Reproducing Statistical Results

Victoria Stodden

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Champaign, IL 61820-6211; email: vcs@stodden.net

Annu. Rev. Stat. Appl. 2015. 2:1–19

First published online as a Review in Advance on February 20, 2015

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-010814-020127

Keywords

reproducible research, statistical reproducibility, replication, data sharing, code sharing, open data, open code, open science, open licensing

Abstract

The reproducibility of statistical findings has become a concern not only for statisticians, but for all researchers engaged in empirical discovery. Section 2 of this article identifies key reasons statistical findings may not replicate, including power and sampling issues; misapplication of statistical tests; the instability of findings under reasonable perturbations of data or models; lack of access to methods, data, or equipment; and cultural barriers such as researcher incentives and rewards. Section 3 discusses five proposed remedies for these replication failures: improved prepublication and postpublication validation of findings; the complete disclosure of research steps; assessment of the stability of statistical findings; providing access to digital research objects, in particular data and software; and ensuring these objects are legally reusable.

1. INTRODUCTION

A fundamental goal of statistics is to ensure the reproducibility of scientific findings. Distinguishing between signal and noise is the primary occupation of the field, from collaboration with data generators to experimental design through to power calculations, tests of significance, and validation of findings. If discoveries are made, it is of great interest to understand whether these findings persist in different samples, which may be drawn from the same or different populations, and potentially with different measurement or estimation techniques. The persistence of findings across different samples is the basis upon which scientific claims are evaluated. However, numerous reports over the past few years lament the inability of others to replicate published scientific results (Begley & Ellis 2012; Jasny et al. 2011; Nat. Publ. Group 2012, 2013; Peng 2011; Prinz et al. 2011); some reports have appeared in the popular press (Economist Staff Writer 2013a,b; Hiltzik 2013; Johnson 2014). This article is intended to provide an overview of issues of reproducibility and how statistical research has been and could be addressing these concerns. It outlines some key issues, rather than providing a comprehensive account of the entire body of research on reproducibility. These key issues can be roughly grouped into five categories, which are discussed in Section 2: power and sampling issues; misapplication of statistical tests; stability of findings under reasonable perturbations of data or models; lack of access to methods, data, or equipment; and cultural barriers such as researchers' incentives. Section 3 discusses five possible remedies for these issues: improved prepublication and postpublication validation of findings; the complete disclosure of research steps; assessment of the stability of statistical findings; providing access to digital research objects, in particular data and software; and ensuring that these research objects are legally reusable.

Elsewhere I have suggested refining the word reproducibility, proposing the terms "empirical reproducibility," "computational reproducibility," and "statistical reproducibility" to disambiguate an overloaded term (Stodden 2011, 2013). Empirical reproducibility refers to the use of appropriate reporting standards and documentation associated with physical experiments and can be traced back to Robert Boyle's [Boyle 2007 (1661); Shapin & Schaffer 1989, p. 59] exhortations in the late seventeenth century "that the person I addressed them to might, without mistake, and with as little trouble as possible, be able to repeat such unusual experiments." In contrast, computational reproducibility refers to changes in scientific practice and reporting standards to accommodate the use of computational technology occurring primarily over the past two decades, in particular whether the same results can be obtained from the data and code used in the original study (Anderson et al. 2013, Bailey et al. 2013, Donoho et al. 2009, King 1995, Nosek et al. 2012, Peng 2009, Sandve et al. 2013, Stodden 2012, Stodden et al. 2013). Finally, statistical reproducibility refers to the failure to replicate an experiment owing to flawed experimental design or statistical analysis. I recently became aware of the term "ethical reproducibility," which refers to standards of transparently reporting research ethics methods used in biomedical research (Anderson et al. 2013). Although these definitions are not mutually exclusive, this article is concerned primarily with statistical reproducibility and is divided into two parts: (*a*) enumerating sources of statistical irreproducibility and identifying gaps in our understanding of how replications can fail and (*b*) providing recommendations to improve the statistical reproducibility of scientific findings. Although results may fail to replicate owing to fraud or falsification of data or methods in the original study, this article does not consider those issues (Panel Sci. Responsib. Conduct Res. et al. 1992).

1.1. Motivating Examples

The following selected examples serve to motivate and concretize the discussion, emphasizing statistical reproducibility issues and their interplay with reporting standards and gaps in science policy. Many such examples exist, and these were selected in part to allow for concise exposition. They should not be considered isolated cases.

1.1.1. Example 1: The misapplication of statistical methods can cause irreproducibility. Shortly after publication, questions emerged concerning the application of statistical methods in "The Consensus Coding Sequences of Human Breast and Colorectal Cancers," by Sjöblom et al. (2006). The authors attempted to identify genes mutated in breast or colorectal cancer tumors using a statistical model. Their model sought to identify candidate genes by calculating the likelihood that observed mutations would occur by chance given an estimated background mutation rate. They applied the false discovery rate (FDR) thresholding method developed by Benjamini & Hochberg (1995) to select 122 breast cancer genes and 69 colorectal cancer genes estimated as having a 90% chance of being true cancer genes. Within a year, three "Comments" on this paper were published in the same journal (Forrest & Cavet 2007, Getz et al. 2007, Rubin & Green 2007), noting methodological flaws preventing the reproducibility of these findings. The authors had applied the FDR method to the observed likelihoods, rather than to *p*-values, for which the algorithm was devised. This use of the FDR method had the effect of falsely amplifying the significance level, leading to an overidentification of candidate genes. In addition, the authors estimated the background mutation rate from a smaller data set derived from a different tumor population, thereby reducing the background mutation rate used in the gene identification process to artificially low levels. Finally, the authors assumed a constant mutation rate across the genome, in contradiction to known regional variations. This assumption had the effect of amplifying the apparent significance of genes that happened to be in regions with high background mutation rates. When these adjustments were made to the methodology and then reapplied to the breast and colorectal cancer data, only 1 gene for breast cancer and 11 for colorectal cancer were significant. Of the 11 significant genes for colorectal cancer, 8 were known prior to this study. As noted in one of the comments, "[a]fter correcting the statistical analysis and using a background mutation rate that better fits the data, one cannot conclude that the ∼200 candidate genes reported in Sjöblom et al. have >90% probability of being cancer-related. The issue is simply one of statistical power: Much larger sample sizes are required to detect cancer genes" (Getz et al. 2007).

1.1.2. Example 2: How "small" decisions can affect significance. In an article entitled "Prediction of Survival in Follicular Lymphoma Based on Molecular Features of Tumor-Infiltrating Immune Cells," Dave et al. (2004) derive a model using gene expression data to predict survival among patients with follicular lymphoma. Using a training data set, they identify two gene clusters and fit a Cox model using the average value for each cluster. They then apply this model to a test data set with significant results. Within a year, two "To the Editor" notes were published in the same journal suggesting that the significant results presented in this article were not reproducible (Hong et al. 2005, Tibshirani 2005). Specifically, the significance of the results disappeared when the test and training data sets were swapped and when the allowable cluster size was changed slightly (Tibshirani 2004). As I discuss below, the authors (Staudt et al. 2005, p. 1496) of the original report defend their validation methods by saying that their "validation method is the accepted standard for supervised analyses of microarray data to create a survival predictor."

A second example stems from the social sciences. In 2006, Donohue & Wolfers published an article entitled "Uses and Abuses of Empirical Evidence in the Death Penalty Debate," in which the authors assessed the robustness and validity of the then-current statistical evidence regarding the deterrent effect of the death penalty in order to reconcile conflicting published accounts. They found that the "effects of the death penalty [are] extremely sensitive to very small changes in econometric specifications" (Donohue & Wolfers 2006, p. 794) such as functional form or sample used. This indicates a lack of power, and their "estimates suggest . . . profound uncertainty" about any deterrent effects of the death penalty. Donohue & Wolfers (2006) proffer publication bias, the "file drawer" problem (reporting only statistically significant results), and reporting bias as possible explanations for the differing conclusions in the literature.

