



Irreproducibility of published bioscience research: Diagnosis, pathogenesis and therapy



Jeffrey S. Flier was named the 21st Dean of the Faculty of Medicine at Harvard University on July 11, 2007, at which point he became the Caroline Shields Walker Professor of Medicine at Harvard Medical School. Flier is an endocrinologist and an authority on the molecular causes of obesity and diabetes. Prior to becoming dean, he served from 2002 to 2007 as Harvard Medical School Faculty Dean for Academic Programs and Chief Academic Officer for Beth Israel Deaconess Medical Center (BIDMC), a Harvard teaching affiliate. Born in New York City, Dr. Flier received a BS from City College of New York in 1968, and an MD from Mount Sinai School of Medicine in 1972, graduating with the Elster Award for Highest Academic Standing. Following residency training in internal medicine at Mount Sinai Hospital from 1972 to

1974, Flier moved to the National Institutes of Health as a Clinical Associate. In 1978, he joined the Faculty of Medicine at Harvard Medical School, serving as Chief of the Diabetes Unit at Beth Israel Hospital until 1990, when he was named chief of the hospital's Endocrine Division. Dr. Flier is one of the country's leading investigators in the areas of obesity and diabetes and has authored over 200 scholarly papers and reviews. His research has produced major insights into the molecular

mechanism of insulin action, the molecular mechanisms of insulin resistance in human disease, and the molecular pathophysiology of obesity. An elected member of the Institute of Medicine and a fellow of the American Academy of Arts and Sciences, Flier's honors include the Eli Lilly Award of the American Diabetes Association, the Berson Lecture of the American Physiological Society, and an Honorary Doctorate from the University of Athens. He was the 2003 recipient of the Edwin B. Astwood Lecture Award from the Endocrine Society, and in 2005 he received the Banting Medal from the American Diabetes Association, its highest scientific honor. In 2008, Dr. Flier was awarded the Albert Renold Award from the American Diabetes Association for outstanding achievements in the training of diabetes research scientists and the facilitation of diabetes research. In 2010, Flier was awarded an Honorary Doctor of Science Degree from the University of Edinburgh. In 2011, Dr. Flier received the Rolf Luft Award for Metabolic Research from the Karolinska Institute in Sweden. In July of 2016, after nine years as Dean of Harvard Medical School, Dr. Flier stepped down from that position, rejoining the HMS faculty, based in the Neurobiology Department, as the George Higginson Professor of Physiology and Medicine, and Harvard University Distinguished Service Professor. Dr. Flier is married to Eleftheria Maratos-Flier, MD, who is also on the faculty of Harvard Medical School and with whom he has collaborated on research in the area of neuroendocrine control of body weight. They have two daughters, Lydia and Sarah, and live in Newton, Mass.

Jeffrey S. Flier^a

During a 40 year career as a biomedical researcher and academic leader, my primary professional goal has been to discover and disseminate new knowledge relevant to biology and health, with my own efforts focused on metabolic physiology and disease. I have done this during a period of dynamic growth of the bioscience enterprise, which has produced remarkable discoveries to illuminate our understanding of human biology and disease while creating numerous benefits for the health and welfare of society. The bioscience research ecosystem that supports this effort is large and complex. Physicians and PhD scientists in academia and industry conduct basic and clinical research, spending over 100 billion dollars yearly in the US alone [1]. The results of this research eventually appear in over one million scientific papers per year [2], in more than 5000 journals [3] of varying focus, standards, and impact. These publications are the bricks on which the edifice of scientific progress is built. The worldwide scientific community reads, discusses, assesses, and wherever possible builds upon the results reported in these papers. The overall arc of progress from this activity is evident, and there is little doubt that the future will continue to bring discoveries of profoundly important impact. But the direction of scientific progress is not exclusively forward. Much research is exploratory in nature, and tentative conclusions are both expected and beneficial. Research publications will contain errors, despite procedures designed to avoid them. Fortunately, a fundamental attribute of science is its capacity for "self-correction", through published ideas and claims being reviewed and tested by others [4]. Although scientists should and most often do seek to publish reliable results, to expect a standard of certainty before publication, and/or to excessively stigmatize or penalize claims later found honestly to be in error, would diminish progress by replacing a spirit of scientific

excitement and daring with professional fear of error. The key question, therefore, is to define an optimal balance, surely weighted in the direction of reliability, but appropriately tolerant of tentative conclusions and honest errors, while continuously seeking to reduce the latter.

But today we face claims, from a variety of sources, that published bioscience research is far less reproducible than anyone previously imagined [5–7]. If the most extreme of these claims are true, they challenge the integrity of the research enterprise, threatening the public support and funding that sustain it. Indeed, we might be required to seriously reconsider our approach to conducting and publishing research. Consequently, the bioscience research community, and those committed to its success, must take these claims seriously.

I write from a perspective that begins with forty years in the trenches of metabolic research. I have published many papers, and have played diverse roles in scientific publishing, both as peer reviewer and editor. As Dean of Harvard Medical School, I oversaw both academic appointment and promotion processes that emphasize the evaluation of published work, and the investigation of research misconduct and fraud by our faculty. I have both extolled and defended biomedical research to the general public. Research reproducibility is of paramount importance in each of these realms.

