Abb. 1: Gerade in den Lebenswissenschaften ist eine Reproduzierbarkeit von publizierten Forschungsergebnissen unabdingbar.

Analyse

Zeitverträge schaden der wissenschaftlichen Qualität

Publish or perish: Verwaltungstechnische Anreize führen heute in Deutschland dazu, dass bei wissenschaftlichen Veröffentlichungen die Quantität und ein hoher Impact-Faktor im Mittelpunkt des Interesses stehen, auch in den biomedizinischen Wissenschaften. Dadurch gerät jedoch ein zentrales Kriterium für die Qualität einer Veröffentlichung völlig außer Blick: die Reproduzierbarkeit eines Forschungsergebnisses. Gerade in den Lebenswissenschaften sind die Reproduzierbarkeitsquoten sehr niedrig, wie Studien belegen.

Von Victor I. Spoormaker

DAS ZIEL JEDES wissenschaftlichen Systems sollte es sein, hohe Qualität zu liefern. Was aber ist eigentlich wissenschaftliche Qualität? Diese Frage wird auch im Rahmen der Auswahl der besten Köpfe für Professuren oder Direktorenposten heiß diskutiert. Einen klaren Konsens gibt es jedoch nicht – unterschiedliche Bewertungsgrundlagen dafür umso mehr: Manager und Wissenschaftler orientieren sich mehr und mehr an Rankings, basierend auf Zitationen und High-Impact-Veröffentlichungen, und zwar sowohl für die Bewertung einzelner Wissenschaftler als auch ganzer Institutionen. Ohne zu tief in diese Diskussion hineinzusteigen, steht außer Frage, dass keines dieser Kriterien etwas darüber aussagt, ob die veröffentlichten Ergebnisse korrekt sind oder nicht – und dies sollte ja das Hauptkriterium wissenschaftlicher Qualität sein. Das einzige Merkmal, das in diesem Zusammenhang aussagekräftig ist, ist die Reproduzierbarkeit eines Ergebnisses.

Im Sinne einer Signal-zu-Rausch-Überlegung sind reproduzierbare Ergebnisse das Signal, nicht-reproduzierbare Ergebnisse das Rauschen (Abb. 2). Investiert man nun in die Quantität und nicht in die Qualität von Veröffentlichungen, kann statt des gewünschten Signals auch nur die Menge des Rauschens vermehrt werden. Dies ist im derzeitigen Anreizsystem mit der starken Betonung auf „publish or perish“ leider der Fall – es ist egal, ob die gefundenen Forschungsergebnisse real sind oder nicht. Dies hat wiederum zur Folge, dass das wahre Signal zu oft untergeht, und öffentliche Gelder werden an vermeintlich interessante, jedoch kurzlebige Forschungsprojekte verschwendet.

Nicht-reproduzierbare Ergebnisse können zwar auch in böswilliger Absicht verursacht werden, sie sind jedoch in den meisten Fällen einfach eine Folge methodischer und statistischer Fehler: zu viel Flexibilität im Studienentwurf und in Datenanalysen, inkorrekte statistische Tests oder Schlussfolgerungen, a posteriori Festlegung der Hypothesen und post hoc „Data-Fishing“. Besonders die biomedizinischen Wissenschaften scheinen von einer niedrigen Reproduzierbarkeit betroffen zu sein. In der Zeitschrift „Nature“ haben Forscher von Bayer über interne Studien berichtet, in denen sie versucht haben, veröffentlichte präklinische (Tier-)Experimente zu wiederholen. Die Reproduzierbarkeitsquote lag bei sehr mageren 20 bis 25 Prozent. Ein ähnliches internes Reproduzierbarkeitsprojekt der Firma Amgen erreichte eine noch schlechtere Quote von 11 Prozent.

