Einsichten eines Wissenschaftsnarren (1) NEU!

Wie originell sind eigentlich Ihre Hypothesen?

Je tiefer Forschung ins Unbekannte vorstößt, desto mehr nicht-replizierbare Resultate muss sie erzeugen. Und in vielen Fällen ist das gut so!

Schon mal darüber nachgedacht, wie hoch 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 an Hypothesen, die von anderen bestätigt wurden – oder die am Ende ein tatsächlich wirksames Medikament postuliert hatten.

Leider werden heutzutage die wenigsten Resultate unabhängig überprüft (davon gleich mehr!). Und selbst seit Jahren etablierte Therapien werden irgendwann später als unwirksam oder gar schädlich aus dem Verkehr gezogen. Man kann sich so einer „Erfolgs"-Quote, wenn überhaupt, also lediglich annähern – was ich im Folgenden tun will.

Vielleicht wundern Sie sich, warum ich Ihnen diese scheinbar esoterische Frage stelle. Weil die Antwort auf die Frage, wie hoch in etwa der Prozentsatz ist, mit dem sich Hypothesen als richtig erweisen, weitreichende Konsequenzen für die Bewertung von Forschungsergebnissen hätte – sowohl der eigenen, als auch derjenigen von anderen. Und weil diese Frage einen überraschenden, aber direkten Bezug zur gegenwärtigen Krise der biomedizinischen Wissenschaften hat. Es geht nämlich ein Gespenst um!

Momentan verdichtet sich die Gewissheit, dass die meisten Studienergebnisse in Biomedizin und Psychologie sich nicht bestätigen lassen. Nach einer aktuellen *Nature*-Umfrage glauben mittlerweile gar 90 Prozent der Wissenschaftler, dass wir uns mitten in einer „Reproduzierbarkeitskrise" befinden. Auch ich bin davon 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 zehn bis zwanzig Prozent der Studien, die Wissenschaftler von Amgen und Bayer nachgekocht hatten, konnten sie die meist hochrangig publizierten Ergebnisse aus akademischen Laboren replizieren.

Nicht ganz zu unrecht wurden die Autoren damals 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. In der Akademia war deshalb manchem schnell klar, warum die Replikationen scheitern mussten: Postdocs, die nicht gut genug für eine akademische Karriere sind, wandern in die Industrie ab – klar, dass die dort dann auch nichts bringen. Nicht-Replikation als Folge von Kompetenzmangel also.

Mittlerweile konnten allerdings eine Reihe von sehr gut geplanten, systematischen Initiativen der Akademia (beispielsweise in der Psychologie oder der Krebsforschung) ebenfalls nur einen enttäuschend geringen Teil der Resultate aus ausgewählten, hochrangig publizierten Arbeiten nachvollziehen. Sogar in der Zeitung kann man seitdem lesen, dass die Wissenschaft in einer Krise sei. Und der Kommentar eines hochrangigen Mitarbeiters der DFG hierzu: „Klar, 80 Prozent der Befunde sind nicht reproduzierbar, aber die restlichen 20 Prozent wurden durch uns gefördert!"

„Ein Gespenst geht um. Wir befinden uns in einer Reproduzierbarkeitskrise."

Wenn es nur so einfach wäre. Denn wie viele Ergebnisse müssten eigentlich reproduzierbar sein, damit wir zufrieden wären? 80, 90, oder gar 100 Prozent? 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.

Ioannidis' erstes Argument: eine niedrige Qualität in Studiendesign, Analyse und Berichterstattung – und dadurch Verzerrung (Bias) der Ergebnisse. Die Liste dieser Probleme ist lang und schließt unter anderem fehlende Verblindung und Randomisierung, selektive Datenauswahl, sowie Nicht-Publikation von negativen Resultaten ein. Sein zweites Argument: zu niedrige statistische Power durch zu geringe 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-Studien 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 aus dem Tiermodell in wirksame Therapien beim Menschen.

Allerdings, wie viele Resultate sollten denn überhaupt reproduzierbar sein? Wären in einem wissenschaftlichen Utopia, in dem Bias komplett beseitigt und statistische Power bei 100 Prozent läge, tatsächlich alle Studien reproduzierbar? Ganz sicher nicht! Schreitet Erkenntnis nicht durch zumindest teilweise Widerlegung von bisher Anerkanntem fort? Der Berliner Wissenschaftshistoriker Hans-Jörg Rheinberger spricht gar von der „differentiellen Replikation von Experimentalsystemen" als wesentliches Moment des

Fortschrittes in der Wissenschaft. Danach wird im Laufe der Zeit jedes Ergebnis nur „differentiell“, also teilweise, replizierbar sein. Zumal auch Wissenschaftler sich irren.

