Einsichten eines Wissenschaftsnarren (7)

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

Viele erheben den p-Wert zum Non-plusultra, um zwischen falschen und richtigen Hypothesen zu unterscheiden. Oftmals hilft er hierbei aber nur wenig – oder gar nicht.

In der letzten Ausgabe nahm sich der Narr die Wissenschaftskultur in der Physik vor, und fand gar einiges, was wir Lebenswissenschaftler von dort abschauen könnten. Überhaupt ist die Physik – insbesondere die Teilchenphysik – eine Fundgrube von Lehrstücken. Zwei besonders aktuelle 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 und auf eine 730 Kilometer lange Reise geschickt. Dann registrierte deren Ankunft ein Detektor, der durch Tausende von Metern Stein in die Dolomiten gesprengt wurde. Und siehe da: Die Neutrinos kamen schneller an, als Photonen dies über dieselbe Strecke geschafft hätten!

Auch dem Nichtphysiker wird sofort klar, was mit dem Ergebnis dieses sogenannten OPERA-Experiments alles auf dem Spiel steht (Spezielle Relativitätstheorie) – oder dann vielleicht möglich würde (beispielsweise 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 $p < 3x10^{-7}$!) auf 6 Sigma. Außerdem wiederholten sie das Experiment mehrmals. Trotzdem, kein Zweifel, die Neutrinos machten sich nichts aus der Lichtgeschwindigkeit, und das Signifikanzniveau lag bei unerreichten 6,2 Sigma.

Also wurde flugs die Weltpresse informiert, und ein Paper geschrieben. Allerdings hatten die Autoren trotz rekordverdächtigem p-Wert weiterhin 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 bisher nicht über das Kino-Stadium hinausgekommen sind. Genauso wie wir wissen, dass Photonen immer noch den absoluten Geschwindigkeitsrekord halten. In den Wochen nach dem Medienrummel nahmen sich die Physiker ihren Versuchsaufbau also nochmals genau vor. Und fanden, dass das zur Entfernungsmessung genutzte GPS nicht korrekt synchronisiert war. Außerdem, man glaubt es kaum: Ein Kabel war locker!

»Hat der Versuchsaufbau einen systematischen Fehler, nutzt ein niedriger p-Wert gar nichts.«

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 extrem niedriger p-Wert etwas noch eine Replikation am selben Versuchsaufbau.

Wir Lebenswissenschafter können daraus natürlich das Gleiche lernen. Ein p-Wert kann einem bei der Beantwortung der Frage, ob unsere Hypothese richtig ist – etwa, dass ein Medikamentenkandidat wirkt, oder ähnliches –, recht wenig und oftmals sogar gar nichts nützen. Und: Eine Replikation eines Experiments im selben Labor ist sowieso von sehr bedingtem Wert (siehe auch, was der Wissenschaftsnarr hierzu in *Laborjournal* 4/2017 auf den Seiten 24 bis 25 schrieb).

Diese ganze Sache ist unter anderem deshalb so aktuell, weil gerade ein *All-Star*-Team aus Statistik, Epidemiologie und Psychologie in *Nature Human Behavior* (doi: 10.1038/s41562-017-0189-z) einen aufsehenerregenden Vorschlag gemacht hat: Nämlich das von Ronald A. Fisher in den 1920er Jahren eingeführte Signifikanzniveau um eine Größenordnung abzusenken. Von dem von uns fast wie eine Naturkonstante behandelten Wert $p < 0,05$ auf $p < 0,005$! Die Autoren haben natürlich recht, dass damit die Rate der falsch positiven Resultate, unter der wir alle zu leiden haben, deutlich reduziert werden könnte. Und damit auch die Anzahl publizierter Studien, denn an der 0,005-Hürde würden viele Veröffentlichungen scheitern.

Ich halte den Vorschlag, auch mit Blick auf die OPERA-Schlappe, dennoch für einen Fehler. Den Experten, die diese Absenkung vorschlagen, ist klar, was ein p-Wert ist – und was nicht. So wissen sie, dass nicht nur *alpha*, also der Fehler 1. Art, für die Frage wichtig ist, ob ein Ergebnis falsch positiv ist. Dies hängt nämlich auch von *beta*, also dem Fehler 2. Art, beziehungsweise der *Power* ab – genauso wie von der Wahrscheinlichkeit, mit welcher die Hypothese richtig ist. Die Autoren verwechseln also den p-Wert nicht mit dem positiv prädiktiven Wert, wie so viele von uns. Indem sie aber die Aufmerksamkeit in dieser Weise auf den p-Wert – ja, konkret auf 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 eben nur den *richtigen* Wert annehmen. Wer den Artikel indes aufmerksam liest, wird alles Richtige dazu erfahren. In der Berichterstattung zu diesem Vorschlag ging es aber einzig und allein um die neue Schwelle – und damit um die „Rettung des p-Werts".

Nicht zuletzt deshalb hier gleich noch ein für uns Lebenswissenschaftler lehrreiches Beispiel aus der Physik. Bei OPERA ging es um eine sehr unwahrscheinliche Hypothese – und am

Ende war das Resultat trotz exorbitant niedrigem p-Wert falsch positiv. Der Grund: Im experimentellen Aufbau steckte ein systematischer Fehler. Beim LIGO-Experiment, mit dem man vor kurzem endlich die lange gesuchten Gravitationswellen nachweisen konnte, war es umgekehrt: Hier glaubte man das Ergebnis schon vorher zu kennen. Die von Einstein 1919 vorausgesagten Gravitationswellen musste es einfach geben, denn alle Voraussagen der Allgemeinen Relativitätstheorie hatten sich bisher experimentell belegen lassen. Zudem gab es kein ernsthaftes Argument, warum Gravitationswellen nicht existieren sollten.

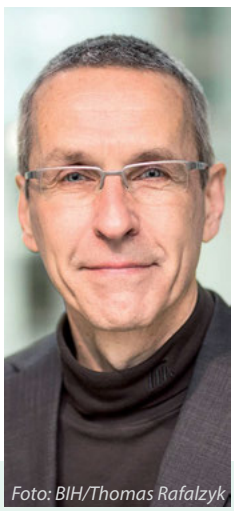

Foto: BIH/Thomas Rafalzyk

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.

Das Problem hierbei war nur, dass man praktisch seit 1919 nonstop 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, haben es zu Recht auf die mangelnde Sensitivität ihres Experiments geschoben – und an deren Verbesserung gearbeitet.

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

Die nicht-signifikanten p-Werte – das heißt, 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 von wissenschaftlichen Resultaten und noch mehr bei deren Publikation stärker auf die Effektstärken, die Varianzen und vor allem die Güte des experimentellen Designs zu stützen – als auf p-Werte und statistische Signifikanz.

»Die nicht-signifikanten p-Werte bedeuteten nicht, dass es das untersuchte Phänomen nicht gibt.«

Weiterführende (und hier teils ohne Angabe zitierte) Literatur findet sich unter http://dirnagl.com/lj.