A Guide to Reproducibility in Preclinical Research
Greg Samsa, PhD, and Leigh Samsa, PhD

Abstract

Many have raised concerns about the reproducibility of biomedical research. In this Perspective, the authors address this “reproducibility crisis” by distilling discussions around reproducibility into a simple guide to facilitate understanding of the topic.

Reproducibility applies both within and across studies. The following questions address reproducibility within studies: “Within a study, if the investigator repeats the data management and analysis, will she get an identical answer?” and “Within a study, if someone else starts with the same raw data, will she draw a similar conclusion?”

Contrastingly, the following questions address reproducibility across studies: “If someone else tries to repeat an experiment as exactly as possible, will she draw a similar conclusion?” and “If someone else tries to perform a similar study, will she draw a similar conclusion?”

Many elements of reproducibility from clinical trials can be applied to preclinical research (e.g., changing the culture of preclinical research to focus more on transparency and rigor). For investigators, steps toward improving reproducibility include specifying data analysis plans ahead of time to decrease selective reporting; more explicit data management and analysis protocols; and increasingly detailed experimental protocols, which allow others to repeat experiments. Additionally, senior investigators should take greater ownership of the details of their research (e.g., implementing active laboratory management practices, such as random audits of raw data [or at least reduced reliance on data summaries], more hands-on time overseeing experiments, and encouraging a healthy skepticism from all contributors). These actions will support a culture where rigor + transparency = reproducibility.

Some have argued that biomedical research is not reproducible. We would argue that this “reproducibility crisis” instead falls somewhere between urban legend and established fact.

The notion of a reproducibility crisis started with the widely cited 2005 article by Ioannidis entitled “Why Most Published Research Findings Are False.”

As not everyone has read this article closely, popular belief has essentially become: “I think someone published something somewhere demonstrating that most published research findings are false. That paper was published, so it must be true.”

Some of the original research suggesting this reproducibility crisis comes from the fields of psychology and oncology drug development. The psychology literature is particularly remarkable: With the cooperation of the original investigators, 100 representative studies reported in major journals were repeated using (as close to as possible) their original protocols. Yet, only 36% of these replications had statistically significant findings, the average effect size was halved, and fewer than half of the studies were subjectively judged to have been successfully replicated. A large portion of the replications produced weaker evidence than the original studies.

Similarly worrisome results have been observed in oncology drug development. Begley and Ellis attempted to confirm the preclinical findings published in 53 “landmark” studies but only succeeded in confirming the findings in 6 studies. Although the level of consistency between the original and confirmatory experiments was not as high as that of the work in psychology, an attempt was made to contact the original authors, discuss the discrepant findings, exchange reagents and repeat experiments under the authors’ direction, occasionally even in the laboratory of the original investigator. Prinz and colleagues similarly reported that only 20% to 25% of validation studies (mostly in the field of oncology drug development) were “completely in line” with the original reports.

In 2014, the National Institutes of Health (NIH) held a workshop with the editors of more than 30 basic and preclinical science journals to discuss the issue of reproducibility and the rigor of research findings. They published their initial recommendations, which likely will evolve over time as additional evidence accumulates. The NIH also dedicates substantial efforts to enhancing reproducibility.

According to a 2016 survey of the readers of Nature, more than half of respondents (predominately laboratory scientists) felt there was a “significant crisis.” Putative causes of this crisis included (1) selective reporting, (2) pressure to publish,
In sum, at least some of the general public, the largest funding agency of biomedical research in the United States, and investigators themselves all have called into question the reproducibility of biomedical research. This “reproducibility crisis” has the potential to erode public trust in biomedical research, potentially leading to additional unfortunate consequences. In this Perspective, we distill the discussions around reproducibility into a simple guide to facilitate understanding of the topic. We contrast clinical trials, for which approaches to increase reproducibility have been systematized and tested over a number of years, with preclinical research, which remains a work in progress. We conclude with additional considerations and action items for investigators as well as others involved in biomedical research.