1.1.3. Example 3: The importance of understanding the data-generation mechanism. A collaboration between scientists at UC Berkeley and Harvard anticipated spending a few months obtaining comparable data measurements from their labs in each of the two locations. In the article "Sorting Out the FACS: A Devil in the Details," Hines et al. (2014) explain that it unexpectedly took them two years to understand and reconcile differences in the data produced by each lab. Both labs have decades of experience with the task at hand, isolating cells from breast tissues and flow-sorting primary cells. Initial explanations for the differences were ruled out, such as differing instrumentation, antibodies, reagents, or tissue sources. Eventually, the two groups discovered a difference in how the collagenase digests were incubated. When a similar technique was employed at both labs, a similar yield was produced, and the fluorescence-activated cell sorting (FACS) profiles were identical. The number of differing data sets thought to be similar, such as those described above, that underlie published research findings is left to the reader's imagination.

1.1.4. Example 4: Omitting crucial information can result in irreproducibility. In 2013, Herndon and colleagues published an article entitled "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff," in which they questioning the methodology in a published paper by Reinhart & Rogoff (2010) called "Growth in a Time of Debt." Herndon et al. (2013, p. 21) claimed to have found "exclusion of available data, spreadsheet errors, and an inappropriate weighting method" in the original research which, when corrected, reduce the significance of the original findings.

As a second example, in 2010, Duke University suspended clinical trials based on a series of publications by Joseph Nevins & Anil Potti that attempted to show how using microarray profiles with chemotherapeutic drug sensitivity data could improve a cancer patient's sensitivity to particular drugs (Reich 2011). Researchers at the M.D. Anderson Cancer Center at the University of Texas at Austin found numerous mistakes in the research when trying to replicate the results in their articles. These mistakes ranged from mislabeled and duplicated observations in treatment and control groups to reversing labels indicating whether a group was sensitive to or resistant to a particular chemotherapy. The researchers at M.D. Anderson were able to find these mistakes because some of the original genome data had been made openly available. This case is quite complicated, and I have omitted many details in this encapsulation (see, e.g., Baggerly & Coombes 2009), including assertions that the authors tampered with the data. The discussion in this article is not focused on fraud, however, but on statistical causes of irreproducibility [for a discussion of fraud, see the NAS report (Panel Sci. Responsib. Conduct Res. et al. 1992)].

As a final example, in 2013, Feizi et al. published "Network Deconvolution as a General Method to Distinguish Direct Dependencies in Networks," which came under fire for omitting two scaling parameters [one was subsequently disclosed in a correction to the original supplement posted on the *Nature Biotechnology* website on August 26, 2013, and the other was discovered in a parsing of the authors' released code (Pachter 2014)]. Pachter (2014), who attempted to replicate the results in the article by Feizi et al. (2013), claims he found little guidance regarding how to set these data-dependent parameters and could not replicate the published results.

2. CAUSES OF STATISTICAL IRREPRODUCIBILITY

A published finding may not reproduce in independent replications of the original experimental design (i.e., reimplementing the experiment) for any of several statistical reasons. We assume an independent researcher who is perfectly able to follow instructions given in the publication. In other words, we assume that the replicator does not introduce error. I discuss issues of replication with sources of error arising from experimental conditions in Section 3.2. In the following subsections, I begin with the most obvious failings and end with consideration of more subtle issues. These potential problems also compose a checklist of sorts for reporting standards when publishing statistical findings, which I discuss in Section 3.2.

2.1. Low Power and Sampling Issues

If by chance a study with relatively low statistical power, perhaps owing to a flawed experimental design or insufficient data, reveals significant findings, we expect that an independent replication would not show the same results (Ioannidis 2005). All else remaining equal, increasing the amount of data available would, a priori, increase the power of statistical tests. However, the rise of so-called big data may paradoxically reduce experimental power because the researcher may be more removed from the data-generation mechanism than (s)he would be if the data had been generated specifically for the study in question (Siegfried 2013). This lack of awareness of the data-generation mechanism increases the possibility of inappropriate sampling techniques, which may thereby reduce the power of a study. In their recent book chapter, Kreuter & Peng (2014) refer to designed data collection—a term they use to describe data collected with a specific research question in mind and chosen to best answer that question. This concept contrasts with what they refer to as organic or accidental data—data for which research is often a by-product of the original collection purpose. The vast majority of the current data deluge comprises these latter organic data. Kreuter & Peng (2014) also note that reduced knowledge of the underlying data-generation mechanism makes both undercoverage and overcoverage more likely: Undercoverage refers to the exclusion of units that belong to the appropriate population, and overcoverage refers to the inclusion of units that do not belong to the appropriate target population. Undercoverage and overcoverage may lead to bias and loss of power, especially if the magnitude of these effects is not well understood.

Overcoverage could include repetition of records within the data, for example, highlighting another potential source of irreproducibility. Data provenance, or the history or changes made to the data set (by the researcher and before the researcher obtained the data, if applicable), is crucial for replicating findings accurately. To replicate a result, one must know the answers to the following questions, among others: How were duplicate entries dealt with? How were outliers identified and managed? Were missing values imputed? If so, how? Small changes in data-filtering decisions and preparation steps can dramatically affect the outcome of statistical analysis (Simmons et al. 2011), highlighting the need to accurately record these steps in software or text, including the input parameters used. If a researcher claims duplicate records were deleted, for example, how do we know whether the code he or she used to identify and delete such records did so correctly? These issues are discussed further in Section 3.2.

Undercoverage may mean that entire variables that should be included in the model are omitted, possibly introducing omitted variable bias. A researcher may be aware of important omitted variables, or (s)he may not be aware of such variables, as in the recent discovery that the gender of the animal handler can affect the outcomes of experiments that involve live rodents (Katsnelson

2014). This kind of gender information has traditionally not been collected, potentially causing an omitted variable problem in such studies.

2.2. Misapplication of Statistical Tests

The misapplication of statistical tests, for example, using a one-tailed *t*-test when a two-tailed *t*-test is appropriate, can cause an independent replication to yield results that are not similar to the original (except for the degenerate case in which the same flawed test is applied to the same data). However, several less trivial and common misapplications also occur, including overreliance on *p*-values, problems of multiplicity, and application of methods for which necessary assumptions are not satisfied.

Using a statistical test that results in a *p*-value has become a standard way to report scientific findings in many fields. This approach was widely adopted in the wake of Karl Popper's (2002) well-known conceptualization of hypothesis testing and falsifiability as defining concepts in scientific progress. Although this approach—the Fisher/Neyman–Pearson approach to hypothesis testing—was developed prior to Popper's falsification criterion, it institutionalizes Popper's ideas in a tractable way (Neyman & Pearson 1933). Fisher generalized the two-sample test, and recommended using 5% as the cutoff level for deciding whether or not to reject the hypothesis (Grove 1930, Lehmann 1993, Masicampo & Lalande 2012). This cutoff level is now used nearly ubiquitously in this testing framework.

The use of the 5% cutoff level for rejecting the null hypothesis has led to *p*-values being interpreted as something of a litmus test for publication; research resulting in *p*-values of greater than 0.05 is often considered unpublishable (Masicampo & Lalande 2012, Simmons et al. 2011). As a result researchers may cherry pick results, running many tests until one of them yields a *p*-value of less than 0.05 and then reporting this result as if the other tests had never been run. This is also called the file drawer or multiplicity problem, and the typical solution is to adjust the interpretation of the *p*-value for the total number of tests run (Benjamini & Hochberg 1995, Johnson 2013, Rosenthal 1979). Indeed, recent research in the psychology literature suggests a striking prevalence of *p*-values just under the 0.05 threshold level (Masicampo & Lalande 2012).

