Is Most Published Research Really False?

Jeffrey T. Leek[1,2] and Leah R. Jager[1]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205; email: jtleek@gmail.com

[2]Center for Computational Biology, Johns Hopkins University, Baltimore, Maryland 21205

Annu. Rev. Stat. Appl. 2017. 4:109–22

First published online as a Review in Advance on October 5, 2016

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

This article's doi:
10.1146/annurev-statistics-060116-054104

Keywords

science-wise false discovery rate, reproducibility, replicability, false discoveries, meta-analysis, reliability research

Abstract

There has been an increasing concern in both the scientific and lay communities that most published medical findings are false. But what does it mean to be false? Here we describe the range of definitions of false discoveries in the scientific literature. We summarize the philosophical, statistical, and experimental evidence for each type of false discovery. We discuss common underpinning problems with the scientific and data analytic practices and point to tools and behaviors that can be implemented to reduce the problems with published scientific results.

1. INTRODUCTION

Most published research findings are false.

This statement seems absurd on the first reading. Scientific research is carried out by highly trained and skilled scientists, vetted through peer review, and publicly scrutinized once it appears in journals. The entire scientific publishing infrastructure was originally conceived to prevent the publication of incorrect results and provide a forum for correcting false discoveries (Csiszar 2016). It seems inconceivable that most of the findings that pass through this process are false.

But this system was invented before modern computing, data generation, scientific software, email, the Internet, and social media. Each of these inventions has placed strain on the scientific publication infrastructure. These modern developments have happened during the careers of practicing scientists. Many laboratory leaders received their training before the explosion of cheap data generation, before the widespread use of statistics and computing, and before there was modern data analytic infrastructure (Irizarry 2012). At the same time, there has been increasing pressure from review panels, hiring committees, and funding agencies to publish positive and surprising results in the scientific literature. These trends have left scientists with a nagging suspicion that some fraction of published results are at minimum exaggerated and at worst outright false.

This suspicion was codified in a widely read paper titled "Why most published research findings are false" (Ioannidis 2005b). The paper focuses on how application of the hypothesis testing framework could result in a preponderance of false positive results in the medical literature. The key argument is analogous to population-based screening using a biomarker. A very sensitive and specific biomarker may still have a low positive predictive value if the disease is not prevalent in the population. Analogously, suppose that the null hypothesis is true for 99% of the hypotheses being considered by scientific investigators. If scientists test 1,000 hypotheses with a statistical power of 80%, then $1{,}000 \times 1\% \times 80\% = 8$ true alternative hypotheses should be correctly detected. Even though the type I error rate is much lower, the prevalence of null hypotheses is much higher. With a type I error rate of 5%, we expect $1{,}000 \times 99\% \times 5\% = 49.5$ null hypotheses will be incorrectly detected. In this situation, $1 - \frac{8}{8+49.5} = 86\%$ of rejected hypotheses will actually be null. Assuming selective reporting of positive results, a lower prevalence of true alternatives in hot fields, or other sources of bias in the publication process makes this estimate even worse.

Under this argument, a false research finding is defined as a finding where the null hypothesis is true but is incorrectly called statistically significant by a researcher. At its core is an argument about the way statistical evidence is assessed across the medical literature. But this is hardly the only difficulty with the modern scientific process. There are a variety of other ways that a statistical result can be considered false or suspect. It is possible to confuse correlation with causation, a predictive model may overfit the training data, a study may be underpowered, and results may be overinterpreted or misinterpreted by the scientific press (Leek 2015, Leek & Peng 2015). Each of these scenarios may undermine the credibility of a scientific result without satisfying the definition of false from a hypothesis testing perspective.

The added complexity of the scientific process also contributes to the difficulties in verifying the results from scientific publications. It used to be sufficient to describe scientific results using text and figures, but it is increasingly necessary to describe a set of scientific results through scientific code, by releasing a data set, or through a series of complex protocols that cannot be completely described in text (Peng 2011). Because this change has happened during the careers of many active scientists, there has been a slow realization that publications frequently do not provide sufficient detail to describe the scientific and computational protocols included in a study.

This complexity raises the question of how often it is possible to reproduce or replicate the results of a particular scientific study. Reproducibility is defined as the ability to recreate all of the

figures and numbers in a scientific publication from the code and data provided by the authors. Replicability is defined as the ability to reperform the experiments and computational analyses in a scientific publication and arrive at consistent results. Both of these ideas are related to the rate of false discoveries in the medical literature. If a study cannot be reproduced, then it is impossible to fully assess whether the evidence supports any claims of statistical significance. If a study does not replicate, then it raises the question of whether the original report was a false discovery.