1. REPRODUCIBILITY IN DISTINCT DOMAINS OF BIOSCIENCE RESEARCH

Bioscience research covers a broad spectrum from basic science in a variety of disciplines, to translational research that links basic science

^aDepartment of Medicine and Neurobiology, Harvard Medical School, 220 Longwood Ave, Goldenson 542, Boston, MA, 02115, USA

E-mail: Jeffrey_flier@hms.harvard.edu.

Available online 21 November 2016

<http://dx.doi.org/10.1016/j.molmet.2016.11.006>

and animal research to therapeutic implications, to clinical research involving human subjects. The need for reproducibility applies to all such research, though some issues are specific to specific domains. The dominant focus of this paper is on reproducibility in basic and translational research, but I will reference, where appropriate, specific issues related to clinical research reproducibility.

2. WHAT DOES RESEARCH REPRODUCIBILITY MEAN?

To understand this problem, we must first define it. Unfortunately, the definition and criteria for considering research reproducible are far from unambiguous [8,9]. In the most restrictive and precise use of the term, experimental results are replicated when an independent group performs the same experiment, under the same conditions and using the same reagents, then finding the same results. Most published research is never replicated in this manner, and I don't believe that a healthy scientific ecosystem requires it to be. Certainly, the academic and financial incentives to routinely replicate the work of others do not exist. Rather than being replicated, published results are typically built upon and extended by subsequent research, perhaps carried out under somewhat different conditions from the original work, or with different reagents, but with results that are seen as consistent with and supportive of the earlier claims. Thus, though the first results were not formally replicated, they provided a sound foundation for subsequent work, thereby advancing our understanding.

In some fields, replication may be accomplished by reanalyzing a published data set and reaching the same conclusions, rather than generating a new data set [9].

Problems arise when new research or a new analysis of prior research is, or at least seems to be, inconsistent with previously published findings. This may have many explanations, related to differences in design or execution of either the initial or the subsequent work [10,11]. One often cannot conclude whether one of the studies is at fault, or whether they are just different, and the answer often requires additional experimentation. In cases where a new result appears inconsistent with a prior report, if the new study employed experimental conditions and reagents distinct from the first, it should not be taken as a failure to replicate it. It is possible that the experiment would have been successfully replicated had its precise conditions been employed. On the other hand, a failure to replicate under slightly different conditions may suggest that the initial finding is at least less generalizable than initially claimed.

Reproducing research employing molecules, cells, and model organisms is practically less daunting than attempting to reproduce clinical research, especially complex and expensive clinical studies involving hundreds or thousands of patients. This fact, plus the immediate importance of human studies, renders it especially important that clinical research experiments be well designed, pre-registered, have data available at publication for analysis by others, and be committed to publishing results whether positive or negative in relationship to the initial hypothesis. Of course, similar arguments can be made for optimal design of pre-clinical research as well, though there are many practical obstacles to all of these practices becoming routine.

There are at least two ends of the spectrum of reproducibility. At one end, a huge number of the one million papers published each year are minimally read or referenced [12]. Like a tree falling unobserved in the forest, whether such papers are reproducible is unknown. There are undoubtedly some gems buried within that largely unexamined and untested pool of papers, just as there may be many among the

unread and unreferenced papers whose truth would be challenged if more closely scrutinized. But either way, these papers now have limited impact on the world of science. Perhaps in the future, natural language processing and machine learning algorithms will permit the under-examined literature to be more effectively explored.

At the other end of the spectrum are papers of high impact with results of potential importance and great interest to the community. Sometimes, and thankfully quite rarely, the conclusions of such papers are rapidly shown to be false. In such cases, other scientists have been actively involved in the same research question, and at the time of publication already have data bearing on the new papers' claims. They view the success of their own research as requiring the new claims to be rapidly verified or rejected. If their work fails to replicate it, they would be motivated to quickly publish the negative results. Two prominent examples of high profile studies whose major findings were quite rapidly shown to be false, were the claims about the putative beta cell stimulatory hormone betatrophin [13–15] and STAP cells as a facile approach to creating totipotential stem cells [16–18]. In such cases, though the corrections and/or retractions may be quickly published, it is important to understand why the erroneous findings came to be believed by the scientific teams and then published in high impact journals to public acclaim. Such outcomes can result from honest errors and/or incompetence, or more nefarious causes involving research misconduct or even outright fraud.

Although high impact papers whose results are rapidly retracted garner major attention, we should perhaps be more concerned about another situation that appears to be more common: papers of potential importance where many in the community have difficulty building on or reproducing the results, but despite much informal discussion of these failures, whether at meetings or at the water cooler, papers documenting this fact do not get published.

3. HOW COMMON IS IRREPRODUCIBLE RESEARCH?

What fraction of the published bioscience research is not reproducible? The most straightforward answer is that we really don't know. The answer would require examining data from a sufficiently large and representative sample of studies where replication was attempted, with all of the caveats about what true replication means. The empirical data from which to draw such conclusions are generally not available.

Some papers addressing this topic make claims about the high rate of irreproducibility of the published literature [5]. These are theoretical examinations, based on sampling of the literature and assessing such factors as appropriateness of sample sizes, statistical deficiencies, experimental and publishing bias, and other factors, and these have led to dire conclusions about the truthfulness of the larger literature. In another approach, several articles have reported that pharmaceutical labs are unable to reproduce the core findings of a large fraction of academic papers published in high impact journals [6,7]. These pharma groups are highly motivated to verify the results of publications reporting new therapeutic targets, and they have the resources and highly trained scientists to do the work. On the other hand, despite their claims of failures to reproduce, these papers present no actual data on which to evaluate their claims [6,7]. Such concerns appear to be quite prevalent in the biopharma community, are clearly worrisome to the academic community, and very likely reflect real problems. But as for all scientific claims, their conclusions cannot be accepted as true without the opportunity to examine the underlying data. It would be very helpful to the scientific enterprise if

biopharmaceutical scientists would publish more of their findings, including those that are negative. While the narrow business case for such publications may be limited, the salutary effect on the scientific community on which the companies ultimately depend for most of their insights should justify this. The Reproducibility Project: Cancer Biology, is one effort to get a better, quantitative estimate of the reproducibility of important published work, with results to be published in the journal *eLife* [19,20].

Outside the pharmaceutical industry, many academic scientists report difficulties reproducing the work of others. Nature conducted an online survey of 1576 scientists, and the results showed that 52% of those surveyed across all fields believed there was a “crisis” of reproducibility [21]. 73% of all respondents believed that at least half of published papers can be trusted, with a somewhat lower number in biology and medicine than in physics and chemistry. In all fields, most scientists have doubts about the validity of a substantial fraction of the published literature [21]. If this survey is representative of the broader bioscience research community, the prevalence of skepticism about the work of others would be a matter of great concern, given that science is so powerfully dependent on trust in the work of others.

4. RETRACTIONS

Retractions are one direct indicator of erroneous published research results. It is clear that retractions, which can be easily tracked, have been increasing over the past 20 years [22]. Several aspects of this topic are worthy of discussion. First, though retractions are rising in frequency, they represent a tiny fraction of all published papers, perhaps several hundred of the one million published yearly. Though it is likely that these are the “tip of the iceberg”, the most important question is the size of the iceberg, and that is unknown. One reason that retractions are increasing is that it is far easier today than in the past to search for and access papers online, and previously unavailable tools now exist for identifying image manipulation [23] and plagiarism. There are also online sites on which to present such claims, and people and organizations, such as Retraction Watch, who see their role as seeking out and publicizing such cases [24].

A retraction serves to alert readers that the findings in a published paper, or at least its major elements, are no longer supported by the authors. While this outcome can have many explanations, the majority of formal retractions today are related to scientific misconduct [25]. In those situations, institutional authorities conducting investigations mandate retractions at the conclusion of their investigatory process. In some instances where retractions appear to be called for because data are no longer considered valid, authors resist this, perhaps out of fear they will be misconstrued as being guilty of misconduct rather than simply being wrong. So it should be clear that retractions do not imply research misconduct. Perhaps different naming conventions should be employed to clarify specific causes. Another question that has received much attention is whether retractions are uniformly distributed among journals. One study found the frequency of retractions is greater in journals with the highest impact factors [26]. However, this does not define the causal sequence for the association. Do journals with high impact factors more frequently publish papers requiring retractions, or does their being highly read increase the opportunity for irregularities to be identified? Another important question is whether journals create unnecessary barriers to publishing retractions, perhaps to limit unwelcome damage to their reputations [27].

5. CAUSES OF RESEARCH IRREPRODUCIBILITY

Whatever its frequency, the causes of research irreproducibility are complex. Research is needed to clarify the role of various factors, so appropriate counter measures can be designed.

5.1. Poor experimental methodology

Most instances of irreproducibility likely result from the failure to properly design, execute, and evaluate experimental data — activities at the core of what is required to conduct excellent research. Such failures include poor design of control and experimental groups, insufficient attention to reducing observer bias, inappropriate statistical approaches, and many factors relevant to the use and evaluation of specific experimental technologies [5,8–10]. While the scientific community should be concerned by the prevalence of these issues, the good news is that they are, in theory, remediable. Because the scientific community is highly decentralized, and is really a community of communities with scientists who greatly prize their independence, designing an effective and actionable approach to addressing these problems is challenging.

5.2. Poorly characterized reagents

To produce experimental results that are both true and reproducible, key reagents must be well described and they must be capable of producing the claimed measurement outcomes. Problems with reagents, including antibodies, cell lines, chemical agents, and experimental animals, are the source of many issues in irreproducible studies. Antibodies that claim to have specificity for antigens, but lack such, commonly continue to be employed [28] even after the demonstrated lack of specificity has been published. There are well documented examples of widely employed cell lines being misidentified as having a specific lineage or properties [29]. There are also many complexities in the use of animal models, especially mouse models. How these may lead to poor reproducibility as well as failure to predict outcomes when applied to humans has recently been reviewed [30].

5.3. Deficient oversight, training and mentorship

In most modern biomedical labs, the responsible leader is the principal investigator, but most or all of the work is carried out by students, post docs or technicians. One of the responsibilities of the PI is to assure that their trainees receive proper education in experimental conduct and data analysis. Although provision of formal curricula is necessary, education must go beyond lectures on research methodologies and responsible conduct, the impact of which, even if well designed, may be limited [31]. There must also be increased attention to behaviors and attitudes in group and individual meetings that reinforce correct approaches. The “hidden curriculum”, e.g. how a lab leader approaches issues of appropriate research conduct in real time, may be more important than the formal curricula in these areas.