**Types of Reproducibility**

Reproducibility can apply both within and across studies. Investigators can directly affect the former. The latter, on the other hand, pertains to inconsistency in results across studies. The term “reproducibility crisis” often refers to this latter type of reproducibility. Typically, a discussion about reproducibility includes at least one of the following questions:

1. Within my study, if I repeat the data management and analysis, will I get an identical answer?

2. Within my study, if someone else starts with the same raw data, will she or he draw a similar conclusion?

3. If someone else tries to repeat my study as exactly as possible, will she or he draw a similar conclusion?

4. If someone else tries to perform a similar study, will she or he draw a similar conclusion?

In the sections that follow, we explore each of these questions in more depth.

**Question 1: Within my study, if I repeat the data management and analysis, will I get an identical answer?**

Data management is the process by which the original data are restructured and prepared for analysis. Data cleaning is one element of data management and is the process by which potential errors in the data are identified and addressed.

The process of data management moves from the original raw data file to an analysis file, where all changes to the original database are documented and auditable. At a minimum, the rigorous execution of data management requires that investigators keep copies of the original raw data file, the final analysis file, and all the data management programs.

Data cleaning typically involves flagging and then addressing values that are unusual. For example, a systolic blood pressure value of 1,505 mm Hg is physically impossible, so recoding this value to 150.5 mm Hg would be reasonable. A systolic blood pressure value of 310 mm Hg at baseline is physically possible but unlikely; if the value at the end of the intervention is 135 mm Hg, a good case can be made that a typographical error turned a value of 130 into 310. Some editorial changes to the database can be considered permanent, such as changing 1,505 to 150.5. To maintain analytical flexibility, other changes are best considered provisional, such as the 310 mm Hg example, where the potential revised data values include 130; 310 (if a review of the patient’s medical history suggests that an extreme blood pressure spike is plausible); and missing (e.g., if 310 is considered implausible but the investigators are uncomfortable imputing another value such as 130).

Data cleaning is best performed in a blinded fashion before data analysis. Once the assignment of a patient to a study group is known, it can potentially bias subsequent decisions. Although sensitivity analyses are an important component of statistical assessment, trying multiple approaches to data cleaning runs the danger of ultimately deciding on the one that best supports an investigator’s hypothesis.

Data analysis proceeds by applying statistical analysis programs to the analysis file. Such a program might make temporary changes to the database (e.g., for the purposes of one analysis, all systolic blood pressure readings above a certain value might be set to missing) before the statistical analyses are performed. Reproducibility in this step is achieved by maintaining the final version of all the analysis programs used to produce the results (e.g., to generate all the numbers reported in a manuscript).

Version control is helpful, for example, to ensure that the correct version of an analysis program is applied to the correct version of a dataset.

Standards for all of the above actions have been developed and tested. For large studies extending across multiple sites over long periods of time with many data elements and/or databases, the implementation of the above practices can become technically complex. However, answering our first question above primarily requires care and attention to the details of the process.

For preclinical research, the functional specifications of reproducible data management are essentially identical to those of a clinical trial—an auditable record of the original raw data, including the rationale for any data cleaning in the data management programs, and retaining the analysis programs. Indeed, many of the components of reproducible data management were once embedded in a physical laboratory notebook. Changes were made by drawing a line through the original data (thus retaining a record of the original data value), dating the change, initialing it, and providing a note justifying the change.

For small to moderately sized experiments, upgrading this paper record-keeping system to an electronic equivalent is relatively straightforward. The key is to replace cutting and pasting from Excel spreadsheets with data management programming, because the latter is easily auditable while the former is not. The goal should be to create a formal software-based system for data cleaning and management that is analogous to the systems associated with clinical trials. Once an analysis file is created, reproducible documentation of the data analysis then follows the same principles as for clinical trials (and, indeed, for research in general).
Some preclinical experiments generate large amounts of electronic data, which are then processed, with the resulting summary values used as the basis for the analysis file. (For example, a machine might generate 1,000 readings, and what is ultimately analyzed is the mean of those 1,000 values.) In that case, a computerized archiving system might be required. If so, the implementation of these data management principles is more complex, but the principles themselves are the same.