This review tackles each of these challenges separately and attempts to summarize what we know about the current state of false discoveries in science. We also point to ongoing and potential initiatives to reduce the statistical problems within the medical and scientific literature.

2. DEFINING FALSE DISCOVERIES IN THE MEDICAL LITERATURE

Here we consider a range of potential definitions of issues in the medical literature, starting with computational issues and concluding with false discoveries at the completion of an analysis. Any single scientific or medical study consists of a question of interest, an experimental design, an experiment, data sets, analysis code, and conclusions.

We will consider the following issues with the medical literature (Patil et al. 2016a):

- **Reproducibility.** A study is reproducible if all of the code and data used to generate the numbers and figures in the paper are available and exactly produce the published results.
- **Replicability.** A study is replicable if an identical experiment can be performed like the first study and the statistical results are consistent.
- **False discovery.** A study is a false discovery if the result presented in the study produces the wrong answer to the question of interest.

Reproducibility is the easiest of these problems to both define and assess. Assessing reproducibility involves checking the published manuscript, looking for published data and code, then comparing the results of that data and code to the published results. If they are the same, the study is reproducible, and if they are not, then the study is not.

Replicability is a more challenging concept to both define and measure. A study replicates if the same experiment can be performed a second time with consistent results. If the data collected during the study are subject to sampling variability, even in the best-case scenario the results of a replication will not be identical to the original study. However, we would expect that the results would be within the range of values predicted by the parameter estimates and variability estimates from the original study. The difficulties in assessing replicability are compounded by potential for publication bias, regression to the mean, fragility of scientific results to a particular context, and imperfect replication.

A false discovery is the most challenging of these three problems to assess. A false discovery means that the reported parameter or answer to a scientific question is not consistent with the underlying natural truth being studied. A false discovery is the most difficult to assess because we rarely know the true state of nature for any particular scientific study. Single replications are not sufficient to separate true discoveries from false discoveries because both the original study and the replication are subject to sampling error and other potential difficulties with replication studies. Repeated replications or near replications that all point to a similar conclusion are the best way to measure false discoveries in the medical literature. However, repeated replication or near replication of identical studies is very expensive and tends to only occur for highly controversial ideas—such as the claim that vaccines cause autism.

3. WHAT IS THE RATE OF REPRODUCIBILITY OF THE SCIENTIFIC LITERATURE?

We begin by considering the rate of reproducibility of scientific studies. Computational reproducibility was first identified as a key component of the scientific process more than two decades ago within the computational and statistical community (Buckheit & Donoho 1995). But the role of reproducibility in the broader scientific process was not highlighted until the mid-2000s across a variety of fields from biostatistics (Peng 2009) and epidemiology (Peng et al. 2006) to physics (Buckheit & Donoho 1995). Ultimately the reproducibility of scientific results became a baseline standard by which the data analysis in any particular study should be judged.

Prior to the widespread knowledge of the importance of reproducible research, many papers did not provide data and code. Part of this issue was cultural—it was not well known that providing data and code was a key component of the scientific publication process. This issue was compounded because most scientists were not trained in statistics and computation. Moreover, the tools for creating and sharing reproducible documents were often difficult to use for people without sufficient computational training.

There have been multiple studies and papers evaluating reproducible research across different disciplines. We list some of them below.

- **Study:** "Repeatability of published microarray gene expression analyses" (Ioannidis et al. 2009)

  Main idea: This paper attempts to collect the data used in published papers and to repeat one randomly selected analysis from each paper. For many of the papers, the data were either not available or available in a format that made it difficult or impossible to repeat the analysis performed in the original paper. The types of software used in the original papers were also not clear.

  Important drawback: This paper focused on 18 data sets published in 2005–2006. This paper was published early in the era of reproducibility and so is potentially driven by cultural change in genomics studies 10 years ago.

- **Study:** "Toward reproducible computational research: an empirical analysis of data and code policy adopted by journals" (Stodden et al. 2013)

  Main idea: The authors evaluated code and data sharing policies for 170 journals. They found that 38% and 22% of these journals had data and code sharing policies, respectively.

  Important drawback: The chosen journals were more computationally focused, although they also included high-impact journals such as *Nature Genetics*, *Cell*, and *The Lancet*.

- **Study:** "Believe it or not: how much can we rely on published data on potential drug targets?" (Prinz et al. 2011)

  Main idea: 67 studies that focus primarily on oncology produce relevant code and data only 20% of the time.

  Important drawback: The data, code, and methods used to perform this study are not available and so it is not a scientific study.