Observational studies present an additional issue: Bias (often due in part to omitted variables) and confounding effects can obscure the interpretation of *p*-values. Schuemie et al. (2014) recently tested these effects in observational drug safety studies by replicating three exemplar studies and applying their experimental designs to sets of negative controls for whom the drug in question is known to be ineffective. They found (Schuemie et al. 2014, p. 209) that "at least 54% of findings with $p < 0.05$ are not actually statistically significant and should be re-evaluated."

2.3. Robustness and Lack of Generalizability

A great amount of energy in statistics has been expended to understand the circumstances under which a particular statistical method will work as expected. However, many methodological assumptions are not met in practice, although the experimental conditions may be close. For example, ordinary linear regression requires a fairly rigid set of assumptions to be met that are rarely fully satisfied in applied settings, possibly affecting the reliability of the finding. In the big data context of organic or accidental data, this problem may be exacerbated because the researcher is typically removed from the data-generation process. This disconnect thereby makes evaluation of the assumptions upon which statistical tests are based even more challenging. Model error terms may not be handled appropriately if the data-generation mechanism is not known.

The discussion above points to at least two deeper issues: First, how do results change if the underlying data are "reasonably" perturbed? Second, how do the results change if the model is altered slightly? In the replication context, the former question arises naturally, as an independent replication could be seen as another sampling from the population. In such a framework, carrying out the same statistical analysis would not be expected to produce an identical outcome, but it would presumably produce an outcome that is substantially similar. The degree of similarity depends on the characteristics of the population, the sampling mechanism, and the nature of the stability of statistical methods applied. Yu (2013) addresses both of these concerns and proposes the technique of estimation stability to improve stability in the case of the lasso (see also Lim & Yu 2013). Kleiner et al. (2014) investigate the case of data sets that are too large for direct implementation of the bootstrap algorithm, as well as the effectiveness of the *m* out of *n* subsampled bootstrap. This latter procedure is akin to resampling the population as carried out in a replication study if the data set can be effectively considered a population. Finding the *m* out of *n* bootstrap unstable, they propose what they call the bag of little bootstraps (BLB) "which incorporates features of both the bootstrap and subsampling to yield a robust, computationally efficient means of assessing the quality of estimators" (Kleiner et al. 2014).

Li et al. (2011) propose a method for evaluating reliability in replicated experiments. They rank the results of replications by significance and seek consistency across these curves. The curves are fit with a copula model that produces a graphical assessment of reproducibility, thereby offering a researcher greater knowledge of the variability between replicates, along with its impact on results. Li et al. (2011) assign a reproducibility index by jointly modeling the significance of the results on a given replicate and the consistency of results between replicates. They then define the irreproducible discovery rate (IDR) and a selection procedure for significant results. Results may be *p*-values or other scores associated with significance. Madigan et al. (2014) have taken an empirically driven approach and implemented thousands of epidemiological drug effectiveness study designs over perturbed data. These authors used very large medical claims databases instead of the data associated with the original study (which typically has a much smaller *n*) to characterize bias, accuracy of confidence intervals, and discrimination between positive and negative controls. They found that their "results suggest that bias is a significant problem in many contexts, and that statistical measurements, such as confidence intervals and *p*-values, are substantially invalid" in that they indicate greater significance than is warranted (Madigan et al. 2014, p. 12). The authors have developed data-driven approaches to control for this bias and return confidence intervals and *p*-values with the expected characteristics.

As mentioned above, the gender of a researcher may present a bias in an animal study if it is not accounted for in the model design or reported in the description of the experiment (Katsnelson 2014). Indeed, by not recording this information, the experimenter can introduce error into the replication process that can interfere with reproducibility and reliability of findings. This occurs in tandem with the amount of tacit knowledge associated with an experiment—in other words, the amount of presumed knowledge, unspecified in the article, that a researcher has when reimplementing another researcher's experiment may affect reliability. Some aspects of an experiment can be completely specified, such as cases in which scripts that comprise the entire experiment are made available and run by a downstream researcher in a substantially similar computing environment. But it is more challenging to specify other aspects, such as complicated manual procedures carried out in a laboratory setting by individuals with years of training or assumptions about what is common knowledge in their field. In such cases, large amounts of knowledge, tacit knowledge, may not be reported as standard operating procedure (Polanyi 1962, 1967). Ioannidis (2005) lists several additional reasons why empirical findings may not replicate (some of which I have already discussed): small sample size, small effect size, the number of statistical tests carried out relative to the number of possible relationships, overly flexible experimental design, conflicts of interest, and the level of competition in an area of investigation. Section 2.4 discusses problems in reproducibility that stem from a lack of access to materials, equipment, data, or code.

2.4. Lack of Access to Data, Software, and Tools of Analysis

In the context of computational reproducibility—an exact implementation of the same software steps as the original research, using the original fixed data set—a lack of access to the data and code presents a challenge to one seeking to the verify the findings (Bailey et al. 2013, Donoho et al. 2009, Stodden et al. 2012). The challenge posed by lack of access to the data underlying a study is not a new suggestion (Fienberg et al. 1985, Nat. Res. Counc. 2003). As LeVeque states (LeVeque 2007, p. 7 of preprint):

> Even brilliant and well-intentioned computational scientists often do a poor job of presenting their work in a reproducible manner. The methods are often very vaguely defined, and even if they are carefully defined they would normally have to be implemented from scratch by the reader in order to test them. Most modern algorithms are so complicated that there is little hope of doing this properly. . .
>
> The idea of "reproducible research" in scientific computing is to archive and make publicly available all of the codes used to create the figures or tables in a paper in such a way that the reader can download the codes and run them to reproduce the results. The program can then be examined to see exactly what has been done. The development of very high level programming languages has made it easier to share codes and generate reproducible research. . .These days many algorithms can be written in languages such as MATLAB in a way that is both easy for the reader to comprehend and also executable, with all details intact.

A primary difficulty in reporting and reproducing computational aspects of statistical analysis stems from the increase in the sheer number of computational steps carried out in modern research. **Figure 1** shows the remarkable increase in the number of lines of code submitted to the journal *ACM Transactions on Mathematical Software* (*TOMS*) from 1960 to 2012. The total number of lines

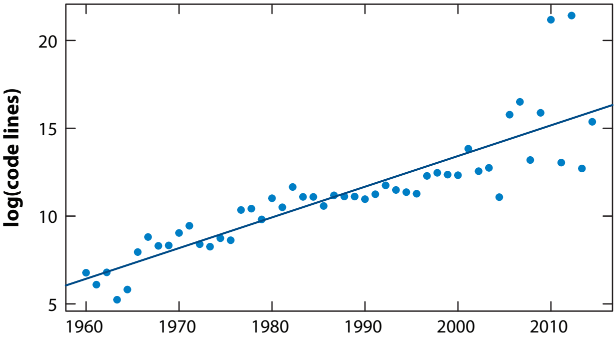


Figure 1

The increase in log(lines of code) submitted to *ACM Transactions on Mathematical Software* (*TOMS*) from 1960 to 2013. The proportion of publications whose authors submitted their code remained roughly constant at approximately 1/3, with a standard error of approximately 0.12, and *ACM TOMS* consistently published approximately 35 articles associated with software submissions each year throughout the period of analysis. The ordinary least squares best fit line is superimposed. These data also appear in figure 6 of a recent article by Stodden et al. (2015).

of code is increasing on a logarithmic scale; 875 lines were submitted in 1960, and approximately 5 million lines were submitted in 2012 (including libraries). The number of articles in *ACM TOMS* that are associated with software submissions has been roughly constant over this period, and the conclusions are clear: Algorithms are requiring increasing amounts of code, even given the increasing sophistication of modern computing languages. Without access to the software, comparing results to understand different methodological implementations or to create extensions to published findings can be nearly impossible (Collberg et al. 2014).

