Einsichten eines Wissenschaftsnarren (58)

Zen und die Kunst, Forschungsqualität zu bewerten

„Forschungsqualität" ist das Hauptkriterium in der Beurteilung von Wissenschaft. Folglich entscheidet sie wesentlich über Wohl und Wehe bei Anträgen, Publikationen und Karriere. Nur, wie definiert man Forschungsqualität?

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, hochkarätig besetzte Jury jährlich eine halbe Million Euro an Individuen, Gruppen, Institutionen und auch junge Wissenschaftler vergibt, die herausragend zur Verbesserung von Qualität in der Wissenschaft beigetragen haben – nämlich des „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 jedoch keineswegs allein, auch ist das Thema nicht wirklich neu. Bereits 1830 griff etwa Charles Babbage, ein Multitalent und Erfinder des Vorläufers der modernen Computer, in seinen „Reflections on the Decline of Science in England, and on Some of Its Causes" das wissenschaftliche Establishment, die Universitäten und die britische Royal Society scharf an. Unter anderem prangerte er darin selektive Datenanalyse („Trimming"), unsaubere Statistik („Cooking") und Wissenschaftsbetrug („Forging", „Hoaxing") an. Die Qualität von Wissenschaft steht also schon recht lange auf dem Prüfstand.

Doch nicht nur Kritiker des Wissenschaftssystems interessieren sich für „Forschungsqualität". So muss Forschung von hoher Qualität sein, um überhaupt gefördert oder publiziert zu werden – wobei der Peer Review als weithin akzeptierte Qualitätskontrolle fungiert. Qualität bildet folglich zusammen mit Originalität und Exzellenz die Trias der wesentlichen Kriterien, die über Antragserfolg, Publikation oder Karriere entscheiden.

Aber wüssten Sie, „Forschungsqualität" zu definieren? Falls Sie damit Schwierigkeiten haben, sind Sie nicht allein. Erst kürzlich hat sich beispielsweise ein Gremium hochkarätiger Wissenschaftler aus den verschiedensten Disziplinen auf Einladung der Einstein Stiftung (die auch den oben erwähnten Preis vergibt) mit der Frage auseinandergesetzt, wie man Forschungsqualität definieren oder gar messen kann und ob es hierfür womöglich disziplinäre Standards gibt. In Bezug auf diese Fragen herrschte am Ende jedoch keineswegs Einigkeit.

»Qualität kann man nur erkennen, wenn man eine Definition hat, was das sein soll.«

Dabei ist die Definition von „Forschungsqualität" keine rein theoretische Angelegenheit. Schließlich ist sie *das* Hauptkriterium in der Beurteilung von Wissenschaftlern und deren Produkten. Folglich sollte man eigentlich ziemlich genau wissen, mit welchem Maßstab man da urteilt. Denn sonst wird das Urteil willkürlich beziehungsweise 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, jedoch ohne 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 tatsächlich, wenn man sie hat. Wie also könnten wir Qualität in der Forschung stattdessen definieren?

»Rankings machen die Gutachter in der Wissenschaft zwar auch. Nur auf welcher Grundlage?«

Wenn wir über Gegenstände des täglichen Lebens nachdenken, haben wir damit meist keine Schwierigkeiten. Welche Qualität hat eine Matratze? Eine Kaffeemaschine? Profis in solchen Fragen sind Organisationen wie etwa die Stiftung Warentest. Sie definieren Qualität anhand klarer Kriterien wie Güte der Verarbeitung, Haltbarkeit, heutzutage auch Nachhaltigkeit – sowie natürlich dem Nutzen für den Gebrauch (Wie schläft sich's drauf? Wie lange dauert es, bis die erste Tasse Kaffee rauskommt? ...) oder dem Preis-Leistungs-Verhältnis. Auf die-

Ulrich Dirnagl

ist experimenteller Neurologe an der Berliner Charité und ist Gründungsdirektor des QUEST Center for Responsible 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.

Foto: BIH/Thomas Rafalzyk

Sämtliche Folgen der „Einsichten eines Wissenschaftsnarren" gibt es unter
www.laborjournal.de/rubric/narr

ser Basis kann man letzlich sogar quantifizieren, verschiedene Produkte vergleichen und für diese ein Ranking erstellen.

Rankings machen die Gutachter in der Wissenschaft zwar auch. Nur auf welcher Grundlage? Wenn Qualität eines der Kriterien dabei ist, wie fassen wir sie? Und warum tun wir uns da so schwer?

»Qualität kann viele Dimensionen und Bedeutungen haben. Der einzige gemeinsame Nenner: Sie haben alle eine positive Konnotation.«

Die Definition von „Qualität", etwa bei Wikipedia, wonach Qualität die Summe beziehungsweise Güte aller Eigenschaften eines Objektes, Systems oder Prozesses sei, hilft uns angesichts ihrer Allgemeinheit nicht weiter. Robert M. Pirsig dagegen neutralisiert in seinem Buch „Zen und die Kunst, ein Motorrad zu warten", einem Klassiker der philosophischen Betrachtung des Qualitätsbegriffs, seine rationale Betrachtung des Gegenstandes am Ende wieder mit einem subjektiven, Zen-artigen „Im-Moment-Sein" – und dies wollten wir in der Forschungsbewertung ja gerade hinter uns lassen.

