ANNUAL REVIEWS

Annual Review of Public Health

Reproducible Research: A Retrospective

Roger D. Peng and Stephanie C. Hicks

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA; email: rdpeng@jhu.edu, shicks19@jhu.edu

Annu. Rev. Public Health 2021. 42:79–93

First published as a Review in Advance on January 19, 2021

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

https://doi.org/10.1146/annurev-publhealth-012420-105110

Keywords

reproducibility, replicability, data analysis

Abstract

Advances in computing technology have spurred two extraordinary phenomena in science: large-scale and high-throughput data collection coupled with the creation and implementation of complex statistical algorithms for data analysis. These two phenomena have brought about tremendous advances in scientific discovery but have raised two serious concerns. The complexity of modern data analyses raises questions about the reproducibility of the analyses, meaning the ability of independent analysts to recreate the results claimed by the original authors using the original data and analysis techniques. Reproducibility is typically thwarted by a lack of availability of the original data and computer code. A more general concern is the replicability of scientific findings, which concerns the frequency with which scientific claims are confirmed by completely independent investigations. Although reproducibility and replicability are related, they focus on different aspects of scientific progress. In this review, we discuss the origins of reproducible research, characterize the current status of reproducibility in public health research, and connect reproducibility to current concerns about the replicability of scientific findings. Finally, we describe a path forward for improving both the reproducibility and replicability of public health research in the future.

INTRODUCTION

Scientific progress has long depended on the ability of scientists to communicate to others the details of their investigations. The exact meaning of "details of their investigations" has changed considerably over time and in recent years has been nearly impossible to describe precisely using traditional means of communication. Rapid advances in computing technology have led to large-scale and high-throughput data collection coupled with the creation and implementation of complex statistical algorithms for data analysis. In the past, it might have sufficed to describe the data collection and analysis using a few key words and high-level language. However, with today's computing-intensive research, the lack of details about the data analysis in particular can make it impossible to recreate any of the results presented in a paper (36). Compounding these difficulties is the impracticality of describing these myriad details in traditional journal publications using natural language. To address this communication problem, a concept has emerged known as reproducible research, which aims to provide for others far more precise descriptions of an investigator's work. As such, reproducible research is an extension of the usual communications practices of scientists, adapted to the modern era.

The notion of reproducible research, which was popularized in the early 1990s, was ultimately designed to address an emerging and serious issue at the time (46). Results of published findings were increasingly dependent on complex computations done on powerful computers, often implementing sophisticated algorithms on large data sets. Given the importance of computing to the generation of these results, it was surprising that consumers of scientific results had no ability to inspect or examine the details of the original computations. Traditional forms of scientific publication allowed for extended descriptions of study design and high-level analysis approaches, but low-level details about computer code, data processing pipelines, and algorithms were not prioritized and were generally left in an appendix or, with the wider availability of the Internet, an online supplement.

Jon Claerbout, a geophysicist at Stanford University, wrote down many of the original ideas concerning reproducibility of computational research. His concern focused largely on developing a software system whereby the research produced by his lab could be passed on to others, including the original authors, the authors' colleagues, students, research sponsors, and the general public. He noted in particular the benefits of reproducibility to the original authors: "It may seem strange to put the author's own name at the top of the list to whom we wish to provide the reproducible research, but it often seems that the greatest beneficiary of preparing the work in a reproducible form is the original author!" (10, p. 1). It is equally notable that the public was listed last; all the other constituencies mentioned would likely exist within the small orbit of an individual investigator (10). In Claerbout's discussion, the primary focus is on improving the transparency and productivity of the lab itself, given that much time can be lost attempting to recreate past findings for the sole purpose of understanding what was previously done.

Buckheit & Donoho introduced much of the statistical community to the concept of reproducibility with an influential paper in 1995 detailing their WaveLab software for implementing wavelets for data analysis (9). Citing Claerbout as a strong influence, Buckheit & Donoho's rationale for promoting reproducible research produced a useful summary of Claerbout's ideas that has since been repeated many times:

> An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

The general conclusion was that delivering a research end product such as a figure or table was no longer sufficient. Rather, the software environment and the means to create the end product

must also be delivered, as those additional elements represent the actual scholarship. To satisfy this requirement, one would have to make available the data and the computer code used to generate the results.

A searing example of the complexity of data analysis and the critical need to verify data and computations can be taken from the coronavirus disease 2019 (COVID-19) pandemic. Early in the pandemic, researchers hypothesized that the drug hydroxychloroquine may be useful for treating COVID-19, but there was little evidence for or against the hypothesis. In May 2020, a large observational study was published in the *Lancet*, which claimed that there was no benefit to using hydroxychloroquine (30). Soon after publication, numerous questions were raised about the data analysis and the nature of the data themselves (27). Although the original authors commissioned an independent audit of the analysis, the company that owned the data refused to turn over the data set to the independent auditors. Without the data set, there was no way to reproduce the original analysis, and ultimately the paper was retracted (31). This example demonstrates the importance of having a clear understanding of how data analysis was conducted when critical life and death decisions must be made.

REPRODUCIBLE RESEARCH

The definition of reproducible research generally consists of the following elements. A published data analysis is reproducible if the analytic data sets and the computer code used to create the data analysis are made available to others for independent study and analysis (36). This definition is sufficiently vague that it ultimately raises more questions than it answers. What is an "analytic data set"? What does it mean to be "available"? What is included with the "computer code"?