Technological solutions to the problem of the reproducibility of data management and analysis are becoming increasingly popular and sophisticated, but as yet there is no one-size-fits-all solution for all biomedical research. For many laboratories, electronic lab notebooks are an important component of improved data management; these software packages vary widely and can offer sophisticated features ranging from edit tracking to integrated genome browsing. Although such tools are certainly helpful, especially within the context of an overall data management system, it is also important to recognize that reproducibility is a broader construct than a technical software requirement and that the responsibility for ensuring reproducibility continues to reside with the investigator.

As the functional specifications for reproducible data management are fundamentally similar across fields, one might ask about the barriers to achieving reproducible data management in the preclinical environment. These barriers are both technical and cultural. For example, a technical barrier is that preclinical data can potentially require significant amounts of preprocessing (which often requires subject matter expertise), and adequately documenting all these preprocessing steps is an involved process. A key to success is changing, whenever realistically feasible, the data management process from “point, click, drag, and drop” to the formal application of auditable computer code. Another technical barrier is that preclinical experiments often have more moving parts than clinical trials, as we will discuss later. A cultural barrier, on the other hand, is that the culture of preclinical research has tended to focus on reproducibility across studies rather than reproducibility both within and across studies.

**Question 2: Within my study, if someone else starts with the same raw data, will she or he draw a similar conclusion?**

Here, three potential sources of variation exist: (1) Data management decisions can differ (e.g., during the data-cleaning process, one investigator might retain a data value that another deleted); (2) analytical techniques can differ (e.g., one investigator might use different statistical tests than another); and (3) conclusions can differ (e.g., one investigator might use different thresholds for statistical and/or clinical significance or focus on different results than another).

This question touches on the issues of robustness and selective reporting. Generally, a statistical result is robust if different analyses lead to similar conclusions. If a result is statistically robust, it will not matter if another investigator conducts the analyses somewhat differently so long as both sets of analyses are performed reasonably. On the other hand, if results have been selectively reported, another investigator with the same raw data is unlikely to draw similar conclusions. The potential effects of selective reporting are part of the rationale for requiring investigators to archive their data and computer code and to specify their approach before data analysis begins.

Returning to the issues of data cleaning and sensitivity analysis, performing separate analyses using the systolic blood pressure (mm Hg) values of 310, 130, and missing (or, alternatively, retaining the value of 310 and applying statistical techniques that are either more or less affected by outliers) contributes to the robustness of the study and its scientific soundness. Problems may arise not with sensitivity analyses per se but in performing such analyses and then preferentially reporting the desired results. Requiring that detailed analysis plans be specified ahead of time in a publicly accessible place (e.g., clinicaltrials.gov) is one way that clinical trials differ from preclinical work.

In the preclinical environment, a significant barrier to robustness is selective reporting. Selective reporting can include performing multiple statistical analyses and limiting reporting to those that are most favorable (this approach is unsound regardless of the context). Analysis plans for preclinical studies do not necessarily need to be specified ahead of time with the same level of detail required for clinical trials, but a general analytical approach should nevertheless be reported and archived publicly.

Another form of selective reporting—the tendency to delete or otherwise modify data values before analysis—is a particular risk for preclinical research. For example, an experiment can be technically difficult to perform, may not always work properly, and may yield ambiguous data. Some of these ambiguous data reflect an “experimental failure” or a “finicky protocol” (e.g., sick mice, incorrect levels of the reagents, or some other technical failure in one of the many complex steps of the experiment) and could legitimately be dropped from the analysis. Other ambiguous data simply reflect variation in the results, and a scientifically plausible, albeit self-interested, argument could be made for dropping these data points or otherwise editing them to support the desired conclusion (e.g., stretching the rationale of a finicky protocol). An investigator may analyze the edited data (with data points from “finicky protocol days” dropped) rather than the original raw data, and the audience may never be aware. The investigator’s colleagues may have similar incentives and challenges and may not provide sufficiently critical comment. Indeed, in this scenario, their critique might be limited to the macrolevel question of whether the results appear plausible (e.g., consistent with previous results or hypotheses—i.e., reproducible across studies) rather than the question of precisely how the experiment was conducted and analyzed.