There are many reasons that data sharing may not be possible; for example, patient privacy protections may be in place (Lane et al. 2014). Specialized hardware, such as some very large-scale computer systems, or specialized experimental devices, such as the Large Hadron Collider, make fully independent replication impossible or nearly impossible for some experiments (Shi 2014).

2.5. Ineffective Cultural Incentives

Among the pressures exerted on modern scientists, most would agree the strongest is the pressure to produce high-quality research publications (Fanelli 2010). Alberts et al. (2014) describe some of the reasons for increasing hypercompetitivity in biomedical research, such as the slower growth of funding relative to the growth in the number of scientists, and these incentives show no sign of abating (Kelly & Marians 2014). The trendiness of the research topic has also been suggested as a factor contributing to irreproducibility in the life sciences (Neyman & Pearson 1933). Other cultural reasons for irreproducibility include publication bias toward positive findings or established authors or ineffective peer review (Groves & Lyberg 2010).

Currently, with some exceptions, tenure and promotion committees and research managers at research labs do not recognize reproducibility of studies or the generation of data sets and software products as important for promotion and hiring decisions. Software and data set contributions should be rewarded as part of expected research practices (Stodden 2009a). Steps taken by the National Science Foundation (NSF) and the National Institutes of Health (NIH) to recognize software and data set contributions in the biographical sketch of a grant applicant encourage greater data and code production and disclosure (NSF 2013, Rockey 2014).

3. STEPS FORWARD

Responses to statistical irreproducibility can be grouped into two broad categories: procedural changes and improvement of reporting standards. This section describes five recommendations: three procedural and two focusing on reporting standards.

3.1. Remedy 1: Improved Prepublication and Postpublication Validation of Findings

Although much care is often taken in model selection and fitting, verification of the effectiveness of the model after publication is not a standard practice. Indeed, only a subset of publications employ verification methods prior to publication; among those that do, such methods include model testing on data that were not used in model fitting or validation of model results against other independent results. Standardizing expectations to include both forms of verification would be a step toward improving reliability of results. Some fields, such as simulation or machine learning, employ one or both of these methods as standard practice, and these practices could be widely adopted. The reliability of published findings would be improved if (*a*) the replication of previous work was standard for new contributions and replicated findings were reported in publications and (*b*) the publication of replication studies was facilitated. The Association for Psychological Science is attempting a novel approach: This organization is publishing *Registered Replication Reports*, which are announced prior to publication (Assoc. Psychol. Sci. 2013). The community is invited to participate in producing these reports, with the promise of publication in *Perspectives on Psychological Science*. Similar short replication reports could be published as a new category in statistical journals.

Validation of results within and across studies will permit increased investigation into what evidence is needed to have certain levels of confidence in findings. This idea builds on the field of meta-analysis not only by learning from the collation of similar studies testing the same hypothesis, but also by extending existing models to new data, testing predictions, and systematizing independent checks against other data sources when possible. This approach is similar to the concept of validation in scientific computing, a concept arising from the well-known area of verification, validation, and uncertainty quantification (Nat. Res. Counc. 2012). This suggested form of validation also follows the spirit of the total survey error framework in survey research (Groves & Lyberg 2010). Madigan et al. (2014) point to the seriousness of this issue when they validate published claims on much larger data sets and find widespread overestimation of the effect size by the original study, including in confidence interval estimation and *p*-value reporting.

Adjustments to *p*-values to control for the multiple comparisons problem is an ongoing area of research (Heller et al. 2013). In a seminal paper, Benjamini & Hochberg (1995) developed the false discovery rate (FDR), a statistic that controls for the expected proportion of errors among a set of independent and significant tests (Abdi 2007, Holm 1979). The problem of multiple comparisons is distinct from, and not resolved by, the use of meta-analysis. Heller, Bogomolov & Benjamini address the situation of reproducibility with the *r*-value (Bogomolov & Heller 2013, Heller et al. 2013). In contrast to a meta-analysis, an investigator using the FDR is interested in evaluating the significance of a new hypothesis test, given that a significant *p*-value was found in a previous study. The *r*-value is the lowest FDR at which the finding can be considered replicated, where replicated means the new test achieved a level of significance identical to or lower than that found in the original study. In their setting the follow-up study has been carried out because the null hypothesis was rejected in the original study.

Replication studies and increased validation of the original findings will also help uncover any errors in the original analysis, for example, modeling or estimation errors or errors in the software implementation of the statistical analysis. The potential for such errors to go undiscovered leads to the next remedy, the complete disclosure of the steps taken prior to publication.

3.2. Remedy 2: Complete Disclosure of Research Steps

Prior to the digitization of scholarly communication, journals typically enforced page limits for articles and restricted the length of methods sections (and often still do). Thus, for computational research it is almost impossible to include all of the details relevant to understanding, contextualizing, and potentially replicating the study in the article itself. All steps in data collection, preprocessing, and filtering should be disclosed, along with decisions regarding the treatment of outliers, imputation of missing values, and those made when combining data sets, as such decisions may often have a large potential impact on the statistical results. As discussed in Section 3.4 (Remedy 4), it may be most effective to communicate these steps by making the source code that implemented them available.

Reporting all of the steps carried out prior to publication would help address the file drawer problem (the problem of cherry picking the significant tests among many and publishing only those results). Disclosing steps that simply allow another researcher to achieve the same results if followed faithfully, will not adequately address the file drawer problem, however. Additional measures to rectify this problem may be helpful, such as the preregistration of hypotheses before data are collected or accessed, as is now done for clinical trials. Improved software tools that permit a researcher to more easily track the various tests (s)he has done will also help researchers accurately report all steps taken prior to publication (Stodden et al. 2014). Tracking all tests carried out implies the publication of negative results. which can be reinterpreted as workflow publication. This issue is discussed in Section 3.4 (Remedy 4).

Best practices for communicating computational statistical findings are not yet standardized. The workshop on "Reproducibility in Computational and Experimental Mathematics," held at the Brown University Institute for Computational and Experimental Research in Mathematics (ICERM) in 2012, made several recommendations for computational mathematics papers, some of which are instructive for computational statistical research publications (Stodden et al. 2012). The recommended inclusions are as follows:

1. A precise statement of assertions made in the article.
2. A statement of the computational approach and why it constitutes a rigorous test of the hypothesized assertions.
3. Complete statements of or references to every algorithm employed.
4. Salient details of software used in the computation.
5. Salient details of data reduction and statistical analysis methods.
6. A full statement (or at least a valid summary) of experimental results.
7. Verification and validation tests performed by the author(s).
8. Availability of computer source code, input data and output data, with some reasonable level of documentation.
9. Curation: Where are code and data available? With what expected persistence and longevity? Is there a way to access future updates, for example, a version-control repository of the code base?
10. Instructions for repeating computations described in the paper.
11. Terms of use and licensing. Ideally code and data default to open; that is, they are published under a permissive reuse license.
12. Avenues of exploration examined throughout development, including information about negative findings.
13. Proper citation of all code and data used, including that generated by the authors.

Some of these recommendations, in particular those relating to code and data sharing, are discussed in Section 3.5 (Remedy 5). In addition to code and data citation, full disclosure also includes systemized and standard disclosure of all funding sources for the study. These suggestions should be adapted to different research contexts, but the ultimate goal is to ensure that readers have the information needed to independently verify computational statistical results. Examples of such information are metadata, including parameter settings and workflow documentation, the data themselves, and the code used.

Within the past two years, the journals *Science* and *Nature* have both implemented checklists for authors proposing to publish statistical studies (Lane et al. 2014, Nat. Publ. Group 2013). Building on the recommendations of a workshop report from the US National Institute of Neurological Disorders and Stroke (NINDS), *Science* implemented the following publications requirements in January 2014 (Landis et al. 2012; McNutt 2014b, p. 229): "Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment."

These requirements have been discussed previously in this article, with the exception of the final point, which can be considered a form of tacit knowledge.

3.3. Remedy 3: Assessing the Stability of Statistical Findings

Development of a research agenda around understanding the sensitivity of estimates to both model choice and perturbations in the underlying data is crucial. Lim & Yu (2013) have investigated how the characteristics of the population, the sampling mechanism, and the statistical method affect outcome stability. As mentioned above, they propose the technique of estimation stability to improve stability relative to reasonable perturbations in the data in the case of the lasso (see also Yu 2013). Evaluating the stability of findings would bolster their potential reproducibility.

In practical empirical research, power calculations are seldom carried out, presumably owing to their mathematical complexity. A simplified set of approximate power calculations that became widely used could also make great strides in improving the reliability of the scholarly record.

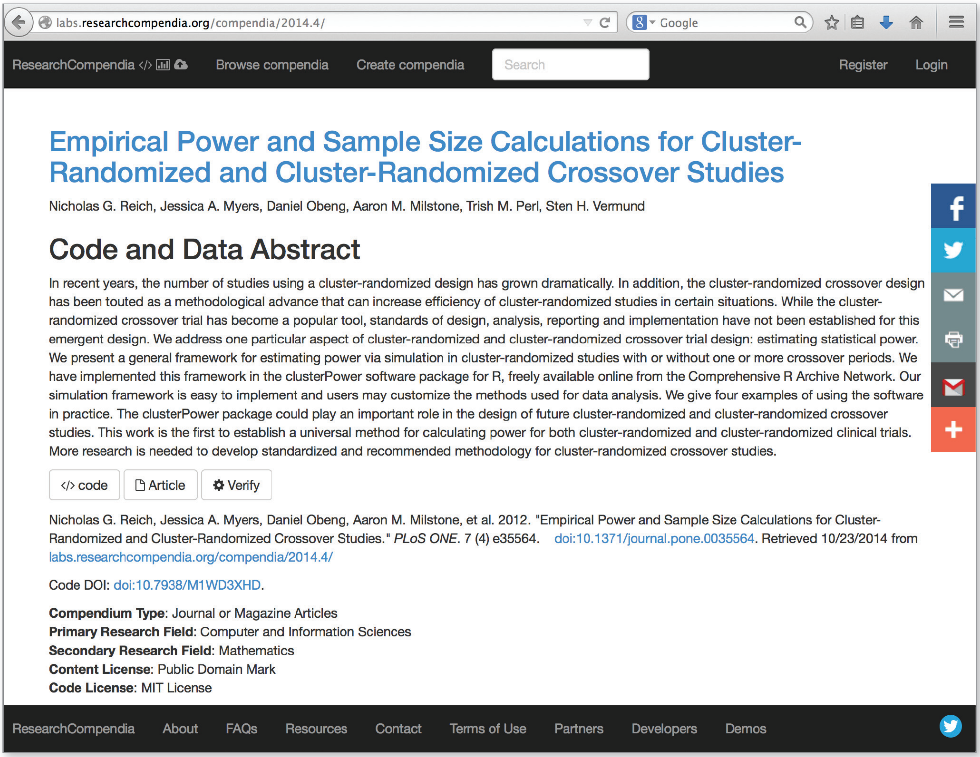
The increased use of sensors, the value of data collection to industry, and the increased availability of data collected by the government, among other influences, have resulted in increased measurement of the world around us. In observational studies, omitted variables occur frequently and can introduce bias into estimates. The impact of this increase in data availability on omitted variable bias, along with a better understanding of how omitted variables and proxies for omitted variables affect estimation stability, would also be helpful in increasing the reliability of statistical findings in observational studies (Stock & Watson 2011). Eventually, such stability assessments could become a best practice for statistical inference.

3.4. Remedy 4: Access to Digital Research Objects I: Tools for Reproducible Research

The use of computer systems is now central to much statistical research. For example, Donoho et al. (2009) document the need for access to the code and data that underlie published results, and Gentleman &Temple Lang (2007) propose publishing what they call the "research compendium"—a triple including data and code along with the article. It is common to submit `R` packages that implement new statistical methods to the Comprehensive `R` Network (CRAN) (R Core Team 2014), but a persistent linkage between the published results and the broad utility software in CRAN does not yet exist.

Many tools are emerging that seek to facilitate reproducibility in computational science (LeVeque et al. 2012, Stodden et al. 2014). I have been developing one such tool, ResearchCompendia (**http://www.researchcompendia.org**), which extends the ideas of Gentleman & Temple Lang (2007) to facilitate reproducibility in computational science by persistently linking the data and code that generated published findings to the article and by executing the code in the cloud to validate or certify those findings. ResearchCompendia is a website that houses a collection of compendium pages. Each compendium page is associated with an externally hosted article that has been published in a journal or made available in a preprint repository such as arXiv or SSRN. **Figure 2** gives an example of a compendium webpage from ResearchCompendia.

A compendium page links to the webpage where the publication is available, and if the publication is open access, users can download it directly from ResearchCompendia. As seen in **Figure 2**, data and code provided by the author are available for download by clicking the appropriately labeled button. ResearchCompendia links to larger data sets or code hosted in external repositories, and it hosts smaller files itself. To encourage proper citation, a suggested citation appears whenever a user clicks to download code or data. As ResearchCompendia gathers more data sets, each linked



Empirical Power and Sample Size Calculations for Cluster-Randomized and Cluster-Randomized Crossover Studies

Nicholas G. Reich, Jessica A. Myers, Daniel Obeng, Aaron M. Milstone, Trish M. Perl, Sten H. Vermund

Code and Data Abstract

In recent years, the number of studies using a cluster-randomized design has grown dramatically. In addition, the cluster-randomized crossover design has been touted as a methodological advance that can increase efficiency of cluster-randomized studies in certain situations. While the cluster-randomized crossover trial has become a popular tool, standards of design, analysis, reporting and implementation have not been established for this emergent design. We address one particular aspect of cluster-randomized and cluster-randomized crossover trial design: estimating statistical power. We present a general framework for estimating power via simulation in cluster-randomized studies with or without one or more crossover periods. We have implemented this framework in the clusterPower software package for R, freely available online from the Comprehensive R Archive Network. Our simulation framework is easy to implement and users may customize the methods used for data analysis. We give four examples of using the software in practice. The clusterPower package could play an important role in the design of future cluster-randomized and cluster-randomized crossover studies. This work is the first to establish a universal method for calculating power for both cluster-randomized and cluster-randomized clinical trials. More research is needed to develop standardized and recommended methodology for cluster-randomized crossover studies.

Nicholas G. Reich, Jessica A. Myers, Daniel Obeng, Aaron M. Milstone, et al. 2012. "Empirical Power and Sample Size Calculations for Cluster-Randomized and Cluster-Randomized Crossover Studies." *PLoS ONE*. 7 (4) e35564. doi:10.1371/journal.pone.0035564. Retrieved 10/23/2014 from labs.researchcompendia.org/compendia/2014.4/

Code DOI: doi:10.7938/M1WD3XHD.

Compendium Type: Journal or Magazine Articles
Primary Research Field: Computer and Information Sciences
Secondary Research Field: Mathematics
Content License: Public Domain Mark
Code License: MIT License

Figure 2

An example compendium page within the ResearchCompendia.org website. A compendium page links to the published paper and allows the user to download the code and data associated with that publication. It also provides information and metadata about the code and data, permits commenting, and suggests citations for reusing both code and data.

to specific results, a natural validation data set will be created for models trained on similar data. These models can then be tested on the much larger data sets gathered by ResearchCompendia.

ResearchCompendia seeks to solve an immediate problem: linking research code and data with articles containing the results they generate. However, algorithms are being combined in increasingly complex research pipelines. The individual algorithms that are typically documented in published articles represent only a small, and shrinking, fraction of the code involved in a computational science project. Exploiting modern computing resources, including large-scale computing and the cloud, requires the scaling of complicated workflows that pipeline together numerous methods and software packages. Thus, any given computational project may involve much more infrastructure than is explicitly described in the associated published article. In the modern computational context, journal articles necessarily become mere advertisements that point to a complex body of software development, experimental outcomes, and analyses, and nonspecialists will be at a disadvantage in understanding the full meaning of those summaries. Structured sharing of these research workflows, including their component parts, will become a best practice of reproducible computational statistical research. Consider the dream applications mentioned in an article by Gavish & Donoho (2012), in which robots crawl research projects, reproducing and varying results. Reproducible computational research can be more easily extended and generalized, and optimized. Code and data sharing necessitate the development of standards for dissemination, including documentation, metadata, best practices for code development, and data dissemination—an emerging area of research (WSSSPE 2013, Stodden & Miguez 2013).

3.5. Remedy 5: Access to Digital Research Objects II: Legal Barriers

Evolving community standards and peer review cannot be relied upon to solve all dissemination issues, as some, such as licensing for code and data, require coordinated action to ensure that goals such as interoperability are met. Owing to the default nature of copyright, I suggest that scholarly objects be made openly available using the *Reproducible Research Standard* (Stodden 2009a,b; 2014a), which recommends attribution-only open licensing for code, data, and the research article. Such licensing allows one to maximize downstream reuse and enable reproducibility, while ensuring alignment with scientific norms such as attribution. Examples of attribution-only open licensing include The Creative Commons CC-BY License for text and figures, the MIT License or the BSD 2-Clause License for code, and the CC0 Creative Commons Public Domain Dedication for data (Creative Commons 2013a,b; Open Source Initiative 2013a,b).

There may be a conflict between openness for the replication of computational results and traditional methods of privacy protection via data sequestration. I believe that whenever possible, middle ground solutions need to be established that will allow researchers to verify computational findings while taking into account any legal and ethical barriers, such as privacy and confidentiality. For example, permitting authorized researchers access to confidential data within a "walled garden" could increase the ability of others to independently replicate the findings. In other work (Stodden 2014b), I suggest two principles to help guide thinking regarding reproducibility given constraints on code or data: the Principle of Scientific Licensing and the Principle of Scientific Data and Code Sharing. That is, I believe that legal encumbrances to data and code sharing for purpose of independent verification should be minimized wherever possible, and that access to the data and code associated with published results should be maximized subject to legal and ethical restrictions (Stodden 2014b).

It is not atypical for data-producing entities, particularly those in the commercial sphere, to require researchers with access to the data to sign a nondisclosure agreement to prevent data access. But consider the results when research based on these protected data sets is published and questioned. How can others independently verify a researcher's findings without access to the data? Many legitimate reasons for not publicly releasing such data exist, for example, protection of subjects' privacy, but consideration of how to maximize access and reproducibility given these constraints is important. Awareness of the access issue is vital among researchers and institutions (including their technology transfer offices), publishers, funders, and data- and software-producing entities.

4. CONCLUSION

The issue of reproducibility has recently garnered significant attention within both the scientific community and the mainstream press. Changes have arisen in part from mandates from the White House, and Congressional requirements of federal funding agencies to ensure data arising from federal grants are openly available [for example, the COMPETES Act (Pub. L. 111-358, H.R. 5116, 111 U.S. C.), Fair Access to Science and Technology Act (FASTR) (see Am. Lib. Assoc. 2014), and various directives (e.g., Obama 2013, Stebbins 2013)]. Funding agencies began implementing their own data access initiatives several years ago (NSF 2011). Journals are implementing statistical requirements on empirical articles they publish to address issues of reproducibility (Nat. Publ. Group 2013, Lane et al. 2014). *Science* recently instituted a Statistics Board of Reviewing

Editors who are intended to assess the statistical methodology in papers referred to them by their Board of Reviewing Editors (the traditional reviewers) (McNutt 2014a, p. 9). Marcia McNutt, Editor-in-Chief of *Science*, stated that "[t]he creation of the [statistics board] was motivated by concerns broadly with the application of statistics and data analysis in scientific research and is part of *Science*'s overall drive to increase reproducibility in the research we publish." Concerted efforts are important because widely accepted standards for research evaluation and publishing are relied upon by individual researchers and can be insufficient. Recall the example in Section 1.1.2 in which authors defended their validation methods by asserting that their "validation method is the accepted standard. . ." (Dave et al. 2005, p. 1496).

In this article, I have sought to (*a*) unpack the notion of reproducible science, distinguishing between failures in established reporting standards for empirical science and new standards needed in computational research, (*b*) outline and address key statistical issues that may lead to irreproducibility, and (*c*) suggest remedies to improve the reliability of the scholarly record. Statistical sources of irreproducibility include underpowered studies, lack of knowledge about the sampling mechanism, overcoverage/undercoverage in sampling, and omitted variable bias. Other sources of irreproducibility include the incorrect application of statistical tests, the misuse and misinterpretation of *p*-values, or a lack of model robustness, among others. Understanding sources of variability can be more complex in cases of multiple combined data sets, where the statistician is typically removed from the data collection process. Finally, a lack of access to experimental equipment, data, software, and tools of analysis creates barriers to understanding and replicating statistical findings. I have suggested five possible remedies as a roadmap for resolving irreproducibility: improved prepublication and postpublication validation of findings; the complete disclosure of research steps; assessment of the stability of statistical findings; providing access to digital research objects, in particular data and software; and ensuring that these objects are legally reusable. Developing a dedicated research agenda within the statistical community to directly address issues surrounding reproducibility is imperative.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This research was supported by Alfred P. Sloan Foundation award number PG004545 for "Facilitating Transparency in Scientific Publishing." The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Pakwesi Taylor provided outstanding research assistance.

LITERATURE CITED

Abdi H. 2007. The Bonferroni and Šidák corrections for multiple comparisons. *Encycl. Meas. Stat.* 3:103–7

Alberts B, Kirschner MW, Tilghman S, Varmus H. 2014. Rescuing US biomedical research from its systemic flaws. *PNAS* 18:111(16):5773–77

Am. Libr. Assoc. 2014. *The Fair Access to Science and Technology Research Act* (*FASTR*). Chicago, IL: ALA. **http://www.ala.org/advocacy/access/legislation/fastr**

Anderson JA, Eijkholt M, Illes J. 2013. Ethical reproducibility: towards transparent reporting in biomedical research. *Nat. Methods* 10(9):843

Assoc. Psychol. Sci. 2013. *Registered Replication Reports*. Washington, DC: APS. **http://www.psychologicalscience.org/index.php/replication**

Baggerly KA, Coombes KR. 2009. Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Stat.* 3:1309–34

Bailey DH, Borwein JM, Stodden V. 2013. Set the default to "open." *Not. Am. Math. Soc.* 60(06):1

Begley CG, Ellis LM. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483(7391):531–33

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57(1):289–300

Bogomolov M, Heller R. 2013. Discovering findings that replicate from a primary study of high dimension to a follow-up study. *J. Am. Stat. Assoc.* 108(504):1480–92

Boyle R. 2007 (1661). *The Sceptical Chymistor Chymico-Physical Doubts & Paradoxes, Touching the Spagyrist's Principles Commonly Call'd Hypostatical; As They Are Wont to Be Propos'd and Defended by the Generality of Alchymists. Whereunto Is Præmis'd Part of Another Discourse Relating to the Same Subject.* Salt Lake City, UT: Proj. Gutenberg. **http://www.gutenberg.org/ebooks/22914**

Collberg C, Proebsting T, Moraila G, Shankaran A, Shi Z, Warren AM. 2014. *Measuring reproducibility in computer systems research*. Tech. Rep., Dep. Comput. Sci., Univ. Ariz., Tucson. **http://reproducibility.cs.arizona.edu/tr.pdf**

Creative Commons. 2013a. *CC0 1.0 Universal* (*CC0 1.0*) *Public Domain Dedication*. Mountain View, CA: Creative Commons. **https://creativecommons.org/publicdomain/zero/1.0/**

Creative Commons. 2013b. *Creative Commons—Attribution 4.0 International* (*CC BY 4.0*). Mountain View, CA: Creative Commons. **https://creativecommons.org/licenses/by/4.0/**

Dave SS, Wright G, Tan B, Rosenwald A, Gascoyne RD, et al. 2004. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N. Engl. J. Med.* 351:2159–69

Donoho DL, Maleki A, Rahman IU, Shahram M, Stodden V. 2009. Reproducible research in computational harmonic analysis. *Comput. Sci. Eng.* 1:8–18

Donohue JJ, Wolfers JJ. 2006. Uses and abuses of empirical evidence in the death penalty debate. *Stanford Law Rev.* 58:791–846

Economist Staff Writer. 2013a. Problems with scientific research: how science goes wrong. *The Economist*, Oct. 19. **http://www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong**

Economist Staff Writer. 2013b. Unreliable research: trouble at the lab. *The Economist*, Oct. 19. **http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble**

Fanelli D. 2010. Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLOS ONE* 5(4):e10271

Feizi S, Marbach D, Médard M, Kellis M. 2013. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* 31:726–33

Fienberg SE, Martin ME, Straf ML. 1985. *Sharing Research Data*. Washington, DC: Nat. Acad. Press

Forrest WF, Cavet G. 2007. Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers." *Science*. 317(5844):1500

Gavish M, Donoho D. 2012. Three dream applications of verifiable computational results. *Comput. Sci. Eng.* 14(4):26–31

Gentleman R, Temple Lang D. 2007. Statistical analyses and reproducible research. *J. Comput. Graph. Stat.* 16(1):1–23

Getz G, Höfling H, Mesirov JP, Golub TR, Meyerson M, et al. 2007. Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers." *Science* 317(5844):1500

Grove CC. 1930. Review of *Statistical Methods for Research Workers* by RA Fisher. *Am. Math. Mon.* 37(10):547–50

Groves RM, Lyberg L. 2010. Total survey error: past, present, and future. *Public Opin. Q.* 74(5):849–79

Heller R, Bogomolov M, Benjamini Y. 2013. Deciding whether follow-up studies have replicated findings in a preliminary large-scale "omics' study." arXiv:1310.0606 [stat.AP]

Herndon T, Ash M, Pollin R, 2013. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Camb. J. Econ.* 38:257–79

Hines WC, Su Y, Kuhn I, Polyak K, Bissell MJ. 2014. Sorting out the FACS: a devil in the details. *Cell Rep.* 6:779–81

Hiltzik M. 2013. Science has lost its way, at a big cost to humanity. *Los Angeles Times*, Oct. 27. **http://articles.latimes.com/2013/oct/27/business/la-fi-hiltzik-20131027**

Holm S. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6(2):65–70

Hong W-J, Warnke R, Chu G. 2005. Immune signatures in follicular lymphoma. *N. Engl. J. Med.* 352:1496–97

Ioannidis JPA. 2005. Why most published research findings are false. *PLOS Med.* 2(8):e124

Jasny BR, Chin G, Chong L, Vignieri S. 2011. Again, and again, and again . . . . *Science* 334(6060):1225–25

Johnson G. 2014. New truths that only one can see. *New York Times*, Jan. 20. **http://www.nytimes.com/2014/01/21/science/new-truths-that-only-one-can-see.html**

Johnson VE. 2013. Revised standards for statistical evidence. *PNAS* 110(48):19313–17

Katsnelson A. 2014. Male researchers stress out rodents. *Nat. News*, 28 Apr. **http://www.nature.com/news/male-researchers-stress-out-rodents-1.15106**

Kelly T, Marians K. 2014. Rescuing US biomedical research: some comments on Alberts, Kirschner, Tilghman, and Varmus. *PNAS* 111:E2632–33

King G. 1995. Replication, replication. *PS Polit. Sci. Polit.* 28:443–99

Kleiner A, Talwalkar A, Sarkar P, Jordan MI. 2014. A scalable bootstrap for massive data. *J. R. Stat. Soc. B.* 76:795–816

Kreuter F, Peng RD. 2014. Extracting information from Big Data: issues of measurement, inference and linkage. See Lane et al. 2014, pp. 257–75

Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, et al. 2012. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* 490(7419):187–91

Lane JI, Stodden V, Bender S, Nissenbaum H, eds. 2014. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. New York: Cambridge Univ. Press

Lehmann EL. 1993. The Fisher, Neyman–Pearson theories of testing hypotheses: One theory or two? *J. Am. Stat. Assoc.* 88:1242–49

LeVeque RJ. 2007. Wave propagation software, computational science, and reproducible research. *Proc. Int. Congr. Math., Madrid, Aug. 22–30*, pp. 1227–54. Zurich, Switz.: Eur. Math. Soc. Preprint URL: **http://faculty.washington.edu/rjl/pubs/icm06/icm06leveque.pdf**

LeVeque RJ, Mitchell IM, Stodden V. 2012. Reproducible research for scientific computing: tools and strategies for changing the culture. *Comput. Sci. Eng.* 14(4):13–17

Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* 5(3):1752–79

Lim C, Yu B. 2013. Estimation stability with cross validation (ESCV). arXiv:1303.3128 [stat.ME]

Madigan D, Stang PE, Berlin JA, Schuemie M, Overhage JM, et al. 2014. A systematic statistical approach to evaluating evidence from observational studies. *Annu. Rev. Stat. Appl.* 1:11–39

Masicampo EJ, Lalande DR. 2012. A peculiar prevalence of *p* values just below .05. *Q. J. Exp. Psychol.* 65(11):2271–79

McNutt M. 2014a. Raising the bar. *Science* 345(6192):9

McNutt M. 2014b. Reproducibility. *Science* 343(6168):229

Nat. Publ. Group. 2012. Must try harder. *Nature* 483(7391):509

Nat. Publ. Group. 2013. Announcement: reducing our irreproducibility. *Nature* 496(7446):398

Nat. Res. Counc. (Nat. Res. Counc. Comm. Math. Found. Verif. Valid. Uncertain. Quantif.). 2012. *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. Washington, DC: Nat. Acad. Press. **http://www.nap.edu/catalog/13395/assessing-the-reliability-of-complex-models-mathematical-and-statistical-foundations**