Published research can be thought of as living on a continuum up until the point of publication (37), starting from question formulation and study design, proceeding to data collection, data processing, and data analysis, and finally to presentation. Along this journey, various elements are introduced to aid in executing the research, such as computing environments, measurement instruments, and software tools. One could choose to make available to others any aspect of this sequence, depending on the practicalities of doing so and the relevance to the final published results. It is challenging to develop a universal cut point for determining which aspects of an investigation should be disseminated and which are not required. However, within various research communities, internal standards have been developed and are continuously evolved to keep pace with technology (6, 44).

The analytic data set generally contains all the data that can be directly linked to a published result or number. For example, if a paper publishes an estimate of the rate of hospitalization for heart attacks, but the overall study also collected data on hospitalizations for influenza, the influenza data may not be part of the analytic data set if it makes no appearance in the published result and is not otherwise relevant. While outside investigators may be interested in seeing the influenza data (and the original authors may be happy to share them), these data are not needed for the sake of reproducible research. The analytic computer code is any code that was used to transform the analytic data set into results, which may include some data processing (such as variable transformations) as well as modeling or visualization. Generally, the software environment in which the analysis was conducted (e.g., R, Python, Matlab) does not need to be distributed if it is easily obtainable or open source. However, niche software, which may be unfamiliar to many readers, may need to be bundled with the data and code.

Many see reproducibility as a nonissue, upon first consideration. How could it be that applying the original computer code to the original data set would not produce the original results? The practical reality of modern research, though, is that many, even simple, results depend on a precise record of procedures and processing and the order in which they occur. Furthermore, many

statistical algorithms have many tuning parameters that can dramatically change the results if they are not inputted exactly the same way as was done by the original investigators (21). If any of these myriad details are not properly recorded in the code, then there is a significant possibility that the results produced by others will differ from the original investigators' results.

A Sidebar on Terminology

The terminology of reproducible research can be bewildering to some in the scientific community because there is little agreement about the meaning of the phrase with regard to other related concepts (3, 20, 38). In particular, one related concept with which all scientists are concerned is what we refer to here as replication. In this review, we define replication as the independent investigation of a scientific hypothesis with newly collected data, new experimental setups, new investigators, and possibly new analytic approaches. In a thorough investigation of the terminologies of reproducible research, Barba found that some fields of study made no distinction between "reproducible" and "replicable," whereas some fields used those terms to mean the exact opposite of how we define them here (3, 32). However, a significant plurality of fields, including epidemiology, medicine, and statistics, appear to adopt the definitions we use here.

A key distinction between reproducibility and replication is that reproducibility does not allow for any real variation in the results. If an independent investigator were to reproduce the results of another investigator with the original data and code, there should not be any variation between the two investigators' results, except for some allowance for differences in machine precision. Thus, exact reproducibility is sometimes referred to as "bitwise reproducibility" (32). However, replication generally allows for differences in results that arise from statistical variability. Two independent investigators conducting the exact same experiment should, in theory, differ only by an amount quantified by the standard deviation of the data. More generous definitions of replication allow for slightly different study designs, analytic populations, or statistical techniques (32). In those cases, differences in results may arise beyond simple statistical variation. Patil et al. (35) have devised a useful visualization of what may or may not differ when reproducing or replicating a published study.

It is difficult to argue that interest in exactly reproducing another investigator's work is anything but a modern phenomenon (13). Interest in reproducibility prior to the computer and Internet age was likely low or nonexistent, given that there was generally no expectation that investigators would share data in papers—there was simply no practical way to do that except for very small data sets. In the past, other investigators could resort to independently replicating a published study only by using their own data collection and whatever high-level description of the methods that was available in the paper. In this setting, detailed descriptions of the methods of analysis were critical if other investigators were to execute the same approach. If the process of conducting the experiment or analysis were simple enough or were sufficiently standardized, then it could be reasonably described in the confines of a journal article. Suggesting that analyses be described with data and code is a departure from previous ways of communicating scientific results, which relied on describing experiments and analyses in more general terms to give readers the highlights of what was done. A more abstract approach could not be taken with this new form of computational research because the proper abstractions for communicating ideas and standardization of approaches were not yet available.

Newer Developments

The concept of reproducible research was developed to achieve arguably modest goals. Its original aims focused on providing an approach to better communicate the details of computationally

intensive research to one's collaborators, colleagues, students, and oneself. But two key developments over the past 30 years have changed the context around which reproducible research lives. Although the definition of reproducibility has not changed much since the 1990s, almost everything else about scientific research has.

In much of the early literature on reproducible research, the focus is on "computationally intensive" research, which, because of its reliance on complex computer algorithms, was considered perhaps more impenetrable than other research. Fast-forward 30 years and the use of computing in scientific research is ubiquitous. It is no longer the domain of niche geophysical scientists or mathematical statisticians using obscure computer packages. Now, all scientific research involves the use of powerful computers, whether it is for the data collection, the data analysis, or both. Furthermore, the increase in complexity of statistical techniques over this time period has resulted in the need for detailed descriptions of analytic approaches and data processing pipelines. We are all computational scientists now, and thus the concept of reproducibility is relevant to all scientists.