Although others may be physically present in the laboratory, in contrast to those performing clinical trials, the preclinical investigator really stands alone. One way to remedy this situation and the potential consequences is to design data management systems that store all raw data instead of just a version that has been selectively analyzed and to insist that a consideration of all data become a common part of the culture of preclinical science. Explicitly stating criteria for editing data can help too.

**Question 3: If someone else tries to repeat my study as exactly as possible, will she or he draw a similar conclusion?**

Repeating a study as exactly as possible is sometimes termed “direct replication.”
Of course, this name implies that other investigators have access to a sufficiently detailed version of the original protocol so they can repeat the essential components of the study. This question, then, might reasonably have as a preamble the following: “Within my study, is the protocol described in sufficient detail that others could repeat it if they desired?” For example, the protocols associated with randomized drug trials tend to be described with a high level of specificity, not only because of government regulations but also because this information is needed by others to implement the interventions in question should the trials yield positive results. Such specificity has now become embedded in the culture associated with randomized drug trials.

Direct replication is unusual in the clinical trial literature, however, if for no other reason than that obtaining funding to do such experiments is difficult. In comparison with Question 4 below (If someone else tries to perform a similar study, will she or he draw a similar conclusion?), direct replication is a weaker condition. (That is, it should be easier for an identical study to draw a similar conclusion than a study that is similar yet not identical.) Indeed, this is what makes the psychology study we cited earlier so worrisome. (One possible explanation for the results of that study is selective reporting.)

In contrast to clinical trials, direct replication is an integral part of the culture of preclinical research and is a natural first step for an investigator who is expanding her or his work into a new area of study. To replicate a complex preclinical experiment, the protocol must be described with a high degree of specificity and transparency. Manuscripts and similar reports tend to be high-level summaries about the research rather than fully actionable descriptions of how the research was performed. Previously, limited journal space was a significant barrier to reporting such detailed descriptions of study protocols. Now, however, complete protocols and similar information can be archived online without taking up space in print. Journals that publish peer-reviewed methods/protocols, such as *Nature Methods* and the *Journal of Visualized Experiments*, are additional resources intended to address the need for the reporting of specific and transparent protocols.

Another technical barrier to such reporting is the sheer number of variables that would need to be documented to exactly repeat a preclinical experiment. Investigators do not always know ahead of time which variables are important (e.g., whether a buffer must be made fresh before use) and usually recognize that there is a trade-off between precision of documentation and research pace.

**Question 4: If someone else tries to perform a similar study, will she or he draw a similar conclusion?**

This question pertains to reproducibility across studies, and indeed conceptual replication is usually the construct of interest when multiple, related studies are being discussed. For example, investigators often assess whether a treatment that works well in one circumstance will perform similarly in another. Reasons for variability in study results include bias in the original protocol, artifacts, and real differences in the outcomes being measured.

Bias can apply to the study design, reporting of results, or both. A biased study design likely leads to untrustworthy conclusions. Similarly, if the results of the original study have been exaggerated, subsequent studies are likely to produce weaker evidence in favor of the original findings. Finalizing data analysis plans ahead of time and making them publicly available can limit selective reporting. Creating registries of clinical trials is meant to limit investigators’ tendency to only publish positive findings. Indeed, in most areas of medicine, protections against selective reporting are relatively strong.

One example of an artifact that can affect the outcomes of a study is the influence of chance. Even if it were possible to perform two identical studies, the results are unlikely to be identical. Chance plays a stronger role in small studies than in larger ones, but good instrumentation (e.g., outcome variables with excellent measurement properties) can limit the level of noise in the data and thus the effects of chance.

Apart from the influence of bias and artifacts, the efficacy of an intervention might simply differ from study to study. For example, the level of efficacy might depend on the precise formulation and delivery of the intervention, the specific patient population, or the instruments used to assess the outcomes. Moreover, investigators often begin by testing interventions in ideal circumstances, as a cost-effective way to determine whether to proceed further. However, the efficacy of an intervention in ideal circumstances will often be less than that in a real-world scenario, thus motivating the distinction between efficacy and effectiveness.

Ultimately, so long as the study is conducted properly, Question 4 really revolves around the question of what constitutes a similar study. Meta-analysis is helpful not only as a statistical methodology but also as a conceptual framework. Making the design of individual studies as rigorous and transparent as possible offers the best hope that conceptual replication will be achieved.