- **Paper:** "Next-generation sequencing data interpretation: enhancing reproducibility and accessibility" (Nekrutenko & Taylor 2012)

  Main idea: As part of an opinion piece, the authors randomly selected 50 of 378 papers in 2011 that used the same alignment technique. Of these papers, only 7 provided

all necessary details for reproducibility; 19 provided enough information on software implementation, and almost half provided access to data.

Important drawback: The set of papers examined focuses on a relatively small area of science, high-throughput biology.

- **Study:** "Public availability of published research data in high-impact journals" (Alsheikh-Ali et al. 2011)

  Main idea: In a sample of 500 research articles from 50 high-impact journals in 2009, only 9% of articles made raw data fully available. Of the 50 journals surveyed, 44% had a policy explicitly requiring materials to be available as a condition of publication and 12% had no specific policy for data availability. Even when data-sharing policies were in place, more than half of the sampled articles did not meet the journal's stated requirements for data availability.

  Important drawback: Data availability is only one piece of reproducibility; to be fully reproducible code must be available as well.

- **Study:** "Case studies in reproducibility" (Hothorn & Leisch 2011)

  Main idea: 100 randomly sampled papers from *Bioinformatics* were sampled and evaluated for code and data availability. Code for simulations was available more than 80% of the time, data were available approximately 50% of the time when used, and code was available slightly more than 60% of the time when used.

  Important drawback: This paper represents a relatively solid evaluation of the rate of reproducibility, but only in the bioinformatics world, where the rate of reproducibility might be artificially high.

In addition to these empirical evaluations, there has been a large literature dedicated to philosophical and opinion-based pieces on reproducibility in the sciences. There is likely some truth to these opinion pieces, but they tell us little about the actual rate of reproducibility. Certainly, the rate at which studies are reproducible varies by discipline. The most extensive studies of reproducibility have occurred within the bioinformatics community, where the rate of computational and statistical sophistication is arguably higher than other disciplines. Research areas with a mature view of statistics and computation likely produce more reproducible research than areas newly introduced to computational and empirical research.

Lack of reproducibility itself does not certainly imply that a particular scientific result is false. A study may be fully correct including the reporting of a true positive and not be fully reproducible. A fully reproducible study may be riddled with errors and could be wrong. Without full access to the data and code describing a particular scientific result it is difficult to assess how credible that result should be. Regardless, most of the effort to improve the scientific literature has so far focused on reproducibility because it is very difficult to evaluate replicability or true and false discoveries without data and code.

4. WHAT IS THE RATE OF REPLICATION IN THE MEDICAL LITERATURE?

Reproducibility is concerned with obtaining the exact results of a published study once the code and data are available. A higher standard is the ability to replicate a published study. Replication involves following the published protocol and repeating the entire experiment, data collection, and data analysis process. The goal is to determine if the result of the replication is consistent with the original study.

Replication studies include several levels of subtlety. The first is defining what it means to be a successful replication. Definitions of successful replication have included consistent effect sizes, consistent distributions, and consistent measures of statistical significance. The difficulty is that both original and replication experiments involve multiple sources of variation. Sampling variation alone may explain why a result may be significant in one study and not another. In other cases, regression to the mean and publication bias may lead to reductions in the estimated effects in replication studies. Finally, replication studies are extremely hard to perform well because variations in measurement technology, population characteristics, study protocols, and data analyses may lead to nonreplication.

Despite the expense of performing replication of scientific experiments, there is increasing energy and attention being focused on these studies. Typically only the most important and widely publicized studies have been the focus of replication, but that is changing with larger-scale studies of replication and increased incentives to perform this type of research. Here we focus on a subset of important replication studies.

- **Study:** "Estimating the reproducibility of psychological science" (Open Science Collaboration 2015)

  Main idea: The goal of this Reproducibility Project: Psychology study is to re-perform 100 experiments in psychology and evaluate the replicability of results in psychological science. A widely reported claim is that many of the results are not replicable; the study found 36% of the experiments were significantly replicated.

  Important drawback: The study only performs a single replication of each study with potential issues due to power, bias, and study design.

- **Study:** "What should researchers expect when they replicate studies? A statistical view of replicability in psychological science" (Patil et al. 2016b)

  Main idea: A re-analysis of the psychological science replication data shows that 77% of the replication effects fall within the 95% prediction interval for the effect size based on the original study.

  Important drawback: This paper does not model the potential biases in either the original or replication studies. Consistency of effect estimates does not say anything about whether the results are true or false discoveries.