Fragen wir konkreter: Wie steht es 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 stelle, mit einem Prozentsatz weit über Fünfzig. Man ist ja schließlich ein guter Wissenschaftler.

Aber wäre es nicht tragisch, wenn sich ein hoher Prozentsatz unserer Hypothesen als richtig herausstellte? Schließlich läge dann der Verdacht nahe, dass man überwiegend triviale Hypothesen untersuchte! Dass vorher schon so viel bekannt war, dass der nächste kleine Erkenntnisschritt mit großer Sicherheit vorhergesagt werden konnte. Wie langweilig!

Zum Glück sind wir mit unseren Hypothesen offenbar weit weniger treffsicher. Wo dies formal untersucht wurde, lag die Quote eher bei zehn Prozent. Dies hätte weitreichende Konsequenzen. Es würde zum Beispiel bedeuten, dass beim gängigen Signifikanzniveau von fünf Prozent ($p \leq 0{,}05$) und einer statistischen Power, wie sie in klinischen Studien gefordert (achtzig Prozent), aber in den meisten präklinisch-experimentellen Studien nicht annähernd erreicht wird, mehr als ein Drittel aller statistisch signifikanten Befunde falsch positiv sind!

„Wir sind doch keine unfehlbaren Langweiler!“

Die meisten Experimentatoren jedoch wiegen sich in trügerischer Sicherheit, da sie glauben in nur maximal fünf Prozent 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 A-priori-Wahrscheinlichkeit der Hypothese. Nun kennen wir die A-priori-Wahrscheinlichkeit unserer Hypothese aber gar nicht – und sie ist ganz sicher deutlich unter 100 Prozent, 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 achtzig Prozent durchgeführt werden. Deshalb liegt die Falsch positiv-Rate vermutlich deutlich über fünfzig Prozent. 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 klar, dass explorative Forschung – sofern sie nicht Banalitäten untersucht, keinen Bias aufweist und mit ausreichender Power ausgestattet ist – schon ihrem Wesen nach nicht-replizierbare Befunde erzeugen *muss*.

Ließe sich dann die Treffsicherheit nicht durch Wiederholung des „positiven“ Experiments deutlich verbessern? Leider nein – es sei denn, die Fallzahl wird deutlich erhöht. Auch diese unangenehme Wahrheit ist den Wenigsten bekannt: Die Wahrscheinlichkeit, einen auf Fünf-Prozent-Niveau signifikanten Befund (also etwa $p=0{,}049$), der auf einer richtigen Hypothese beruht, mit dem selben experimentellen Set-up samt Fallzahl auf dem gleichen statistischen Signifikanzniveau zu reproduzieren, liegt bei fünfzig Prozent. 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!

„Veröffentlichen Sie Ergebnisse, die Ihre Hypothese nicht bestätigt haben?“

Könnte es also paradoxerweise so sein, dass insbesondere dort, wo Experimente ins wahrlich Unbekannte vordringen oder vielleicht sogar bisheriges Wissen in Frage stellen, eine niedrige Replikationsrate ein Zeichen für besonders „heiße“ und spannende Wissenschaft wäre? Dass 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 Experimente mit zu geringen Fallzahlen, mit mangelnder Verblindung und Randomisierung, mit selektiver Datenauswahl, mit fehlerhafter Statistik oder mit fehlender Publikation von neutralen oder negativen Ergebnissen den Garaus machen.

Aus dem oben Gesagten lassen sich ein paar einfache Schlussfolgerungen ziehen, deren Umsetzung recht dramatische Wirkung hätte:

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, auf 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 müssten. Diese wäre sogar dort am höchsten, wo Wissenschaft richtig *cutting edge* ist. Das heißt aber im Nebenschluss und ganz zwingend, 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 die Konfirmation gleich mitliefern!

Der neben der notwendigen Qualitätsverbesserung 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, und so weiter. Sie darf nicht als langweilige Fleißarbeit abgetan werden. Qualitativ hochwertige Konfirmation – und nur diese wird uns aus der Reproduktionskrise herausführen – ist nicht nur methodisch aufwendig und Ressourcen-intensiv, sondern auch eine intellektuelle Herausforderung.

(Unter https://dirnagl.com/lj *findet sich eine Auswahl einschlägiger Literatur zum Thema.)*

Foto: BIH/Thomas Rafalzyk

Ulrich Dirnagl leitet die Experimentelle Neurologie an der Berliner Charité und ist Gründungsdirektor des Center for Transforming Biomedical Research am Berlin Institute of Health. Für seine Kolumne schlüpft er in die Rolle eines „Wissenschaftsnarren“ – um mit Lust und Laune dem Forschungsbetrieb so manche Nase zu drehen.