5.4. Complex collaborative arrangements

For many good reasons, research today increasingly extends across scientific boundaries, requiring collaborations to bring together disparate techniques and approaches. As a result, a single PI may be more reliant than ever before on the contributions of collaborators, whether geographically local or distant. While the diversity of research approaches and the collaborations required to accomplish them do advance research progress, they create new challenges for those responsible for the accuracy and integrity of the published work. The veracity of an entire publication is assumed, at the very least, to be supported by the senior and corresponding author. But if that author is

either unaware of key experimental issues addressed by a collaborator, or cannot understand them at a sufficiently critical level, the opportunity for unacknowledged flaws is enhanced. This doesn't imply that each author must have equal insight into all aspects of a complex investigation. But on the other hand, claiming ignorance about details of a paper conducted by collaborators may not succeed as a defense if flaws or misconduct are later identified.

5.5. Inappropriate responses to incentives distinct from those directly related to the conduct of science

Scientific research typically begins with a scientist who desires to understand the workings of the world. Objectivity and honesty are essential tools and values throughout the process, as is the ability to generally trust the work of others. But scientists are human, and, of course, fallible. They live in a world filled with incentives that conflict with behaviors required to find the truth. When such incentives collide with a diverse community of scientists with different life situations and values, it should not be surprising that some will exhibit behaviors inconsistent with the dispassionate pursuit of the truth.

What are the key external incentives that may oppose the simple search for truth in research? Science is a career as well as a calling. So concerns about academic appointments, grant funding, compensation, and fame are ever-present during the conduct of research and, as choices are made, about its publication and dissemination. The highest professional standards for the ethical conduct of research require that such exogenous incentives not affect the integrity of research conduct, analysis, and reporting, and I believe this is most often the case. But this line is too often crossed, perhaps more so today when researchers, having been drawn into research during a period of relatively plentiful funding, are experiencing increased difficulty obtaining funding to support their research and compensation.

Much more needs to be done to explore this issue. We must understand how and why scientists sometimes cross the line between rigorous objectivity and practices that violate it, tempted by what they may see as short term benefits, though in the end these decisions are self-defeating. A second line that may be crossed is hard to delineate in some cases: the point where "questionable practices" transform into research misconduct, against which strong sanctions are required. The official NIH definition of misconduct includes fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results [32]. Fabrication and plagiarism are quite clear in their meanings. But what about falsification? The NIH definition for falsification is manipulating research materials, equipment, or processes, or changing or omitting data or results *such that the research is not accurately represented in the research record* [32]. Frequently, the question is: where do sloppiness, wishful thinking, and, perhaps, morally innocent selectivity in data presentation transition into falsification? A key element in making the distinction is whether there is intention to deceive, as opposed to errors, self-deception, or honest differences of opinion. It should be obvious that such distinctions are very difficult to make in individual cases. Those individuals required by circumstance and institutional role to make these difficult distinctions face an onerous task, since intention is often hard to assess, and careers typically hang in the balance.

The above discussion of conflicts between scientific objectivity and incentives that may oppose it has stressed those conflicts potentially related to career advancement, research funding, employment compensation, and even fame. Another class of incentives that could promote unscientific behavior relates to financial incentives from a

researcher having links to commercial/pharmaceutical development. Though these conflicts of interest surely exist, and must be disclosed, their contribution to the problem of research irreproducibility writ large, especially as relates to basic and translational research, is likely quite small compared to the more prevalent incentives related to career considerations.

5.6. Ethical lapses and sociopathy

There are many well documented and reported cases of scientists engaging in the willful fabrication and/or falsification of data that makes its way to publication, only to be retracted after the duplicity is discovered [33]. Such cases are not new in the history of science. In some cases, the extent and duration of such fabrication/falsification is astounding. An individual committed to and highly skilled in such deception can make it difficult or impossible for colleagues, and certainly reviewers and journals, to identify the fraud. On the other hand, in many such cases retrospective analysis revealed signs that should have provided vigilant colleagues reason to suspect, well before publication, that all was not well with the perpetrator and the data.

Apart from cases of misconduct that reach public awareness, the data on frequency of such events is limited. In 2010, the NIH Office of Research Integrity (ORI) received 155 allegations of research misconduct, and closed 31 cases of which 9 led to findings of research misconduct [34]. This order of magnitude is consistent with my experience overseeing a highly professional and organized process for adjudicating misconduct inquiries at Harvard Medical School for nine years. Another quite different approach has used survey data, where scientists were asked how often they themselves, or others they knew, had committed acts of fabrication, falsification or cooking of data. In one meta-analysis of such surveys [35], 1.97% of scientists admitted to having personally fabricated, falsified or modified data at least once, and up to 33% admitted to other "questionable research practices". When surveyed about the behavior of their colleagues, 14% said they believed colleagues had falsified data, and 72% said colleagues had committed other questionable practices.

5.7. Incentives produced by the major funders of bioscience research

NIH and other funding agencies place great emphasis on, and often have requirement for, research proposals with well delineated hypotheses and "expected findings", some of which are expected to have been demonstrated at the time of grant submission. It has been persuasively argued that this approach to both conducting and funding science has major conceptual flaws [36]. As relates to the issue of reproducibility, a key flaw of this requirement is its' placing the scientist in a position where data is filtered through the lens of the stated hypothesis, in a way that promotes expectation of a particular result, and biases against, or promotes rejection of, contradictory evidence. It is easy to see how this construct puts great pressure on a scientist to avoid falsifying the hypothesis upon which their grant was funded, even when evidence suggests this is the most rational approach.

5.8. Incentives produced by the publishing system

The process for reviewing and accepting manuscripts for publication may also create incentives to publish irreproducible results. There are several possible ways this may occur. Some journals, referred to as "predatory journals", are a small subset of open access journals that charge authors publication fees and claim to have rigorous peer review, but publish papers with minimal or shoddy peer review, and are

magnets for shoddy or even fraudulent science [37]. On the other end of the spectrum, many highly prestigious and sought-after journals seek to publish papers that make exciting claims that will generate buzz in the scientific community and garner press attention, goals that are easy to understand. To accomplish this goal, their decisions to accept papers are sometimes contingent on authors producing specific results suggested by reviewers or editors. Such an approach is distinct from criteria for acceptance that primarily demand well conducted studies where data support the conclusions and findings advance the field. The highest impact journals can easily afford to pass on such papers. Sometimes these editorial requests occur late in what is too often a lengthy review process, causing authors to fear loss of priority, or inability to use the acceptance in support of grant submissions or promotions. The power of such journals, reflecting in part the ability of their decisions to influence funding and promotions, creates a strong, if misplaced, incentive to fulfill the editorial requests. We may surmise that some investigators make decisions in data selection that they would not make absent the high stakes editorial exchanges. To describe this sequence is in no way to justify it.

There is also a more prevalent problem of publication bias [38], wherein new and positive findings are far more easily published than are confirmatory or negative findings. Very often, the latter go unpublished, to the detriment of scientific knowledge.

6. POTENTIAL RESPONSES TO ADDRESS THE PROBLEM OF RESEARCH IRREPRODUCIBILITY

The first, and a necessary step in addressing this issue, is to accept that it is real and important, even if its prevalence is uncertain, and requires concerted responses by and on behalf of all elements of the bioscience research community. Despite its importance, we should recognize that there are many countervailing forces working against this. First, many scientists believe that the problem is exaggerated, either in its prevalence or impact. These people see the system as working well, and/or they remain unconvinced of the problems' salience based on the literature or their personal experience. In fact, the extent of the problem may vary widely between fields and sub-fields, so the academic response might not be uniform across fields. Second, some believe that despite the problems, science will in due course sort out these issues, as it always has. Some of these people believe that what they see as "bureaucratic" efforts to counter irreproducibility will not succeed, which is likely true, and by placing additional burdens on already stressed researchers, could make the situation worse. Third, some are concerned that public discussions of these issues would provide ammunition to forces opposed to science and science funding. While this is indeed a risk, inaction seems to be a greater risk to the reputation of the field over the long term. Finally, some individuals and institutions fear diminished support from donors and organizations inclined to support research in response to public airing of these problems. It is my view that though many aspects of the prevalence and trajectory of irreproducible research are unknown, there is now sufficient evidence that this is a problem requiring attention and remediation. For the leaders of biomedicine to deny or ignore the problem today would be a strategic mistake. But, as discussed below, no one speaks for all of biomedicine, and no organization or component of the ecosystem can solve this on its own. And we must all be cognizant that intended remedies can have unintended negative consequences. I will outline some of the approaches that I believe should be considered. Others have commented on this important topic as well [39,40].

6.1. Enhance training in experimental design, statistics, proper use of reagents, data management, research and publication ethics, and clarify expectations for how to respond to concerns about these

That this is necessary is obvious — such knowledge is at the core of what is required to do quality research. Most institutions have research training programs designed to address these issues, but it seems clear that additional focused efforts are needed. Institutions should consider how best to ensure that skilled educators and appropriate materials are available and effectively employed. Education is needed at all levels of training, including for faculty, who may have deficiencies in key realms of which they are unaware. In the end, faculty bear the greatest responsibility for both applying proper methodologies and educating their trainees. Leaders of institutions that conduct substantial research programs must make this area a priority.

6.2. Place a greater emphasis on the reproducibility and importance of published research by faculty and a reduced emphasis on the number of publications and journal in which they are published

In an age when quantitative metrics are widely employed, including measures of article citations [41] and impact factor of specific journals [42], we should be reminded that such metrics can hide major defects in the quality and reproducibility of published research. Though academic organizations seek to go beyond these metrics in their judgment of scientific quality and impact, and often succeed in this endeavor, my experience as dean suggests that this approach not infrequently fails. As difficult as it might be to accomplish, there should be a major effort to identify better metrics of research reproducibility and durable scientific impact. If developed, such metrics could support academic appointments and promotions and the response to cases where poor reproducibility is an issue. It would also be important for faculty review processes to accord greater credit to well-conducted studies that confirm or contradict published work. Although quantifiable metrics are a goal, it would be hard to have a better approach than one based on a deeply informed assessment by objective experts in the field.

6.3. Develop increased expectations for open data as the standard approach in all publications

This should apply not only to clinical investigations and "big data" assemblies where such expectations are already prevalent, but as much as possible to standard laboratory work. This would likely require use of quality lab data management tools, now lacking in many labs. Data being publicly available after publication would allow additional analyses to be performed and would facilitate identification of errors in published work [43,44]. The FAIR guiding principles for scientific data management and stewardship (findable, accessible, interoperable and reusable) are a major global effort in this regard [45].

6.4. Promote changes in scientific publishing to facilitate research reproducibility

6.4.1. New procedures by journals to enhance quality of manuscripts and reviews

Such changes might include rigorous checklists to promote appropriate design features, enhanced statistical assessment by journals, and encouragement or requirement that raw data be provided during submission to be available online. As one example, Nature Press recently established such guidelines [46], and Cell Press will soon be introducing a new approach to more structured, transparent, accessible reporting of methods in their publications [47]. How such

guidelines are now and will be used and what their effects will be will require research, and publishers should be willing to share their findings.

6.4.2. Increased willingness to publish negative and confirmatory studies

Efforts should be made to increase the ease of publishing well conducted studies with negative results, and studies that confirm prior research. Though such studies offer fewer incentives to both scientists and journals, for science as a whole they are of critical importance, and we should find ways to celebrate the best among them. Having said this, negative or confirmatory studies can be poorly done, sloppy, fraudulent, or defective in other ways, so high standards are important for publishing these studies as well. Although all journals should see the importance of doing this, their failure to date has led to new venues arising that specifically seek to accommodate negative studies [48]. High profile journals are beginning to take up the call to recognize and publish replication studies, sometimes creating new venues in which to do this [49].

6.4.3. Clarify use of retractions

Develop an agreed upon and consistent approach for journals to manage retractions. This would include clear criteria for retractions versus corrections, how they are linked to published papers, and how the reason for the retraction or correction is explained, whether due to misconduct or a wide array of more innocent explanations.

6.4.4. Changes to the peer review process

The peer review process is a key gatekeeper for scientific publication, and, as such, how it operates is pertinent to the reproducibility of published work. This is a much larger topic than can be dealt with here, but a few key points are relevant to this discussion. If the major reasons for irreproducibility relate to poor experimental design, statistics, reagent reliability, and data selection, then reviewers (and editorial staff) could become more capable, collectively, of competently dealing with these topics in the course of reviews, realizing that the authors are ultimately responsible for the quality and integrity of the work. Apart from use of programs to identify image manipulation or plagiarism, reviewers and editors cannot be expected to identify most instances of wilful and skillfully executed fraud. And the fact that reviewers are almost never paid for their efforts limits the incentive to be maximally attentive to the task.

What changes to peer review would have the greatest opportunity to positively impact the reproducibility of research over the long term? I believe the most important changes would be to increase the trend of making the reviews themselves and various versions of the manuscripts available online, and more controversially, to strongly encourage or require reviewers to sign their reviews. There are many reasons to consider supporting these practices [50], now employed by a limited number of both venerable and more recently launched publications [51,52]. In addition to the inherent value of transparency in scientific understanding, this approach would disincentivize superficial and ill-informed reviews, and those based on personal or professional conflicts, while incentivizing reviews of the greatest insight and quality. Indeed, the ability to cite high quality reviews and reviewers would for the first time permit the academic community to properly recognize them, impossible today since the relevant data are hidden. The major argument against reviewer identification is a concern that critical but honest reviewers could be subjected to consequential retribution by unhappy but influential authors. This is a real concern. But to the extent it is true, this brings dishonor to the scientific community and would need to be actively resisted, rather than be deployed as an argument for

continuing a dysfunctional approach. Specific remedies must be designed. Another major benefit of transparent reviews would be to provide new data enabling research in the field of peer review. This could allow us to determine which review and editorial policies actually promote reproducibility and quality of published papers [53]. Though journals may be in possession of such data, there is a very limited tradition for their contributing research in this area [51]. I strongly believe that valuable insights into optimization of the review process to enhance reproducibility and quality would emerge.

Two additional elements of today's publishing ecosystem merit brief comment as they may influence, over time, the reproducibility of published work. The first is the practice of placing manuscripts on "pre-print servers" for public comment prior to peer review and publication in traditional journals [54]. There are potential advantages and some risks of this practice, which has existed for quite a while in physics and mathematics. What impact its growing use might have on reproducibility of published research is currently unknown. A second new element relates to new venues for "post publication peer review", such as PubMed Commons [55], PubPeer [56] and other sites, where participants discuss published data in variably moderated online communities. The potential ability of such venues to enhance scientific communication seems obvious. Although many discussions on PubPeer have questioned the validity of published data and some have even led to retractions, the fact that most discussants are anonymous has been challenged and is a topic of ongoing debate [57]. Overall, it seems likely that a robust capacity for extended online discussion of published research will eventually advance scientific progress, and may hasten discovery of problems with some papers, while creating unfortunate opportunities for anonymous and misdirected harassment in some cases.

6.5. Changes to the culture of scientific research

It is well known that efforts to change culture, apart from changing rules, are exceptionally difficult and often unsuccessful. But we can start by identifying the cultural elements where change would be welcome. First among these would be to focus more attention on linking the respect we accord scientists to our assessment of the integrity of their work, attention to detail, the durability of their discoveries, and their ability to mentor junior colleagues along these dimensions. Unfortunately, these desirable characteristics do not always align with a scientists' academic rank, fame, recognition, level of funding, number of publications, or in which journals they are published. There is no simple approach to achieving this, but awareness and articulation of its importance by scientific leaders, and the broader research community, are necessary first steps.