- **Study:** "Contradicted and initially stronger effects in highly cited clinical research" (Ioannidis 2005a)

  Main idea: This paper looks at studies that attempted to answer the same scientific question when the second study had a larger sample size or more robust (e.g., randomized trial) study design. Some effects reported in the second study do not match the results exactly from the first.

  Important drawback: The title does not match the results. 16% of studies were contradicted (meaning the effect was in a different direction). 16% reported smaller effect size, 44% were replicated and 24% were unchallenged. So $44\% + 24\% + 16\% = 84\%$ were not contradicted. Lack of replication is also not proof of error.

- **Study:** "A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic" (Mobley et al. 2013)

  Main idea: Researchers surveyed faculty/trainees at MD Anderson Cancer Center about their ability to replicate published results. Of those who responded, 55% reported that they had tried and failed to replicate a result from a published paper. For those who

contacted the authors of the original paper, less than 40% received a positive/helpful response in reply.

Important drawback: The study was based on a voluntary sample with a low response rate, and participants were all from a single institution.

These replication efforts suggest that some studies replicate and some studies do not. The best estimates of replication are available for the 13 studies in the Many Labs Replication Project, which estimated the impact of differing study designs, experimental procedures, and personnel. In that case, 10 of 13 studies replicated; however, this represents only the most high-profile results of a relatively narrow area of psychology. The replication rate is, again, likely field specific. In psychology, where most of the replication efforts have been concentrated, the true rate of replication likely lies between the 36% of studies with statistically significant results in the same direction in the Reproducibility Project: Psychology and the 77% of studies that replicated in the Many Labs Replication Project.

The reality is that consistent replication of all scientific studies is not a viable approach. Currently, only papers that garner sufficient interest and attention in the form of high-profile publication or press will be subject to replication. An alternative strategy would be to perform random audits of specific disciplines or sub-disciplines, like the Many Labs project, to provide an overall estimate of the replication rate for a particular field.

5. WHAT IS THE RATE OF FALSE DISCOVERIES IN THE MEDICAL LITERATURE?

Replication represents an attempt to confirm knowledge that we think we have garnered from a scientific study. But a single replication of a scientific result is not sufficient to measure the true state of the world. If only a single replication is performed, there is variation in both the original study and the replication study. So if the results are not consistent, we can't be sure if that is because either study has produced a false result or if it is simply expected owing to natural variation.

The problem of the rate of false discoveries in the medical literature has primarily been addressed through theoretical arguments and discussion. Only recently have efforts been made to empirically estimate the rate of false discoveries in the scientific or medical literature. Some of the key studies in the area include the following.

- **Paper:** "Why most published research findings are false" (Ioannidis 2005b)

  Main idea: Researchers use hypothesis testing to determine whether specific scientific discoveries are significant. This significance calculation is used as a screening mechanism in the scientific literature. Given certain assumptions about the way people perform these tests and report them, it is possible to construct a universe where most published findings are false positive results.

  Important drawback: The paper contains no real data; it is purely based on a theoretical argument about the behavior of scientists and how they choose hypotheses to study.

- **Study:** "Drug development: raise standards for preclinical research" (Begley & Ellis 2012)

  Main idea: Many drugs fail when they move through the development process. Amgen scientists tried to replicate 53 high-profile basic research findings in cancer and could only replicate 6.

  Important drawback: This is not a scientific paper. The study design, replication attempts, selected studies, and the statistical methods to define "replicate" are not defined. No data are available or provided.

- **Study:** "An estimate of the science-wise false discovery rate and application to the top medical literature" (Jager & Leek 2014)

  Main idea: The paper collects *p*-values from published abstracts of papers in the medical literature and uses a statistical method to estimate the false discovery rate proposed by Ioannidis (2005b). This paper estimates the rate of false discoveries at 14%.

  Important drawback: The paper only collected data from major medical journals and the abstracts; *p*-values can be manipulated in many ways that could call into question the statistical results in the paper.

- **Study:** "Investigating variation in replicability: a 'Many Labs' replication project" (Klein et al. 2014)

  Main idea: The paper considered 13 published high-profile results and used multiple labs to replicate the results. They successfully replicated 10 out of 13 results, and the distribution of results allows the measurement of the potential effect of regression to the mean.

  Important drawback: The paper covers only 13 high-profile experiments in psychology. This is by far the strongest, most comprehensive, and most reproducible analysis of replication among all the papers surveyed here.

Only the study of published *p*-values (Jager & Leek 2014) and the Many Labs Replication Project (Klein et al. 2014) empirically evaluate the rate of false discoveries in the medical literature. The study of published *p*-values suffers from potential issues with publication and other biases (Gelman & O'Rourke 2014), but it is one of the only comprehensive estimates of the science-wise false discovery rate that is available. This study suggested that the rate of false discoveries in the medical literature was inflated, but only to 14%.