Nat. Res. Counc. (Nat. Res. Counc. Comm. Responsib. Authorship Biol. Sci.). 2003. *Sharing Publication-Related Data and Materials Responsibilities of Authorship in the Life Sciences*. Washington, DC: Nat. Acad. Press. **http://www.nap.edu/catalog/10613/sharing-publication-related-data-and-materials-responsibilities-of-authorship-in**

Neyman J, Pearson ES. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. A* 231:289–337

Nosek BA, Spies JR, Motyl M. 2012. Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* 7(6):615–31

NSF (Nat. Sci. Found.). 2011. *NSF Data Management Plan Requirements*. Arlington, VA: Nat. Sci. Found. **http://www.nsf.gov/eng/general/dmp.jsp**

NSF (Nat. Sci. Found.). 2013. *GPG Summary of Significant Changes*. Arlington, VA: Nat. Sci. Found. **http://nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_sigchanges.jsp**

Obama B. 2013. *Executive order—making open and machine readable the new default for government information*. White House, Off. Press Secr., Washington, DC, May 9. **http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government**

Open Source Initiative. 2013a. *The BSD 2-Clause License*. Palo Alto, CA: Open Source Initiative. **http://opensource.org/licenses/BSD-2-Clause**

Open Source Initiative. 2013b. *The MIT License* (*MIT*). Palo Alto, CA: Open Source Initiative. **http://opensource.org/licenses/MIT**

Pachter L. 2014. The network nonsense of Manolis Kellis. *Bits of DNA: Rev. Comment. Comput. Biol. Blog*, Feb. 11. **https://liorpachter.wordpress.com/2014/02/11/the-network-nonsense-of-manolis-kellis/**

Panel Sci. Responsib. Conduct Res., Comm. Sci. Eng. Public Policy, Nat. Acad. Sci., Nat. Acad. Eng., Inst. Med. 1992. *Responsible Science*, Volume I: *Ensuring the Integrity of the Research Process.* Washington, DC: Nat. Acad. Press. **http://www.nap.edu/catalog/1864/responsible-science-volume-i-ensuring-the-integrity-of-the-research**

Peng RD. 2009. Reproducible research and biostatistics. *Biostatistics* 10(3):405–8

Peng RD. 2011. Reproducible research in computational science. *Science* 334(6060):1226–27

Polanyi M. 1962. *Personal Knowledge: Towards a Post-Critical Philosophy*. London: Routledge

Polanyi M. 1967. *The Tacit Dimension*. Garden City, NY: Anchor Books

Popper KR. 2002. *The Logic of Scientific Discovery*. London: Routledge

Prinz F, Schlange T, Asadullah K. 2011. Believe it or not: How much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* 10(9):712–12

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna: R Found. Stat. Comput. **http://www.R-project.org/**

Reich ES. 2011. Cancer trial errors revealed. *Nature* 469:139–40

Reinhart CM, Rogoff KS. 2010. Growth in a time of debt. *Am. Econ. Rev.* 100:573–78

Rockey S. 2014. Changes to the biosketch. *Extramural Nexus*, May 22. **http://nexus.od.nih.gov/all/2014/05/22/changes-to-the-biosketch/**

Rosenthal R. 1979. The “file drawer problem” and tolerance for null results. *Psychol. Bull.* 86(3):638–41

Rubin AF, Green P. 2007. Comment on “The Consensus Coding Sequences of Human Breast and Colorectal Cancers.” *Science* 317(5844):1500

Sandve GK, Nekrutenko A, Taylor J, Hovig E. 2013. Ten simple rules for reproducible computational research. *PLOS Comput Biol.* 9(10):e1003285

Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. 2014. Interpreting observational studies: why empirical calibration is needed to correct *p*-values. *Stat. Med.* 33(2):209–18

Shapin S, Schaffer S. 1989. *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton, NJ: Princeton Univ. Press

Shi J. 2014. Seeking the principles of sustainable software engineering. arXiv:1405.4464 [cs.DC]

Siegfried T. 2013. Science’s significant stats problem: Researchers’ rituals for assessing probability may mislead as much as they enlighten. *Nautilus*, Aug. 22. **http://nautil.us/issue/4/the-unlikely/sciences-significant-stats-problem**

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22(11):1359–66

Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, et al. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* 314:268–74

Staudt LM, Wright G, Dave S. 2005. Immune signatures in follicular lymphoma. *N. Engl. J. Med.* 352:1496–97

Stebbins M. 2013. Expanding public access to the results of federally funded research. *OSTP Blog*, Feb. 22. **http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research**

Stock JH, Watson MW. 2011. *Introduction to Econometrics*. Boston: Addison-Wesley. 3rd ed.

Stodden V. 2009a. Enabling reproducible research: licensing for scientific innovation. *Intl. J. Comm. Pol.* 13:1

Stodden V. 2009b. The legal framework for reproducible scientific research: licensing and copyright. *Comput. Sci. Eng.* 11(1):35–40

Stodden V. 2011. Trust your science? Open your data and code. *Amstat News*, July 1. **http://magazine.amstat.org/blog/2011/07/01/trust-your-science/**

Stodden V. 2012. Reproducible research: tools and strategies for scientific computing. *Comput. Sci. Eng.* 14(4):11–12

Stodden V. 2013. Resolving irreproducibility in empirical and computational research *IMS Bull*. Nov. 17. **http://bulletin.imstat.org/2013/11/resolving-irreproducibility-in-empirical-and-computational-research/**

Stodden V. 2014a. Intellectual property and computational science. In *Opening Science: The Evolving Guide on How the Internet Is Changing Research, Collaboration and Scholarly Publishing*, ed. S Bartling, S Fiesike, pp. 225–35. New York: Springer Open

Stodden V. 2014b. Enabling reproducibility in Big Data research: balancing confidentiality and scientific transparency. See Lane et al. 2014, pp. 112–32

Stodden V, Bailey DH, Borwein J, LeVeque RJ, Rider W, Stein W, eds. 2012. *Setting the default to reproducible: reproducibility in computational and experimental mathematics*. Collab. Rep., ICERM (Inst. Computat. Exp. Res. Math.), Brown Univ., Dec. 10–14, Providence, RI. **http://icerm.brown.edu/html/programs/topical/tw12_5_rcem/icerm_report.pdf**

Stodden V, Borwein J, Bailey D. 2013. "Setting the default to reproducible" in computational science research. *SIAM News*, Jun. 3. **http://www.siam.org/news/news.php?id=2078**

Stodden V, Leisch F, Peng RD. 2014. *Implementing Reproducible Research*. Boca Raton, FL: CRC Press

Stodden V, Miguez S. 2013. Best practices for computational science: software infrastructure and environments for reproducible and extensible research. *J. Open Res. Softw.* 2:e21

Stodden V, Miguez S, Seiler J. 2015. ResearchCompendia.org: cyberinfrastructure for reproducibility and collaboration in computational science. *Comput. Sci. Eng.* 17:12–19

Tibshirani R. 2004. *Re-analysis of Dave et al., NEJM Nov. 18, 2004*. Rep., Stanford Univ., Stanford, CA. **http://statweb.stanford.edu/~tibs/FL/report/**

Tibshirani R. 2005. Immune signatures in follicular lymphoma. *N. Engl. J. Med.* 352:1496–97

WSSSPE (Work. Sustain. Softw. Sci. Pract. Exp.). 2013. Contributions. *WSSSPE Workshop 1, Denver, CO*, Nov. 17. **http://wssspe.researchcomputing.org.uk/wssspe1/contributions/**

Yu B. 2013. Stability. *Bernoulli* 19(4):1484–500

Annual Review of Statistics and Its Application

Volume 2, 2015

Contents