Leider beschränken sich solche Ergebnisse nicht auf die Krebsforschung oder medizinische Forschung im Allgemeinen. Andere Fachgebiete haben ähnliche Probleme, verfügen jedoch nicht über konkrete Daten zur tatsächlichen Reproduzierbarkeit. So hat eine Analyse in den Neurowissenschaften gezeigt, dass es eine klare Diskrepanz gibt zwischen den veröffentlichten Effektstärken und den Stichprobengrößen der individuellen Studien, die diese Effekte dokumentiert haben. Anders gesagt: Die zitierten Studienpopulationen waren viel zu klein, um Effekte jener Größe überhaupt entdecken zu können. Studien sollten theoretisch darauf abzielen, mit der geplanten Stichprobengröße eine Wahrscheinlichkeit von 70 bis 80 Prozent für einen wahren Effekt bestimmter Größe zu erlangen; dies bezeichnet man als die statistische Power des Tests. In den genannten Veröffentlichungen lag diese Zahl jedoch nur bei 8 bis 31 Prozent, mit Relevanz sowohl für Human- als auch für Tierstudien. Psychologische Experimente haben ähnliche Probleme mit einer Verzerrung der veröffentlichten Studien in Richtung positiver Ergebnisse (Abb. 3), was zusammen mit zu niedriger statistischer Power auch ein Problem mit falsch positiven Ergebnissen indiziert. Deswegen wird auch in diesem Fachgebiet langsam mehr Nachdruck darauf gelegt, Standard-Experimente zu reproduzieren, u. a. in einem Konsortium mit dem Namen „Many Labs Replication Project“.

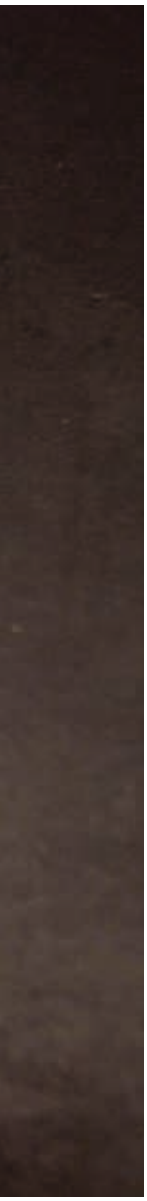


Abb. 2: Ein gutes (A) und schlechtes Signal-Rausch-Verhältnis (B): Wenn Experimente mit zu niedriger Stichprobengröße, mangelhafter Statistik und/oder post hoc „Data-Fishing“ durchgeführt werden, ist unklar, ob der berichtete Effekt tatsächlich real ist oder nur ein statistisches Rauschen wie in (B). Fehlentscheidungen sind dann wahrscheinlicher.

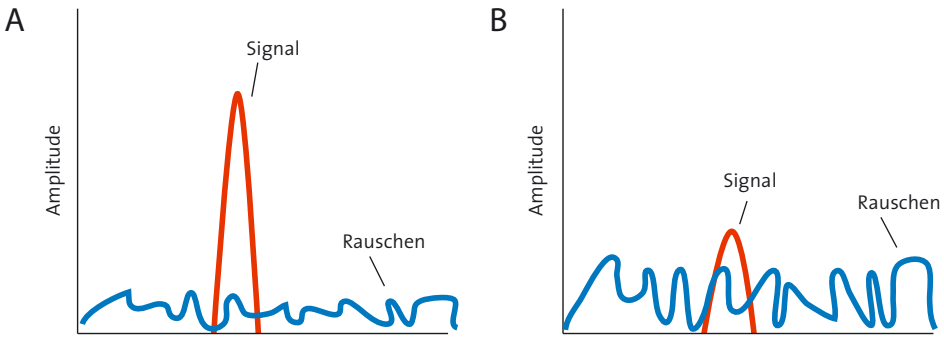


ABB.: PICTURE ALLIANCE / DPA; V. I. SPOORMAKER; GRAFIK: TAUSENDBLAUWERK.DE

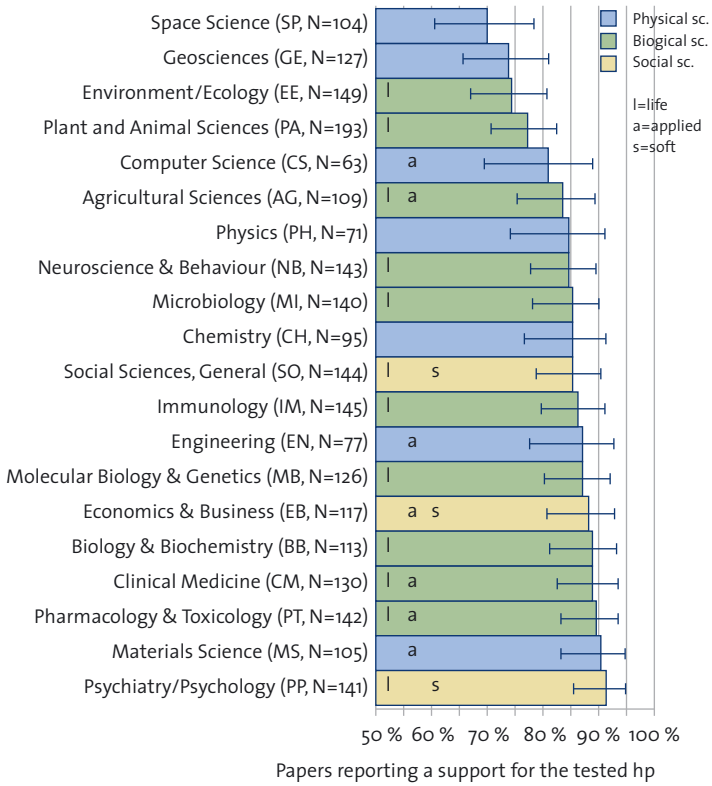
Abb. 3: Positive Veröffentlichungstendenzen in unterschiedlichen Fachgebieten.

Derart niedrige Reproduzierbarkeitsquoten, verursacht von methodischen und statistischen Fehlern, weisen auf eine armselige Praxis in den Lebenswissenschaften hin. Die Effekte sind in den Naturwissenschaften vermutlich kleiner, da die untersuchten Systeme und angewendeten Read-outs viel robuster und weniger rausch-behaftet sind und die Stichprobengröße einfacher erhöht werden kann. Für die Geisteswissenschaften dürfte eine Replizierbarkeit ohnehin weniger relevant sein. Aber 30 bis 40 Prozent der Gelder der Deutschen Forschungsgemeinschaft fließen in die biomedizinischen Wissenschaften (Stand 2013) und 65 Prozent der Postdocs in den USA arbeiten in diesem Feld. Die genannte Problematik trifft also nicht nur ein kleines Forschungsgebiet. Wahrscheinlicher ist, dass die Lebenswissenschaften mit ihrem rasanten Wachstum und der ständigen technischen und methodischen Innovation einfach empfindlicher sind für den Druck auf das System („publish or perish").

Es wurde argumentiert, dass niedrige Reproduzierbarkeitsquoten u. a. durch einen Mangel an Expertise und Erfahrung mit den verwendeten Methoden und Techniken verursacht werden könnten. Neuartige Methoden benötigen in manchen Fällen mehrere Jahre, um optimiert zu werden. Allerdings sind methodische und statistische Fehler in den Lebenswissenschaften sehr verbreitet und auch vorhanden in Studien, die einfache oder bereits lange erprobte Techniken anwenden. So konnte z. B. gezeigt werden, dass ungefähr die Hälfte der hoch veröffentlichten neurowissenschaftlichen Artikel, die eine Intervention untersuchten, falsche statistische Tests angewendet haben, obwohl dies Lehrbuchwissen sein sollte. Man kann daher feststellen, dass es in den biomedizinischen Wissenschaften bei der methodischen und statistischen Ausbildung auf der prädoktoralen Ebene noch Luft nach oben gibt. Das eigentliche Problem ist jedoch, dass die methodische und statistische Ausbildung einfach viel Zeit kostet und nicht immer im Masterstudium oder den ersten Jahren der Promotion unterzubringen ist – neben allen neuen Informationen zu Theorien, Hypothesen und Techniken.

Dass solche Trainings mittlerweile in Graduate Schools fest etabliert werden, ist eine positive Entwicklung und ein großer Schritt in die richtige Richtung. Jedoch sind Graduate Schools notwendigerweise sehr breit orientiert, da die Doktoranden aus unterschiedlichen Fachgebieten kommen. Deswegen bleibt weiterhin eine Anpassung der methodischen und statistischen Ausbildung auf den individuellen Doktoranden und das individuelle Projekt notwendig, sowohl für eine gute Ausbildung als auch für den Erfolg des Forschungsprojekts. Wie auch immer – am Ende ist und bleibt der Principal Investigator für die Qualität der Forschung im Labor zuständig. Die Hauptfrage, die sich stellt, ist: Wie viele Doktoranden kann ein Principal Investigator individuell betreuen und ausbilden?

ABB. AUS: D. FANELLI, „POSITIVE" RESULTS INCREASE DOWN THE HIERARCHY OF THE SCIENCES. PLOS ONE 2010;5(4):E10068; GRAFIK: TAUSENDBLAUWERK.DE



Das Verhältnis von Doktoranden zu Arbeitsgruppenleitern hat sich in den vergangenen Jahren nicht verbessert – es wird schlechter (Bild 4A). Die derzeit meist populäre und billige Lösung ist, einfach mehr Postdocs anzustellen (Bild 4B), so dass die Auszubildenden von einem gerade erst Ausgebildeten lernen. Aber ein Postdoc hat auch ein eigenes Projekt und kann

schon vor dem Ende des Projektes des Doktoranden wieder eine neue Stelle angetreten haben. Zudem hat man als Postdoc einfach noch nicht so viel Erfahrung und Expertise sammeln können wie ein Principal Investigator oder methodisch stark entwickelte wissenschaftliche Mitarbeiter. Da es für die letzte Gruppe an deutschen Universitäten kaum Dauerstellen oder seriöse Laufbahnperspektiven gibt, ist die gängige Praxis, dass die Hauptmasse der Forschung von Doktoranden und Postdocs, also Auszubildenden und gerade ausgebildeten Forschern, geleistet wird. Das hat logischerweise Folgen für die Qualität der Studien-Entwürfe, Messungen und Analysen. Natürlich können Doktoranden und Postdocs aufgrund ihrer geringen Erfahrung eine Studienkonzeption meist nicht selbständig durchführen: Die Ausbildungszeit sollte ja auch ein Lernen mit der Trial-and-Error-Methode ermöglichen. Wie viel Raum bleibt aber für Versuche und Irrtümer, wenn man nur zwei bis drei Jahre für ein Projekt hat und in dieser Zeit bereits ein bis zwei Veröffentlichungen vorweisen muss? Hierdurch entsteht die Ursache für den Löwenanteil der oben genannten Probleme: eine Analyse der Daten für andere Zwecke als geplant, Data-Fishing und a posteriori Festlegung der Studien-Hypothesen, die dann natürlich besser von den Daten bestätigt werden.

Die Botschaft ist eindeutig: Unser derzeitiges Wissenschaftssystem ist kein winning System. Wissenschaftliche Studien sind am Ende des Tages so gut (und replizierbar) wie die Forscher, die sie ausführen. Ein besseres zahlenmäßiges Verhältnis von Doktoranden zu Principal Investigators ist absolut notwendig. Da es aber finanziell gesehen nicht realistisch scheint, dass die Zahl der Principal Investigators stark steigen wird, sollte man in Deutschland dem Beispiel anderer Wissenschaftssysteme folgen und stärker auf die erfahrenen wissenschaftlichen Mitarbeiter setzen. Natürlich kosten erfahrene wissenschaftliche Mitarbeiter auch mehr, allerdings kann ein einzelner erfahrener Mitarbeiter die Forschungsaufgaben von zwei bis drei Auszubildenden oder gerade ausgebildeten Forschern betreuen. Nicht-reproduzierbare Ergebnisse und damit verschwendetes Geld sind in diesem Modell geringer. Eine Verschiebung der Beurteilung individueller Wissenschaftler bei Anträgen weg von Zitationen und High-Impact-Veröffentlichungen und hin zu einer Mischung dieser Elemente und einer individuellen Reproduzierbarkeitsquote (wie vom Stanforder Statistikprofessor John P. A. Ioannidis vorgeschlagen) würde bereits andere Anreize schaffen. Die Flexibilität des Systems bleibt erhalten, indem die wissenschaftlichen Mitarbeiter nicht einem einzelnen Principal Investigator zugeordnet wären, sondern einer Abteilung oder Fakultät zugehörten, für welche sie die wissenschaftlichen Kerndienstleistungen erbringen. So wie jetzt kann es jedenfalls nicht länger weitergehen. Das derzeitige Wissenschaftssystem produziert zum Großteil wissenschaftliches Rauschen und wird damit auf lange Sicht zum Milliardengrab.

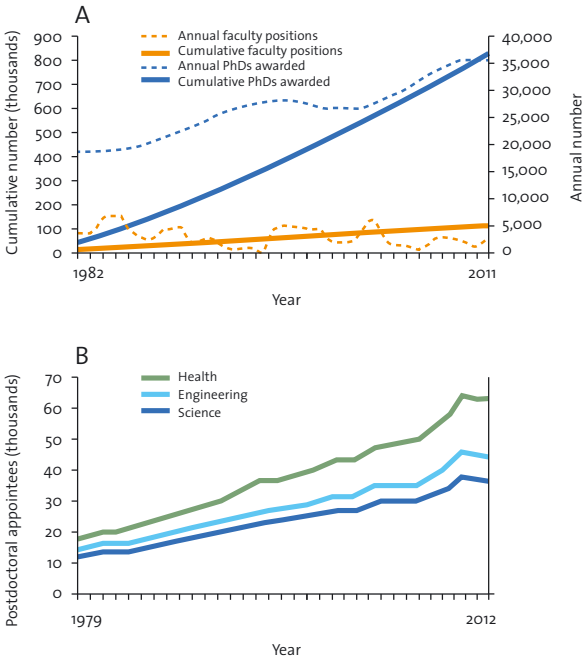


ABB. AUS: M. SCHILLEBEECKX, B. MARICQUE, C. LEWIS: THE MISSING PIECE TO CHANGING THE UNIVERSITY CULTURE. NAT BIOTECHNOL 2013;31(10):938-41; K. POWELL, THE FUTURE OF THE POSTDOC. NATURE 2015;520:144-147; MIT GENEHMIGUNG DES HERAUSGEBERS MACMILLAN PUBLISHERS; GRAFIK: TAUSENDBLAUWERK.DE

Abb. 4: Zunahme von Doktoranden (A) und Postdocs (B) in den vergangenen drei Jahrzehnten in den USA, bei weitgehend gleichbleibenden Karriereaussichten (Tenure-Stellen, A).

Literatur

M. Bissell, Reproducibility: The risks of the replication drive, in: Nature 2013, 503(7476), 333–334.

K. S. Button, J. P. A. Ioannidis, C. Mokrysz et al., Power failure: why small sample size undermines the reliability of neuroscience, in: Nature Reviews Neuroscience 2013, 14(5), 365–376.

S. Nieuwenhuis, B. U. Forstmann, E. J. Wagenmakers, Erroneous analyses of interactions in neuroscience: a problem of significance, in: Nature Neuroscience 2011, 14(9), 1105–1107.

C. G. Begley, L. M. Ellis, Drug development: Raise standards for preclinical cancer research, in: Nature 2012, 483(7391):531-3.

F. Prinz, T. Schlange, K. Asadullah, Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery 2011, 10(9), 712.

DER AUTOR

Victor I. Spoormaker, Ph. D., ist wissenschaftlicher Mitarbeiter am Max-Planck-Institut für Psychiatrie. Seit 2011 ist er Mitglied im Jungen Kolleg der Bayerischen Akademie der Wissenschaften, wo er mit dem Forschungsvorhaben „Die Verknüpfung zwischen Gehirnregionen während des ‚rapid eye movement´ (REM)-Schlafes" gefördert wird.