The Many Labs project represents a much higher standard, involving complete replication of individual studies in multiple labs, for evaluating the rate of false discoveries in the literature. With repeated replications by multiple groups, it is possible to distinguish effects that are consistent from those that are not. This study suggested that 24% of results were false discoveries. However, this approach to evaluating the veracity of the literature is time consuming and expensive, and we only have this estimate for a small sample of psychology experiments.

Overall, the empirical rate of false discoveries in the medical literature is unknown. Our only empirical estimates—however shaky or specific they may be—suggest that most published research is not false. A key challenge for the future is designing studies that are rigorous, cover a broad fraction of important scientific disciplines, and are inexpensive and efficient enough to be performed regularly.

6. WHAT IS THE RATE OF INCORRECTLY ANALYZED DATA IN THE MEDICAL LITERATURE?

Among all studies that are replicable and reproducible, the primary source of false discoveries is likely incorrect experimental design or data analysis (Leek 2015). There have recently been two very public examples of incorrect analysis producing questionable or false results. The first involves a series of papers that claimed to identify genomic signatures that could predict the response of patients to chemotherapy (Potti et al. 2006). These studies were originally not reproducible. The irreproducibility of these studies has been a major source of discussion (Baggerly & Coombes 2009, Baggerly 2010, Micheel et al. 2012). Reproducible versions of the analyses in the studies revealed major flaws in the statistical methods and computational software used to make the

predictions (Baggerly & Coombes 2009). These flaws included problems with study designs, incorrect probabilistic statements, and predictions subject to random error.

A second public example, in economics, involves a paper that suggested that a high debt-to-GDP ratio was related to decreased economic growth (Rogoff & Reinhart 2010). The paper was ultimately used by politicians and economists as justification for austerity measures around the globe. A graduate student discovered an error in the spreadsheet used by the authors, and this error received both widespread academic and press attention. However, the major issues with the analysis were the extremely limited sample size, choices about which data points to exclude, and unconventional weighting schemes.

These cases highlight the issues that arise with data analysis in nearly every paper in the scientific literature. As data become increasingly common and complicated, so too do the software and statistical methods used to analyze the data. There is increasing concern about the use of p-values, underpowered studies, potential confounders, the winner's curse, robustness of analytic pipelines, researcher degrees of freedom, and p-value hacking in scientific studies. There have now been studies focused on evaluating the extent of these issues in scientific and medical studies (Aschwanden & King 2015).

There is a clear connection between poorly designed or analyzed studies and both a lack of replication and an increased risk of false discoveries. But we know little about the data analytic practices used across individual studies. There is some indication that certain fields tend to have low-powered and observational studies (Button et al. 2013). Other fields have focused on the use of well-designed randomized trials. In either case, as data become more complicated, the role of data analytic pipelines on the ultimate validity of statistical results is not well understood. Understanding and improving the data analytic process still remains the surest way to reduce false discoveries in the medical literature.

7. HOW CAN WE IMPROVE THE SCIENTIFIC AND MEDICAL LITERATURE?

We do not believe science is in the midst of a crisis of reproducibility, replicability, and false discoveries. But there are a large number of scientific studies that do suffer from the underlying computational and statistical issues (Aschwanden & King 2015). Addressing the real statistical problems in science requires efforts to improve the incentive system for scientists, to produce tools that simplify computation and statistics, and to increase training in computation, software development, and data analysis (Ioannidis 2014).

7.1. Improving Incentives

One of the major barriers to reproducibility, replicability, and true discoveries in science is an imbalanced incentive system. There is strong pressure to publish only positive results and to not publish replication studies. There is pressure against reusing published data, with strong advantages for data generators over data analysts. To improve the statistical side of science, we need incentive systems to be aligned with solid statistical efforts.

Some incentive programs that are underway are the following.

- **The preregistration challenge (https://cos.io/prereg/).** This challenge provides researchers with an incentive of $1,000 to preregister their hypotheses before performing a study. These studies will avoid p-hacking, outcome switching, and other potential problems with discovery based research.

- **Transparency and openness promotion guidelines (Nosek et al. 2015).** These guidelines, adopted by hundreds of journals, commit these journals to promoting standards of openness, data sharing, and data citation.
- **National Science Foundation (NSF) and National Institutes of Health (NIH) research output definitions (Collins & Tabak 2014, NIH 2015, NSF 2015).** Both the NSF and NIH have recently made efforts to legitimize research outputs beyond academic publications including software production, data generation, and scientific outreach activities.

These initiatives represent excellent efforts to improve the incentive system for scientists who wish to perform reproducible, replicable, and accurate research. However, the major remaining impediment to the incentive system is the value placed on specific research activities by hiring, promotion, and grant-evaluation committees. These committees hold incredible sway over the careers of all scientists, but particularly the junior scientists who are most likely to adopt practices that avoid the issues we have discussed here.

The highest-impact potential incentive would be for these committees to publicly and formally recognize the value in activities that take away time from publication but contribute to good science (Allen & Leek 2013). Specifically, by placing emphasis on behaviors such as software maintenance, data generation and sharing, and peer review, these committees could have an immediate and dramatic impact on the rate of reproducibility, replicability, and accuracy in the scientific literature.

7.2. Improving Tools

One of the key impediments to performing reproducible and replicable research in the past was the lack of availability of tools that make it easy to share data, share code, and perform correct analyses. Fortunately there has been an explosive growth in the availability of tools for these purposes over the past ten years. For general purpose analysis, the growth of the R programming language and the Python for Data Analysis community has led to a host of free software for performing a wide range of analyses.

There are now many tools that facilitate the distribution of reproducible analysis code and pipelines, including the following.

- **`knitr` and `rmarkdown` (Xie 2015, Allaire et al. 2015).** These are R packages that have been developed for creating documents that interweave text, code, and figures. These documents are fully reproducible within the R environment and can be easily created by students with only basic R training.
- **Jupyter Notebooks (Pérez & Granger 2007).** These create interactive and static documents that interweave text, code, and figures. These documents are fully reproducible and are supported by Github (a popular code-sharing repository) so that they can be easily viewed online.
- **Galaxy (Goecks et al. 2010).** This is an infrastructure for creating reproducible work flows by combining known tools. Galaxy workflows are reproducible and can be shared with other researchers.

There are similarities and differences across these different platforms, but the commonality is that they add negligible effort to an analyst's data analytic workflow. Both `knitr` and Jupyter Notebooks have primarily increased reproducibility among scientists who have some scripting experience. A major reason for their popularity is that code is written as usual but is simply embedded in an easy-to-use document. The workflow is not changed for the analysts, because

they were going to write the code anyway. The platform just allows it to be included into an easily shareable document.

Galaxy has increased reproducibility for a range of scientists, but the target user is someone who has less scripting experience. The Galaxy project has worked hard to make it possible for everyone to analyze data reproducibly. The reproducibility is almost incidental—a user would have to stitch pipelines together anyway, and Galaxy provides an easy way to do so. Reproducibility comes as an added bonus.

Data can now also be easily shared. Researchers used to post data to their personal websites, but they were largely impermanent in that state. When researchers changed institutions or web servers, data often were lost—link rot has been one of the most fundamental sources for lack of reproducibility in science. Now, however, there are a variety of permanent places data can be made publicly available. General purpose sharing sites include the following.

- **Figshare (https://figshare.com/):** If it is designated as publically available, an unlimited amount of data can be posted on this site free of charge. Private data storage requires a fee. Figshare accepts all data types and provides a digital object identifier (doi) so data can be cited.
- **Open Science Framework (https://osf.io/):** This site can be used to post all types of data sets and can be integrated with Figshare.
- **Dataverse (King 2007):** This is a free data storage hosted by Harvard.

Data can also be hosted in a variety of field-specific data repositories. Specific repositories have also been developed for handling sensitive data, such as personally identifiable information. As a result, there are resources available for all but the largest data-sharing efforts.

7.3. Improving Data Analysis

Even with the improvements in incentives and tools, a major obstacle to improving the issues with statistics in science involves providing sufficient training to the students and postdocs who perform most data analyses. It is no longer feasible to expect that all data analysis will be performed by a person with advanced training in statistics and data analysis. Every lab across every scientific discipline is now engaged in the generation of abundant and cheap data. There are simply not enough data analysts with advanced training to keep pace.

It is clear that improving data analysis skills in the scientific community requires an investment in training, both to improve statistical skills and to educate analysts about good statistical practices. Recently, for example, the American Statistical Association released a statement on the use of p-values (Wasserstein & Lazar 2016), hoping to educate researchers outside of the statistical community about the appropriate interpretation and use of p-values in science.

Several educational initiatives have risen to meet this demand for training. Some are short, intensive workshops in software engineering and data analysis. Others are online courses that can be used to dramatically scale basic training in these areas.

- **Software carpentry (Wilson 2006):** workshops on reproducible research, software design, and software use that are run at locations around the world.
- **Data carpentry (Teal et al. 2015):** workshops on data cleaning, management, and analysis that are run at locations around the world.
- **Massive open online courses (Gooding et al. 2013):** a range of data science, data analysis, and reproducible research courses that are available to all researchers around the world.

Table 1 Strategies for assessing and improving three major problems with the scientific literature

	Reproducibility	Replicability	False discoveries
Strategies for assessment of a single study	■ Are code and data both readily available? ■ Does running the provided code exactly produce the study results?	■ Does the paper provide enough information that the experiment can be exactly duplicated? ■ Upon duplication of the study, are the results within the range of values predicted by estimates from the original study?	■ Is the experimental design appropriate for the research question? ■ Are analysis methods properly applied? ■ Are statistical results properly interpreted?
Strategies for assessment across science	Random audits of papers from specific disciplines		
Strategies for improvement	■ Journal requirements for data/code sharing ■ Easy-to-use tools for distributing data and code ■ Incentives for sharing data and code	■ Publication incentives for performing replication studies	■ Statistical training for all scientists ■ Incentives for good statistical practice

Materials and courses are now available to learn these critical techniques and improve the scientific process. Unfortunately, however, many scientific programs, including all medical education programs, do not require a single statistics or data analysis class. Given how frequently data and statistical analysis play a role in research that is published in the medical literature, the key next step is to integrate these training modules into the formal education of all scientists.

8. CONCLUSIONS

We have summarized the empirical evidence that most published research is not reproducible, not replicable, or false. Our summary suggests that the most extreme opinions about these issues likely overstate the rate of problems in the scientific literature.

These problems may be exacerbated by the onslaught of increasingly large data sets. However, it seems that the areas most inundated with new data—genomics, neuroimaging, wearable computing and social networks—have been the quickest to adapt to large data sets and reproducible practices. It has been pointed out that no matter the size of the data, good experimental design, clear communication, and open code and data have the largest potential impact on all of the issues we have discussed in this review (Leek 2015, Kass et al. 2016).

There is clearly still work to be done. **Table 1** summarizes the three types of problems, how they can be assessed, and strategies for improvement. It is clear that statistics and data analysis now play a central role across all sciences, including medical science. We need to work to encourage the adoption of best practices and the implementation of available tools to improve the accuracy of published scientific results.

DISCLOSURE STATEMENT

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

ACKNOWLEDGMENTS

Parts of the text of this review have appeared in the blog posts "A summary of the evidence that most published research is false" (Leek 2013) and "Why the three biggest positive contributions to reproducible research are the iPython Notebook, knitr, and Galaxy" (Leek 2014) and the book *How to Be a Modern Scientist* (Leek 2016).

LITERATURE CITED

Allaire JJ, Cheng J, Xie Y, McPherson J, Chang W, et al. 2015. `rmarkdown`: dynamic documents for R. **http://rmarkdown.rstudio.com/**

Allen G, Leek J. 2013. Changing our culture: perspectives from young faculty. *Amstat News*, Dec. 1. **http://magazine.amstat.org/blog/2013/12/01/changing-our-culture/**

Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JP. 2011. Public availability of published research data in high-impact journals. *PLOS ONE* 6(9):e24357

Aschwanden C, King R. 2015. Science isn't broken. *FiveThirtyEight Science*, Aug. 19. **http://fivethirtyeight.com/features/science-isnt-broken/#part1**

Baggerly K. 2010. Disclose all data in publications. *Nature* 467(7314):401

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

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

Buckheit JB, Donoho DL. 1995. WaveLab and reproducible research. In *Wavelets and Statistics*, ed. A Antoniadis, G Oppenheim, pp. 55–81. New York: Springer

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, et al. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14(5):365–76

Collins FS, Tabak LA. 2014. NIH plans to enhance reproducibility. *Nature* 505(7485):612

Csiszar A. 2016. Peer review: troubled from the start. *Nat. News* 532(7599):306

Gelman A, O'Rourke K. 2014. Discussion: difficulties in making inferences about scientific truth from distributions of published *p*-values. *Biostatistics* 15(1):18–23

Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11(8):R86

Gooding I, Klaas B, Yager JD, Kanchanaraksa S. 2013. Massive open online courses in public health. *Front. Public Health* 1:59

Hothorn T, Leisch F. 2011. Case studies in reproducibility. *Brief. Bioinform.* 12(3):288–300

Ioannidis JP. 2005a. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 294(2):218–28

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

Ioannidis JP. 2014. How to make more published research true. *PLOS Med.* 11(10):e1001747

Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, et al. 2009. Repeatability of published microarray gene expression analyses. *Nat. Genet.* 41(2):149–55

Irizarry R. 2012. People in positions of power that don't understand statistics are a big problem for genomics. *Simply Statistics Blog*, Apr. 27. **http://simplystatistics.org/2012/04/27/people-in-positions-of-power-that-dont-understand/**

Jager L, Leek J. 2014. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* 15(1):1

Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, Reid N. 2016. Ten simple rules for effective statistical practice. *PLOS Comput. Biol.* 12(6):e1004961

King G. 2007. An introduction to the dataverse network as an infrastructure for data sharing. *Sociol. Methods Res.* 36(2):173–99

Klein RA, Ratliff KA, Vianello M, Adams RB Jr., Bahník Š, et al. 2014. Investigating variation in replicability: a "Many Labs" replication project. *Soc. Psychol.* 45:142–52

Leek JT. 2013. A summary of the evidence that most published research is false. *Simply Statistics Blog*, Dec. 16. **http://simplystatistics.org/2013/12/16/a-summary-of-the-evidence-that-most-published-research-is-false/**

Leek JT. 2014. Why the three biggest positive contributions to reproducible research are the iPython Notebook, knitr, and Galaxy. *Simply Statistics Blog*, Sep. 4. **http://simplystatistics.org/2014/09/04/why-the-three-biggest-positive-contributions-to-reproducible-research-are-the-ipython-notebook-knitr-and-galaxy/**

Leek JT. 2015. *The Elements of Data Analytic Style*. **https://leanpub.com/datastyle**. Victoria, Can.: Leanpub

Leek JT. 2016. *How to Be a Modern Scientist*. **https://leanpub.com/modernscientist**. Victoria, Can.: Leanpub

Leek JT, Peng RD. 2015. Statistics: *P* values are just the tip of the iceberg. *Nature* 520(7549):612

Micheel CM, Nass SJ, Omenn GS, eds. 2012. *Evolution of Translational Omics: Lessons Learned and the Path Forward*. Washington, DC: Natl. Acad. Press

Mobley A, Linder SK, Braeuer R, Ellis LM, Zwelling L. 2013. A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. *PLOS ONE* 8(5):e63221

Nekrutenko A, Taylor J. 2012. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.* 13(9):667–72

NIH (Natl. Inst. Health). 2015. Update: new biographical sketch format required for NIH and AHRQ grant applications submitted for due dates on or after May 25, 2015. Not. No. NOT-OD-15-032, Natl. Inst. Health, Bethesda, MD. **http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-032.html**

Nosek B, Alter G, Banks G, Borsboom D, Bowman S, et al. 2015. Promoting an open research culture: author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science* 348(6242):1422

NSF (Natl. Sci. Found.). 2015. Dear colleague letter—supporting scientific discovery through norms and practices for software and data citation and attribution. Doc. No. 14-059. Natl. Sci. Found., Arlington, VA. **http://www.nsf.gov/pubs/2014/nsf14059/nsf14059.jsp**

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349:6251

Patil P, Peng RD, Leek JT. 2016a. A statistical definition for reproducibility and replicability. *Cold Spring Harb. Labs J.* **http://biorxiv.org/content/early/2016/07/29/066803**

Patil P, Peng RD, Leek JT. 2016b. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.* 11:539–44

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

Peng RD, Dominici F, Zeger SL. 2006. Reproducible epidemiologic research. *Am. J. Epidemiol.* 163(9):783–89

Pérez F, Granger BE. 2007. IPython: A system for interactive scientific computing. *Comput. Sci. Eng.* 9(3):21–29

Potti A, Dressman HK, Bild A, Riedel RF, Chan G, et al. 2006. Genomic signatures to guide the use of chemotherapeutics. *Nat. Med.* 12(11):1294–300

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

Rogoff K, Reinhart C. 2010. Growth in a time of debt. *Am. Econ. Rev.* 100(2):573–8

Stodden V, Guo P, Ma Z. 2013. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. *PLOS ONE* 8(6):e67111

Teal TK, Cranston KA, Lapp H, White E, Wilson G, et al. 2015. Data carpentry: workshops to increase data literacy for researchers. *Int. J. Digit. Curation* 10(1):135–43

Wasserstein RL, Lazar NA. 2016. The ASA's statement on *p*-values: context, process, and purpose. *Am. Stat.* 70:129–33

Wilson G. 2006. Software carpentry. *Comput. Sci. Eng.* 8:66–69

Xie Y. 2015. *Dynamic Documents with R and knitr*. Boca Raton: Chapman & Hall/CRC

Annual Review of Statistics and Its Application

Volume 4, 2017

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics