SCIENTIFIC INTEGRITY

What does research reproducibility mean?

Steven N. Goodman,* Daniele Fanelli, John P. A. Ioannidis

The language and conceptual framework of "research reproducibility" are nonstandard and unsettled across the sciences. In this Perspective, we review an array of explicit and implicit definitions of reproducibility and related terminology, and discuss how to avoid potential misunderstandings when these terms are used as a surrogate for "truth."

Concern about the reproducibility of scientific research has been steadily rising recently with reports that the results of experiments in numerous domains of science could not be replicated (*1*, *2*). Whereas problems in biomedical research have garnered most of the attention, concerns have touched almost every field in the biological and social sciences and beyond (*3*) (Fig. 1). As the movement to examine and enhance the reliability of research expands, it is important to note that some of its basic terms—reproducibility, replicability, reliability, robustness, and generalizability—are not standardized. This diverse nomenclature has led to confusion, both conceptual and operational, about what kind of confirmation is needed to trust a given scientific result. Here, we dissect this vocabulary, explore the reasons for the confusion, and offer a framework to improve both communication and understanding.

DEFINING THE TERMS

Although the importance of multiple studies corroborating a given result is acknowledged in virtually all of the sciences (Fig. 1), the modern use of "reproducible research" was originally applied not to corroboration, but to transparency, with application in the computational sciences. Computer scientist Jon Claerbout coined the term and associated it with a software platform and set of procedures that permit the reader of a paper to see the entire processing trail from the raw data and code to figures and tables (*4*). This concept has been carried forward into many data-intensive domains, including epidemiology (*5*), computational biology (*6*), economics (*7*), and clinical trials (*8*). According to a U.S. National Science Foundation (NSF) subcommittee on replicability in science (*9*), "*reproducibility* refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results…. Reproducibility is a minimum necessary condition for a finding to be believable and informative."

Documenting this kind of reproducibility thus requires, at minimum, the sharing of analytical data sets (original raw or processed data), relevant metadata, analytical code, and related software. Reproducibility defined in this way mainly addresses issues of trust that data and analyses are as represented. The definition does not specify to what extent deviations are acceptable. Such reproducibility does not add new evidential weight, although greater subjective weight is often accorded to evidence that is more highly trusted. New evidence is provided by new experimentation, defined in the NSF report as "replicability," which refers to "the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected."

Although the preceding conceptual distinctions might seem clear, the definitions do not provide clear operational criteria for what constitutes successful replication or reproduction. Furthermore, the terminology is not universally used, and sometimes the meanings above are reversed. Consider the language of Francis Collins, director of the U.S. National Institutes of Health (NIH), in his commentary on plans to enhance research reproducibility (*10*):

> "… a complex array of other factors seems to have contributed to the lack of reproducibility. Factors include poor training of researchers in experimental design, increased emphasis on making provocative statements rather than presenting technical details, and publications that do not report basic elements of experimental design. Some irreproducible reports are probably the result of coincidental findings that happen to reach statistical significance, coupled with publication bias. Another pitfall is overinterpretation of creative 'hypothesis-generating' experiments, which are designed to uncover new avenues of inquiry rather than to provide definitive proof for any single question. Still, there remains a troubling frequency of published reports that claim a significant result, but fail to be reproducible."

This short passage covers a wide range of issues subsumed under the rubric of reproducibility: design, reporting, analysis, interpretation, and corroborating studies (that is, replication, as previously defined). If one looks at the terminology being used across the scientific literature, one finds similar variation and intermingling of concepts. For example, the largest-scale attempt to replicate experiments in psychology was published with the title "Estimating the reproducibility of psychological science," (*2*) clearly allying the term "reproducibility" with the conduct of new studies.

One notable absence from this diverse lexicon is the word "truth." The fundamental concern of Collins and others is, in fact, not reproducibility per se, but whether scientific claims based on scientific results are true. Below, we discuss how treating reproducibility as an end in itself—rather than as an imperfect surrogate for scientific truth—is partly responsible for the current terminological and operational morass, and suggest how we can benefit by refocusing on cumulative evidence and truth.