Along with computing power, another key advance over the past 30 years has been the development of the Internet. Claerbout's original scheme for distributing data and code to others was via CD-ROM discs, which was a perfectly reasonable approach at the time. However, the need for a physical medium greatly limited the transferability of information to a large audience. With the development of the Internet, it became possible for academics to distribute data and code to the entire world for seemingly minimal cost. This increase in distribution reach changed the nature and importance of reproducible research from primarily improving the internal efficiency of the lab to allowing others to anonymously build on another researcher's work. The Internet dramatically grew the size of an investigator's personal scientific community to include many members beyond their immediate circle of collaborators. This phenomenon has provided significant benefits to science, but some implications could threaten the viability of maintaining and supporting reproducible research in the long run. Before we consider these implications, we must first identify the goals of reproducible research and the problems we want reproducibility to solve.

THE GOALS OF REPRODUCIBLE RESEARCH

Beyond communicating the details of an investigation, what are the goals of making research reproducible? The stated goals achieved by making research reproducible have evolved over time since the early 1990s and have become somewhat more elusive. The original goal was to better reveal the process of doing the research. Computational research added a new complexity in the form of software code and high-dimensional data sets, and that complexity made understanding the research process more difficult for a reader to infer. Therefore, the solution was to simply publish every step in the process along with the data. Claerbout and colleagues were concerned that others (including themselves!) would not be able to learn from what they had done if they did not have the details. The easiest way to make this information available was via the literal computer code that executed the steps. Any less precise format could risk omitting a key step that affected downstream results (21).

Reproducible research comes with a few side benefits. In addition to being able to fully understand the process by which the results were obtained, readers also have access to the data and the computer code, both of which are valuable to the extent that they can be reused or repurposed for future studies or research. Some have suggested that making data and computer code available to others is a per se goal of reproducible research because both can be built on and leveraged to further scientific knowledge (17). However, such an interpretation is an extension of the original ideas of reproducibility. The former view saw data and code as a medium for communicating ideas,

whereas the latter view sees data and code essentially as products or digital research objects to be used by others (48). While converting a data set into a data product and packaging computer code into usable software may seem like nominal tasks, given that the underlying data and code already exist, there are non-negligible costs associated with the development, maintenance, and support of these products.

Another goal of reproducible research is to provide a kind of audit trail, should one be needed. In fact, one could suggest a definition of reproducible research as "research that can be checked." Desiring an audit trail for data analyses raises the question of when such an audit trail might be used. In general, one might be interested in seeing the details of the data and the code for an analysis when a researcher is curious about how a specific result was reached. Sometimes that curiosity is raised because of suspicion of an error in the analysis, but other times there is a desire to learn the details of new techniques or methodologies (21). Thus, reproducibility concerns primarily the integrity and transparency of the data analysis for an investigation. Unlike replication, reproducibility allows for an internal check on the results and is not immediately connected to the context of the outside world.

REPRODUCIBILITY AND DATA ANALYSIS

One could summarize the goal of reproducible research as providing a means to answer the question, "Do I understand and trust this data analysis?" With the computational nature of today's research, we cannot hope to answer that question without being able to look at the data and the code. In addition, we may wish to know things about the experimental or study design as well as the hypothesis being examined (22). Given the claimed results, the data, and code, one can theoretically determine the reproducibility status of a data analysis. Reproducibility gives us the means by which we can assess our confidence and trust in an analysis, but it is important to reiterate that reproducing an analysis is not a check on the validity of the analysis.

The notion of reproducibility as a binary or perhaps multilevel "state" of a data analysis is a useful characterization in part because it is one of a few qualities of a data analysis that can be immediately verified. Unlike with replication, we do not need to wait for future studies to be conducted in order to determine the reproducibility of an analysis. However, this notion suggests that reproducibility's usefulness is limited. What do we ultimately learn from merely reproducing the results of an analysis? For example, it may be possible to execute code on a data set without ever looking at the code or the data. In that case, the original goal of reproducibility—to learn about the details of an investigation—has been thwarted. We have simply learned that the code produces what the authors claim the code produces. In general, executing a process and seeing that process produce the results exactly as they were expected produce very little new information.

The answer to the question "Do I understand and trust this data analysis?" depends critically on the perspective of the person asking the question. If the person asking is an expert in the area, they might be able to glance at the code and data and understand immediately what is going on. A nonexpert in the field might be able to execute the code and produce results without ever understanding the operations of the analysis. An adjacent question that might be worth asking is, "Is this data analysis understandable and trustworthy?" However, this question is not any easier to answer because it hypothesizes underlying objective qualities of a data analysis. But opinions may still vary widely about what these underlying qualities should be, depending on who is asking the question. To answer either question, one needs to look carefully at the data and the code to learn exactly what was done. But ultimately, the data and code represent only part of the answer. Whether an analysis can be understood or trusted depends critically on many aspects outside of the analysis itself, including the perspective of the person reading the analysis.

Nevertheless, one hope is that reproducibility can lead to higher-quality data analyses. The logic infers that requiring all analyses to provide data and code would put investigators on notice that their work would be scrutinized. However, one high-profile example suggests that enforcing reproducibility is unlikely to result in better data analyses.