Nützlicher fand ich da Peter Dahler-Larsens Abhandlung „Quality – from Plato to performance". Darin listet er die vielen Dimensionen und Bedeutungen auf, die „Qualität" haben kann – und findet nur einen einzigen gemeinsamen Nenner all dieser Definitionen: dass sie nämlich alle eine positive Konnotation haben. Qualität will man haben, sie ist etwas Gutes; wenn es dagegen 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, Qualität als Exzellenz und so weiter.

»Von der Stiftung Warentest können wir lernen, dass Qualitätskriterien transparent sein müssen.«

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-Faktor für ein gutes Kriterium für die Relevanz einer Publikation? Dann haben Veröffentlichungen in Journalen wie *Nature*, *Cell* oder dem *New England Journal of Medicine* logischerweise 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 also 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, die 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 schließlich 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 hinsichtlich 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 (zum Beispiel kleine Studien von zweifelhafter Qualität oder große randomisierte kontrollierte Studien) wird die Stärke beziehungsweise die Verbindlichkeit der Empfehlungen auch noch quantifiziert. Diese sogenannte GRADE-Methodik (Grading of Recommendations, Assessment, Development and Evaluation) wird von über hundert Organisationen (einschließlich der Weltgesundheitsorganisation WHO) befürwortet und/oder verwendet, um die Qualität von Evidenz und die Stärke von Empfehlungen im Gesundheitswesen zu bewerten.

»Zunächst müsste man sich einigen, was man unter Forschungsqualität verstehen möchte.«

Wie wäre es angesichts dessen also, 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 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,

» Ethik für Mensch und Tier.

Für jede dieser Dimensionen lassen sich einfach und zwanglos Bewertungskriterien ableiten, die selbstverständlich kontextabhängig sein müssen. Also beispielsweise in Teilen anders für Grundlagenforscher und deren Produkte als für Psychologen oder klinische Forscher.

Robust sind Forschungsergebnisse, wenn sie von hoher interner und externer Validität sind. Wenn sie also zum Beispiel in der präklinischen Forschung randomisiert und verblindet durchgeführt wurden, dazu mit hoher methodischer Kompetenz und am besten in verschiedenen Spezies oder experimentellen Settings.

Wenn die Studien- und Analysenprotokolle überdies noch präregistriert werden, sind sowohl das Vorgehen wie auch die Auswertung *transparent* festgelegt. Praktiken wie selektive Datenanalyse („Cherry picking"), Hypothesenbildung nach Erhalt der Ergebnisse („HARKING") oder statistische Tricksereien („p-Hacking") werden damit erschwert.

Nützlich ist Forschung für die Wissenschaft, wenn ihre Daten zur Nachnutzung veröffentlicht werden; und *nützlich* ist sie für die Gesellschaft (in unserem Fall die Patienten), wenn die klinischen Forscher sie in die Planung ihrer Studien involvieren.

»Forschung niedriger Qualität kann keine relevanten Ergebnisse erzielen. Da hilft es auch nichts, wenn sie originell ist.«

Ethisch ist Forschung für Patienten dann, wenn Nutzen und möglicher Schaden für sie in klinischen Studien im Gleichgewicht stehen. Und sie ist es für Tiere, wenn deren Leid minimiert wird oder gleich ganz auf ihre Verwendung verzichtet wird (3R-Regel).

Und schon hätten wir ein Set von Kriterien, mit denen die Forschungsqualität einer Studie oder eines Antrages (in Bezug auf die Vorarbeiten und das Arbeitsprogramm) bewertbar wird. Genauso wie das Oeuvre einer Forscherin oder eines Forschers. Und zwar mit transparenten und überprüfbaren Kriterien, die auf die gesamte Bewerberschaft oder die 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 über „Unklare/fragwürdige Qualität" bis hin zu „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. Denn Forschung niedriger Qualität kann keine relevanten Ergebnisse erzielen. Da hilft es auch nichts, wenn sie originell ist oder wenn die Antragsteller schon mal in ganz tollen Journalen publiziert haben.

»Um Qualität zu bewerten, müsste man den Antrag oder das Paper gelesen haben.«

Überhaupt sind Originalität und Relevanz Kategorien, die sich viel stärker einer Operationalisierung entziehen und somit wesentlich subjektiver bewertet werden müssten als Forschungsqualität. Hinzu kommt, dass Relevanz eine zeitliche Dimension hat, die ihre Bewertung nahezu unmöglich macht. Was heute als irrelevant erscheint, kann schon morgen hochrelevant werden, die Wissenschaftsgeschichte liefert hierfür unzählige Beispiele. Das Gleiche gilt natürlich auch umgekehrt für gehypte Projekte, für die sich nach wenigen Jahren niemand mehr interessiert. Die Bewertung von Originalität ist hingegen sogar noch subjektiver – und diskriminiert letzten Endes Ergebnisse, die rein konfirmativ sind. Nicht zuletzt deshalb bekommen wir viel zu wenige davon.

Die von mir hier angebotene Definition von Forschungsqualität dagegen operationalisiert und objektiviert diese, wodurch sie in eine konkrete und bis zu einem gewissen Grad sogar messbare Form gebracht wird. Es braucht dafür nicht mehr als ein einfaches Formular. Einen gravierenden Nachteil hat mein Vorschlag aber: Um Qualität mithilfe des Formulars zu bewerten, müsste man den Antrag oder das Paper gelesen haben. Querlesen mit Blick auf die Affiliation (Harvard? Stanford?) oder auf die Namen der Journale hilft nicht weiter. Weshalb er wohl leider niemals umgesetzt wird.

Weiterführende Literatur und Links finden sich wie immer unter: http://dirnagl.com/lj