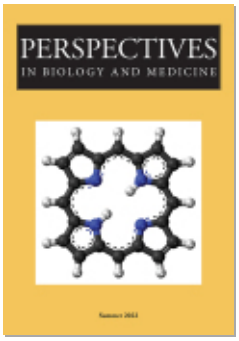
The Problem of Irreproducible Bioscience Research

Jeffrey S. Flier



Perspectives in Biology and Medicine, Volume 65, Number 3, Summer 2022, pp. 373-395 (Article)

Published by Johns Hopkins University Press
DOI: *https://doi.org/10.1353/pbm.2022.0032*

The Problem of Irreproducible Bioscience Research

Jeffrey S. Flier

ABSTRACT Over recent decades, progress in bioscience research has been remarkable, but alongside the many transformative advances is a growing concern that a surprisingly high fraction of published research cannot be reproduced by the scientific community. Though experimental and interpretive errors are unavoidable features of the scientific process, recent evidence suggests that irreproducibility is a serious issue requiring analysis, understanding, and remediation. This article reviews the meaning of research reproducibility, examines ongoing efforts to estimate its prevalence, and considers the factors that contribute to it. Two recent case studies illustrate the disparate responses that researchers may take when facing serious claims that a high-profile research finding is irreproducible and may be false. Finally, the article examines potential interventions to counter the current level of irreproducibility, aimed at increasing the efficiency and impact of society's substantial and critically important investment in bioscience research.

Late in 2010, C. Glenn Begley, an outstanding Australian physician-scientist and oncologist, approached his 10-year anniversary as global head of hematology and oncology at the biotechnology company Amgen. As he contemplated his next career move, he thought about one concerning aspect of his Amgen ex-

Harvard Medical School, 220 Longwood Avenue, Boston, MA 02115.
Email: jeffrey_flier@hms.harvard.edu.
The author wishes to thank the following individuals for their valuable comments on various drafts of this article: David Glass, Len Harrison, John Ioannidis, and Eleftheria Maratos-Flier.

Perspectives in Biology and Medicine, volume 65, number 3 (summer 2022): 373–395.

perience, and how he might address it before leaving for his next gig. At Amgen and other biopharmaceutical companies, scientists typically scour the scientific literature for exciting papers from academic labs that suggest new therapeutic targets for their companies' powerful drug development platforms. Over the previous 15 to 20 years, few areas had generated more papers suggesting potential drug targets than cancer biology, so Begley and his team had many possible targets to pursue. But that produced an unexpected problem. Despite highly motivated efforts, backed by teams of skilled company scientists and deep financial resources, their efforts to reproduce key observations within these high-profile academic papers were overwhelmingly negative. This was true even after reaching out to the authors, and where possible sending company scientists to conduct key experiments in their labs.

Amgen never published these negative results, so the skepticism harbored within the pharma industry was communicated, if at all, via informal conversations. To begin to remedy this, Begley sought and obtained agreement from Amgen leadership to publish a summary of their recent negative experience, and asked a prominent academic scientist, Lee Ellis from MD Anderson in Houston, to join him in doing so. The paper they coauthored, entitled "Drug Development: Raise Standards for Preclinical Cancer Research," was published in *Nature* in March 2012, and it had an immediate and seismic effect on the scientific community. The paper described how, over a period of years, they sought to reproduce key findings in 53 high profile papers, each published in a prestigious journal, before initiating large-scale internal drug development efforts based on those results. The punchline? Despite extensive efforts, they could reproduce the core findings in only six of the 53 studies, or 11%—a claim totally shocking to most readers, including me. Several months earlier, in September 2011, a group from Bayer Pharmaceuticals used a similar approach and reported reproducing only 25% of the preclinical cancer research papers they examined (Prinz, Schlange, and Asadullah 2011). Several similar papers followed. These didn't identify the studies they had attempted to reproduce, the standard employed to assess reproducibility, or specific details of their experimental efforts, so in effect, ironically, their own findings could never be reproduced. Nevertheless, these very public claims by multiple serious pharmaceutical scientists and their companies created a new context wherein the issue of irreproducible academic research publications could not be ignored.

Do scientists generally believe the research published by other scientists to be true? Since the scientific method is rooted in rationality, objectivity, and the pursuit of truth, and progress requires building on the work of others, one might expect they do. But the answer to this question has been changing, and as suggested by the story above, the direction of change has not been positive. Early in my career, in the 1970s, I believed most research papers I read generally reflected reality, and if I scrupulously followed the authors' published methods, I'd most

often be able to reproduce their most important results, at least to a substantial degree, sufficient to convince me that they were more right than wrong. Of course, I knew scientists were fallible, that researchers sometimes made errors requiring correction, and that some scientists were more skilled, judicious, and honest than others. As one of my mentors used to say: "In God I trust. Everyone else show me data."

But in the 1970s, based on my recollection and review of the literature, the veracity and *reproducibility of research as a whole* wasn't seen as an "issue" that threatened trust in the research enterprise. Today, papers analyzing the topic appear regularly in the literature, and a new field of "meta-research"—research about research—has arisen to address the issue. In 2016, the journal *Nature* conducted an online survey of 1,576 scientists, in which across all fields, 52% thought there was a "crisis" of reproducibility; 73% of respondents believed at least half of published papers could be trusted, with a lower percent in biology and medicine (Baker 2016). The survey revealed that most scientists in all surveyed fields had doubts about the validity of a substantial fraction of published scientific results.

Though the exact prevalence of irreproducibility remains unknown for many reasons that will be discussed, the preponderance of evidence suggests the number is disturbingly high. Many scientists today see the problem as one of growing importance to the integrity and efficiency of the research profession. It's necessary to explore research reproducibility—what the term means, why reproducibility isn't optimal today, and why it must be improved to increase trust in and effectiveness of the research ecosystem on which society depends.

What Does Research Reproducibility Mean?

What do we mean when we say that research is reproducible? This question is receiving considerable attention in the philosophical literature, where, unsurprisingly, many conceptual complexities emerge (Fidler and Wilcox 2018). On one level, it's simple: a finding is reproducible when other scientists perform the "same experiment" and get the same result, thereby "confirming" the prior work. For example, a paper claims that administering a certain chemical to a strain of mice causes their blood sugar level to fall by 50% after 24 hours. Another lab obtains the chemical and the same strain of mice, follows the well-defined experimental protocol, and finds a very similar effect on blood sugar.

But the story is often more complex than that. First, it is noteworthy that many of the 4 million plus bioscience papers published yearly are unread or minimally examined. There are no scientists with the interest, resources, or incentives to "repeat" or confirm this vast sea of published work, so whether the findings they report are reproducible will simply never be assessed.

Second, the lexicon of reproducibility employs imperfectly standardized terms to describe the phenomenon, such as *reproducibility* and *replicability*, each with

variably employed and understood meanings (Fidler and Wilcox 2018). One recent recommendation is to employ a new lexicon that distinguishes between three aspects of reproducibility (Goodman, Fanelli, and Ioannidis 2016). The first aspect is *methods* reproducibility, with procedures described so that an expert can faithfully repeat them. The second aspect is *results* reproducibility—often called "replication"—carried out with a technically competent effort. This effort may be "exact," using identical conditions, or "conceptual," using variably altered conditions, which if positive extends the initial claim to additional conditions, revealing its potential generalizability. The third aspect is *inferential* reproducibility, the ability to make a "knowledge claim" of similar strength from a study replication (Dirnagl 2019; Goodman, Fanelli, and Ioannidis 2016). This reminds us that the goal of these inquiries is not primarily a narrow statistical assessment of the reproducibility of an experiment, but an assessment of the *truth claims* of the experiments, which should be the most important goal. Despite this and other efforts to bring clarity to the terminology, definitional variability remains evident in much of the literature, and requires ongoing attention.

In some fields, reproduction means re-analysis, by which a publicly available dataset is independently analyzed to see whether the same conclusions are found. In contrast, in areas such as preclinical biomedical research—the focus of this article—reproduction most often involves conducting a new experiment, ideally with the same (or sufficiently similar) procedures, reagents, experimental conditions, and analytical approaches.

Unfortunately, published methods are often inadequately described, identical reagents may be difficult to acquire, and other technical roadblocks prevent scientists from reproducing a prior study by repeating it exactly. So different results may be the consequence of different experimental conditions, rather than the original work being "irreproducible." It would be important to determine how often what we refer to as irreproducibility results from inadequate methodologic description, variation in reagents, or other technical barriers.

Another issue relevant to the terminology of reproducibility is the distinction between exploratory and confirmatory research (Kimmelman, Mogil, and Dirnagl 2014). *Exploratory* research aims to provide data that might suggest new theories, and typically involves small and flexible experiments using a variety of methodologies. At the outset, details of experimental design are not precisely established and may evolve in response to initial findings. In contrast, *confirmatory* studies are predesigned to test and validate exploratory results and require sufficient size and statistical power to enable this.

The expectations for reproducibility of these two types of research are obviously different, and although it isn't necessary, scientists and journals commonly desire to combine exploratory and confirmatory studies into single papers. In my experience, early exploratory results are less commonly published, and when they are, they receive less attention because the uncertainty of their claims is

understood—though some scientists might be stimulated to follow up with further research. Another key point is that results that are highly innovative and unexpected—and therefore potentially the most exciting—are more likely to be false positives and untrue, as suggested by Bayesian rules of inference. In contrast, less novel research is more likely to be true and reproducible. To quote Dirnagl (2019): "Reproducibility is not an end in itself. Trivial findings may be highly reproducible, and non-reproducible results may be actually true." In other cases, false claims may stimulate others to discover an important truth.

An additional brief philosophical and methodologic digression might be appropriate here. Because we know that science is inherently subject to error (both false positives—type 1 errors—that receive the greatest attention, as well as false negatives—type 2 errors), we must consider what level of reproducibility might be optimal for scientific progress as a whole, or for specific disciplines where this might vary. It seems intuitively likely that an "excessive" demand for reproducibility (and punishment for its absence) would slow scientific progress and deter discovery and publication of true and useful results. This negative outcome would need to be balanced against the opposite type of negative outcome: wherein untrue claims induce unproductive and ultimately wasteful efforts to repeat or build upon them. One recent paper modeled this and concluded that it is possible to have an efficient scientific community in which many findings of unclear replicability are published, and those that attract the community's interest are replicated after publication (Lewandowsky and Oberauer 2020). To determine the point at which excessive demand for reproducibility reduces progress would likely require empirical outcome assessments. Determining an optimal balance likely goes beyond strictly scientific considerations, however, to include normative and related political considerations, including those related to efficient use of resources (Kane and Kimmelman 2021). These latter issues go beyond the scope of this essay and will not be further considered here.

Human clinical research poses special problems for reproducibility, since many human trials are large, complex, and expensive. Who will pay the tens of millions of dollars required to conduct a study of several thousand patients, in dozens of centers, with numerous clinical outcomes over a year or more of study? Repeating clinical studies of drugs and interventions is subject to limited coordination. One consequence is that human subject research has evolved to require greater degrees of oversight and pre-approval, as by human studies committees, and preregistration of goals and endpoints placed on the public websites like clinicaltrials.gov (Zarin et al. 2019).

Apart from these technical barriers to replication, scientists are poorly incentivized to use their limited resources to "simply reproduce" the work of others. In the first place, limited credit is awarded for reproducing (or failing to reproduce) the work of others, and papers describing such results are difficult to publish (Flier 2019). Journals prefer papers that claim flashy and novel discoveries and have far

less interest in papers reporting confirmations or disconfirmations of published observations. Furthermore, unless reproducing the work of others is essential to their own research, scientists greatly prefer to build upon the findings of others rather than repeat them, enabling the claims of novelty generally required to obtain funding, promotion, and other recognition. If efforts to build upon a newly published result fails, scientists are more likely to move on to another question than to invest additional effort into repeating the earlier experiment. Additionally, there are social aspects connected to these issues, including stigmatization of scientists in response to claims their findings are not reproduced, which, depending on how they have responded to such claims, may be undeserved. On the other hand, those attempting to reproduce the work of others may themselves be stigmatized as unoriginal, or as taking undue pleasure in embarrassing a fellow scientist (Dirnagl 2019). These factors complicate the psychology and sociology of research irreproducibility and efforts to address it.

Some claims of discovery receive special scrutiny from the scientific community, however. Consider a finding that purports to have major implications for a field, is carried out by a highly respected scientist, and is published in an elite, selective journal. Even more, consider one that is picked up by the popular press and social media, which often hype its potential health implications. Findings like this are likely important to other scientists, whose research plans may be altered if the "transformative" new observations are indeed true. Such claims attract great scrutiny, as interested labs race to assess their truth. While many such findings are reproducible, a surprising number are not. And when a widely reported and celebrated finding is not confirmed or is suggested to be false, the consequences can have substantial impact on the field. This is true when the disconfirmed finding is eventually formally retracted, but more so if the finding remains uncorrected in the literature, as is often the case, causing confusion and slowing or derailing progress in the field (Tatsioni, Bonitsis, and Ioannidis 2007).

How Prevalent Is Irreproducibility?

How often is published bioscience research irreproducible? The simple and frustrating answer is that we don't know, as this would require taking a large sample of published research (from different fields, since the extent of the problem likely differs between fields) and commissioning well-designed efforts to reproduce them by scientists capable of doing so. This difficult task hasn't been broadly attempted and most likely never will be. Lacking that, other approaches have emerged to estimate the prevalence. Stanford's John Ioannidis has dedicated his career to this endeavor. In his widely cited 2005 paper, provocatively entitled "Why Most Published Research Findings Are False," Ioannidis examined the literature, and by evaluating appropriateness of sample size, statistical approaches, effect sizes, experimental and publishing biases, and other factors, concluded that

most published findings were likely to be false. Subsequent studies continue to support this disturbing conclusion. Surely irreproducibility is far more prevalent than previously thought, even if it's not true that most papers are false.

Another assessment of the reproducibility of high-profile research has come from pharmaceutical researchers (Begley and Ellis 2012). Ironically, early in my career, academics commonly cast aspersions on the quality of biopharma research. Recently, the tables have turned, at least in the realm of basic research, as prominent papers from large biopharmaceutical companies have reported an inability to reproduce a high percentage of high-profile academic cancer research papers, consistent with informal conversations between academics and skeptical biopharma scientists. Though biopharma scientists continue to rely on academic discoveries, and these remain the initial source of most new therapies, they are skeptical of many scientific papers from the academy.

In 2013, the Reproducibility Project: Cancer Biology sought to provide some definitive data on the reproducibility of basic science cancer studies. Funded by a large grant from the Arnold Foundation to the Center for Open Science (COS), it selected a group of 50 high-profile papers and planned to reproduce them in appropriate labs and publish the results in the journal *eLife,* which agreed in advance to do so to limit concerns about publication bias (Davis 2014). They would seek cooperation from the labs that published the original work to ensure that experimental details were properly described and necessary reagents were available and utilized. Once the study design was agreed upon, they would publish a detailed experimental protocol, with approval from the relevant labs. Finally, they would contract with a group possessing the requisite capability to do the research, and then publish the results (after journal review) along with an associated commentary.

Due to greater than anticipated costs and difficulties obtaining the necessary methodologic information to enable properly designed replications, the number of papers was scaled back from 50 to 23. Brian Nosek, director of the COS, told me that of the 50 papers and 197 individual experiments initially considered, they felt that none—repeat *none*—of the experiments were described in sufficient detail to consider reproduction without further input from the authors. Remarkably, 32% of the authors who were contacted to facilitate replication of their work didn't respond or were not helpful. The complexity of the effort is revealed by the fact that the average time from selecting a study for replication to publishing a final report was 197 weeks.

Fifty experiments reported in 23 papers were repeated, and the summary results have now been published (Errington et al. 2021b). A minority of studies reproduced key elements of the original papers, but a majority could reproduce only some results or none, or the results could not be clearly interpreted. Of 97 numerical effects that were statistically significant in the original papers, only 42 were statistically significant and in the same direction, and seven were statistically

significant in the opposite direction. On average, replication lost 85% of the magnitude of originally reported effects.

Several major conclusions can be drawn from this heroic effort. First, it's often extremely difficult to precisely reproduce methods. Second, there is often substantial ambiguity regarding what it means to get the "same results," a fundamentally important consideration in assessing the prevalence of reproducibility. The project's authors hedged a bit by concluding: "A successful replication does not definitively confirm an original finding or its theoretical interpretation. Equally, a failure to replicate does not disconfirm a finding, but it does suggest that additional investigation is needed to establish its reliability" (Errington et al. 2021b). The reality may be more sobering than this account suggests. The architects of the project concluded:

> No single effect, experiment, or paper provides definitive evidence about its claims. Innovation identifies possibilities. Verification interrogates credibility. Progress depends on both. Innovation without verification is likely to accumulate incredible results at the expense of credible ones and create friction in the creation of knowledge, solutions, and treatments. Replication is important for research progress because it helps to separate what we know from what we think we know. (Errington et al. 2021b)

That so many experiments—published in leading journals—couldn't be reproduced despite the extraordinary efforts employed by the Cancer Reproducibility Project renders unsurprising claims of irreproducibility in other fields, in which efforts to reproduce experiments involved far less attention to detail and time and effort expended. There are still unresolved issues about the optimal degree of reproducibility and the trade-offs involved in achieving this. Though some authors whose findings weren't reproduced were, not surprisingly, critical of the effort, it seems clear to most observers that we do have a serious reproducibility problem (Kane and Kimmelman 2021). Efforts to understand and address this issue are now receiving the attention they deserve.

What Causes Irreproducibility?

It would be nice if the problems with reproducibility had a simple, unifying explanation, which would facilitate implementation of effective remedies. Unfortunately, the causes are complex (Flier 2017a, 2017b; NASEM 2019). As I became interested in this problem, the more I sought explanations for irreproducibility, the more I found them. These include inadequate training, oversight and mentorship; deficiencies in experimental design and reagents; and perhaps as an important root cause, misaligned incentives from our approach to funding, publishing, and recognizing research accomplishments. Each of these is real and important and requires assessment, though their interaction within the prevailing

institutional ecosystem and culture of research complicates the design and implementation of potential remedies.

At the heart of research is the design and execution of experiments capable of delivering reliable and interpretable data and valid conclusions. Although general principles for achieving these outcomes have been articulated, their details vary greatly depending on the field and the type of experiment. Valid statistical approaches must play a key role in both proper experimental design and interpretation. Unfortunately, too many scientists have a rudimentary command of statistics, often employing it robotically to claim that findings are "statistically significant" through so-called "t-tests," hoping for values of $p<.05$, thought to indicate that results are unlikely to be due to chance alone. Statistics are often misapplied and misinterpreted, despite statistical input being required at every stage of the process, from initial design to interpretation. For example, a cursory examination of the published literature suggests that sample sizes are often inadequate to permit reliable conclusions (Szucs and Ioannidis 2017). Especially in the era of "big data" science, procedures like "p-hacking" and "data dredging" are too often employed to search within data sets after the fact for "significant" correlations and conclusions (Chavalarias et al. 2016). This leads to erroneous conclusions that will likely be irreproducible when later tested using proper approaches.

Attention must also be paid to the insight from Bayes that the probability of a new finding being true relates to the weight of the new evidence, as well as the prior probability of its being true. Bayesian analysis suggests that if a hypothesis is highly unlikely a priori, as is often the case for highly novel findings, a larger amount of strong evidence is needed to overcome the prior reasons to be skeptical about the new findings.

Another key methodological issue is observer bias. Whether aware of it or not (most certainly are), scientists prefer their experimental hypotheses to be true, especially when their grant proposals are justified by claims that they are. Since there are numerous ways in which confirmation bias can inappropriately influence experimental results, experimenters should wherever possible be blinded to experimental details, such as the identity of treated vs. untreated groups—whether these be molecules, cells, animals, or people. Blinding of the experimenter is now required, with exceptions for the earliest phase studies and some components of others, in human subjects research. Most human subjects research also requires preregistration, by which the design and goals of the research, including what will be considered a positive outcome, are recorded in advance at public sites such as clinicaltrials.gov (Zarin et al. 2019). Preregistration focuses attention on proper experimental design, reducing the possibility that false post-hoc conclusions will be drawn based on chance alone. Preregistration is rarely employed in preclinical research, and it is less relevant to more open-ended preclinical "exploratory studies" that are not linked to formal, restrictive hypotheses (Kimmelman, Mogil,

and Dirnagl 2014). Nosek of the COS still encourages preclinical experiments to be preregistered, providing a web site on which details can be uploaded and archived. On the other hand, some have persuasively argued that overuse of hypothesis-driven approaches has adverse consequences for the research ecosystem (Glass 2010).

Another cause of irreproducibility are research reagents that are unreliable or otherwise problematic. Many reagents are employed in lab-based research, ranging from chemicals, solutions, nucleic acid derivatives, diverse proteins and antibodies, and cell lines, to stocks of "model organisms" ranging from fruit flies to mice. Some are created by scientists to enable their own research, but many are obtained from other scientists or purchased from commercial suppliers. Needless to say, the reproducibility and reliability of research is heavily dependent on the quality of these reagents, and there are many instances where they are insufficient to the task (Bradbury and Pluckthun 2015).

Another major problem are cell lines. Bioscience research frequently utilizes cell lines derived from experimental animals or humans. The first and most infamous are HeLa cells, initially derived (without her consent) from a tumor of the patient Henrietta Lacks (Skloot 2011). Because of their "immortality," these cells were passed around the world for use by thousands of scientists. Though studies of HeLa cells produced many useful insights, their provenance became increasingly suspect, as HeLa cells contaminated many other research cell lines and the original cells underwent genetic changes (Lucey, Nelson-Rees, and Hutchins 2009). The HeLa cell problem led to recognition of a broader issue for research with cell lines: most scientists simply accepted cells based on their "label" from repositories or other scientists, rarely confirming their exact identity and reported phenotypes before conducting and reporting results. In one recent large study, 5% of human cell lines in papers considered for peer review were misidentified (Souren et al. 2022).

Antibodies are another key class of reagents whose use can promote mistaken research results. Used routinely to identify and quantitate specific proteins, many antibodies sold to researchers are far less specific than their purveyors claim them to be. Together, faulty reagents have produced thousands of papers containing erroneous claims, most of which are never corrected.

Mice are a staple of modern bioscience research. Most lines of genetically modified mice are created by scientists and eventually provided as commodities by commercial purveyors to whom they are transferred. There are hundreds of lines of genetically modified mice in which specific genes are deleted, overexpressed, or modified to test innumerable biological hypotheses in the living organism. It was initially a surprise when different labs conducting experiments on the "same" lines of mice observed different findings. This may be due to misidentifying the mice, exposure to different diets, or different house environments with respect to ambient temperature, noise levels, or other factors not

previously thought to be highly consequential (Bailoo et al. 2020). For example, 10 years ago my wife's lab conducted a series of exciting mouse experiments on the effect of a specific diet to delay cancer progression. The results were clear and reproduced on several occasions by different members of the lab. However, when her lab moved to a different building with a different mouse facility, even though both facilities were managed by the same staff, the results could never be repeated in the new facility. Despite substantial effort, the reason for this change could not be identified, and the results were never published. It now appears that one common variable that might account for such outcomes relates to the mouse microbiome (Basson et al. 2020).

In addition to issues of training, experimental design, statistics, and reagents, another major cause of irreproducibility results from scientists' unfortunate responses to incentives and disincentives prevalent in todays' research environment (Flier 2019; Nosek, Spies, and Motyl 2012). When scientists face highly consequential grant deadlines, publication decisions, and impending promotions, some forgo best practices. Instead, they cut corners when selecting data to "present a story" that in the end varies from the truth and will not be reproduced by others. The most selective journals frequently require results to be packaged as excessively tidy stories that claim to lack ambiguity or uncertainty, often promising to transform the field and lead to therapeutic outcomes. Despite many insightful and potentially important findings, most papers don't deliver these outcomes. As my Harvard Medical School (HMS) colleague Bill Kaelin (2017), recently awarded the Nobel Prize for work defining how oxygen is sensed in cells suggested, journals too often promote publishing "mansions of straw rather than houses of brick." To illustrate how things have changed, he stated that the paper cited as the basis for his Nobel Prize likely would not make the grade in elite journals today, because it didn't claim to have defined the whole story, which gradually emerged through several less expansive papers.

I believe many decisions to select and present data in questionable ways are made innocently, the scientists believing that they really do capture the underlying reality. But this isn't always the case, and considerable effort has gone into distinguishing what are referred to as "questionable research practices" from research misconduct, the latter formally composed of plagiarism, fabrication, and falsification (Steneck 2006). Misconduct as formally defined accounts for a small fraction of irreproducibility, but the prevalence of both misconduct and questionable research practices are too high (de Vrieze 2021; Fanelli 2009). Here is the key moral question: at what point do sloppiness, wishful thinking, and morally innocent (if inappropriate) selectivity in data presentation transition into a more serious realm, falsification—an accepted hallmark of research misconduct—which, if judged to be present, may end a scientific career? In cases I examined as Dean of HMS, I often found it exceedingly difficult to distinguish intent to deceive—which requires insight into motivations—from honest errors, different

opinions on the approach to experimentation, or all too commonly, self-deception (Flier 2021). As the brilliant and insightful physicist Richard Feynman quipped on self-deception: "The first principle is that you must not fool yourself and you are the easiest person to fool."

The publishing ecosystem also contributes to the problem. Many journals resist accepting papers that report failure to confirm the work of others, even more so if the work was published in their own pages. Despite the importance of scientific journals to the research enterprise, most give low priority to publishing confirmatory studies, considering them less interesting, and of course, less novel. These publication biases cause the literature to have more findings that are "positive," but false.

In order to address these problems, we need to require better training and mentoring on best practices, in areas ranging from experimental design and statistics to the use of reagents to developing manuscripts and negotiating their acceptance. The National Institutes of Health (NIH) have taken note of this problem, and the NIH now requires students and postdocs they fund to take courses on the "responsible conduct of research"; most institutions comply by developing and running live or online modules (NIH 2020). Many trainees fail to complete these modules, and it's unclear whether those who do acquire the necessary skills. Courses and lectures vary in quality and can only achieve so much.

In fact, the greatest influence on trainee behavior almost certainly derives not from formal curricula, but from daily observations of colleagues, mentors, and the broader community in which they work. This "hidden curriculum" is far more impactful in shaping behavior than exposure to slides listing dos and don'ts. Unfortunately, many mentors are themselves poorly trained in these areas. Few take the opportunity to enhance these skills as their careers progress, and it seems no one is responsible for asking them to do so.

The impact on trainees of observing and adopting their mentors' questionable approaches was assessed by Smaldino and McElreath in a 2016 paper entitled "The Natural Selection of Bad Science." Their thesis goes like this: poor experimental design and data analysis promote false positive findings, publication of which is incentivized by the requirements of publishing for career advancement, and many errors in that published work are never identified. Too many scientists succeed despite (periodically) publishing irreproducible science. Successful labs produce more researchers, and some of them will mimic their mentors' methods when running their own labs. And so, this undesirable trait is propagated through a form of adverse social evolution (Grimes, Bauch, and Ioannidis 2018).

Two Stories

I strongly suspect that nearly all bioscientists can recite stories of irreproducible research in their fields, claims that they and their colleagues believe to be

false, a few of which eventually became known as false through retraction, and many more remaining uncertain and contested, producing confusion in the field. During a 45-year career in research, nine as Dean of HMS, I identified dozens of such cases. These arose both within my own field of metabolic research, and through hundreds of faculty research assessments during evaluations for appointment and promotion. Brief stories of two cases, both involving individuals of high academic rank with substantial prior accomplishments, reveal some of the distinct ways that such stories can evolve.

Betatrophin as a Beta-Cell Growth Factor

In April 2013, a paper was published by Yi and Melton in the elite journal *Cell,* entitled "Betatrophin: A Hormone That Controls Pancreatic Beta Cell Proliferation" (Yi, Park, and Melton 2013). The paper claimed to have identified a new molecule secreted by liver whose expression was induced in mice with extreme resistance to insulin. Remarkably, increasing the levels of this molecule in normal mice caused massive replication of insulin-producing beta cells. This paper was a particular milestone for the principal investigator, Doug Melton. For 30 years, Doug had passionately pursued a quest to cure type 1 diabetes. His passion arose from the noble goal of conquering a disease shared by many others. But in this case, it was also personal. This discovery had the potential to produce new treatments for his own children.

Melton held a University Professorship at Harvard, one of only 15 such titles, limited to the most illustrious faculty at any of Harvard's schools. (Melton recently announced that he was leaving his Harvard position to join the biopharmaceutical company Vertex, to more directly engage in development of therapies for diabetes). Doug was the founding cochair of both the Department of Stem Cell and Regenerative Biology and the Harvard Stem Cell Institute. An investigator of the Howard Hughes Medical Institute, a position coveted by bioscientists for its prestige and robust research funding, he's also an elected member of the National Academy of Sciences. But as this story reveals, even researchers with years of experience and a stellar record of repeatedly confirmed accomplishments can publish a result that turns out to be mistaken.

Soon after publication of the betatrophin paper in *Cell,* things started to unravel. Several colleagues in the field were surprisingly measured in their response to the paper, some raising technical questions, others pointing out that betatrophin was not a newly identified molecule, as the paper implied. The same molecule had previously been described by others who had given it another name, and suggested it had other activities, facts not referenced in the *Cell* paper (Quagliarini et al. 2012). The informal buzz in the diabetes community became increasingly negative. Most often when this happens there is no clear resolution. But in this case, it didn't take long for concerns to become public.

On October 24, 2014, a little more than a year after the Yi and Melton paper, *Cell* published another paper by Gusarova and colleagues, "ANGPTL8/Betatrophin Does Not Control Pancreatic Beta Cell Expansion." Nothing subtle about that. The authors were well-known scientists, led by a team from Regeneron Pharmaceuticals, a highly regarded biopharmaceutical company. The authors included Helen Hobbs, an outstanding physician and geneticist from the University of Texas Southwestern Medical Center, whose lab had previously described ANGPTL8 and suggested it acted on lipid metabolism (Quagliarini et al. 2012). This was the same molecule that Melton named *betatrophin*, to the immediate chagrin of Hobbs and colleagues. Other authors included Susan Bonner-Weir, a longstanding member of the Joslin/HMS faculty who has spent most of her career studying how new beta cells are formed, and George Yancopoulos, head scientist at Regeneron and a National Academy member.

The authors didn't mince words. First, they used gene targeting to create mice completely lacking "betatrophin" and showed that when challenged by insulin resistance, these mice grew beta cells as well as normal mice. So betatrophin couldn't be responsible. But there was more. Unlike the original report, they administered ANGPTL8/betatrophin protein to normal mice, and in their hands "betatrophin" had no effect on beta cell growth.

That same issue of *Cell* published another paper on this topic: an invited Perspective article authored by Melton and colleagues (Yi, Park, and Melton 2014). Melton had reviewed the Gusarova paper for *Cell* (not a common occurrence in situations like this), and to his great credit recommended its acceptance. His Perspective recounted the prior work of his group, then stated that subsequent, and then as-yet-unpublished studies showed a far smaller effect of betatrophin than they had initially reported. Melton and his colleagues opined that an effect on beta cell growth might still exist and noted that they had plans to study this.

The response in social media and science publications was swift. While the Perspective appeared to largely accept the negative conclusion, the original paper wasn't immediately retracted. When pressed by science reporters, Doug argued that a retraction wasn't called for. Retractions, he asserted, are only for cases of scientific misconduct, and that was not the case here. It's true that most retractions are associated with misconduct, but retractions are the only available remedy for papers whose major conclusions are discovered to be wrong, regardless of cause (Fanelli, Ioannidis, and Goodman 2018). Neither authors nor journals are happy to publish retractions, and they typically resist doing so.

To his great credit, Doug and colleagues did take the issue very seriously. In July 2016, they published another paper in the journal *PLoS One,* with Jake Kushner from Baylor the senior author. Its title was "Resolving Discrepant Findings on ANGPTL8 in Beta Cell Proliferation: A Collaborative Approach to Resolving the Betatrophin Controversy" (Cox et al. 2016). In this study, several labs cooperated as few do, using a blinded approach to test beta cell response to

a new preparation of betatrophin in mice. The conclusion was clear: no effect of betatrophin on beta cells was detected. The paper did offer one positive outcome, however: it was a rare example of scientists from different labs cooperating productively to resolve a controversial question (Mellers, Hertwig, and Kahneman 2001).

The discussion in the *PLos One* paper contained the following statement: "One of the two main conclusions of the original paper describing the betatrophin hypothesis needs to be withdrawn." And in 2017, the original *Cell* paper was officially retracted. In my view retraction was the appropriate outcome, given that the key finding causing the paper to be published and bringing attention to it was now found, unambiguously, to be untrue (Yi, Park, and Melton 2017). It's possible that a specific circulating factor that accounts for beta cell replication in this model does exist, but if so it awaits discovery, perhaps by a scientist motivated by reading this retraction.

This case study is an example of a responsible scientist addressing a major error openly. The cause of the erroneous claim that betatrophin/ANGPTL8 stimulates beta cell replication remains unknown, but it most likely resulted from a combination of a technical lab error and confirmation bias, causing initial enthusiasm to override scientific skepticism. Thankfully, following publication of a highly credible disconfirming report, this mistake was followed by an admirable effort to acknowledge it and to provide a definitive scientific answer, something rarely seen. Scientists are fallible and mistakes will on occasion be made—even by the most highly skilled practitioners with the best of intentions. The efficient and definitive correction of this mistake should be celebrated as a positive outcome.

GDF11 as an Anti-Aging Rejuvenation Factor for Skeletal Muscle

There has been substantial interest in the causes of aging at the cellular, tissue, and organismal levels. Evidence that systemic factors may influence tissue aging have emerged from the technique of parabiosis, wherein mice are surgically joined so as to have a shared blood circulation (Finerty 1952). Using this technique with parabiosis between young and old mice, it was shown in 2005 that exposure to "youthful blood" could restore a youthful capacity for skeletal muscle regeneration in old mice (Conboy et al. 2005). In 2013, the labs of Richard Lee and Amy Wagers at Harvard collaborated in a report that made several claims: (1) that age-related heart failure was reversed by parabiosis of old to young mice; (2) that levels of the molecule GDF11 in blood fell with aging and were restored after parabiosis; and (3) that administration of recombinant GDF11 for 30 days reversed the age-related cardiac hypertrophy (Loffredo et al. 2013). One year later, in 2014, the same labs published a paper in the journal *Science*, entitled "Restoring Systemic GDF11 Levels Reverses Age-Related Dysfunction in Mouse Skeletal Muscle" (Sinha et al. 2014) This paper claimed that when GDF11 levels were restored in aged mice by either parabiosis or administration

of recombinant GDF11, the young skeletal muscle phenotype was restored. That key aspects of aging might be reversed by administering and restoring youthful levels of a single circulating factor was exceptionally exciting, and of great interest to scientists, physicians, and the population at large. This finding was reported in the *New York Times* on May 4, 2014 (Zimmer 2014). The research led to the launch of a biotechnology company with Harvard patents and several Harvard faculty as founders. Named Elevian, its initially stated aim was to develop GDF11 as a treatment for aging and age-related disorders.

Shortly after publication of the *Science* paper, some in the field questioned the central claim that GDF11 levels fell with aging. They further questioned whether—even if its level did fall—this could explain skeletal muscle loss with aging. The initial reason for skepticism was that GDF11 was a very close homologue of another molecule, GDF8, also known as myostatin (Lee and McPherron 2001). GDF8 and GDF11 had been shown to induce similar cellular signals, phosphorylation of the SMAD2/3 transcription factors, and paradoxically, deficiency of GDF8 was known to cause *increased* muscle size and function, not the decreased muscle function claimed to be a consequence of low GDF11 in aging mice by the Wagers/Lee groups.

Many labs sought to clarify whether GDF11 was in fact a "rejuvenation factor" whose deficiency caused skeletal muscle dysfunction that was reversed by raising its levels (Egerman and Glass 2019; Egerman et al. 2015; Hinken et al. 2016; Kaiser 2015). As reviewed in a paper in 2019, subsequent evidence strongly indicates that both the *Cell* paper, in which it was shown that GDF11 decreased with age, and the *Science* paper, in which it was shown adding back GDF11 improved skeletal muscle regeneration, were flawed, and that their prime conclusions were likely to be incorrect (Egerman and Glass 2019). First, the techniques that were employed in the *Cell* paper to measure GDF11, a fundamental aspect of that paper, also cross-reacted to GDF8, a much more abundant molecule in mouse serum. With more specific methods, it was later shown that it was actually GDF8/myostatin that decreased with age, not GDF11 (Glass 2016; Semba et al. 2019). Studies with specific techniques for measuring GDF11 revealed its levels remain constant or rise—not fall—with age, refuting the fundamental basis for the subsequent work. Second, as expected from knowledge of GDF8 and GDF11 as muscle regulatory factors, administration of GDF11 impaired muscle mass and function, rather than restoring it (Egerman et al. 2015; Hammers et al. 2017). Several labs published reports showing that the levels used in the *Science* paper had no effect on skeletal muscle, and that greater levels of GDF11 actually blocked skeletal muscle regeneration (Egerman et al. 2015). Further, supraphysiological levels of GDF11 induced frank cachexia, a dramatic muscle-wasting phenotype. GDF11 signals in a very similar way to myostatin and activates an almost identical set of genes. Most strikingly as a refutation of the initial claim that GDF11 treatment improved skeletal muscle (opposite to the action of myostatin), replacing

myostatin with GDF11 in the germline of mice showed them to have essentially indistinguishable—not opposite—actions on this tissue in vivo (Lee et al. 2022).

While there is still much to be learned about the biology of GDF11 and its potential therapeutic utility in one or more human disorders, we must conclude today that the main claim of the 2014 *Science* paper regarding GDF11 and skeletal muscle should be—and for the most part is—viewed as erroneous. As stated earlier, it's not unusual that highly novel and surprising findings are false positives that fail to be reproduced. No scientist wishes to make such mistakes, but there is no shame in doing so on occasion while searching for truth in a new and difficult area. Unfortunately, unlike the case of the betatrophin paper, the authors of this GDF11 paper have taken few if any steps to clarify the issue in the light of subsequent reports. Several of their subsequent papers refer to the factor that falls with age as GDF8/GDF11 (as opposed to GDF11), without directly acknowledging their initial misidentification of the factor as GDF11. This perpetuates confusion, allowing those new to the field to draw erroneous conclusions about the role of GDF11, approaches to measuring it, and the logic of pursuing it as a target for drug development.

As of March 2022, the 2014 paper reporting GDF11 as a rejuvenating factor for skeletal muscle in aging has been cited 334 times, the vast majority not referencing issues about its reproducibility. A web page on the Harvard Stem Cell Institute site accessed on June 27, 2022, entitled "Aging and GDF11: What We Know," repeats the claim that GDF11 falls with aging, and its restoration remains an active approach to reversing aging. I believe the field would be enhanced if the Lee/Wagers groups took a more forthright and balanced approach to describing and assessing research published after their 2014 report, and to considering the implications of this work for the validity of their initial hypothesis. Of course, this in no way precludes the possibility that a yet-unidentified circulating factor may reverse aging in one or more tissues, or that GDF11 may have some therapeutic utility. GDF11 has been reported to have beneficial effects in preclinical models of stroke, a finding unrelated to the claims about muscle, and the startup Elevian is pursuing that application of the molecule.

These stories of irreproducible research with distinct outcomes are but two of thousands that might be told. I am aware of far more instances resembling the second case, wherein excellent and highly regarded scientists publish influential studies whose key elements are not reproduced by others. These negative findings may or may not get published, and over time the original findings become viewed as false, though they are never publicly modified or retracted by the original authors. Instead, the previously published claims gradually fade from scientific discourse, as the scientists who published them move on to other topics, often with little or no negative impact on their reputations. Because most irreproducible research garners little or no attention from the scientific community or the public, stories such as these involving high-profile "discoveries" will never be told.

Addressing the Crisis of Reproducibility

To address excessive irreproducibility of the published literature, the research community must first accept the problem as real, and then, using consistent definitions and terminology, commit to doing something about it. There are, however, many obstacles to the success of such efforts.

First, not everyone in the research community accepts that the problem requires such attention; some believe it is overblown. If the problem exists, these people say, it may be restricted to certain fields or institutions, certainly not their own! Though the extent of the problem does vary from field to field, and like all issues its severity may sometimes be overstated, such defensive postures are too often uninformed or self-interested, and will in any case fail to move us in a positive direction.

Other members of the community are concerned that focusing on the problem and taking concerted actions may have unintended consequences, such as "research bureaucrats" jumping in to limit academic freedom, imposing interventions that will increase administrative burdens while failing to improve the quality of research. These concerns, though plausible and requiring attention, shouldn't prevent reasonable actions from being designed and implemented, as we seek to limit ineffective or harmful responses.

Another concern, which I share, is that public attention to irreproducible research could be weaponized by anti-science activists to reduce funding and tarnish the reputation of the research community. This is one reason why many academic leaders are hesitant to take up this issue, preferring (not surprisingly) to highlight the quality and impact of the research at their institutions, which they surely believe, rather than alerting others, including funders, donors, and the general public, of systemic flaws in the ecosystem. But I believe that failing to identify and take responsibility for this problem poses greater risks. In fact, failing to recognize a quite evident problem provides ammunition for zealots to pursue their irrational anti-science agenda.

An additional reason why institutional leaders rarely take the lead on this issue may be seen as a collective action problem, wherein effectively addressing the problem requires cooperation between many institutions. Such coordination and agreement is difficult to achieve, and local exigencies typically incentivize leaders to leave such difficult problems for others to solve. While Dean of HMS, I became increasingly aware of the problem of irreproducibility, as related to my own faculty and of course more broadly. I thought about the issue and consulted others about potential responses. But this topic attained lower priority on the list of urgent issues requiring my attention, and I decided to take it up as a serious project after leaving the dean's office, which I have done, through this essay and others (Flier 2017a, 2017b, 2021).

Once past these objections and limitations to action, corrective steps fall into several broad categories (Ioannidis 2014). These are education and training;

adoption of best practices by academic institutions, funders, and publishers; and shifting academic culture to focus more on quality and reproducibility of research, and less on the number of papers and the impact factors of the journals in which they are published (Moher et al. 2020).

We must enhance the requirement for training on the fundamentals of design, interpretation, and statistical analysis of experimental data, the proper use of diverse reagents, and research and publication ethics. The content of existing programs is of variable quality, and their delivery to the end users is incomplete. Training is insufficiently prioritized by many leaders of the research community, who in addition to having other problems to deal with are heavily incentivized to communicate—to the public and donors—the success of their programs rather than their problems. Perhaps a response from a consortium of leaders might avoid concerns about adverse effects of a single institutional leader getting out front on this issue.

We must seek greater consensus on best practices. Though consensus is lacking in some areas, there are many areas of broad agreement that should receive concerted attention. As one example, institutions should expect their faculty to adopt minimal standards for data archiving and record-keeping, and to demonstrate compliance with open-data standards, to ensure their data is findable, accessible, interoperable, and reusable (so-called FAIR principles; Wilkinson et al. 2016). This may require modest institutional investments to provide both necessary infrastructure and access to trained individuals to facilitate its use. Routinely archiving data according to FAIR principles incentivizes better organization of data and facilitates reproduction of results both within labs and by outside scientists.

Importantly, institutions should review such behaviors and assess the reproducibility of a scientist's work at the time of appointment and promotion, taking appropriate actions if clearly articulated standards are not met (Flier 2017a). During promotion reviews, some people asked to provide confidential comments raise issues of reproducibility, but many with such concerns don't mention them, concerned about consequences for the faculty being reviewed or fearful that confidentiality of their comments might be breached. Funders like the NIH, which support open access practices and research rigor by grantees and their institutions, should make funding decisions dependent on meeting such requirements. The NIH has recently required grant proposals to include specific comments on the "rigor of research," addressing many of the issues discussed here. It is uncertain how such welcome requests will translate into beneficial changes in behavior, but since the NIH funds most bioscience research, the organization might consider allocating a small fraction of its budget to a new intramural program tasked with identifying and prioritizing externally funded research whose finding are contested, and then attempting to reproduce a small number in a manner similar to the Cancer Reproducibility Project. The practical and political challenges of such a program are obvious, but such an effort could nevertheless be helpful to the

scientific community, while sending a strong signal about the NIH's view of the importance of the problem.

The crisis of reproducibility that we face today involves errors that far too often go unacknowledged and uncorrected. Too many scientists and academic institutions sidestep questions about reproducibility rather than address them head on. When other scientists mistakenly accept their validity during their own search for the truth, these errors generate wasted effort. The rate of scientific progress is slowed, and, not surprisingly, mistrust—sometimes justified and sometimes not—of the scientific enterprise grows. To reverse this unfortunate assault on the integrity of biomedical science and to enhance the efficiency of our incredible scientific ecosystem, the culture of irreproducibility must be acknowledged and addressed by participants at every level of the enterprise.

REFERENCES

Bailoo, J. D., et al. 2020. "Effects of Weaning Age and Housing Conditions on Phenotypic Differences in Mice." *Sci Rep* 10 (1): 11684.

Baker, M. 2016. "1,500 Scientists Lift the Lid on Reproducibility." *Nature* 533 (7604): 452–54.

Basson, A. R., et al. 2020. "Artificial Microbiome Heterogeneity Spurs Six Practical Action Themes and Examples to Increase Study Power-Driven Reproducibility." *Sci Rep* 10 (1): 1–19.

Begley, C. G., and L. M. Ellis. 2012. "Drug Development: Raise Standards for Preclinical Cancer Research." *Nature* 483 (7391): 531–33.

Bradbury, A., and A. Pluckthun. 2015. "Reproducibility: Standardize Antibodies Used in Research." *Nature* 518 (7537): 27–29.

Chavalarias, D., et al. 2016. "Evolution of Reporting P Values in the Biomedical Literature, 1990–2015." *JAMA* 315 (11): 1141–48.

Conboy, I. M., et al. 2005. "Rejuvenation of Aged Progenitor Cells by Exposure to a Young Systemic Environment." *Nature* 433 (7027): 760–64.

Cox, A. R., et al. 2016. "Resolving Discrepant Findings on ANGPTL8 in Beta-Cell Proliferation: A Collaborative Approach to Resolving the Betatrophin Controversy." *PLoS One* 11 (7): e0159276.

Davis, R. 2014. *Reproducibility Project: Cancer Biology*. *eLife*, Dec. 10. https://eLifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology.

de Vrieze, J. 2021. "Large Survey Finds Questionable Research Practices Are Common." *Science* 373 (6552): 265.

Dirnagl, U. 2019. "Rethinking Research Reproducibility." *EMBO J* 38 (2): e101117.

Egerman, M. A., and D. J. Glass. 2019. "The Role of GDF11 in Aging and Skeletal Muscle, Cardiac and Bone Homeostasis." *Crit Rev Biochem Mol Biol* 54 (2): 174–83.

Egerman, M. A., et al. 2015. "GDF11 Increases with Age and Inhibits Skeletal Muscle Regeneration." *Cell Metab* 22 (1): 164–74.

Errington, T. M., et al. 2021a. "Challenges for Assessing Replicability in Preclinical Cancer Biology." *eLife* 10.

Errington, T. M., et al. 2021b. "Investigating the Replicability of Preclinical Cancer Biology." *eLife* 10.

Fanelli, D. 2009. "How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data." *PloS One* 4 (5): e5738.

Fanelli, D., J. P. A. Ioannidis, and S. Goodman. 2018. "Improving the Integrity of Published Science: An Expanded Taxonomy of Retractions and Corrections." *Eur J Clin Invest* 48 (4).

Fidler, F., and J. Wilcox. 2018. "Reproducibility of Scientific Results." *Stanford Encyclopedia of Philosophy,* ed. E. N. Zalta. https://plato.stanford.edu/entries/scientific-reproducibility/.

Finerty, J. C. 1952. "Parabiosis in Physiological Studies." *Physiol Rev* 32 (3): 277–302.

Flier, J. 2017a. "Faculty Promotion Must Assess Reproducibility." *Nature* 549 (7671): 133.

Flier, J. S. 2017b. "Irreproducibility of Published Bioscience Research: Diagnosis, Pathogenesis and Therapy." *Mol Metab* 6 (1): 2–9.

Flier, J. S. 2019. "Credit and Priority in Scientific Discovery: A Scientist's Perspective." *Perspect Biol Med* 62 (2): 189–215.

Flier, J. S. 2021. "Misconduct in Bioscience Research: A 40-Year Perspective." *Perspect Biol Med* 64 (4): 437–56.

Glass, D. J. 2010. "A Critique of the Hypothesis, and a Defense of the Question, as a Framework for Experimentation." *Clin Chem* 56 (7): 1080–85.

Glass, D. J. 2016. "Elevated GDF11 Is a Risk Factor for Age-Related Frailty and Disease in Humans." *Cell Metab* 24 (1): 7–8.

Goodman, S. N., D. Fanelli, and J. P. Ioannidis. 2016. "What Does Research Reproducibility Mean?" *Sci Transl Med* 8 (341): 341ps12.

Grimes, D. R., C. T. Bauch, and J. P. A. Ioannidis. 2018. "Modelling Science Trustworthiness Under Publish or Perish Pressure." *R Soc Open Sci* 5 (1): 171511.

Gusarova, V., et al. 2014. "ANGPTL8/Betatrophin Does Not Control Pancreatic Beta Cell Expansion." *Cell* 159 (3): 691–96.

Hammers, D. W., et al. 2017. "Supraphysiological Levels of GDF11 Induce Striated Muscle Atrophy." *EMBO Mol Med* 9 (4): 531–44.

Harvard Stem Cell Institute. 2022. "Aging and GDF11: What We Know." https://hsci.harvard.edu/aging-and-gdf11-what-we-know.

Hinken, A. C., et al. 2016. "Lack of Evidence for GDF 11 as a Rejuvenator of Aged Skeletal Muscle Satellite Cells." *Aging Cell* 15 (3): 582–84.

Ioannidis, J. P. 2005. "Why Most Published Research Findings Are False." *PLoS Med* 2 (8): e124.

Ioannidis, J. P. 2014. "How to Make More Published Research True." *PLoS Med* 11 (10): e1001747.

Kaelin, W. G., Jr. 2017. "Publish Houses of Brick, Not Mansions of Straw." *Nature* 545 (7655): 387.

Kaiser, J. 2015. "Regenerative Medicine: 'Rejuvenating' Protein Doubted." *Science* 348 (6237): 849.

Kane, P. B., and J. Kimmelman. 2021. "Is Preclinical Research in Cancer Biology Reproducible Enough?" *eLife* 10.

Kimmelman, J., J. S. Mogil, and U. Dirnagl. 2014. "Distinguishing Between Exploratory and Confirmatory Preclinical Research Will Improve Translation." *PLoS Biol* 12 (5): e1001863.

Lee, S.-J., and A. C. McPherron. 2001. "Regulation of Myostatin Activity and Muscle Growth." *Proc Natl Acad Sci* 98 (16): 9306–11.

Lee, S.-J., et al. 2022. "Functional Replacement of Myostatin with GDF-11 in the Germline of Mice." *Skeletal Muscle* 12 (1): 1–12.

Lewandowsky, S., and K. Oberauer. 2020. "Low Replicability Can Support Robust and Efficient Science." *Nat Commun* 11 (1): 358.

Loffredo, F. S., et al. 2013. "Growth Differentiation Factor 11 Is a Circulating Factor That Reverses Age-Related Cardiac Hypertrophy." *Cell* 153 (4): 828–39.

Lucey, B. P., W. A. Nelson-Rees, and G. M. Hutchins. 2009. "Henrietta Lacks, HeLa Cells, and Cell Culture Contamination." *Arch Pathol Lab Med* 133 (9): 1463–67.

Mellers, B., R. Hertwig, and D. Kahneman. 2001. "Do Frequency Representations Eliminate Conjunction Effects? An Exercise in Adversarial Collaboration." *Psychol Sci* 12 (4): 269–75.

Moher, D., et al. 2020. "The Hong Kong Principles for Assessing Researchers: Fostering Research Integrity." *PLoS Biol* 18 (7): e3000737.

National Academies of Sciences Engineering and Medicine (NASEM). Committee on Reproducibility and Replicability in Science, et al. 2019. *Reproducibility and Replicability in Science: A Consensus Study Report of the National Academies of Sciences, Engineering, and Medicine*. Washington, DC: National Academies Press.

National Institutes of Health (NIH). 2020. *Responsible Conduct of Research Training*. NIH Office of Intramural Research, Dec. 1. oir.nih.gov/sourcebook/ethical-conduct/responsible-conduct-research-training.

Nosek, B. A., J. R. Spies, and M. Motyl. 2012. "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability." *Perspect Psychol Sci* 7 (6): 615–31.

Prinz, F., T. Schlange, and K. Asadullah. 2011. "Believe It or Not: How Much Can We Rely on Published data on Potential Drug Targets?" *Nat Rev Drug Discov* 10 (9): 712.

Quagliarini, F., et al. 2012. "Atypical Angiopoietin-Like Protein That Regulates ANGPTL3." *Proc Natl Acad Sci USA* 109 (48): 19751–56.

Semba, R. D., et al. 2019. "Relationship of Circulating Growth and Differentiation Factors 8 and 11 and Their Antagonists as Measured Using Liquid Chromatography–Tandem Mass Spectrometry with Age and Skeletal Muscle Strength in Healthy Adults." *J Gerontology A* 74 (1): 129–36.

Sinha, M., et al. 2014. "Restoring Systemic GDF11 levels Reverses Age-Related Dysfunction in Mouse Skeletal Muscle." *Science* 344 (6184): 649–52.

Skloot, R. 2011. *The Immortal Life of Henrietta Lacks*. New York: Broadway Books.

Smaldino, P. E., and R. McElreath. 2016. "The Natural Selection of Bad Science." *R Soc Open Sci* 3 (9): 160384.

Souren, N. Y., et al. 2022. "Cell Line Authentication: A Necessity for Reproducible Biomedical Research." *EMBO J* e111307.

Steneck, N. H. 2006. "Fostering Integrity in Research: Definitions, Current Knowledge, and Future Directions." *Sci Eng Ethics* 12 (1): 53–74.

Szucs, D., and J. P. A. Ioannidis. 2017. "When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment." *Front Hum Neurosci* 11: 390.

Tatsioni, A., N. G. Bonitsis, and J. P. Ioannidis. 2007. "Persistence of Contradicted Claims in the Literature." *JAMA* 298 (21): 2517–26.

Wilkinson, M. D., et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Sci Data* 3: 160018.

Yi, P., J. S. Park, and D. A. Melton. 2013. "Betatrophin: A Hormone That Controls Pancreatic Beta Cell Proliferation." *Cell* 153 (4): 747–58.

Yi, P., J. S. Park, and D. A. Melton. 2014. "Perspectives on the Activities of ANGPTL8/Betatrophin." *Cell* 159 (3): 467–68.

Yi, P., J. S. Park, and D. A. Melton. 2017. "Retraction Notice to: Betatrophin: A Hormone that Controls Pancreatic Beta Cell Proliferation." *Cell* 168 (1–2): 326.

Zarin, D. A., et al. 2019. "10-Year Update on Study Results Submitted to ClinicalTrials.gov." *N Engl J Med* 381 (20): 1966–74.

Zimmer, C. 2014. "Young Blood May Hold Key to Reversing Aging." *NY Times,* May 4.