Example: Forensic Bioinformatics

In a now-retracted 2006 study by Potti et al. (40), the investigators claimed to have identified genomic signatures using microarrays that could predict whether an individual responded well to chemotherapy. The analysis was conducted using data from publicly available cell lines, and so the data were in a sense available. However, subsequent attempts to reproduce the findings failed and reproducibility was achieved only when errors were deliberately inserted into the analysis code (2, 11). Baggerly, Coombes, and Wang meticulously reconstructed the error-prone analysis and laid out all the details in both text and code. Ironically, they were ultimately able to reproduce the analysis of Potti et al. (40) after significant reverse engineering and forensic investigation. In fact, we might never have learned what mistakes were made if Baggerly and his colleagues were not able to reproduce the analysis.

The Potti study is a pathological example of a reproducible analysis (after much forensic investigation) being profoundly incorrect. However, it is worth asking what role reproducibility might have played in this case? If Potti et al. had released the code and data that were clearly linked together, perhaps as a research compendium (17), then the errors could have been found more quickly. However, given the sheer number and complexity of the problems, it still likely would have taken some time to understand them all. Coombes et al. (11) published their letter only a year after the initial publication, so the timeline might have been advanced by a few months. However, a key fact would remain: The flawed analysis was already completed. Furthermore, once the truth was ultimately revealed to the authors, it took years of further investigation by many others before the original paper was retracted.

Reproducibility and Quality: A Prevention Model?

Examples such as the Potti paper raise the question of whether demanding or requiring reproducibility of a study beforehand can preemptively improve its quality. Evidence of this connection between reproducibility and quality is lacking, which is not surprising given that the question is somewhat ill-posed. What exactly are we looking for in a "high-quality" data analysis? One could hypothesize that if an investigator knew in advance that the data and the code would be publicly available for scrutiny, then they would take the extra effort to make sure that the analyses were properly done. Perhaps if Potti et al. had been forced to make their code publicly available, they would have checked it first.

In the case of Potti et al., we now know that requiring reproducibility or even just code sharing would not have made much difference. Reporting done by *The Cancer Letter* showed definitively that the investigators were aware of numerous statistical and coding errors with the analysis but did not think that the problems were serious (19). Rather, they were considered "differences of opinion." The notion that requiring reproducibility can lead to improved data analyses relies on the critical assumption that the investigators are able to recognize what is an error in the first place. If they do recognize the error and hide it, then that is fraud. If they do not recognize the error and publish the paper anyway, then that is at best careless. However, in both cases, forcing the data and code to be published would not have made any difference.

Replication and Reproducibility

Reproducibility does not provide a useful route to preventing poor data analyses from occurring, but it does provide the basis for a meaningful discussion about whether there might be problems in the analysis and how such problems might be fixed. Replication differs from reproducibility primarily because it addresses a different goal. Replication answers the question, "Is this scientific claim true?" Reproducibility addresses the integrity of the data analysis that generated the evidence for a scientific claim, whereas replication addresses the integrity of the claim itself in the context of the outside world. Fundamentally, reproducible research has little to say on the question of external validity. Claims resulting from reproducible results can be both correct and incorrect (28). Claims resulting from irreproducible results are less likely to be true, but that may depend on the reasons for the lack of reproducibility. For example, evidence generated via random algorithms may not be exactly reproducible if random number generator seeds are not saved, but the underlying evidence may still be sound. Ultimately, claims made by irreproducible studies may in fact be true, but irreproducible studies simply do not provide evidence for such claims.

Example: Reanalysis of Air Pollution Studies

In the mid-1990s, two large studies of ambient air pollution and mortality—The Six Cities Study (12) and the American Cancer Society (ACS) Study (39)—were published, presenting evidence that differences in air pollution concentrations between cities were significantly associated with rates of mortality in these cities. Both studies came under intense scrutiny when the US Environmental Protection Agency (EPA) cited the results in its revision of the National Ambient Air Quality Standards for fine particles. In particular, there were demands from numerous corners that the data used in the studies should be made available. However, the data in these studies, as with most health-related studies, included personal information about the subjects and thus the original investigators argued that promises of confidentiality had to be kept. To address the impasse of making the data available, the original investigators engaged the Health Effects Institute (HEI) to serve as a kind of trusted third party to broker a reanalysis of the studies. Ultimately, HEI recruited a research team led by investigators at the University of Ottawa to obtain the original data for both the Six Cities Study and the ACS Study, reproduce the original findings, and conduct additional sensitivity analyses to assess the robustness of the original findings (26).

The extensive reanalysis found that the original studies were largely, if not perfectly, reproducible. For the Six Cities Study, the key result was a mortality relative risk of 1.26, which the reanalysis team computed to be 1.28. For the ACS Study, the original mortality relative risk was 1.17, close to the reanalysis value of 1.18. While one could argue that these studies were strictly speaking not reproducible, such small differences are not likely to be material. In fact, we now know, after numerous follow-up studies and independent replications, that the core findings of both studies appear to be true (8) and that the US EPA itself rates the evidence of a connection between fine particles and mortality to be "likely causal" (15). The reanalysis team ran many other analyses, including variables that had not been considered in the original studies. Overall, they found that the sensitivity analyses did not change any of the major conclusions. Interestingly, one of the key conclusions of the final HEI report stated that at the end of the day "[n]o single epidemiologic study can be the basis for determining a causal relation between air pollution and mortality" (26, p. iv).

The HEI reanalysis of the Six Cities Study and the ACS Study highlights the role of trust in data analysis. Prior to the reanalysis, many parties simply did not trust that the analysis was done properly or that all reasonable competing hypotheses had been considered. While making the data available might have allowed others to build that trust for themselves, allowing a neutral third party

to examine the data and reproduce the findings at least ensured that one other group had seen the data. In addition, HEI's role in organizing the expert panel, conducting public outreach, and managing an open process played an important role in building trust in the community. Although not all parties were completely satisfied with the process, the reanalysis allowed fellow scientists to learn from the original studies and gain insight into the process that led to the original findings. The key goals of reproducible research were ultimately achieved.

In hindsight, another lesson learned from the HEI reanalysis is that the importance of reproducibility of a given study can fade with time. More than 25 years later, scores of follow-up studies and replications have come to largely similar conclusions as those in the Six Cities Study and the ACS Study. Although both studies remain seminal in the field of air pollution epidemiology, they could be deleted from the literature at this point and have little impact on our understanding of the science. This is not to say that the data and ongoing analyses do not have value, but rather to say that the original results have been subsumed by later studies. Reproducibility was critical when the studies were first published only because of the paucity of large studies at the time.

REPRODUCIBILITY AND BETTER DATA ANALYSIS

Recent work has focused on the quality and variability of data analyses published in various fields of study (23, 24, 33, 34), with some researchers claiming the existence of a "replication crisis" due to the wide variation among studies examining the same hypotheses (45). The causes of this variation among studies are myriad, but one large category includes various aspects of the data analysis. Because of the increasing complexity of data analyses, many choices and decisions must be made by analysts in the process of obtaining a result. With these increasing complexities, we also increase the risk of human error and bias in data analysis. These choices and decisions often have an unknown impact on the final estimates produced and therefore may or may not be recorded by the investigators (48). These "research degrees of freedom" allow investigators to unknowingly, or perhaps knowingly, steer data analyses in directions that may support specific hypotheses rather than represent all the evidence in the data (47).

What role can reproducible research play in improving the quality of data analyses across all fields? The answer can be found in part with the experience of the HEI reanalysis of the Six Cities Study and the ACS air pollution study. Because the HEI-sponsored studies were reanalyses, one could expect that the results would be confirmed to some reasonable degree. If there was a significant deviation from the published results, then we would have to dig into the original analysis to discover why. Because the results were largely reproduced, one could argue that little was learned. However, additional analyses were done and sensitivity analyses were conducted. As a result, we learned much about the data analysis process. The reanalysis thus produced valuable knowledge about how to analyze air pollution and health data.

For example, the reanalysis team noted that both mortality and air pollution were highly spatially correlated, a feature that was not considered in the original analysis. They noted,

> If not identified and modeled correctly, spatial correlation could cause substantial errors in both the regression coefficients and their standard errors. The Reanalysis Team identified several methods for dealing with this, all of which resulted in some reduction in the estimated regression coefficients. (26, p. iii)

In addition, reproducibility helps free up time for the analysts who are interested in reanalyzing the data to focus on parts of the data analysis that require more human interpretation. For example, if an independent data analyst knew that an analysis was already reproducible, then more time

and resources would be available to understand why a specific model was chosen, instead of what version of software was used to run this model. In the reanalysis of the data from Potti et al. (40), Baggerly and Coombes noted that they had spent thousands of hours re-examining the data attempting to reproduce the original results (1, 18).

There are also different degrees of reproducibility when building a data analysis and differences in audiences that may or may not be allowed to have access to these components. For example, a data analyst may choose to make the data available but not the code (or the opposite). Others may make both the code and data available for only one audience (Audience A) but not for another audience (Audience B). There are valid reasons why an analyst might choose to do this, such as if the data analysis uses data with protected health information in a hospital setting or if the data analyst works at a business or company and cannot share the code or data with others outside of the company. It is important to note that just because an analysis is not fully reproducible to one audience (Audience B) does not mean that it is an invalid analysis with incorrect conclusions. Although the access limitations do make it more difficult for Audience B to trust the results, the analysis can still be valid or correct. However, the lack of reproducibility to this audience may mean that the evidence supporting any claims is weaker. Despite these potential differences in degrees of reproducibility, as demonstrated in the HEI reanalysis, efforts made to make a data analysis more reproducible is a step in the right direction for making it a better data analysis.

The reproducibility of research, when possible, ultimately allows us a significant opportunity to (*a*) learn from others about how best to analyze certain types of data; (*b*) reduce human errors and bias as data become larger and more complex; (*c*) free up time for reanalyzers to focus on parts of a data analysis that require more human interpretation; (*d*) have discussions about what makes for a good data analysis in certain areas of study; and (*e*) improve the quality of future data analyses. When teaching data analysis to students, it is common to talk in abstractions and theories, describing statistical methods and models in isolation. When real data are shown, they are often in the form of toy examples or in short excerpts. Increasing the reproducibility of all studies presents an opportunity to dramatically expand instruction on the craft of data analysis so that core set of elements and principles for characterizing high-quality analyses can be established within a field (22).

REFINING REPRODUCIBILITY

In the 30 years since the idea of reproducible computational research was brought to the forefront of the research community, we have learned much about its role and its value in the research enterprise. The original goal of providing a transparent means by which researchers can communicate what they have done and allow others to learn remains a primary motivator. Reproducibility has a secondary role to play in improving the quality of data analysis in that it serves as the foundation on which people can learn how others analyze data. Without code and data, it is nearly impossible to fully understand how a given data analysis was actually performed. But much about computational research has changed in the past 30 years, and we can perhaps develop a more refined notion about what it means to make research reproducible. The two key ideas about reproducibility—data and code—are worth revisiting in greater detail.

Data

The sharing of data is valuable in and of itself. Data sharing, to the extent possible, reduces the need for others to collect similar data, allows for combined analyses with other data sets, and can create important resources for unforeseen future studies. Data sets can retain their value for a

considerable period of time, depending on the area and field of study. One example of the value of data sharing comes from the National Mortality, Morbidity, and Air Pollution Study, a major air pollution epidemiology study conducted in the late 1990s and early 2000s (42, 43). The mortality data for this study were shared on a website and then later updated with new data. A systematic review found that 67 publications had used the data set, often to demonstrate the development of new statistical methodology (4). In addition, the release of the data at the time allowed for a level of transparency and trust in air pollution research that was novel for its time.

Today, many data sharing Web repositories exist, which allow easy distribution of data of almost any size. Whereas in the past an investigator interested in sharing data had to purchase and set up a Web server, now investigators can simply upload to any number of services. The Open Science Framework (16), Dataverse Project (25), ICPSR (50), and SRA (29) are only a handful of public and private repositories that offer free hosting of data sets. The major benefit of repositories such as these is to absorb and consolidate the cost of hosting data for possibly long periods of time.

The view of data sharing as inherently valuable is not without its challenges. Indeed, stripping data from their original context can be problematic and lead to inappropriate off-label reuse by others. Stodden has argued that data have value only in their explicit connection to the knowledge that they produce and that we must be careful to preserve the connections between the data and the knowledge they generate (49).

Best practices for sharing data have recently been developed. Some of these practices are specific to areas of study, while others are more generic. In particular, the emergence of the concept of tidy data has provided a generic format for many different types of data that serves as the backbone of a wide variety of analytic techniques (52). Practical guidance on sharing data via commonly used spreadsheet formats (7) and on providing relevant metadata to collaborators is now widely applicable to many kinds of data (14).

Code

The primary purpose of sharing code is to communicate what was done to transform the data into scientific results. Today, almost all actions relevant to the science will have been carried out on the computer, so we must have a precise way to document those actions. Computer code, via any number of programming and data analytic languages, is the most precise way to do that. The sharing of code generally represents less of a technical burden than the sharing of data. Code tends to be much smaller in size than most data sets and can easily be served by code-sharing services such as GitHub, BitBucket, SourceForge, or GitLab.

While the benefits of code sharing tend to focus on the code's usability and potential for repurposing in other applications, it is important to reiterate that code's primary purpose is to communicate what was done. In short, code is not equivalent to software. Software is code that is designed and organized specifically for use by others in a wide variety of scenarios, often abstracting away operational details from the user. The usability of software depends critically on aspects such as design, efficiency, modularity, and portability—factors that should not generally play a role when releasing research code. Sharing research code that is poorly designed and inefficient is far preferable to not sharing code at all. That said, this notion does not preclude the possibility for best practices in developing and sharing research code.

Software is often a product of research activity, particularly when new methodology is developed. In those cases, it is important that the software is carefully considered and designed well for its intended users. However, it should not be considered a requirement of reproducible research that software be a product of research. For software that is developed for distribution, there is increasing guidance for how such software should be distributed. Software package development has

become easier for programming languages such as R, which have robust developer and user communities (41), and numerous tools have been developed to make incorporating code into packages more straightforward for nonprofessional programmers (53). In addition, the concept of testing and test-based development has been a useful framework for setting expectations on how software should perform and identifying errors or bugs (51).

FUTURE DIRECTIONS

Technological trends over time generally favor a more open approach to science as the costs of sharing, hosting, and publishing have gone down. The continuing rapid advancement of computing technology, Internet infrastructure, and algorithmic complexity will likely introduce new challenges to reproducible research. As the scientific community expands its sharing of data and code, some important issues should be considered going forward.

The rapidly evolving nature of scientific communication serves to highlight the role of reproducibility in advancing science. Without reproducibility, countless hours could be wasted simply trying to figure out what was done in a study. In situations where key decisions must be made on the basis of scientific results, it is important that the robustness of the findings can be assessed quickly without the need for guessing or inferences about the underlying data. A stark example can be drawn from the COVID-19 pandemic. In April 2020, little was known about the disease, and a study was published on medRχiv that produced an estimate of the prevalence of COVID-19 in the population (5). At the time, important public health decisions had to be made in response to the pandemic, and any information about the disease would have been highly relevant. Upon publication of the study, numerous criticisms about the study's design and analysis appeared on social media and the Web. However, the aspect most relevant to this review is that in many of the critiques, substantial time was taken to simply guess at what the researchers had done. Although a written statistical appendix was provided with the paper, no data or code was published along with the study. As a result, independent investigators had little choice but to infer what was done.

The urgency of decision-making based on scientific evidence can exist in a variety of situations, not just on the minute-by-minute timescale of a worldwide pandemic. Many regulatory decisions in environmental health have to be made on the basis of only a handful of studies. Often, public health officials and policy makers cannot wait years for another large cohort study to replicate (or not) existing findings. In such situations where decisions need to be made, the more code and data that can be made available to assess the evidence, the more informed those decisions will be. In the interim, follow-up studies can be conducted and revisions to the evidence base can be made in the future if needed. The reanalyses of the Six Cities Study and the ACS Study provide a clear example of this process, and history has shown those results to be highly consistent across a range of replication studies.

The maintenance of code and data is generally not a topic that is discussed in the context of reproducible research. When a paper is published, it is sent to a journal and is considered finished by the investigators. Unless errors are found in the paper, one generally need not revisit a paper after publication. However, both code and data need to be maintained, to some degree, in order to remain useful. Data formats can change and older formats can fall out of favor, often making older data sets unreadable. Code that was once highly readable can become unreadable as newer languages come to the fore and practitioners of older languages decrease in number. Maintenance of data and code is not a question of paying for computer hardware or services. Rather, it is about paying for people to periodically update and fix problems that may be introduced by the constantly changing computing environment.

Unfortunately, funding models for scientific research are aligned with the mechanism of paper publication, where one can definitively mark the end of a project (and also the end of the funding). However, with data and code, there is often no specific end point because other investigators may reuse the data or code for years into the future. Term-based project funding, which is the structure of almost all research funding, is simply not designed to provide support for maintaining materials on an uncertain timeline.

The first 30 years of reproducible research centered largely on discussions of the validity of the idea and what value it provided to the scientific community. Such discussions are largely settled, and both data sharing and code sharing are practiced widely in many fields of study. However, we must now engage in a second phase of reproducible research, which focuses on the continued development of infrastructure for supporting reproducibility.

DISCLOSURE STATEMENT

R.D.P. serves on the Health Review Committee of the Health Effects Institute and is paid for his work. S.C.H. is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Baggerly K. 2010. Disclose all data in publications. *Nature* 467(7314):401–1
2. Baggerly KA, Coombes KR. 2009. Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Stat.* 3(4):1309–34
3. Barba LA. 2018. Terminologies for reproducible research. arXiv:1802.03311 [cs.DL]
4. Barnett AG, Huang C, Turner L. 2012. Benefits of publicly available data. *Epidemiology* 23(3):500–1
5. Bendavid E, Mulaney B, Sood N, Shah S, Ling E, et al. 2020. COVID-19 antibody seroprevalence in Santa Clara County, California. MedRxiv 2020.04.14.20062463
6. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* 29(4):365–71
7. Broman KW, Woo KH. 2018. Data organization in spreadsheets. *Am. Stat.* 72(1):2–10
8. Brook RD, Rajagopalan S, Pope CA 3rd, Brook JR, Bhatnagar A, et al. 2010. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. *Circulation* 121(21):2331–78
9. Buckheit JB, Donoho DL. 1995. Wavelab and reproducible research. In *Wavelets and Statistics*, ed. A Antoniadis, pp. 55–81. New York: Springer-Verlag
10. Claerbout J, Schwab M. 2001. *CD-rom versus the Web*. Rep. 84, Stanford Explor. Proj., Stanford, CA. **http://sepwww.stanford.edu/public/docs/sep84/jon1.pdf**
11. Coombes KR, Wang J, Baggerly KA. 2007. Microarrays: retracing steps. *Nat. Med.* 13:1276–77
12. Dockery DW, Pope CA 3rd, Xu X, Spengler JD, Ware JH, et al. 1993. An association between air pollution and mortality in six U.S. cities. *N. Engl. J. Med.* 329(24):1753–59
13. Drummond C. 2018. Reproducible research: a minority opinion. *J. Exp. Theor. Artif. Intel.* 30(1):1–11
14. Ellis SE, Leek JT. 2018. How to share data for collaboration. *Am. Stat.* 72(1):53–57
15. EPA (Environ. Prot. Agency). 2009. *Integrated science assessment (ISA) for particulate matter*. Rep., EPA Natl. Cent. Environ. Assess., Washington, DC. **https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=216546**
16. Foster ED, Deardorff A. 2017. Open science framework (OSF). *J. Med. Libr. Assoc.* 105(2):203–6
17. Gentleman R, Temple Lang D. 2007. Statistical analyses and reproducible research. *J. Comput. Graph. Stat.* 16(1):1–23
18. Goldberg P. 2014. Duke scientist: I hope NCI doesn't get original data. *The Cancer Letter*, Jan. 16. **https://cancerletter.com/articles/20150116_1/**

19. Goldberg P. 2015. Duke officials silenced med student who reported trouble in Anil Potti's lab. *The Cancer Letter*, Jan. 9. **https://cancerletter.com/articles/20150109_1/**
20. Goodman SN, Fanelli D, Ioannidis JPA. 2016. What does research reproducibility mean? *Sci. Transl. Med.* 8(341):341ps12
21. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, MAQC Soc. Board, et al. 2020. The importance of transparency and reproducibility in artificial intelligence research. arXiv:2003.00898 [stat.AP]
22. Hicks SC, Peng RD. 2019. Elements and principles of data analysis. arXiv:1903.07639 [stat.AP]
23. Ioannidis JPA. 2005. Why most published research findings are false. *PLOS Med*. 2(8):e124
24. Jager LR, Leek JT. 2014. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* 15(1):1–12
25. King G. 2007. An introduction to the dataverse network as an infrastructure for data sharing. *Sociol. Methods Res.* 36:173–99
26. Krewski D, Burnett RT, Goldberg MS, Hoover K, Siemiatycki J, et al. 2000. *Reanalysis of the Harvard Six Cities Study and the American Cancer Society study of particulate air pollution and mortality*. Rep., Health Effects Inst., Cambridge, MA. **https://www.healtheffects.org/publication/reanalysis-harvard-six-cities-study-and-american-cancer-society-study-particulate-air**
27. Lancet Editors. 2020. Expression of concern: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of covid-19: a multinational registry analysis. *Lancet* 395:E102
28. Leek JT, Peng RD. 2015. Opinion: Reproducible research can still be wrong: adopting a prevention approach. *PNAS* 112(6):1645–46
29. Leinonen R, Sugawara H, Shumway M, Int. Nucleotide Seq. Database Collab. 2010. The sequence read archive. *Nucleic Acids Res*. 39(Suppl. 1):D19–21
30. Mehra MR, Desai SS, Ruschitzka F, Patel AN. 2020. Retracted: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of covid-19: a multinational registry analysis. *Lancet*. **https://doi.org/10.1016/S0140-6736(20)31180-6**
31. Mehra MR, Ruschitzka F, Patel AN. 2020. Retraction—Hydroxychloroquine or chloroquine with or without a macrolide for treatment of covid-19: a multinational registry analysis. *Lancet* 395:P1820
32. Natl. Acad. Sci. Eng. Med. 2019. *Reproducibility and Replicability in Science*. Washington, DC: Natl. Acad. Press
33. Open Sci. Collab. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251):aaac4716
34. Patil P, Peng RD, Leek JT. 2016. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.* 11(4):539–44
35. Patil P, Peng RD, Leek JT. 2019. A visual tool for defining reproducibility and replicability. *Nat. Hum. Behav.* 3(7):650–52
36. Peng RD. 2011. Reproducible research in computational science. *Science* 334(6060):1226–27
37. Peng RD, Dominici F, Zeger SL. 2006. Reproducible epidemiologic research. *Am. J. Epidemiol.* 163(9):783–89
38. Plesser HE. 2017. Reproducibility vs. replicability: a brief history of a confused terminology. *Front. Neuroinform.* 11:76
39. Pope CA 3rd, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, et al. 1995. Particulate air pollution as a predictor of mortality in a prospective study of US adults. *Am. J. Respir. Crit. Care Med.* 151(3):669–74
40. 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
41. R Core Team. 2020. R: A language and environment for statistical computing. *The R Project for Statistical Computing*. **https://www.r-project.org/**
42. Samet JM, Dominici F, Zeger SL, Schwartz J, Dockery DW. 2000. *The National Morbidity, Mortality, and Air Pollution Study. Part I: methods and methodological issues*. Res. Rep. 94-I, Health Effects Inst., Cambridge, MA. **https://www.healtheffects.org/publication/national-morbidity-mortality-and-air-pollution-study-part-i-methods-and-methodologic**
43. Samet JM, Zeger SL, Dominici F, Curriero F, Coursac I, et al. 2000. *The National Morbidity, Mortality, and Air Pollution Study. Part II: morbidity and mortality from air pollution in the United States*.

Res. Rep. 94-II, Health Effects Inst., Cambridge, MA. **https://www.healtheffects.org/publication/national-morbidity-mortality-and-air-pollution-study-part-ii-morbidity-and-mortality-air**
44. Sandve GK, Nekrutenko A, Taylor J, Hovig E. 2013. Ten simple rules for reproducible computational research. *PLOS Comput. Biol.* 9(10):e1003285
45. Schooler JW. 2014. Metascience could rescue the 'replication crisis.' *Nature* 515(7525):9
46. Schwab M, Karrenbach N, Claerbout J. 2000. Making scientific computations reproducible. *Comput. Sci. Eng.* 2(6):61–67
47. 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
48. Stodden V. 2015. Reproducing statistical results. *Annu. Rev. Stat. Appl.* 2:1–19
49. Stodden V. 2020. Beyond open data: a model for linking digital artifacts to enable reproducibility of scientific claims. In *P-RECS '20: Proceedings of the 3rd International Workshop on Practical Reproducible Evaluation of Computer Systems*, pp. 9–14. New York: ACM. **https://dl.acm.org/doi/10.1145/3391800.3398172**
50. Swanberg SM. 2017. Inter-University Consortium for Political and Social Research (ICPSR). *J. Med. Libr. Assoc.* 105(1):106
51. Wickham H. 2011. testthat: Get started with testing. *R. J.* 3:5–10
52. Wickham H. 2014. Tidy data. *J. Stat. Softw.* 59(10):1–23
53. Wickham H, Hester J, Chang W. 2020. devtools: tools to make developing R packages easier. *devtools*. **https://devtools.r-lib.org/**