, dass bei multiplen Vergleichen, nicht vorab definierten Analyseverfahren und sehr niedrigen Fallzahlen Teststatistiken in den meisten Fällen gar keine Rolle spielen sollten. Stattdessen sollte man die Ergebnisse – die man dann auch klar als explorativ kennzeichnen muss – lieber mit sauberer deskriptiver Statistik, ordentlichen Varianz- bzw Präzisionsmaßen und einer anschaulichen Darstellung aller Datenpunkte präsentieren.

Wer mit sowas in den Review Prozess geht – der Narr hat das mehrfach gemacht – muss sich warm anziehen. Die Gutachter sind im besten Fall verwirrt, meistens aber ziemlich ablehnend. Weil sie's einfach nicht besser wissen. Schon an dieser Stelle würde die geballte Expertise eines Statistikers im Team – beim Schreiben und als Koautor – enorm viel bringen, sich gegen die Ignoranz der Gutachter und Editoren durchzusetzen.

Das Gleiche gilt für die Auseinandersetzung mit den Genehmigungsbehörden. Wenn man einmal verstanden hat, dass man mit n=6 nur Effekte nachweisen kann, die biologisch völlig unrealistisch groß sind – und die meistens das Ergebnis von cleverem „Tuning" der Studienbedingungen und anschließendem Rosinenpicken in den Daten sind – dann müsste man eigentlich die Fallzahl erhöhen. Abgesehen vom Ressourcenproblem: Was würde die Behörde dazu sagen? Bisher hat man seine Anträge immer mit Gruppengrößen von 8 durchbekommen, und jetzt will man plötzlich 20? Ohne eine vernünftige Begründung und vorherige Diskussion mit der Behörde läuft da natürlich erstmal gar nichts.

Es ist aber möglich – und auch hier spricht der Narr aus eigener Erfahrung. Denn man hat gute Argumente: Zu kleine Gruppengrößen führen zu falsch-positiven und falsch-negativen Ergebnissen, zu überschätzten Effektgrößen, zu mangelnder Reproduzierbarkeit – kurz: zu Ergebnissen, auf die man sich wenig verlassen kann. Und genau deshalb sind solche Studien auch unethisch. Mit der Unterstützung von Biostatistikern, die

übrigens oft auch beratend in den entsprechenden Kommissionen sitzen – kann man sich mit solchen Argumenten durchaus durchsetzen. Oder sogar, wie es in Berlin passiert ist, nach engem Austausch mit Vertretern der Behörde gemeinsam zu einer sauberen Fallzahlabschätzung kommen und hierzu sogar Empfehlungen zu veröffentlichen.

Damit ist das Grundübel aber noch lange nicht beseitigt. In der Community der Experimentatoren gilt es nämlich als völlig akzeptabel, von Statistik keine Ahnung zu haben. Die meisten haben nie eine fundierte Ausbildung darin genossen – höchstens mal einen Kurs (und auch nur die Jüngeren), in dem man gelernt hat, wie man mit einem Statistikprogramm aus den eigenen Daten p-Werte rausholt. Nach dem bekannten Muster der im Labor eh schon lange genutzten Analysemethoden – nur dass man's jetzt eben selbst machen kann und dabei auch gleich noch einen hübschen Graphen erzeugt. Aber eben ohne wirklich zu verstehen, was man da tut – und mit einem fast ausschließlichen Fokus auf die Auswertung, nicht auf das Studiendesign.

Auch wenn man sich irgendwie daran gewöhnt hat – es ist eigentlich kompletter Wahnsinn. Wie kann es akzeptiert sein, dass Wissenschaftler, die an hochkomplexen Geräten arbeiten, mit aufwändigen Modellen jonglieren, dabei enorme Ressourcen verbrauchen und nicht selten das Leben von Tieren opfern, keine fundierten Kenntnisse in einer der grundlegendsten wissenschaftlichen Fähigkeiten besitzen? Nämlich: Wie plant man ein Experiment statistisch kompetent? Wie wertet man Daten korrekt aus? Und was bedeutet ein p-Wert unter einer bestimmten Schwelle – *wirklich*? Und das eigentlich Schlimmste: Niemand scheint sich darüber ernsthaft aufzuregen.

Was es bräuchte, wäre eine verpflichtende, umfassende Ausbildung für alle, die in der biomedizinischen Forschung arbeiten. Nicht nur, weil sie selbst Studien planen, Daten analysieren und interpretieren – sondern auch, weil sie als Wissenschaftler die Studien anderer lesen, bewerten und daraus ihre (deshalb oft falschen) Schlüsse ziehen. Ohne ein tiefergehendes Verständnis von Statistik – eines, das deutlich über die praktische Durchführung einer ANOVA in Excel hinausgeht – kann man die Qualität der Evidenz und die Tragfähigkeit der Schlussfolgerungen in wissenschaftlichen Arbeiten im eigenen Fachgebiet schlicht nicht richtig einschätzen.

So eine Ausbildung müsste sich stark aufs Studiendesign konzentrieren – und auf die grundlegenden Prinzipien der Statistik, weniger auf die technischen Details der Durchführung. Es würde zum Beispiel schon helfen, nicht nur etwas über Nullhypothesen-Statistik (NHST) zu lernen – also das, was die meisten von uns machen – sondern auch über Bayes'sche Statistik. Die meisten wissen gar nicht, dass es das überhaupt gibt, und halten NHST für gottgegeben. Dabei würde man gerade im Vergleich mit dem Bayes'schen Ansatz merken, dass es auch ganz anders geht. Und plötzlich versteht man, was eine a-priori-Wahrscheinlichkeit ist – und in der Verlängerung, was ein p-Wert *wirklich* ist. Nämlich keine Falsch-Positiv-Rate. Nicht, dass man danach sofort zur Bayes'schen Statistik überläuft. Aber man sähe das eigene methodische Vorgehen wie in einem Spiegel – und hätte ein ganz anderes Verständnis dafür, was man da eigentlich tut.

Statistikern muss man Anreize bieten, damit ihr Eigeninteresse geweckt wird, sich intensiver mit präklinischem Studiendesign und Datenanalyse auseinanderzusetzen. Schon allein der Umstand, dass ihnen besser ausgebildete Experimentatoren (siehe oben) mit mehr Sachverstand und echter Wertschätzung begegnen, könnte positiv wirken. Außerdem kann es wissenschaftlich durchaus reizvoll und karrieremässig ergiebig sein, gemeinsam neue, effektivere Designs und Analyseverfahren zu entwickeln und anzuwenden. Zusätzliche Ressourcen wie etwa die Möglichkeit, dafür Mittel in den Förderantrag einzustellen, oder ausreichende Beratungskapazität durch die Biostatistik-Abteilungen würden ihr Übriges tun. Vielleicht wäre die „Biostatistik kleiner Fallzahlen" –

und zwar nicht nur in der Präklinik – sogar ein richtig spannendes Thema für die Biostatistik-Community, das noch viel wissenschaftliche Potenzial bietet, auch in der Karriereentwicklung.

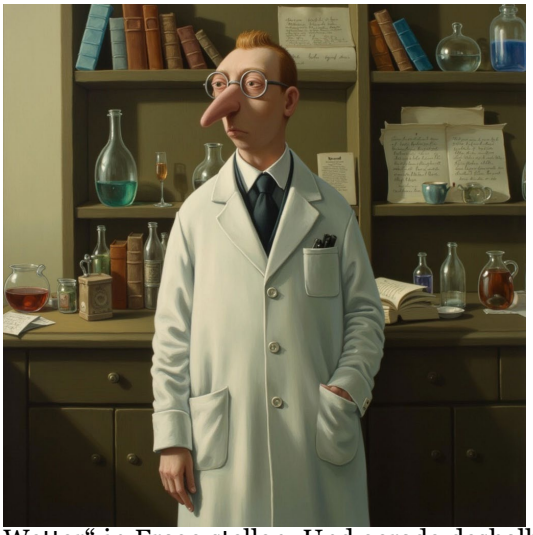
Das Gleiche gilt für das Design und die Biostatistik von Replikationsexperimenten – besonders im Kontext von Team-Science-Ansätzen, von denen man zum Glück in letzter Zeit immer mehr sieht. Das liegt nicht zuletzt daran, dass auch Fördergeber (wie etwa das BMBF) inzwischen erkannt haben, dass der Replikationskrise und der nur selten gelingenden Übertragung vielversprechender präklinischer Ergebnisse in tatsächlich wirksame Therapien aktiv entgegengewirkt werden muss.

Was muss außerdem passieren? Fördergeber und Fachzeitschriften müssen den Beschreibungen von Versuchsdesign und Statistik deutlich mehr Aufmerksamkeit schenken. Diese Abschnitte gehören von Biostatistikern gründlich geprüft. Gleichzeitig sollten Forschungsförderer es nicht nur erlauben, sondern ausdrücklich fördern, dass Antragsteller gezielt Mittel für die Planung und Auswertung ihrer Experimente beantragen können. Besonders in kollaborativen Projekten wie SFBs oder Forschergruppen wäre das sinnvoll. Das würde ganz nebenbei auch die „Teambildung“ zwischen Experimentatoren und Biostatistikern stärken – was für sich genommen schon ein echter Gewinn wäre.

John W. Tukey – Pionier der explorativen Datenanalyse, Erfinder des Boxplots und Namensgeber zahlreicher statistischer Eponyme – schrieb schon 1964: „Die meisten Anwendungen klassischer statistischer Methoden wurden, werden und werden auch in Zukunft von Menschen durchgeführt, die nicht wissen, was sie tun.“ Etliche Galaxien von $p \leq 0.05$-Sternchen später lohnt es sich immer noch darüber nachzudenken, wie man Experimentatoren und Biostatistiker wieder zusammenbringen kann.

Personalisierte Medizin, oder: Warum die Nase der biomedizinischen Forschung immer länger wird.

LJ 6/2025



Personalisierte Medizin, oder wie sie sich heute lieber nennt: Präzisionsmedizin (PM) – wird nun schon seit einiger Zeit als Heilsversprechen gehandelt. Der große Fortschritt auf dem Weg zu einem gesünderen, zufriedeneren und längeren Leben. PM, das klingt so gut, da kann man nicht dagegen sein. So überzeugend, dass es offenbar keiner weiteren Belege mehr bedarf, dass es was wirklich Neues ist, und den Königsweg für die Medizin der Zukunft darstellt.

Denn wer würde sich schon gegen eine „präzisere“ oder „personalisiertere“ Medizin aussprechen? Das wäre in etwa so, als würde man den Wunsch nach „sicherem Straßenverkehr“ oder „besserem Wetter“ in Frage stellen. Und gerade deshalb ist es bemerkenswert, mit welcher Überzeugung Wissenschaft und Politik diese Begriffe öffentlichkeitswirksam ins Schaufenster hängen – und dafür auch massiv finanzielle Mittel mobilisieren. Sogar der Koalitionsvertrag der neuen Bundesregierung formuliert die Selbstverpflichtung: „Wir stärken die Gesundheitsforschung auch mit Fokus auf personalisierte Medizin“. Ein Schelm, der Böses dabei denkt, und gar frägt: Cui bono?

Was Heilsversprechen angeht, liegt PM gleich hinter der Künstlichen Intelligenz (KI), die sich erst kürzlich auf den Spitzenplatz der biomedizinischen Hypes geschoben hat. Beide wetteifern darum, nichts weniger als die „Zukunft der Medizin“ zu sein.

KI und PM als Propheten einer goldenen Ära von Prävention, Diagnostik und Therapie haben dabei einiges gemeinsam: die Hybris ihrer Versprechen, von denen bisher kaum etwas eingelöst wurde, das auf Hochglanz polierte Marketing, eine erstaunlich lange Vorgeschichte, und eine bemerkenswerte Evidenzlücke, über die man konziliant einfach hinwegsieht.

Dabei reicht die Geschichte der PM deutlich weiter zurück, als die der KI – nämlich bis ins frühe 20. Jahrhundert, als man gerade anfing, Metabolismus und Genetik zu verstehen. Damals entstand das Konzept der „chemischen Individualität“, entwickelt unter anderem von Archibald Garrod und von ihm ausformuliert in *The Inborn Factors in Disease*. Zu dieser Zeit war „Faktor“ ein gängiger Begriff für das, was wir heute als Gen bezeichnen.

Auf die Idee, den eigentlich ziemlich offensichtlichen Einfluss individueller genetischer und daraus abgeleiteter biochemischer Faktoren auf Krankheitsverläufe (und damit auch Therapien) zur Grundlage einer gigantischen Marketingkampagne zu machen, kam man allerdings erst in den 1990ern. Und zwar mit der Erfindung der sogenannten „Molekularen Medizin“, nach der – man staune – Krankheiten auf Basis molekularer Vorgänge verstanden werden sollen.

Nicht zu verwechseln übrigens mit der „Orthomolekularen Medizin“ – einem esoterischen Konzept, das niemand Geringerer als 2-fach Nobelpreisgewinner Linus Pauling propagierte. So viel also zur epistemischen Autorität von Nobelpreisträgern. Heute begegnet uns dieses Gedankengut in jeder Apotheke – in Form von meterlangen Regalen voller ebenso teurer wie nutzloser Placebo-Produkte der Firma Orthomol.

Die akademische „Molekulare Medizin“ – heute eigentlich nur noch in Institutsnamen, Studiengangsbezeichnungen oder Journaltiteln zu finden – wurde schon kurz nach ihrer Erfindung sprachlich von der Marketing-Konkurrenz überrannt: Der Begriff „Personalized Medicine“ war einfach überlegen. Jeder Landarzt kann zu Recht von sich behaupten, seine Patientinnen und Patienten individuell zu behandeln – aber „personalisiert“ klingt einfach origineller und massentauglicher als „molekular“. Letzteres wirkt doch eher unpersönlich und technokratisch – dabei wünschen wir uns doch Empathie bei den Ärzten, eben eine persönliche Medizin.

Sehr schön auf den Punkt gebracht hat das kein geringerer als Barack Obama: *„Ärzte wussten schon immer, dass jeder Patient einzigartig ist – und sie haben auch stets versucht, ihre Behandlungen so gut wie möglich auf den Einzelnen abzustimmen. Man kann zum Beispiel eine Bluttransfusion auf die Blutgruppe abstimmen – das war eine bahnbrechende Entdeckung. Aber was wäre, wenn es genauso einfach und selbstverständlich wäre, eine Krebstherapie auf unseren genetischen Code abzustimmen? Was wäre, wenn die richtige Medikamentendosis zu finden so simpel wäre wie Fiebermessen?*“ Heute, im Zeitalter von Trump kaum mehr vorstellbar, aber 2015 kündigte der damalige US-Präsident tatsächlich so die milliardenschwere „Precision Medicine Initiative“ (PMI) der amerikanischen Regierung an.

Interessanterweise war die akademische PR-Maschinerie in den USA da schon wieder einen Schritt weiter. Aus „Personalized Medicine“ war „Precision Medicine“ geworden. Ein genialer Schachzug. Denn personalisiert war Medizin eben im Grunde schon immer irgendwie, und der Begriff ließ viele glauben, es gehe darum, dass dabei wirklich jeder eine ganz individuell zugeschnittene Behandlung bekommen würde.

Die Effekte dieser sogenannten „Personalisierung“ ließen sich wissenschaftlich ohnehin kaum nachweisen. Dafür wären lauter n=1-Studien – ohne Kontrollgruppe, ohne Verblindung, ohne Randomisierung nötig. Reine Anekdoten also. Behandlungen, bei denen Patienten im Übrigen häufig sowieso eine Besserung zeigen, die dann fälschlicherweise der „Personalisierung“ zugeschrieben wird – und nicht, wie es korrekt wäre, der Regression zum Mittelwert oder schlicht dem Placeboeffekt (vgl. dazu auch der Narr in *LJ* 1–2/2018). Oder, wie Voltaire es treffend formulierte: *„Die Kunst der Medizin besteht darin, den Kranken so lange abzulenken, bis die Natur die Krankheit geheilt hat.“*

Die Zukunft der Medizin gehört nun also – mindestens seit Obama – der Präzision! Hierzulande sind die feinen Unterschiede in der Begriffswahl allerdings noch nicht so richtig angekommen. Viele Kollegen sprechen nach wie vor von „personalisierter Medizin“, obwohl der Begriff marketingtechnisch eigentlich überholt ist.

PM berücksichtigt also individuelle Unterschiede in Genen, Umwelt und Lebensstil. Klingt erstmal nicht gerade bahnbrechend – im Gegenteil, das ist so offensichtlich, dass es eigentlich als banales und damit selbstverständliches Konzept sein sollte. Na klar, Klappern gehört zum Handwerk, schließlich brauchen wir Fördergelder für unsere Forschung und die Pharmaindustrie muss ihre Shareholder bei Laune halten. Aber warum regt sich der Narr dann überhaupt so auf?

Weil genau da der Haken liegt: Diese überzogene Verheißung, man könne damit Krankheiten verhindern, heilen, das Leben Kranker verbessern – und sogar verlängern.

Angeblich könnten wir mit PM unglaubliche Effekte erzielen. Das sollte schon deshalb hinterfragt werden, weil es doch viel einfachere und wirksamere Wege gäbe, die Bevölkerungsgesundheit zu verbessern – vorausgesetzt, man meint es damit wirklich ernst. Nur eben ohne milliardenschwere Förderprogramme für die Biomedizin und ohne fette Gewinne für die Pharmaindustrie. Und die Versprechen der PM? Die sind nicht nur maßlos übertrieben, sondern auch oft wissenschaftlich nicht haltbar.

Weniger Krankheit, substanzielle Verlängerung des Lebensalters bei Verbesserung seiner Qualität, und das Evidenz-basiert, mit massiven Effekten, ohne präzisere oder persönlichere Medizin, und das hier und jetzt? Wie soll das denn gehen?

Zunächst einmal sind da die sogenannten behandlungsvermeidbaren Todesfälle. Diese Todesfälle wären durch eine rechtzeitige und wirksame medizinische Versorgung vermeidbar, z. B. die Früherkennung und Behandlung von Krebs, die Versorgung bei Herzinfarkten oder Schlaganfällen, die Behandlung chronischer Erkrankungen (z. B. Diabetes), und zwar alles nach existierendem, medizinischem Schulwissen. Ganz ohne PM, KI, oder anderen Ansätzen, welche bisher nur auf dem Papier existieren.

Der Narr möchte niemand zu nahetreten, aber Raucher sind für ihn als Proponenten von PM disqualifiziert. Rauchen ist nämlich neben unzureichender Bewegung und ungesunder Ernährung ein wesentlicher Risikofaktor für schwere chronische Erkrankungen wie Herz-Kreislauferkrankungen, Atemwegserkrankungen oder Krebs. Jedes Jahr sterben deutschlandweit schätzungsweise 143.000 Menschen an den Folgen des Rauchens - weltweit sind es über 7,6 Millionen Menschen. Damit ist knapp jeder siebte Todesfall oder gut 15 Prozent aller Todesfälle auf direkte Folgen des Rauchens zurückzuführen, weitere zwei Prozent entfallen auf die Folgen von Passivrauchen (Quellen, Zitate und weiterführende Literatur wie immer unter https://dirnagl.com/lj). Minimale Reduktionen in der Zahl der Raucher hätten deutliche Wirkungen.

Noch ein Beispiel: Bluthochdruck ist der Hauptrisikofaktor für Schlaganfall, und die Todesursache von mindestens 10 Millionen Menschen weltweit. Bluthochdruck ist leicht zu erkennen, und effektiv zu behandeln. Mit einfachsten Instrumenten und günstigen Medikamenten ist da eine Menge an Lebenserwartung bei länger anhaltender Gesundheit zu holen.

Damit wären wir bei der Struktur von Gesundheitssystemen. Wieviel wird für gesundheitliches Bildung, Screening, Verfügbarkeit von und Zugang zu Behandlungen, etc. getan? Wieviel Potential in der Verbesserung dieser Strukturen steckt, zeigen einfache Vergleiche von Gesundheitssystemen in verschiedenen Ländern. Also der Vergleich von Gesundheitsausgaben pro Kopf am Bruttoinlandsprodukt (BPI), durchschnittlicher Lebenserwartung, Morbiditätsraten, sowie dem QUALY -basierten Effizienzindex. Da tun sich extreme Kluften auf. Südkorea, Spanien, Japan, haben eine hohe Lebenserwartung, relativ gute Gesundheit im Alter und vergleichsweise geringe Systemkosten. Das ist ganz anders in Ländern wie Deutschland und den USA. Das Gesundheitssystem der USA schluckt 5 Billionen US $ entsprechend 17% BPI, dabei ist die Lebenserwartung 10 Jahre geringer als z.B. Südkorea, das nur ein Viertel pro Bürger ausgibt.

Aber man braucht gar nicht weit zu reisen, um ein immenses Potential für die Verbesserung der Bevölkerungsgesundheit zu finden. Bereits ein Blick auf die Lebenserwartungsunterschiede in den deutschen Bundesländern sollte einen aufhorchen lassen. Z.B. leben Männer in Baden-Württemberg im Schnitt 4 Jahre länger als solche in Sachsen, und Männer im Berliner Bezirk Steglitz-Zehlendorf durchschnittlich 3 Jahre länger (und mit mehr Krankheit) als im 10 km entfernten Berlin Lichtenberg.

Es geht, wie könnte es anders sein, in den USA, dem Land der Precision Medicine Initiative, noch extremer. Von einer zur anderen Endhaltestelle der Red Line Subway in Chicago sinkt die durchschnittliche Lebenserwartung um 30 Jahre. Wäre es nicht sinnvoll hier prioritär anzusetzen, bevor man Millionen von Steuergeldern auf die versprochenen Segnungen der Präzisionsmedizin setzt? Der Narr hält es also für zynisch, die Versprechungen von PM und KI nicht der Wirklichkeit gegenüberzustellen.

Es geht dabei natürlich nicht nur um die oben genannten Unterschiede im Gesundheitssystem, sondern ganz prinzipiell um die sozioökonomischen Strukturen – ein Euphemismus für die einfache Tatsache, dass es viele Menschen gibt, die nicht genug verdienen, um gesund zu leben zu können. Wie steht es also um Gesundheitsverhalten (Rauchen, Alkohol, Bewegung), Bildung, Zugang zu medizinischer Versorgung (der trotz astronomischer Ausgaben im Gesundheitssystem offensichtlich nicht richtig funktioniert), Einkommen, Arbeits- und Wohnbedingungen und so weiter. Rudolf Virchow formulierte es 1848 bereits so: „Medizin ist eine soziale Wissenschaft, und Politik ist weiter nichts als Medizin im Großen."

Aber warum kommt - außer dem Narren- niemand auf die naheliegende Idee, immens teure, in der Zukunft liegende, nur Wenige erreichende und vermutlich wenig ausgeprägte Effekte einer „Präzisierung" oder gar „Personalisierung" der medizinischen Möglichkeiten den Maßnahmen gegenüberzustellen, welche unmittelbar umsetzbar, kausal, und hochgradig effektiv wären?

Weil es allesamt Maßnahmen wären, welche keine Forschungsgelder ausschwitzen würden, und für die pharmazeutische Industrie komplett uninteressant sind. Wir sind doch Wissenschaftler und Ärzte, mit sozioökonomischen Faktoren und der Struktur des Gesundheitswesens haben wir doch nichts zu tun, sollen sich doch bitte andere drum kümmern.

Aber es geht uns was an. Präzisionsmedizin fokussiert nicht ohne Grund auf seltene Erkrankungen. Oft sind diese monogenetisch – und Gentherapie ist die hohe Messe von Präzision und Personalisierung. Viele Länder bieten spezielle Anreize für die Entwicklung von sog. Orphan Drugs, bei denen sich die Entwicklung für Pharmaunternehmen normalerweise wirtschaftlich nicht lohnt. Dazu zählen Steuervergünstigungen, Zuschüsse für Forschung, verkürzte Zulassungsverfahren, und Marktexklusivität. Und es dürfen extreme Preise pro Patienten verlangt werden. Ein paar Hunderttausend Euro und Jahr sind da eher die Regel als die Ausnahme.

PM sucht hochspezialisierte Lösungen für kleine Patientengruppen, während bevölkerungsweite Strategien (z. B. Impfprogramme, Lebensstilprävention) kosteneffizienter sind und mehr Menschen erreichen. PM fragmentiert damit die Gesundheitsversorgung und steuert die Ressourcen in teure Tests und hochpreisige Nischenmedikamente. Klar ist, auf welche Einkommensgruppen das zielt.

Dazu kommt der bisher sehr begrenzte medizinische Nutzen. Obwohl genetische Tests schon lange existieren, und viele Biomarker identifizieren, haben sich nur wenige mit Biomarkern personalisierte Therapien als tatsächlich wirksam erwiesen. Die meisten Krankheiten, insbesondere komplexe wie Diabetes oder Alzheimer, sind nicht monogenetisch bedingt, sondern sind multigenetisch und multifaktoriell – Umwelt- und Lebensstilfaktoren spielen eine große Rolle.

Das ganze Gerede um die durch mehr Präzision zu erwartenden Segnungen für die Bevölkerungsgesundheit lässt einen glatt übersehen, dass die großen Durchbrüche in der medizinischen Versorgung auch in jüngster Zeit nicht durch mehr Präzision erreicht wurden. Ganz im Gegenteil. Man denke nur an die GLP-1 Agonisten – mittlerweile

nehmen mehr als 10 % der Amerikaner einen, Tendenz steigend -, oder die mRNA-Technologien, welche zuletzt in der Impfstoffentwicklung bahnbrechend waren und vermutlich in die Krebstherapie einziehen werden.

Und außerdem: Erinnern Sie sich an die Cytochrom-P450-Enzyme (z. B. CYP2D6, CYP2C19, CYP3A4)? Die sind zentral für den Metabolismus zahlreicher Medikamente. Genetische Varianten dieser Enzyme führen zu unterschiedlichen Metabolisierungstypen – und ermöglichen eine Therapieanpassung basierend auf dem genetischen Profil. Hochrelevant bei Antidepressiva, Clopidogrel, Tamoxifen und vielen anderen Wirkstoffen. Das hat der Narr schon im Medizinstudium gelernt. Und das ist eine ganze Weile her. Spielt die Testung auf CYP-Polymorphismen mithilfe eines simplen pharmakogenetischen Tests bei der Medikamentenverschreibung eine Rolle? Absolut nicht – dieses evidenzbasierte, ausgesprochen effektive und für viele Verordnungen relevante Paradebeispiel der personalisierten Medizin kommt nur in den seltensten Fällen zum Einsatz.

Ist der Narr damit vollends zum Luddisten mutiert, der in seiner Suada die aktuellesten Triumphe der PM unterschlägt, wie z.B. die CAR-T-Therapie? Werden da doch dem Patienten eigene T-Zellen entnommen, genetisch so verändert, dass sie gezielt seine eigenen Tumorzellen erkennen und zerstören können, und dann wieder zurückgegeben. Das ist die doch die Essenz von Präzisionsmedizin: die richtige Therapie für den richtigen Patienten zur richtigen Zeit – basierend auf biologischen Merkmalen, nicht pauschal.

Nun ist die CAR-T-Therapie aber auch in ihren Problemen typisch für andere PM-Ansätze: Sie kann schwere Nebenwirkungen verursachen, ist technisch aufwendig und nicht überall verfügbar, sowie extrem teuer und damit kaum skalierbar. Ihre Wirksamkeit ist begrenzt auf bestimmte Krebsarten, sie kann zu Immunschwäche führen und Langzeitfolgen sind noch unklar.

CAR-T-Therapien sind ein Meilenstein der immunologischen Forschung und ermöglichen einigen Patienten Heilung in zuvor ausweglosen Situationen. Das ist großartig und verdient weitere Forschungsförderung sowie Anstrengungen, diese Therapie viel mehr Patienten zugänglich zu machen. Sie sind jedoch keineswegs ein Beispiel dafür, wie Präzisionsmedizin die Bevölkerungsgesundheit – und das schon gar nicht global – revolutionieren könnte.

Der Hinweis, dass Präzision oder Personalisierung etwas ganz Großartiges sei, ist letztlich nichts weiter als die wenig originelle Erkenntnis, dass nahezu jedem Krankheitsgeschehen eine extrem komplexe Pathophysiologie zugrunde liegt, und dabei auch große interindividuelle Unterschiede existieren. Das ist keine Neuigkeit. Ja, bei manchen Krankheiten haben wir mittlerweile ein gewisses Verständnis entwickelt – aber jeder weitere Tag im Labor oder in der Klinik zeigt uns, dass alles noch komplizierter ist, als wir dachten.

Und als wäre das nicht genug, kommen zu diesen intrinsischen pathophysiologischen Prozessen auch noch (epi)genetische, soziale und Umweltfaktoren hinzu – oft als Auslöser, Treiber oder Modulatoren. Diese sind nicht nur in der tierexperimentellen Grundlagenforschung schwer zu greifen, sondern auch klinisch schwer explorierbar, noch dazu interagieren sie in unterschiedlichen Individuen auf interindividuell unterschiedliche, hoch-komplexe Weise. Dazu gehören Lebensumstände, sogenannter "Lifestyle", früher durchgemachte Erkrankungen, Umweltbelastungen, und obendrein die mehr oder weniger gelungene ärztliche Vor-Behandlung und laufende Medikationen bei unterschiedlicher Therapie-Adhärenz – um nur einige zu nennen.

Präzisionsmedizin klingt elegant, förderwürdig, zukunftsweisend. Würde man sie wenigstens als das verkaufen, was sie im besten Fall ist – ein Positivmarketing für die wenig

glamouröse, aber ehrlichere Einsicht: *Es ist verdammt komplex, jeder „Fall" ist anders, wir brauchen mehr Forschung* –, wäre das zwar nicht besonders hilfreich, aber hinnehmbar.

Doch stattdessen dominiert eine evidenzbefreite, medial und politisch befeuerte Erzählung: Präzisionsmedizin sei „besser", halte uns gesünder, verlängere das Leben signifikant, und sei dabei auch noch kosteneffizienter als die vermeintlich grobe und überholte „unpräzise" Medizin. Dieses Narrativ ist nicht nur naiv, sondern eine interessengeleitete, wohlfeile und massive Übertreibung.

Zynisch wird es dort, wo diese Verheißung die Tatsache überdeckt – oder bewusst ignoriert –, dass es längst effektivere, einfachere und evidenzbasierte Wege gäbe, Gesundheit zu fördern und Lebenszeit zu verlängern: etwa durch Armutsbekämpfung, eine gerechtere Verteilung von Gesundheitsressourcen, oder den flächendeckenden Zugang zu guter medizinischer Grundversorgung.

Der Narr fühlt sich unweigerlich an die Geschichte von Pinocchio erinnert. Mit jeder überzogenen Erwartung, jedem voreiligen Presseartikel über den „Durchbruch gegen Alzheimer", jedem Techno-Utopismus rund um KI oder Präzisionsmedizin – wächst die Nase der modernen Medizin. Die Verführer am Wegesrand heißen heute nicht Fuchs und Katze, sondern Forschungsförderer, Big Pharma, Medien und Politik.

Die große Prüfung, die bei Collodi im Bauch des Wals stattfindet, steht der Medizin allerdings noch bevor. Es ist die Prüfung der Evidenz ihrer Behauptungen – und damit letztlich auch die der Effizienz unseres biomedizinischen Forschungssystems.

Ist Trump Laborjournal-Leser?

LJ 9/2025

Bald wird sich der Narr zur Ruhe setzen – sein akademisches Haltbarkeitsdatum läuft mit Ende dieses Sommersemesters ab. Außerdem hält er es mit „Darwin's Bulldog" Thomas Huxley: „A man in science past 60 does more harm than good." Als er begann, sich aufs Laubsägen und Sudoku-Lösen einzustimmen, erschütterte ihn eine ungeheuerliche Nachricht aus dem Oval Office.

Am 23. Mai 2025 unterzeichnete US-Präsident Donald Trump ein Dekret mit dem verheißungsvollen Titel *Restoring the Gold Standard of Science*. Wissenschaft soll damit unter seiner Regentschaft wieder zu ihrem goldenen Standard zurückfinden – den sie, wie er darin behauptet, ganz besonders unter Joe Biden, verloren habe. Und mit ihr das Vertrauen der Bevölkerung.

In dieser Executive Order – und im kurz darauf veröffentlichten Memo des „US Office for Science and Technology Policy" (OSTP) – diktiert er, wie staatlich geförderte Wissenschaft künftig auszusehen habe: Reproduzierbar, transparent, Fehler- und

Unsicherheiten kommunizierend, kollaborativ und interdisziplinär, kritisch gegenüber den eigenen Annahmen, unvoreingenommen von Fachkollegen begutachtet, offen für negative Ergebnisse und frei von Interessenkonflikten.

Da rieb sich der Narr die Augen. Hat Trump etwa über all die Jahre am Pool von Mar-a-Lago das *Laborjournal* gelesen? Ist er gar Fan der Kolumne des Wissenschaftsnarren? Will er tatsächlich umsetzen, was dieser die ganze Zeit gefordert hat? Bricht jetzt, von der (noch) führenden Wissenschaftsnation aus, ein Zeitalter an, in dem sich Forscher wieder auf das Wesentliche konzentrieren: Qualität, Nachhaltigkeit, Relevanz, robuste Ergebnisse statt simpler aber ungeeigneter Leistungsbewertungs-Metriken?

Doch halt – gleich musste sich der Narr daran denken, dass Trump doch gerade massiv in die Forschungsfreiheit eingreift: Themen wie Gender Studies, kritische Gesellschaftsanalyse oder grundlagenorientierte Wissenschaft ohne unmittelbaren Nutzenbezug fliegen aus der Förderung. Er praktizierte eine radikale ideologische Säuberung der Wissenschaft. Unliebsamen Universitäten und Forschenden entzieht er massiv Mittel, manches davon wird in Richtung konservativer Think Tanks und regierungstreuer Projekte umgeleitet. Eine erbarmunglose Disziplinierung jener Institutionen, die er als oppositionell ausgemacht hat. Ein umfassender Angriff auf die „Eliten".

Auch die Leitungsposten in Behörden wie National Institutes of Health (NIH), National Science Foundation (NSF), Food and Drug Administration (FDA) oder Environmental Protection Agency (EPA) werden neu besetzt – nicht mit wissenschaftlicher Expertise, sondern mit Wissenschaftsverweigerern und Verschwörungstheoretikern. Missliebige Mitarbeiter werden in Scharen auf die Straße gesetzt. Loyalität ersetzt Qualifikation. Forschungsergebnisse, etwa zu Klima oder Gesundheit, werden systematisch delegitimiert, ganze Bereiche wie Bevölkerungsgesundheit und Umweltschutz damit dereguliert.

Trump's Vizepräsident formuliert das Ziel öffentlich so: *"We have to honestly and aggressively attack the universities in this country. Professors are the enemy."* Der Staat ruft zum offenen Kulturkampf gegen die Wissenschaft als Institution.

Besonders perfide daran ist, dass diese autoritäre Attacke im Gewand der Reform und unter Berufung auf deren Standards auftritt.Was also auf den ersten Blick wie ein Manifest für bessere Forschung wirkt, 1:1 kopiert aus dem Laborjournal, entpuppt sich bei näherer Betrachtung als strategische Machtverschiebung – eine gezielte Instrumentalisierung wissenschaftlicher Prinzipien und selektiv ausgewählter Evidenz zur Durchsetzung einer politischen Agenda. Im Namen von Reproduzierbarkeit, Transparenz und Integrität verordnet das Weiße Haus Kriterien für staatlich geförderte Wissenschaft – definiert diese nicht nur, sondern, und das ist die Krux, entscheidet, wie sie angewendet werden sollen.

Die Executive Order verlangt nämlich, dass fortan Bundesbehörden wissenschaftliche Evidenz nur berücksichtigen dürfen, wenn sie diesem Kanon vermeintlich objektiver Qualitätsmerkmale genügt. Auf den ersten Blick ist nichts daran auszusetzen – spiegeln diese Forderungen doch zentrale Anliegen einer internationalen Reformbewegung von vielen Akteuren im Wissenschaftssystem wider, darunter viele Forschende, Fördergeber, und akademische Institutionen. Doch die Definitions- und Anwendungshoheit liegt künftig nicht bei unabhängigen, wissenschaftlichen Gremien, sondern bei politisch ernannten Behördenleitungen. Ein Ministerium kann also selektiv entscheiden, ob und wann es den „Goldstandard" anlegt und wie dieser interpretiert werden soll – ein ideales Instrument, um missliebige Studien auszuschließen und genehme Ergebnisse aufzuwerten.