7. CONCLUSION

I hope this commentary makes it clear that while more research into the nature and causes of irreproducible bioscience research is needed, we know enough today about the relevant facts to initiate remediating actions in many areas. Since so many institutions and cultural domains are involved, multiple approaches must be tried, with as much communication and, where possible, coordination among them. But we should not underestimate the difficulty of having these combined efforts produce clear benefits over a reasonable period of time.

It should also be stressed that the goal of the entire bioscience research enterprise is increased knowledge and progress, not the highest achievable level of reproducibility, however defined. If the latter were to be taken as the dominant goal, it is very likely that the former

would suffer. And a world in which highly reproducible papers combine to produce insights and discoveries of limited impact would not be a desirable one. So balance and judgment in addressing this issue are in the end critical, and these have not always been evident in much of the public discussion of this subject.

One final cautionary note also seems necessary. While engaging in this effort, it is important to avoid casting the research community in an unnecessarily negative light. Every profession must work continuously to raise its professional standards. This field is no different. In my personal experience, and that of many people whose judgment I trust, the great majority of biomedical scientists are indeed motivated by a desire to discover the truth, and they conduct their research in an exemplary manner. I am immensely proud to be a member of this remarkable community, whose benefits to society will surely increase over time. As we undertake the necessary task of enhancing research reproducibility to accelerate scientific and medical progress, we must continue to accord this critically important community the respect, admiration and support that it deserves.

ACKNOWLEDGEMENTS

I would like to thank Eleftheria Maratos-Flier, Becky Ward, Michael Lederman, Gretchen Brodnicki and David Glass for reading the manuscript and offering helpful comments.

REFERENCES

- [1] Dorsey, E.R., de Roulet, J., Thompson, J.P., Remnick, J.I., Thai, A., White-Stellato, Z., et al., 2010. Funding of US biomedical research, 2003–2008. *Journal of the American Medical Association* 303(2):137–143.
- [2] Ioannidis, J.P.A., Fanelli, D., Dunne, D.D., Goodman, S.N., 2015. Meta-research: evaluation and improvement of research methods and practices. *PLoS Biology* 13(10):e1002264.
- [3] https://www.nlm.nih.gov/bsd/num_titles.html.
- [4] Popper, K., 1963. *Conjectures and refutations: the growth of scientific knowledge*. London: Routledge. p. 293.
- [5] Ioannidis, J.P., 2005. Why most published research findings are false. *PLoS Medicine* 2(8):e124. <http://dx.doi.org/10.1371/journal.pmed.0020124>.
- [6] Prinz, F., Schlange, T., Asadullah, K., 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 10(9):712.
- [7] Begley, C.G., Ellis, L.M., 2012. Raise standards for preclinical research. *Nature* 483:531–533.
- [8] Goodman, S.N., Fanelli, D., Ioannidis, J.P., 2006. What does research reproducibility mean? *Science Translational Medicine* 8(341):341ps12.
- [9] Leek, J.T., Jager, L.R., 2016. Is most published research really false? *bioRxiv* <http://dx.doi.org/10.1101/050575>.
- [10] Ioannidis, J.P., Greenland, S., Hlatky, M.A., Khoury, M.J., Macleod, M.R., Moher, D., et al., 2014. Increasing value and reducing waste in research design, conduct and analysis. *Lancet* 383(9912):166–175.
- [11] Loscalzo, J., 2012. Irreproducible experimental results: causes, (mis)interpretations, and consequences. *Circulation* 125(10):1211–1214.
- [12] Weingart, S., 2012. How many citations does a paper have to get before it's significantly above baseline impact for the field. <https://www.ScottBot.net/HIAL/index.html@P=22108.html>.
- [13] Yi, P., Park, J.S., Melton, D.A., 2013. Betatrophin: a hormone that controls pancreatic β cell proliferation. *Cell* 153(4):747–758.
- [14] Gusarova, V., Alexa, C.A., Na, E., Stevis, P.E., Xin, Y., Bonner-Weir, S., et al., 2014. ANGPTL8/betatrophin does not control pancreatic beta cell expansion. *Cell* 159(3):691–693.
- [15] Cox, A.R., Lam, C.J., Bonnyman, C.W., Chavez, J., Rios, J.S., Kushner, J.A., 2015. Angiotensin-like protein 8 (ANGPTL8)/betatrophin overexpression does not increase beta cell proliferation in mice. *Diabetologia* 58(7):1523–1531.
- [16] Obokata, H., Wakayama, T., Sasai, Y., Kojima, K., Vacanti, M.P., Niwa, H., et al., 2014. Stimulus-triggered fate conversion of somatic cells into pluripotency. *Nature* 505(7485):641–647.
- [17] Aizawa, S., 2016. Results of an attempt to reproduce the STAP phenomenon. *F1000Res*. 5. p. 1056.
- [18] De Los Angeles, A., Ferrari, F., Fujiwara, Y., Mathieu, R., Lee, S., Lee, S., et al., 2015. Failure to replicate the STAP cell phenomenon. *Nature* 525(7570):E6–E9.
- [19] Maher, B., 2015. Cancer reproducibility project scales back ambitions. <http://www.nature.com/news/cancer-reproducibility-project-scales-back-ambitions-1.18938>.
- [20] Morrison, S.J., 2014. Reproducibility project: cancer biology: time to do something about reproducibility. <https://elifesciences.org/content/3/e03981>.
- [21] Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533(7604):452–454.
- [22] Steen, R.G., 2011. Retractions in the scientific literature: is the incidence of research fraud increasing? *Journal of Medical Ethics* 37(4):249–253.
- [23] Newman, Anthony, 2013. The art of detecting data and image manipulation. <https://www.elsevier.com/editors-update/story/publishing-ethics/the-art-of-detecting-data-and-image-manipulation>.
- [24] Retraction Watch. <https://en.wikipedia.org/wiki/Retraction-Watch>.
- [25] Fang, F.C., Steen, R.G., Casadevall, A., 2012. Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences of the United States of America* 109(42):17028–17033.
- [26] Fang, F.C., Casadevall, A., 2011. Retracted science and the retraction index. *Infection and Immunity* 79(10):3855–3859.
- [27] Allison, D.B., Brown, A.W., George, B.J., Kaiser, K.A., 2016. Reproducibility: a tragedy of errors. *Nature* 530(7588):27–29.
- [28] Baker, M., 2015 May 21. Reproducibility crisis: blame it on the antibodies. *Nature* 521(7552):274–276.
- [29] Marcus, A.D., 2012 Apr 20. Lab mistakes hobble cancer studies but scientists slow to take remedies. *Wall Street Journal*. <http://www.wsj.com/articles/SB10001424052970204571404577257513760102538>.
- [30] Drucker, D., 2016. Never waste a good crisis: confronting reproducibility in translational research. *Cell Metabolism* 24(3):348–360.
- [31] Marusic, A., Wager, E., Utrobicic, A., Rothstein, H.R., Sambunjak, D., 2016. Interventions to prevent misconduct and promote integrity in research and publication. *Cochrane Database of Systematic Reviews* 2016:4.
- [32] https://grants.nih.gov/grants/research_integrity/research_misconduct.htm.
- [33] Broad, W., Wade, N., 1983. *Betrayers of the truth: fraud and deceit in the halls of science*. Edbury Press.
- [34] Office of research integrity annual report 2010. <https://www.yumpu.com/en/document/view/42217572/ori-annual-report-2010-the-office-of-research-integrity-us->.
- [35] Fanelli, D., 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One* 4: e5738.
- [36] Glass, D.J., 2010. A critique of the hypothesis, and a defense of the question, as a framework for experimentation. *Clinical Chemistry* 56(7):1080–1085.
- [37] Predatory open access publishing https://en.wikipedia.org/wiki/Predatory_open_access_publishing.
- [38] Publication bias. https://en.wikipedia.org/wiki/publication_bias.
- [39] Collins, F.S., Tabak, L.A., 2014. NIH plans to enhance reproducibility. *Nature* 505(7485):612–613.
- [40] Casadevall, A., Ellis, L.M., Davies, E.W., McFall-Ngai, M., Fang, F.C., 2016. A framework for improving the quality of research in the biological sciences. *mBio* 4 e01256–16.
- [41] Pendlebury, D.A., 2009. The use and misuse of journal metrics and other citation indicators. *Archivum Immunologiae et Therapiae Experimentalis (Warsz)* 57(1):1–11.

- [42] Casadevall, A., Fang, F.C., 2014. Causes for the persistence of impact factor mania. *mBio* 5(2) e00064–14.
- [43] Travis, K., 2011. Sharing data in biomedical and clinical research. *Science*. <http://www.sciencemag.org/careers/2011/02/sharing-data-biomedical-and-clinical-research>.
- [44] National Institutes of Health Plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research Feb 2015. <https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf>.
- [45] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Dumon, O., Groth, P., et al., 2016. The FAIR guiding principles for scientific data management and stewardship. <http://www.nature.com/articles/sdata201618>.
- [46] Reducing our irreproducibility. *Nature* 496, 2013:398.
- [47] Cell Press: Introducing STAR methods. 2016. <http://www.cell.com/star-methods>.
- [48] Journal of Negative Results in Biomedicine. <https://Jnrbm.biomedcentral.com/about>.
- [49] Go forth and replicate. *Nature* 536, 2016:373.
- [50] Smith, R., 1999. Opening up BMJ peer review. *British Medical Journal* 318(7175):4–5.
- [51] van Rooyen, S., Godlee, F., Evans, S., Black, N., Smith, R., 1999. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomized trial. *British Medical Journal* 318(7175):23–27.
- [52] [Blog.f1000research.com/2014/05/21/what-is-open-peer-review/](http://blog.f1000research.com/2014/05/21/what-is-open-peer-review/).
- [53] Rennie, D., 2016. Let's make peer review scientific. *Nature* 535:31–33.
- [54] Bastian, H., 2016. Breaking down pros and cons of preprints in biomedicine. *PLoS Blogs*. <http://blogs.plos.org/absolutely-maybe/2016/05/01/breaking-down-pros-and-cons-of-preprints-in-biomedicine/>.
- [55] PubMed commons: A forum for scientific discourse. <https://www.ncbi.nlm.nih.gov/pubmedcommons/>.
- [56] PubPeer. <https://en.wikipedia.org/wiki/PubPeer>.
- [57] Blatt, M.R., 2015. Vigilante science. *Plant Physiology* 169(2):907–909.