A NEW LEXICON FOR RESEARCH REPRODUCIBILITY

We start the process of clarification by proposing a new terminology to distinguish between the various interpretations of reproducibility. Rather than offer new technical meanings for words whose common language interpretations are nearly identical (such as reproducibility, replicability, and repeatability), we propose to ally the word reproducibility—currently the most widely used single term in this domain—with descriptors for the underlying construct. This yields three terms: methods reproducibility, results reproducibility, and inferential reproducibility. Although we apply these terms mainly to the biomedical field, they have utility across many domains of science, each of which has different conventions and cultures about how

Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA 94305, USA.
*Corresponding author. Email: steve.goodman@stanford.edu

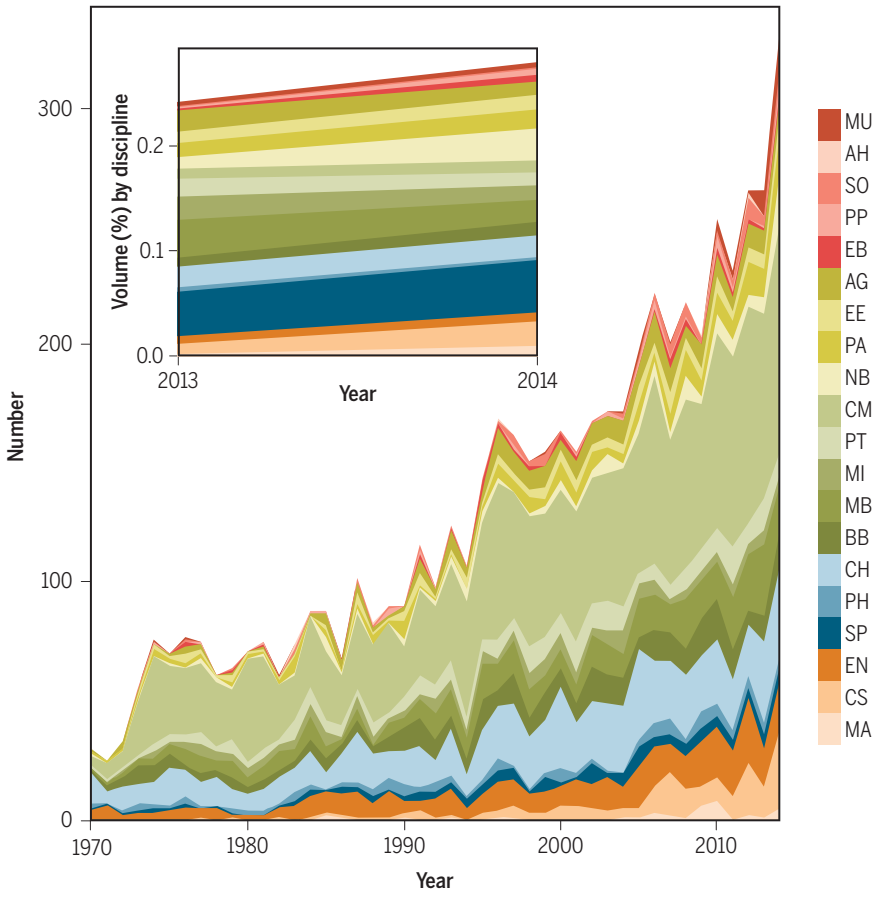

Fig. 1. Reports rising. Number of publications recorded in Scopus that have, in the title or abstract, at least one of the following expressions: research reproducibility, reproducibility of research, reproducibility of results, results reproducibility, reproducibility of study, study reproducibility, reproducible research, reproducible finding, or reproducible result. Papers are classified by discipline on the basis of the journal, following an adaptation and expansion of Thomson Reuters' Essential Science Indicators classification system. Journals not included in the latter database were hand-classified on the basis of their name. The subplot reports the percentage over the total number of records for each discipline, in the last 2 years of the series. Disciplines legend: MA, mathematics; CS, computer sciences; EN, engineering; SP, space science; PH, physics; CH, chemistry; BB, biology and biochemistry; MB, molecular biology; MI, microbiology; PT, pharmacology and toxicology; CM, clinical medicine; NB, neurobiology and behavior; PA, plant and animal sciences; EE, environment and ecology; AG, agricultural sciences; EB, economics and business; PP, psychology and psychiatry; SO, social sciences, general; AH, arts and humanities; MU, multidisciplinary. The time series was truncated at 2014.

to handle the role of chance, the level of certainty required for making published claims, and the adopted criteria for "proof" (Table 1) (*11*).

Methods reproducibility is meant to capture the original meaning of reproducibility, that is, the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results. Results reproducibility refers to what was previously described as "replication," that is, the production of corroborating results in a new study, having followed the same experimental methods. Inferential reproducibility, not often recognized as a separate concept, is the making of knowledge claims of similar strength from a study replication or reanalysis. This is not identical to results reproducibility, because not all investigators will draw the same conclusions from the same results, or they might make different analytical choices that lead to different inferences from the same data. Here, we explore the definitions and operational complexities of each of these concepts.

Methods reproducibility