Politische Einflussnahme auf Wissenschaft war selten so subtil – meist kam sie eher mit dem Vorschlaghammer. Wir erinnern uns an Philipp Lenards „Deutsche Physik“ im Dritten Reich, die moderne, "jüdische" Theorien wie die Relativitätstheorie aus den Universitäten fegte. Oder der sowjetische Lysenkoismus: Pflanzenzüchtung nach ideologischer Doktrin statt nach Genetik – mit katastrophalen Folgen für Millionen.

Aber bereits in den 1960ern berief sich die Tabakindustrie auf „Sound Science“ - um Zweifel an der gesundheitsschädlichen Wirkung des Rauchens zu säen und wissenschaftliche Regulierung zu verzögern. Öl- und Gaskonzerne perfektionierten später das Geschäft mit dem Zweifel: Wo Fakten unbequem waren, forderten sie „mehr Daten“, „bessere Modelle“, „größere Transparenz“. Diese Techniken wurden von Naomi Oreskes und Erik M. Conway in ihrem Klassiker *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming* (2010) eindrucksvoll offengelegt und analysiert. Die darin beschriebenen Grundprinzipien der Instrumentalisierung wissenschaftlicher Normen – insbesondere des von R. Merton formulierten organisierten Skeptizismus – sind heute aktueller denn je.

Fast forward: In der Türkei ließ Erdoğan nach dem gescheiterten Putsch 2016 reihenweise Wissenschaftler entlassen und ganze Unis dichtmachen – unter dem Vorwand der nationalen Sicherheit. Orbán setzte 2019 die Ungarische Akademie der Wissenschaften faktisch außer Kraft. In Polen wurde Forschungsförderung auf „national relevante Themen“ verengt – mit anderen Worten: Bitte nur noch Forschung, die der Regierung passt.

In Brasilien verlangte Präsident Bolsonaro während seiner Amtszeit (2019–2022) immer neue „Audits“ für Satellitendaten zur Abholzung des Regenwalds – nicht um die Daten zu verbessern, sondern um ihre Aussagekraft systematisch zu untergraben.

Und die USA? Trump’s HONEST Act (2017) seiner ersten Amtszeit war der Auftakt: Nur noch Studien mit komplett offenen Patientendaten sollten in der Umweltpolitik berücksichtigt werden – ein juristischer Trick, um unbequeme epidemiologische Forschung einfach auszusortieren. Denn die hochwertigen Studien mit der besten Evidenz durften die individuellen Patientendaten aus Datenschutzgründen gar nicht preisgeben – und müssen daher alle ignoriert werden.

Was ist der gemeinsame Nenner all dieser Fälle? Reformrhetorik und angebliche wissenschaftliche Standards werden missbraucht, um Forschung zu zensieren, Narrative zu kontrollieren und evidenzbasierte Politik zu unterlaufen.

Ziel ist es, wissenschaftliche Erkenntnisse zu unterdrücken, die die politischen Agenden und wirtschaftlichen Interessen – insbesondere im Bereich Gesundheit -, Umwelt- und Klimapolitik – entgegenstehen. Und bei Trump kommt noch ein zusätzliches Motiv dazu, er verfolgt eine anti-elitäre, populistische Strategie, bei der Wissenschaft und Forschung gezielt als Teil eines vermeintlich abgehobenen, liberalen Establishments dargestellt werden. Er hat die Methode neu verpackt – das Muster ist aber alt.

Trumps Executive Order wird als Lehrbuchbeispiel für „Science-Washing“ bzw. „Epistemic Capture“ in die Geschichte eingehen: Politische Akteure inszenieren sich als Vorkämpfer strenger Standards, verbiegen diese aber strategisch. Transparenz wird zur Waffe, Reproduzierbarkeit zum Vorwand. Die Taktik verfängt, weil die Wissenschaft hausgemachte Schwächen hat – fehlender Datenzugang, nicht reproduzierbare Befunde, selektive Veröffentlichung, Interessenkonflikte. Populistische Politiker nutzen dies als als willkommenes Einfallstor: „Wenn ihr eure Hausaufgaben nicht macht, erledigen wir das eben – nach unseren Regeln.“

Nicht einzelne Ergebnisse werden zensiert, sondern die Regeln der Gültigkeit umgeschrieben. Wer festlegt, welche Daten offen genug oder welche Modelle reproduzierbar sind, legt auch fest, was als Wissen zählt. Demokratische Verfahren basieren auf einer klaren Trennung: Politik entscheidet über Werte und Prioritäten, Wissenschaft über Evidenz und Methoden. Wo diese Trennung verwischt, oder gar aufgehoben wird, werden Machtfragen als Qualitätsdebatten verkleidet und verhandelt.

Auch bei uns gibt es mittlerweile beunruhigende Anzeichen dafür, dass sich – häufig populistische – politische Akteure der Sprache wissenschaftlicher Qualität bedienen, um sie für ihre Zwecke zu instrumentalisieren. In den Landtagen Ostdeutschlands sind solche Praktiken bereits voll im Einsatz. Noch zeigen sich unsere Institutionen resilient. Doch diese Resilienz ist keineswegs garantiert. Sie entsteht durch glaubwürdige Selbstregulierung: durch transparente Verfahren, durch unabhängige Standards und durch ein Verständnis von Qualität, das sich nicht an Renommee, sondern an Prinzipien misst.

Wie können wir die akademische Freiheit gegen böswillige politische Einflussnahme verteidigen und sie widerstandsfähig machen gegenüber dem Missbrauch von Reformrhetorik als Instrument autoritärer Kontrolle? Ein Weg besteht darin, unabhängige Begutachtungs- und Förderstrukturen zu stärken, klare Grenzen zwischen Politik und Wissenschaft durchzusetzen und transparente, überprüfbare Verfahren für Datenaustausch und Publikation sicherzustellen.

Ebenso wichtig ist jedoch, dass die Wissenschaft ihre eigenen Reformversprechen einlöst: offene Datensätze, reproduzierbare Protokolle, Replikationsstudien und eine transparente Offenlegung von Interessenkonflikten. Solange Reputation und Publikationsmetriken mehr zählen als Qualität und Verantwortlichkeit, bleibt die Wissenschaft verwundbar. Und dann drohen auch bei uns „närrische Zustände“ in der Wissenschaft.

Dieser Beitrag basiert auf einem Gastbeitrag für die Frankfurter Allgemeine Zeitung, den der Narr zusammen mit seinem Kollegen Prof. Dr. Daniel Strech verfasst hat.

Der Narr tritt ab.

LJ 10/2025

Sie lesen die 77. und letzte Folge der *Einsichten des Wissenschaftsnarren*. Ist im Wissenschaftssystem wirklich nichts mehr zu entlarven? Ist dem Narren die Schärfe abhandengekommen – oder gar die Lust am Spott? Droht er nun, wie der Joker im Batman Comic, in den Wahnsinn zu kippen und das System lieber in Flammen aufgehen zu sehen, als es mühselig weiter zu reformieren – wie es das Bild zur Linken andeutet? Mitnichten!

Die spätmittelalterlichen Rituale und Hierarchien der Akademie, der ungebrochene Drang der Forscher, durch selektive Daten und fragwürdige Statistik spektakuläre Geschichten zu erzählen, ihre Obsession, in „Glamjournals“ wie *Nature*, *Cell* oder *Science* zu publizieren, um anschließend den Impact Factor auf drei Nachkommastellen genau im Lebenslauf zu notieren, die unermüdliche Beschwörung leerer Mantras wie Exzellenz, Translation, Präzision und Personalisierung, deren Gehalt längst verdampft ist, sowie die kleinen – und immer öfter auch kleinen wie großen – Betrügereien: All das würde, leider, Stoff für diese Kolumne bis ins nächste Jahrhundert liefern.

Vielmehr geht der Narr in Rente! Und hat nicht die Absicht, als „Seniorprofessor“ mit Rollator durch die Gänge „seines“ Instituts zu schieben und dem Nachwuchs zu predigen, wie früher angeblich alles besser war. Auch verspürt er keinerlei Lust, noch einmal irgendwo Frühstücksdirektor zu werden oder den Vorsitz einer „wichtigen“ Kommission zu übernehmen – selbst wenn eine Institution so verwegen wäre, ihn darum zu bitten. Und nein: Auch die Gastprofessur in Singapur oder Dubai lockt ihn nicht. Für viele meiner Kollegen kaum vorstellbar – aber wahr: Das Leben endet nicht an den Mauern der Akademia.

Vierzig Jahre Wissenschaft: 400 Publikationen, ein paar Bücher, ein altersbedingt imposanter Hirsch-Faktor von 124. Dazu geschätzt kumulativ ein volles Jahr ununterbrochene Kommissionssitzungen, Hunderte von Drittmittelanträgen bis hin zum Exzellenzcluster, mehr als tausend Gutachten für Artikel und Förderanträge. All das hat am Narren Spuren hinterlassen. Aus dem Saulus ist auf halbem Wege ein Paulus geworden: Die Selbstzweifel am eigenen Tun und am System wuchsen mit den Jahren, während der Enthusiasmus für die Wissenschaft ungebrochen blieb. Doch parallel dazu nahm der Skeptizismus gegenüber ihrem Betrieb – und gegenüber so manchem ihrer Ergebnisse – stetig zu.

Nachdem der Saulus in ihm an der Charité erst eine Abteilung für Experimentelle Neurologie und dann das Centrum für Schlaganfallforschung gegründet hatte, blieb es dem Paulus in ihm überlassen, mit dem QUEST Center for Responsible Research am Berlin Institute of Health das notwendige Korrektiv zu schaffen – eine Einrichtung, die helfen sollte, jenes zu reparieren, was das System, also seine Kollegen und nicht zuletzt er selbst zuvor angerichtet hatten.

In diesen 40 Jahren hat sich das Wissenschaftssystem weltweit massiv verändert. In den 1980er-Jahren war die Biomedizin weniger kompetitiv. Forschung war – im besten Sinne – langsamer, der Druck geringer, es blieb mehr Zeit für Wissenschaft und weniger für Bürokratie. Weil es zudem weniger Forscher gab, die entsprechend weniger Output produzierten, weil die methodische Komplexität noch überschaubar war und das biomedizinische Wissen in ein paar Regalmetern Platz fand, wurden auch weniger Anträge und Paper geschrieben. Ein *circulus virtuosus*, der Zeit freisetzte – für Forschung statt für Schreibtischroutine.

Der eben erst erfundene Impact Factor war damals ein Werkzeug der Bibliothekare, um ihre Zeitschriftenbestände zu sortieren – Herr Hirsch, Namensgeber des späteren Hirsch-Faktor, besuchte noch die Volksschule. Doch die Schattenseite war offensichtlich: Der persönliche Aufstieg (oder Ausstieg) im System hing weniger von wissenschaftlicher Leistung ab als von der Macht des Mentors und des Instituts- oder Klinikdirektors. Zusammen mit dem Fortschritt in Methodik und Wissen – und dem damit einhergehenden gewaltigen Anstieg gesellschaftlicher Investitionen in die Forschung – nährte dies den verständlichen Wunsch nach Objektivierung, nach einer Metrifizierung und damit auch Automatisierung in der Bewertung wissenschaftlicher Leistung.

Fast forward ins Jahr 2025: Publish or perish! Wir leben nun in einer totalen Reputationsökonomie, in der Karrieren weniger über wissenschaftlichen und damit gesellschaftlichen Impact, sondern über Surrogate wie Zitierhäufigkeit des Publikationsorgans (sprich: Impact Factor), Zahl der Publikationen und eingeworbene Drittmittel gesteuert werden. Diese „Objektivierung" der Leistungsbewertung hat nicht nur Transparenz und Verteilungsgerechtigkeit versprochen, sondern vor allem eine Vielzahl unbeabsichtigter Nebenwirkungen hervorgebracht: Hyperkompetition, Matthäus-Prinzip, Mainstream-Forschung, selektives Publizieren, fragwürdige Praktiken bis hin zum Betrug.

Hier liegt die Grundursache des Effizienzproblems unseres Wissenschaftssystems – und der Narr hat darüber, für manche vielleicht allzu oft, auf diesen Seiten lamentiert. Denn die meisten der heute beklagten Auswüchse und Ineffizienzen von Wissenschaft und Forschungsförderung haben genau dort ihren Ursprung. Medizinisch gesprochen: Wer das System mit Appellen und Modifikationen im Kodex guter wissenschaftlicher Praxis, mit Präregistrierungen, dem Publizieren von Nullresultaten und dergleichen verbessern will, betreibt symptomatische Therapie. Notwendig, gewiss – wie in der Medizin. Doch dauerhafte Genesung verspricht nur eine kausale Therapie: die Reform von Karriere- und Bewertungssystem.

Und tatsächlich hat sich einiges getan. Ganze Länder – allen voran die Niederlande – haben sich aufgemacht, in einer konzertierten Aktion von Wissenschaftlern, Universitäten, Forschungsförderern und Politik das Problem an der Wurzel zu packen. Dort gilt nun das Motto „Every talent counts": unsinnige Metriken sind verbannt, inhaltliche Bewertungen rücken in den Vordergrund, Replikationsstudien werden vom staatlichen Forschungsförderer NWO (vergleichbar der DFG) unterstützt – und das ist längst nicht alles. Auch die Schweiz, allen voran ihr wichtigster Forschungsförderer, der Schweizerische Nationalfonds (SNF), schreitet mit Siebenmeilenstiefeln voran: mit modifizierten Förderlotterien, narrativen Lebensläufen und mehr. Die Europäische Union hat die

Coalition for Reforming Research Assessment (COARA) ins Leben gerufen, in der sich inzwischen über 700 Forschungsorganisationen, Förderinstitutionen, Bewertungsstellen, Fachgesellschaften und ihre Verbände auf gemeinsame Leitprinzipien für die Reform der Forschungsbewertung verständigt haben. Und ja – sogar die DFG hat unterschrieben!

In vielen Ländern haben sich Wissenschaftlerinnen und Wissenschaftler in Reproducibility Networks zusammengeschlossen – in Deutschland etwa im German Reproducibility Network (GRN). Die Einstein Stiftung Berlin verleiht inzwischen jährlich so etwas wie den „Nobelpreis" für die Verbesserung der Forschungsqualität: den mit 500.000 € dotierten *Einstein Foundation Award for Promoting Quality in Research*. Meta-Researcher beforschen und reflektieren das Wissenschaftssystem selbst und führen Interventionsstudien durch. Sie prüfen Machbarkeit, Akzeptanz und Wirksamkeit, aber auch mögliche Nebenwirkungen von Veränderungen, die meist plausibel klingen, bislang aber noch nicht evidenzbasiert abgesichert sind. Die VolkswagenStiftung – für den Narren Deutschlands progressivster Förderer in den Lebenswissenschaften – erwies sich erneut als Pionierin, indem sie erprobte, ob Lotterieelemente in die Forschungsförderung integriert werden können. Und das Ministerium *formerly known as BMBF* fördert inzwischen bereits zum zweiten Mal präklinische Replikationsstudien – und legt zudem großen Wert darauf, dass Patientinnen und Patienten in Planung und Durchführung klinischer Studien einbezogen werden, wenn staatliche Fördergelder beantragt werden.

Wo Licht ist, ist auch Schatten. Viele dieser Maßnahmen haben reinen Pilotcharakter – sie wurden mit Bordmitteln oder, wenn es gut lief, mit öffentlichen oder Stiftungsgeldern ausprobiert. Selbst wenn die Evaluation ergab, dass eine Initiative – etwa zur Qualitäts- oder Transparenzsteigerung – erfolgreich war, endet sie oft dort. Denn um solche Maßnahmen „in die Linie" zu bringen, also flächendeckend in einer Institution zu verankern, fehlen die Ressourcen. Drittmittel gibt es nur für Pilotprojekte – danach sind die Einrichtungen auf sich allein gestellt.

Manches hat reinen Alibicharakter – pures Kästchen-Abhaken auf Formblättern. Journals fragen heute routinemäßig, ob die Forscherinnen und Forscher die ARRIVE-Reporting-Guidelines berücksichtigt haben. Wer dort „ja" ankreuzt, rückt ein Feld vor – Nachfragen von Editoren oder Gutachtern gibt es keine.

Auch die Notwendigkeit von Fallzahlabschätzungen vor Studienbeginn ist inzwischen im präklinischen Bereich anerkannt; selbst die Genehmigungsbehörden für Tierversuche verlangen sie mittlerweile. Abgefrühstückt wird das dann von den Forschern mit dem berüchtigten *Sample Size Samba*: Man postuliert eine völlig unrealistisch hohe Effektstärke, und schon reichen n=8, um die geforderten Typ-I/II-Fehlerniveaus zu erfüllen. Weitere Fragen? Fehlanzeige – außer vom Narren, der sich damit als Reviewer nicht selten unbeliebt machte und wohl manchen Leser dieser Kolumne durch wiederholtes Aufspießen solcher statistischen Unarten gelangweilt hat.

Narrative Lebensläufe können zweifellos ein sinnvolles Komplement zur nackten Auflistung von Publikationen, Drittmitteln und Preisen sein: Man beschreibt mit eigenen Worten, worauf man stolz ist und was die wissenschaftliche Community tatsächlich vorangebracht hat. Aber auch hier lässt sich tricksen – und die Schaumschläger stehen längst in der Pole Position.

All dies zeigt: Manipulationen an einem komplexen, gigantischen System, in dem höchst unterschiedliche Interessen regieren, sind alles andere als trivial. Gerade deshalb werden solche Beispiele von den Lordsiegelbewahrern des Status quo – meist jenen an der Spitze des Systems, ausgestattet mit Lebenszeitprofessuren und Sitzen in allen

entscheidenden Kommissionen – gerne zitiert. Nicht etwa als Aufruf, konzertiert, also in gemeinsamer Aktion von Wissenschaftlern, Institutionen, Förderern und Verlagen, systematisch, behutsam und forschungsbegleitet Reformen zu wagen. Sondern einzig, um jede Veränderung abzuwürgen.

Und das macht den Narren traurig. Denn während die jüngeren Wissenschaftler die Notwendigkeit von Reformen durchaus sehen, fehlt ihnen schlicht die Sicherheit und die Position, sich dafür einzusetzen. Ausgerechnet die Arrivierten dagegen wähnen: Weil sie es geschafft haben, kann das System so schlecht doch nicht sein. Dabei wären gerade sie in der Lage, Veränderungen anzustoßen – sie verfügen über Macht, Netzwerke, und sie sind faktisch unantastbar. Doch um dorthin zu gelangen, mussten sie ein engmaschiges „Filtersystem" durchlaufen, das nur wenige durchlässt – und zwar jene mit der stärksten und angepassten Sozialisation.

Wie geht's weiter? Um es mit Karl Valentin zu sagen: „Prognosen sind schwierig, besonders wenn sie die Zukunft betreffen." Fest steht: Die Weltlage ist düster, die Finanzlage der öffentlichen Hand wie auch privater Förderer alles andere als rosig. Soll man in Zeiten von Krise und Unsicherheit also besser nicht am Wissenschaftssystem schrauben – einem System, das ja irgendwie funktioniert und beeindruckende Fortschritte wie CAR-T-Zelltherapien oder mRNA-Technologien hervorgebracht hat? Der Narr widerspricht: Gerade jetzt! Denn letzlich geht es um Effizienz – und die ist in Zeiten knapper Ressourcen wichtiger denn je.

Damit tritt ein Fehler zutage, den der Narr – und manche seiner Mitstreiter – über Jahre möglicherweise gemacht haben: Wir haben für „verantwortungsvollere" Wissenschaft gestritten. Doch dieser moralische Imperativ hat vielleicht so manchen verschreckt, der darin einen Vorwurf wähnte: dass er oder sie bisher *unverantwortlich* geforscht habe. Natürlich dreht sich die Kritik am Wissenschaftsbetrieb auch um Verantwortlichkeit – gegenüber Kollegen, gegenüber der Gesellschaft. Doch diese Verantwortung ist untrennbar verknüpft mit dem Umgang mit Ressourcen. Die Gesellschaft finanziert Wissenschaftler, stellt ihnen den Elfenbeinturm bereit, beteiligt sich als Patient an klinischen Studien – und zahlt in vielerlei Währung, nicht zuletzt im Leid und Tod zahlloser Versuchstiere.

Medizindoktoranden, die Tierversuche machen, um den „Dr. med." zu bekommen. Ärzte, die publizieren, weil die Habilitation eine Mindestzahl an Veröffentlichungen verlangt. Experimentelle Studien, die glanzvolle Papers erzeugen, deren Ergebnisse aber weder reproduzierbar noch generalisierbar sind. Klinische Studien, deren Resultate niemals veröffentlicht werden. Und all die anderen Unarten des Wissenschaftsbetriebes, die der Narr in 76 Folgen aufgespießt hat – sie alle verschwenden Ressourcen, die an anderer Stelle dringend gebraucht würden: für Forschung, die wissenschaftlich und gesellschaftlich relevant und qualitativ hochwertig ist.

Zum Schluss bleibt dem Narren, Dank zu sagen – insbesondere an Ralf Neumann und das *Laborjournal*, die ihm all die Jahre die Bühne überlassen haben, und an Sie, liebe Leserinnen und Leser, darunter einige Fans und selbsterklärte Narr-Ultras, für eine Menge Zuspruch ebenso wie für Kritik. Und zu guter Letzt noch ein Hinweis: Falls Sie noch nicht genug haben und der Weg ins Archiv des *Laborjournals* zu weit ist – es gibt eine digitale, einmalige *Collector's Edition*: alle Folgen, unredigiert („unplugged"), ergänzt durch eigens zu jeder Ausgabe erstellte Abbildungen. Frei verfügbar, kostenlos und unter einer Creative-Commons-Lizenz zur Weiterverwendung zum Download hier: http://dirnagl.com/narr

Index: