Published in final edited form as:
Circulation. 2012 March 13; 125(10): 1211–1214. doi:10.1161/CIRCULATIONAHA.112.098244.

Experimental Irreproducibility: Causes, (Mis)interpretations, and Consequences

Joseph Loscalzo, M.D., Ph.D.
Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

> "[T]he work of science has nothing whatever to do with consensus. Consensus is the business of politics. Science, on the contrary, requires only one investigator who happens to be right, which means that he or she has results that are verifiable by reference to the real world. In science consensus is irrelevant. What is relevant is reproducible results."
>
> --Michael Crichton, from "Aliens Cause Global Warming," a lecture given at California Institute of Technology, Pasadena, California, January 17, 2003
>
> "Trust everybody, but cut the cards."
>
> --Finley Peter Dunne, 19th century American journalist

Experimental reproducibility is the coin of the scientific realm. The extent to which measurements or observations agree when performed by different individuals defines this important tenet of the scientific method. The formal essence of experimental reproducibility was born of the philosophy of logical positivism or logical empiricism, which purports to gain knowledge of the world through the use of formal logic linked to observation (1). A key principle of logical positivism is verificationism, which holds that every truth is verifiable by experience. In this rational context, truth is defined by reproducible experience, and unbiased scientific observation and determinism are its underpinnings.

From a more practical perspective, reproducibility is a means by which to reduce the tentative nature of an initial scientific observation. The implicit assumptions of tests of reproducibility are that if an initial observation is found to be reproducible, then it must be true; and if an initial observation is found not to be reproducible, then it must be false. While the logic of these concepts is unimpeachable, we should not conflate scientific truth and reproducibility. As Jonah Lehrer pointed out recently, "Just because an idea is true doesn't mean it can be proved. And just because an idea can be proved doesn't mean it's true. When the experiments are done, we still have to choose what to believe (2)."

The assumption that objectively true scientific observations must be reproducible is implicit, yet direct tests of reproducibility are rarely found in the published literature. This lack of published evidence of reproducibility stems from the limited appeal of studies reproducing earlier work to most funding bodies and to most editors. Furthermore, many readers of scientific journals — especially of higher impact journals — assume that if a study is of sufficient quality to pass the scrutiny of rigorous reviewers, it must be true; this assumption is based on the inferred equivalence of reproducibility and truth described above.

Address all correspondence to: Dr. Joseph Loscalzo, Department of Medicine, Brigham and Women's Hospital, 75 Francis Street, Boston, Massachusetts 02115, jloscalzo@partners.org.

It is important to distinguish between experimental reproducibility and replicability. Most investigators believe that these two terms are identical; however, Drummond argues that they are distinct and different (3). Replicability is assessed by performing an experiment under exactly the same conditions at different times, while reproducibility is assessed by performing similar, but not identical, experiments at different times, in different locations, and under somewhat different experimental conditions. In this way, replicability reflects the technical stringency or precision of a specific experiment, while reproducibility reflects the fundamental accuracy of an experimental observation. These distinctions are consistent with the view that a precisely conducted experiment can be inaccurate, and an accurate experiment may be performed imprecisely — especially in biomedicine where many factors can account for irreproducible results. It is to a review of these factors that I now turn.

Statistical flaws are a major cause of irreproducible results in all types of biomedical experimentation. These include errors in trial design, data analysis, and data interpretation. Based on statistical simulations, Ioannidis argued that for most study designs and settings, it is more likely that a research outcome is false than true (4). He pointed out that false positive outcomes are more likely to occur with smaller study size, smaller effect size, a greater number and lesser preselection of applied statistical tests, and the presence of bias, among others. These conclusions are intuitively obvious, and are key considerations in designing trials to minimize the type I error. Failure to assume the null hypothesis and selective data presentation are important flaws in experimental design and data handling that can also contribute to the generation of results subsequently found to be irreproducible.

Bias clearly plays an important role in promoting false positive results. Reasons for bias include a lack of experimental equipoise leading to an impassioned belief in one particular experimental outcome clouding objectivity; perceived pressure to publish for academic advancement or to enhance the likelihood of competing successfully for grant funding (5); and the lack of appeal of negative (or neutral) studies in most high-impact journals. This last point has been emphasized recently in a study illustrating the inverse relationship between the scientific hierarchy (physical sciences at the top, social sciences at the bottom, and biological sciences in between) and the publication frequency of ‘positive’ results (6).

Another key statistical aspect of experimental design and execution — especially in biomedicine — is that of biological noise, or random fluctuations in a biological event or measurement. The amplitude and frequency of these fluctuations, when of sufficient magnitude, can contribute to experimental irreproducibility, despite efforts to ascertain these ‘naturally occurring’ stochastic events carefully. Experimental observations in any scientific discipline can be subject to noise; however, in biomedicine, the signal-to-noise ratio is often considerably lower than might be found in other disciplines owing to the greater variability of system determinants than exists, for example, in the physical sciences. Examples of major causes of variability in biological systems include contaminants in ‘purified’ protein preparations (even purified recombinant proteins); phenotypic differences between apparently identical cells in culture owing to microenvironmental differences in gene expression; and subtle differences in metabolic performance caused by modest local changes in temperature or pH, among others. There is simply no way to reduce this level of biological noise experimentally: it must be subsumed into the accepted natural experimental variance of the system under study, and it must be ascertained in order to determine the significance of a signal recorded in any experimental system.

While all experimental studies are theoretically at risk for generating irreproducible results as a consequence of any of these statistical flaws, there are differences between basic and clinical studies in this regard. In basic or early translational experiments, there is a greater degree of control of experimental conditions and a smaller sample size than in clinical trials.

As a result, basic studies may be affected to a greater extent by biological noise than clinical trials, while clinical trials may be at greater risk of bias (and confounding) than basic studies.

One last statistical determinant of experimental irreproducibility is the rare reproducible exception, memorably depicted as the black swan in Taleb's book by the same name (7). This outcome is theoretically reproducible, but so rare that a sufficient number of experiments needed to determine its frequency exceeds feasibility. Owing to its inherent rarity, it is effectively impossible to explain away an irreproducible result on the basis of a rare, 'singular' event; this explanation is one of exclusion for which no supportive evidence can be marshaled.

Technical aspects of the conduct of an experiment represent yet another major source of irreproducible outcomes, especially in basic science. While most authors of scientific publications take great pains to elucidate in detail the methods of an experiment, recognizing that only by so doing can one facilitate experimental reproducibility in the hands of other investigators, subtle technical aspects are generally not described and can be a common cause of inter-laboratory differences in experimental results. When one laboratory cannot reproduce the published results of another, reagents can be shared and, on occasion, personnel may be exchanged in order to observe directly how an experiment is performed in the reporting laboratory. Experimentalists are often unaware of a very nuanced aspect of the execution of an experiment, which can make an extraordinary difference in the outcome and which can often only be sorted out using this collaborative approach. In addition to subtle technical features of the conduct of an experiment, the complexity of experimental design or of the system under study itself may increase the likelihood of irreproducibility.

The last, and most concerning, explanation for irreproducible results is scientific fraud. The precise prevalence of fraud in publications is not known, of course; however, a recent study by Fanelli raises concerning issues about its estimates (8). In this study, the author collated the results of 21 surveys in a systematic review and 18 surveys in a meta-analysis. She found that 1.97% of respondents admitted to having fabricated, falsified, or modified data or published results on at least one occasion; up to 33.7% of respondents admitted to other questionable practices that did not achieve this level of serious misconduct, but are still of concern in a discipline that purports to seek unbiased truth. Clearly, these are troubling statistics that, even if found to be flawed by confounding or bias, warrant further investigation as to cause and prevention.

Fraud in scientific research is certainly not a new phenomenon, with individual cases of varying degrees of notoriety increasingly gaining public recognition through the lay press. Journal editors, ever sensitive to their role as arbiters of scientific truth, proclaim the need to guard against scientific fraud, and propose mechanisms by which to minimize it (9,10). The concept of scientific fraud connotes an assault on the very integrity of the discipline itself, and, for this reason, creates a visceral reaction in any individual who learns of it, scientist, editor, layperson, or policymaker. It is important, however, to avoid equating failure to reproduce a scientific finding with scientific fraud; unfortunately, the conflation of these two phenomena is the norm in the current era.

For example, Prinz and colleagues recently published an analysis from their experience in drug development at Bayer Healthcare in which they surveyed company investigators who were charged with experimentally validating published data on potential new drug targets (11). They observed that a mere 21% of the published observations were reproducible in their hands. While they cautiously described a range of possible explanations for this irreproducibility, and emphasized that they "are not reporting fraud, but a lack of

reproducibility," the hyperbolic title of their report— "Believe it or not: how much can we rely on published data on potential drug targets?"-- can be interpreted to suggest that the authors of the publications whose results cannot be reproduced are not to be believed. An observation that is not to be believed is by definition false, and, therefore, authors of such publications can be viewed as intending to deceive the scientific community, rather than simply generating results that are irreproducible for less nefarious reasons.

An accompanying editorial attempts to put these observations in appropriate context (12). The editorialist points out that the pharmaceutical industry would clearly prefer a better level of reproducibility and more predictability in its efforts to identify viable drug targets. Complete alignment between discovery and development, however, is unrealistic given the uncontrollable determinants of reproducibility discussed above. Yet, improvements in the likelihood of reproducibility and its predictors certainly warrant further consideration.

Over the last decade, the number of journal articles world-wide doubled from 1.1 to 1.9 million. Of these, approximately 0.5 million articles are published in the field of biomedicine, for an output of ~1,400 papers per day. Given this volume alone, is it any wonder that an increasing number of papers are being published that contain irreproducible results? Furthermore, with the growth in search technologies, a greater number of articles is being retracted over time--~40/year in the late 1990s, ~300 in 2010, and ~400 in 2011 (13). The reasons for these retractions include plagiarism, data manipulation (especially in figures), and proven data falsification; however, irreproducibility resulting from 'innocent' causes may also be included in this pool of retracted publications without being recognized as such. With as many as ~50% of all articles listed in PubMed never cited (14,15) at all, one can conclude either that the work is of minimal significance and not worthy of further pursuit or that it has been pursued and could not be reproduced. The extent to which these two explanations account for this statistic has not been (nor cannot easily be) easily determined.

Editors have much at stake in the review process. Peer review is its cornerstone; yet peer review is itself an imperfect process. At *Circulation*, we recognize this imperfection and do so remaining keenly aware of the ideal objective of editorial review: to publish only those papers of maximal impact that will hold up to long-term scrutiny with a high likelihood of reproducibility. To meet this goal, we pride ourselves in the many layers of review a manuscript receives in parallel with and beyond peer review, including discussion at our weekly editorial board meeting, careful review by associate editors, and rigorous statistical review. In addition, we frequently publish accompanying editorials to put the results of a study in proper context and perspective for the interested reader and investigator. We firmly believe that this multistep process, while not eliminating the risk of publishing data that are irreproducible in papers that are later retracted, clearly offers the care necessary to minimize this risk. Careful editorial oversight, thus, has inherent value which all authors should keep in mind. Publishing original research without rigorous review (as occurs in some open access journals and as has been gaining increasing support in some quarters of the scientific community of late) runs a greater risk of irreproducibility than does publishing original research with rigorous review (although objective evidence to support this conclusion is at the current moment limited to editorial experience). The experience, insight, and intuition about what is and is not experimentally and biologically feasible that careful reviews provide limit the extent to which others are led down the path to irreproducibility, preventing wasted time, effort, resources, and energy, as well and importantly as preventing the exposure of patients to useless or risky therapies.

This issue of irreproducibility has clearly caught the attention of the lay press, as well, in which articles on the topic abound (e.g., 16,17). Predictably, these articles overemphasize

the potential for fraud as a cause for irreproducibility, and either describe or insinuate that the problem of irreproducibility is a deception perpetrated widely by the practitioners of the discipline. With titles like, “Lies, Damned Lies, and Medical Science” (16), who can help but be persuaded that irreproducibility and fraud are one and the same?

For a discipline that prides itself on discovering objective truth, anything that threatens this aim threatens the very substance of the discipline. Irreproducible results, no matter the cause, are one such threat; yet, owing to a range of processes and events that are rife in biological systems and their experimental exploration, it should neither be unexpected nor overly concerning when results cannot be reproduced. The scientific community generally responds initially to these situations in a rational, objective way, working with the original reporting authors to understand precisely how the experiments were performed, the data handled, and the results interpreted. Thus, to minimize the harm to reputations and careers that unbridled criticism in the lay press fosters when these stories initially break, the public should be apprised of the scientific process, its uncertainties, and the methods that the scientific community uses to address those uncertainties before accepting an observable truth as such or before rejecting a reported truth as false, or worse, fraudulent.

Equally important, the editorial community must improve its strategies for identifying methods that engender a lower likelihood of irreproducibility; improve the review process to identify those flaws in statistical analysis, experimental methods, or study design that increase the risk of irreproducibility; and encourage the publication of studies that reproduce — or fail to reproduce — previously published work. Efforts by journals of the greatest impact to help assure the accuracy of a reported observation often include requiring that the authors perform a series of related experiments in different systems (different cell types, different animal models, combinations of cell types and animal models), arguing that internal consistency in the results across different experimental platforms renders th conclusion of the study more likely to be believable. However, believable studies are not always truthful, and the truth is not always believable—from an experimental perspective or otherwise. In addition, striving for internal consistency across experimental methods also limits the extent to which one can explore a single method in painstaking detail. This limitation can increase the likelihood of excluding data obtained by one method or in one system when those data do not ‘fit the hypothesis.’ Thus, using different experimental platforms runs the risk of providing a false sense of reassurance, potentially masking by exclusion the true ‘biological noise’ of an experimental system and, as a result, obscuring the ‘truthful’ outcome of a series of experiments. As a result, future attempts at reproducing the reported results may be rendered futile.

Are there mechanisms by which the scientific method can offer greater assurance about reproducibility, regardless of the etiology of the problem? Providing broader access to data sets is one such approach that has been touted of late—in genomic science (18), computational science (19), and drug discovery (20); in the latter case, crowd sourcing may also be a useful and emerging open-innovation approach that promotes comparatively unbiased collaboration and complementarity among interested investigators (20).

In the end, however, it is important to remember that science is an imperfect enterprise, born of the struggle against authority, encouraging its practitioners to question and doubt. As Richard Feynman so clearly put it, “Scientific knowledge is a body of statements of varying degrees of certainty—some most unsure, some nearly sure, but none absolutely certain” (21). As investigators, we must balance healthy skepticism with an acceptance of the stochastic contributors to a natural frequency of experimental irreproducibility in order to decide whether or not an observation is right, whether or not we have struck on objective truth, and whether or not we have done so with a reasonable degree of certainty. This is the

scientific process in practice. This practice should be promulgated to the scientific community to minimize the likelihood that an irreproducible result will be immediately tainted as a fraudulent result, and to the public to ensure that their expectations of the scientific method and its outcomes are realistic.

Acknowledgments

The author wishes to thank Elliott Antman, Anita Loscalzo, Jane Newburger, Marc Pfeffer, and Joseph Vita for helpful comments.

Funding Sources: This work was supported in part by NIH grants HL61795, HL70819, HL48743, HL107192, and HL108630.

References

1. Carnap, R. Der Logische Aufbau der Welt. Leipzig: Felix Meiner Verlag; 1928.
2. Lehrer, J. The truth wears off. The New Yorker; December 13. 2010
3. Drummond, C. Replicability is not reproducibility: nor is it good science. Proc Eval Methods Mach Learn Workshop 26th ICML; Montreal, Quebec, Canada. 2009. http://ww.csi.uottawa.ca/~cdrummon/pubs/ICMLws09.pdf
4. Ioannidis JPA. Why most published research findings are false. PLoS Medicine. 2005; 2:e124. [PubMed: 16060722]
5. Fanelli D. Do pressures to publish increase scientists' bias? An empirical support from US states data. PLoS One. 2010; 5:310271.
6. Fanelli D. "Positive" results increase down the hierarchy of the sciences. PLoS One. 2010; 5:e10068. [PubMed: 20383332]
7. Taleb, NN. The Black Swan: The Impact of the Highly Improbable. New York: Random House; 2007.
8. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One. 2009; 4:e5738. [PubMed: 19478950]
9. Crocker J, Cooper ML. Addressing scientific fraud. Science. 2011; 334:1182. [PubMed: 22144584]
10. De Caterina R, Griffioen AW, Porreca F. Fraud in biomedical research—the role of journal editors. Life Sciences. 2011; 89:755–7566. [PubMed: 21983417]
11. Prinz F, Schlange F, Asadallah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature Rev Drug Disc. 2011; 10:712.
12. The Editors. Nature Med. Vol. 10. 2011. Drug targets slip-sliding away; p. 1155
13. Basken, P. Surge in journal retractions may mask decline in actual problems. Chronicle of Higher Education; January 13. 2012
14. Garfield E. The history and meaning of the journal impact factor. JAMA. 2006; 295:90–93. [PubMed: 16391221]
15. McAlister FA, Lawson FME, Good AH, Armstrong PW. Evaluating research in cardiovascular medicine: citation counts are not sufficient. Circulation. 2011; 123:1038–1043. [PubMed: 21382905]
16. Freedman, DH. Lies, damned lies, and medical science. The Atlantic; November. 2010
17. Naik, G. Scientists' elusive goal: reproducing study results. Wall Street Journal; December 2. 2011
18. Ioannidis JPA, Khoury MJ. Improving validation practices in 'omics' research. Science. 2011; 334:1230–1232. [PubMed: 22144616]
19. Peng RD. Reproducible research in computations science. Science. 2011; 334:1226–1227. [PubMed: 22144613]
20. Lessl M, Bryans JS, Richards D, Asadullah K. Crowd sourcing in drug discovery. Nature Revs Drug Disc. 2011; 10:241–242.
21. Feynman, R. The Value of Science. Address to the National Academy of Sciences; USA, Autumn: 1955.