Methods reproducibility refers to the provision of enough detail about study procedures and data so the same procedures could, in theory or in actuality, be exactly repeated. Operationally, this can mean different things in different sciences. In the biomedical sciences, this means, at minimum, a detailed study protocol, a description of measurement procedures, the data gathered, the data used for analysis with descriptive metadata, the analysis software and code, and the final analytical results. In laboratory science, how key reagents and biological materials were created or obtained can be critical. In theory, these requirements are clear, but in practice, the level of procedural detail needed to describe a study as "methodologically reproducible" does not have consensus. For example, the detection of batch effects, which have been responsible for a number of high-visibility claims and retractions, can require information on exactly which samples were tested on which machine in what order and on what day, together with calibration data. This level of detail is typically not provided in publications and is not always retained by the investigator.

In the clinical sciences, the definition of which data need to be examined to ensure reproducibility can be contentious. The relevant data could be anywhere along the continuum from the initial raw measurement (such as a pathology slide or image), to the interpretation of those data (the pathologic diagnosis), to the coded data in the computer analytic file. Many judgments and choices are made along this path and in the processes of data cleaning and transformation that can be critical in determining analytical results. Last, even if there is consensus on the appropriate analytical data set, methodologic reproducibility requires an understanding of which and how many analyses were performed and how the particular analyses reported in a published paper were chosen. So, whether a particular study is to be considered methodologically reproducible is contingent on whether there is general agreement about the level of detail needed in the description of the measurement process, the degree of processing of the raw data, and the completeness of the analytic reporting.

Results reproducibility

Results reproducibility (previously described as replicability) refers to obtaining the same results from the conduct of an independent study whose procedures are as closely matched

Table 1. Examples of differences that affect the approach to reproducibility in distinct scientific domains.

Degree of determinism
Signal to measurement-error ratio
Complexity of designs and measurement tools
Closeness of fit between hypothesis and experimental design or data
Statistical or analytic methods to test hypotheses
Typical heterogeneity of experimental results
Culture of replication, transparency, and cumulating knowledge
Statistical criteria for truth claims
Purposes to which findings will be put and consequences of false conclusions

to the original experiment as possible. As with methods reproducibility, this might be clear in principle but is operationally elusive. The problem arises in settings where there is substantial random error in any result, making unclear the criteria for considering results to be "the same." The intuition and logic of results reproducibility are derived from systems that are deterministic or for which the signal-to-error ratio is exceedingly high. But, when the same intuition and logic are applied to studies with substantive stochastic components, the paradigm of accumulating evidence might be more appropriate than any binary criteria for successful or unsuccessful replication.

In a deterministic system (for example, computational research), the outcome is determined by the initial conditions. Methods reproducibility is often demonstrated through results reproducibility because the two are linked by determinacy—the signal-to-noise ratio is effectively infinite. A single failure to reproduce the original results with identical inputs casts doubt on the methodology and on any predictions (*12*).

Closely related is a proof-of-principle study, which demonstrates a new phenomenon not previously observed; for example, delivery of the first normal, live-born infant derived from in vitro fertilization or a first case of human limb regeneration would be sufficient to show that such phenomena are possible. That said, a first demonstration will not be accepted without intensive, independent scrutiny of the methods employed and the outcomes claimed, in order to rule out the possibility of misconduct, selective reporting, or procedural compromise. Failure to replicate the phenomenon under circumstances that preclude ancillary causes (for example, mistaken diagnosis, faulty procedures, measurement error, biased design, or fraud) constitutes effective disproof of the original claim. This type of scrutiny helped debunk claims of cold fusion (*13*) and pluripotent stem cell creation (*14*).

The bright-line logic of deterministic and proof-of-principle studies is superficially mimicked through statistical significance testing; findings that are statistically significant are often regarded either as literally true or, at least, as justifying a knowledge claim, and those that aren't are regarded as either confirming the null hypothesis or inconclusive. However, it is inappropriate to combine null hypothesis–significance testing with intuition from fields of science with determinacy or very high signal-to-noise ratios. Statistical significance by itself tells very little about whether one study has "replicated" the results of another. For example, two studies that show identical 10% survival differences between the treatment and control arms would have very different degrees of statistical significance if their sample sizes were substantially different. If one was highly significant and the other far from significance, the two studies might be reported individually as supporting opposite conclusions, in spite of the fact that they are mutually corroborative.

An interpretive error complementary to the one described above involves the assumption that multiple studies that fail to demonstrate statistical significance necessarily confirm the absence of an effect. This fallacy was demonstrated, for example, in a well-known early meta-analysis of the effect of tamoxifen on breast cancer survival (*15*). (Meta-analysis is the mathematical pooling of results of multiple independent studies that investigate the same research question.) In this pooled analysis, 25 of 26 individual studies of tamoxifen's effect were not statistically significant. Naïvely, these nonsignificant findings could be described as having been replicated 25 times. Yet, when properly pooled, they cumulatively added up to a definitive rejection of the null hypothesis with a highly statistically significant 20% reduction in mortality. So the proper approach to interpreting the evidential meaning of independent studies is not to assess whether or not statistical significance has been observed in each, but rather to assess their cumulative evidential weight.

The above example involved randomized experiments without major bias. If major biases are at play, having multiple statistically significant studies and even a statistically significant summary result for a meta-analysis does not guarantee that a genuine effect exists. For example, many studies on single nutrients and even their meta-analyses show significant associations with cancer or death risk, but most reflect confounding and reporting biases (*16*). What matters in such scientific fields is not replication defined by the presence or absence of statistical significance, but the evaluation of the cumulative evidence and assessment of whether it is susceptible to major biases, due to either the study design or the self-selection of subjects in ways that are unknown or not measurable.

It is easier to statistically define nonreplication than replication, through statistical tests of heterogeneity, which can evaluate whether the difference between two or more experimental results might be due to the play of chance. Two or more studies are judged to be statistically heterogeneous when the between-study variance in reported effects is substantially greater than what is expected from sampling error. Such tests, however, are greatly underpowered and therefore unreliable when comparing several studies, particularly when they are small or imprecise (*17*). Conversely, when there are many large studies, tests for heterogeneity might demonstrate statistical heterogeneity (and, therefore, lack of results reproducibility) even if the effect sizes of different studies are close (*17*) and regarded as scientifically equivalent. Therefore, a preferred way to assess the evidential meaning of two or more results with substantive stochastic variability is to evaluate the cumulative evidence they provide vis-á-vis a hypothesis of interest and not whether one contradicts or discredits the other through the lens of statistical significance.

Whether experiments can be pooled to provide cumulative evidence depends further on which features of a study or results are considered scientifically equivalent enough to pool. For example, in a recent replication effort of the anti-Leishmania activity of tested peptides, it was difficult to tell whether replication had been achieved or not; the peptides were found to have anti-Leishmania activity, but at concentrations 10 to 50 higher than in the original experiments and close to the toxicity range of eukaryotic human cells (*18*). Rejection of the null hypothesis in the two sets of experiments was insufficient to garner consensus about results reproducibility when consensus was missing about the operational scientific question, that is, whether the peptides had activity at low (and clinically relevant) con-

centrations or at any concentration. These experiments could be regarded as conflicting on the first question and mutually supportive on the second, so the question of results reproducibility is always dependent on the specificity of the underlying scientific question.

In the absence of a consensus on what constitutes successful results reproduction, investigators employ a range of operational definitions, as occurred in the case of the evaluation of the (results) reproducibility of 100 psychology studies conducted by the Open Science Collaboration (*2*). They acknowledged the lack of an accepted definition and so examined the studies from a variety of perspectives: significance levels, effect sizes, the number of studies whose effect size was within the confidence interval of another selected study, whether the combined estimate of the original and replication studies was statistically significant and finally, a "subjective assessment" of reproducibility. The lack of a single accepted definition opened the door to controversy about their methodological approach and conclusions (*19*).

Robustness and generalizability

We briefly introduce these terms because they are sometimes used in lieu of the term reproducibility. Robustness refers to the stability of experimental conclusions to variations in either baseline assumptions or experimental procedures. It is somewhat related to the concept of generalizability (also known as transportability), which refers to the persistence of an effect in settings different from and outside of an experimental framework. The issue of generalizability arises in clinical trials and other types of studies in which the context of how an intervention is delivered and the types of subjects tested are highly relevant. When a universal property of nature or biology is being explored, generalizability is often assumed, and the concept of robustness of a finding to minor variations in experimental procedures is more frequently invoked. Whether a study design is similar enough to the original to be considered a replication, a "robustness test," or some of many variations of pure replication that have been identified, particularly in the social sciences (for example, conceptual replication, pseudoreplication), is an unsettled question (*12*).

Inferential reproducibility

This dimension of reproducibility, while underrecognized, might be the most important one. It refers to the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study. Inferential reproducibility is not identical to results reproducibility or to methods reproducibility, because scientists might draw the same conclusions from different sets of studies and data or could draw different conclusions from the same original data, sometimes even if they agree on the analytical results. The aforementioned debate about the interpretation of the psychology reproducibility results could be seen as an example of this (*19*). There are many contributors to these differences, including different assessments of the prior probability of the hypotheses being explored—which can only be examined through a Bayesian lens—and different choices about how to analyze and report data, which we will discuss under the general rubric of "multiplicity."

Bayesian perspectives. What scientists and science users are really concerned about when they debate research reproducibility is the truth of research claims. Research reproducibility and other related concepts can be regarded as ways to operationalize truth. To express this informally, if a finding can be reliably repeated, it is likely to be true, and if it cannot be, its truth is in question (*20*). Unfortunately, the standard frequentist approach to statistics does not allow the assigning of a probability of truth to a hypothesis or claim (*21*). However, the philosophy underlying Bayesian statistics does: The probability that a claim is true after an experiment is a function of the strength of the new experimental evidence combined with how likely it was to be true before the experiment. Viewed through this lens, the aim of repeated experimentation is to increase the amount of evidence, measured on a continuous scale, either for or against the original claim.

How much evidence needs to be gathered for effective proof depends on the prior probability of the original hypothesis, which itself depends on prior evidence. If a hypothesis is highly unlikely a priori, such as the presence of extrasensory perception or the therapeutic effect of homeopathy, a large amount of high-quality evidence would have to be gathered to outweigh the very strong prior reasons to view such claims skeptically (*22*, *23*). Conversely, for a hypothesis based on a plausible, coherent, and robust body of prior work, such as the research that preceded the development of imatinib for leukemia (*24*), a claim is more likely to be true both before and after an experiment that supports it. Under the Bayesian paradigm, every study contributes evidence that adds to the prior evidence, represented by the a priori probability of truth of a given claim. Reproducibility plays no formal role except that repeated experiments with similar findings will generate strong cumulative evidence, which can confirm or refute an initial finding.

A hybrid Bayesian-frequentist index that captures the traditional notion of results reproducibility is predictive power: the probability that, given a result in one experiment, the next experiment of specified design will be statistically significant. This probability has been dubbed the replication (*25*) or reproducibility probability (*26*). After a significant result, this probability is typically far lower than most scientists suspect, due to the random variation of the *P* value. This phenomenon shows that the failure to observe a significant result in a second experiment of similar design is to be expected and cannot be used as a criterion to undermine the credibility of the first experiment (*25*–*28*).

Multiplicity. Multiplicity, combined with incomplete reporting, might be the single largest contributor to the phenomenon of nonreproducibility, or falsity, of published claims. Multiplicity can arise in many ways, including testing many hypotheses in one experiment, testing one hypothesis many times or in multiple ways in one or more studies, and other maneuvers that virtually guarantee a chance observation that will appear to strongly support some hypotheses. A diverse vocabulary has developed in various fields for the biases or practices that can mislead because of multiplicity (Table 2). These range from the conduct of multiple experiments (and reporting only "good" ones) to the use of multiple endpoints, multiple predictors, and, perhaps most invisibly, the fitting of many mathematical or statistical models. Coupled with incomplete or selective reporting, these practices are a formula for generating findings unlikely to be supported by further experimentation. However, the adverse effects of multiplicity can be greatly ameliorated through complete reporting of analytical procedures and choices (for example, reporting the total number of associations tested or models considered).

These practices are likely to thrive when there is low consensus on the correct methodology and what is considered sufficiently complete reporting. Many scientific fields have seen an increasing burden of multiplicity,

Table 2. Terminology to describe practices that introduce or hide multiplicity.
Multiple comparisons (many statisticians)
File-drawer problem (*29*)
Pseudoreplication (*32*)
Significance questing (*33*)
Data mining, dredging, torturing (*34*)
Hypothesizing after the results are known (HARKing) (*30*)
Data snooping (*35*)
Selective outcome reporting (*36*)
Silent multiplicity (*37*)
Specification searching (*38*)
P-hacking (*31*)

because they have expanded their capacity to measure more variables and to fit increasingly complex models. Scientific fields that routinely work with multiple hypotheses without correcting for or reporting the occurrence of multiplicity run a higher risk of nonreproducibility of results or inferences.

A variety of old and new practices that are described as specific forms of bias actually result from multiplicity. The classic file-drawer publication bias problem (wherein nonsignificant or "uninteresting" results are not published) (*29*) results in bias under the assumption that multiple studies are being produced independently but a biased sample is published. The acronym "HARKing"—hypothesizing after the results are known—is used in psychology literature to indicate the phenomenon of constructing hypotheses after the data are analyzed, suggesting that only one hypothesis was tested while many were contemplated (*30*). The practice of P-hacking, a term recently coined in psychology literature and applied to a long-recognized phenomenon in modeling, refers to applying multiple statistical analyses and subanalyses until hitting upon and reporting a statistically significant result while not completely reporting how it was obtained (*31*).

Ultimately, inferential reproducibility might be an unattainable ideal, and in some situations not even a desirable one, because differences between scientists and their interpretations of a single or multiple studies are the means through which weaknesses or gaps in the evidence base are identified and science progresses. What is clear, however, is that none of these types of reproducibility can be assessed without complete reporting of all relevant aspects of scientific design, conduct, measurements, data, and analysis. Such transparency will allow scientists to evaluate the weight of evidence provided by any given study more quickly and reliably and design a higher proportion of future studies to address actual knowledge gaps or to effectively strengthen cumulative evidence, rather than explore blind alleys suggested by research inadequately conducted or reported.

CONCLUSIONS

The lexicon of reproducibility to date has been multifarious and ill-defined. The causes of and remedies for what is called poor reproducibility, in any scientific field, require a clear specification of the kind of reproducibility being discussed (methods, results, or inferences), a proper understanding of how it affects knowledge claims, scientific investigation of its causes, and an improved understanding of the limitations of statistical significance as a criterion for claims. Many aspects of the new interest in research reproducibility have been salutary, but we need to move toward a better understanding of the relationship between reproducibility, cumulative evidence, and the truth of scientific claims.

REFERENCES AND NOTES

1. C. G. Begley, An unappreciated challenge to oncology drug discovery: Pitfalls in preclinical research. *Am. Soc. Clin. Oncol. Educ. Book* 466–468 (2013).
2. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
3. C. G. Begley, J. P. A. Ioannidis, Reproducibility in science: Improving the standard for basic and preclinical research. *Circ. Res.* **116**, 116–126 (2015).
4. J. Claerbout, M. Karrenbach, Electronic documents give reproducible research a new meaning, in *Proceedings of the 62nd Annual International Meeting of the Society of Exploration Geophysics*, New Orleans, USA, 25 to 29 October 1992.
5. R. D. Peng, F. Dominici, S. L. Zeger, Reproducible epidemiologic research. *Am. J. Epidemiol.* **163**, 783–789 (2006).
6. E. Afgan, D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Grüning, A. Guerler, J. Hillman-Jackson, G. Von Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko, J. Goecks, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 10.1093/nar/gkw343 (2016).
7. J. Ioannidis, C. Doucouliagos, What's to know about the credibility of empirical economics? *J. Econ. Surv.* **27**, 997–1004 (2013).
8. B. Lo, Sharing clinical trial data: Maximizing benefits, minimizing risk. *JAMA* **313**, 793–794 (2015).
9. K. Bollen, J. T. Cacioppo, R. Kaplan, J. Krosnick, J. L. Olds, *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science* (National Science Foundation, Arlington, VA, 2015).
10. F. S. Collins, L. A. Tabak, Policy: NIH plans to enhance reproducibility. *Nature* **505**, 612–613 (2014).
11. D. Fanelli, W. Glänzel, Bibliometric evidence for a hierarchy of the sciences. *PLOS One* **8**, e66938 (2013).
12. M. A. Clemens, The meaning of failed replications: A review and proposal. *J. Econ. Surv.* 10.1111/joes.12139 (2015).
13. B. V. Lewenstein, W. Baur, A cold fusion chronology. *J. Radioanal. Nucl. Ch.* **152**, 273–297 (1991).
14. A. De Los Angeles, F. Ferrari, Y. Fujiwara, R. Mathieu, S. Lee, S. Lee, H.-C. Tu, S. Ross, S. Chou, M. Nguyen, Z. Wu, T. W. Theunissen, B. E. Powell, S. Imsoonthornruksa, J. Chen, M. Borkent, V. Krupalnik, E. Lujan, M. Wernig, J. H. Hanna, K. Hochedlinger, D. Pei, R. Jaenisch, H. Deng, S. H. Orkin, P. J. Park, G. Q. Daley, Failure to replicate the STAP cell phenomenon. *Nature* **525**, E6–E9 (2015).
15. Early Breast Cancer Trialists' Collaborative Group, Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. An overview of 61 randomized trials among 28,896 women. *N. Engl. J. Med.* **319**, 1681–1692 (1988).
16. J. D. Schoenfeld, J. P. A. Ioannidis, Is everything we eat associated with cancer? A systematic cookbook review. *Am. J. Clin. Nutr.* **97**, 127–134 (2013).
17. R. J. Hardy, S. G. Thompson, Detecting and describing heterogeneity in meta-analysis. *Stat. Med.* **17**, 841–856 (1998).
18. E. Iorns, W. Gunn, J. Erath, A. Rodriguez, J. Zhou, M. Benzinou, The Reproducibility Initiative, Replication attempt: "Effect of BMAP-28 antimicrobial peptides on Leishmania major promastigote and amastigote growth: Role of leishmanolysin in parasite survival." *PLOS One* **9**, e114614 (2014).
19. D. T. Gilbert, G. King, S. Pettigrew, T. D. Wilson, Comment on "Estimating the reproducibility of psychological science". *Science* **351**, 1037 (2016).
20. S. N. Goodman, Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.* **130**, 1005–1013 (1999).
21. S. N. Goodman, Toward evidence-based medical statistics. 1: The *P* value fallacy. *Ann. Intern. Med.* **130**, 995–1004 (1999).
22. J. N. Rouder, R. D. Morey, A Bayes factor meta-analysis of Bem's ESP claim. *Psychon. Bull. Rev.* **18**, 682–689 (2011).
23. S. N. Goodman, J. Gerson, *Mechanistic Evidence in Evidence-Based Medicine: A Conceptual Framework* (Agency for Healthcare Research and Quality, Rockville, MD, 2013); http://www.ncbi.nlm.nih.gov/books/NBK154584/.
24. P. Keating, A. Cambrosio, *Cancer on Trial: Oncology as a New Style of Practice* (University of Chicago Press, Chicago, 2011).
25. S. N. Goodman, A comment on replication, *P*-values and evidence. *Stat. Med.* **11**, 875–879 (1992).
26. D. D. Boos, L. A. Stefanski, *P*-value precision and reproducibility. *Am. Stat.* **65**, 213–221 (2011).
27. A. Gelman, H. Stern, The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.* **60**, 328–331 (2006).
28. L. C. Lazzeroni, Y. Lu, I. Belitskaya-Lévy, Solutions for quantifying *P*-value uncertainty and replication power. *Nat. Methods* **13**, 107–108 (2016).
29. R. Rosenthal, The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**, 638–641 (1979).
30. N. L. Kerr, HARKing: Hypothesizing after the results are known. *Pers. Soc. Psychol. Rev.* **2**, 196–217 (1998).
31. U. Simonsohn, L. D. Nelson, J. P. Simmons, *P*-curve: A key to the file-drawer. *J. Exp. Psychol. Gen.* **143**, 534–547 (2014).

32. S. H. Hurlbert, Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54**, 187–211 (1984).
33. K. J. Rothman, Significance questing. *Ann. Intern. Med.* **105**, 445–447 (1986).
34. J.L. Mills, Data torturing. *N. Engl. J. Med.* **329**, 1196–1199 (1993).
35. H. White, A reality check for data snooping. *Econometrica* **68**, 1097–1126 (2000).
36. A.-W. Chan, A. Hróbjartsson, M. T. Haahr, P. C. Gøtzsche, D. G. Altman, Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA* **291**, 2457–2465 (2004).
37. D. Berry, Multiplicities in cancer research: Ubiquitous and necessary evils. *J. Natl. Cancer Inst.* **104**, 1125–1133 (2012).
38. E. E. Leamer, *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (Wiley, Hoboken, NJ, 1978).

Funding: This work was funded by a grant by from the John and Laura Arnold Foundation. **Competing interests:** The authors declare that they have no competing interests.

10.1126/scitranslmed.aaf5027

Citation: S. N. Goodman, D. Fanelli, J. P. A. Ioannidis, What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341ps12 (2016).

Science Translational Medicine

What does research reproducibility mean?

Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis

Sci. Transl. Med., 8 (341), • DOI: 10.1126/scitranslmed.aaf5027

Science Translational Medicine (ISSN 1946-6242) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Translational Medicine* is a registered trademark of AAAS.