An analogous concept in preclinical science is that of translation. One set of potential barriers to translation includes problems with early studies in the pipeline. Early studies might be poorly designed, or the data might be selectively reported within the study (e.g., by selectively editing data values) or across studies (e.g., with positive experiments submitted for publication and negative experiments not). The archiving and registration of study designs in preclinical science are not as systematic as for clinical trials.

In preclinical research, one of the barriers to reproducibility between studies is real differences in the effect of the intervention across studies. In general, these differences are likely to be larger in the preclinical stage than they are for clinical trials. One reason for this difference is that preclinical science is often mechanistic, the goal being to understand why and how. Unfortunately, at the bench, there are a dizzying array of variables that can influence an experiment. Although laboratory scientists try to control for as many variables as possible, it is simply not possible to control for everything, and it is extremely difficult to accurately identify all the variables involved in an experiment, much less to accurately report all those variables. Part of the reason laboratory scientists train for upwards of five years is that it simply takes that long to instill in trainees...
an intuitive understanding of what is important in an experiment. Properly designed experiments contain internal controls that provide some estimate of technical sensitivity and variability, but the staggering number of possible variables is a major contributor to inconsistent results across studies.

Another barrier to reproducibility is the length of the preclinical pipeline. Clinical trials are all alike at some level, although this does not make them simple to perform. Preclinical studies, on the other hand, extend from experiments in silico, to molecules, to cells, across multiple animal models, to humans. These experimental designs contain more variability than do those of clinical trials. Conclusions from one type of experiment need not generalize to other types (e.g., depending on the characteristics of the model system, an experiment in one type of animal model might or might not be expected to generalize to humans). Achieving reproducibility across the preclinical pipeline is, by its nature, a more difficult challenge than achieving reproducibility across clinical trials.

A Guide to Reproducibility

We do not intend for this Perspective to be a comprehensive discussion of reproducibility in preclinical research. Moreover, many of the questions we asked do not have simple answers (e.g., in a big data application, one of the complications can be that the dataset is too large to store and is not static, making the concept of a single raw data file not feasible). This Perspective is also not intended to provide specific solutions to any of the issues we have raised.

Instead, we tried to provide a straightforward guide to understanding reproducibility. The most fundamental distinction in this discussion is between reproducibility within studies and reproducibility across studies. One case where reproducibility across studies is prominent is the conceptual model used by the NIH: rigor + transparency = reproducibility (and, in particular, that additional rigor and transparency within individual studies should support reproducibility across studies). Elements of transparency in individual studies (e.g., detailed experimental protocols and auditable data management plans) not only contribute to the reproducibility of the individual study in question but also to reproducibility across studies. Although elements of rigor are not always thought of as contributing to reproducibility within studies, they do contribute to reproducibility across studies and thus are important in their own right.

Within an individual clinical trial, transparency, objectivity, and specification of the relevant protocols ahead of time all contribute to reproducibility. Data collection and analysis protocols are finalized and made publicly available before data collection begins. Increasingly, the raw data and computer code are being made publicly available as well. Study measures are as objective as possible, and peer review encourages consistency between results and their interpretation. In general, “good science” is “reproducible science,” and vice versa. Reproducibility can be more difficult to achieve in preclinical studies, but the same core elements of transparency, objectivity, and specification of protocols ahead of time are still critical. Accepting responsibility for within-study reproducibility that meets these criteria is a crucial first step to ensuring reproducibility across studies.

Despite the drama associated with scientific fraud and the evidence that outright scientific fraud does exist, in our experience, intentional and flagrant fraud is not the main problem in either clinical trials or preclinical research. Instead, as Nosek and colleagues argued: We have enough faith in our values to believe that we would rather fail than fake our way to success. Less simple to put aside are ordinary practices that can increase the likelihood of publishing false results, particularly those practices that are common, accepted, and even appropriate in some circumstances.

Improving such questionable research practices is not a simple matter, but it is a crucial one.

Discussion of questionable research practices is more nuanced than that of outright scientific fraud and more difficult to communicate to the general public, who might mistakenly assume that the former is the latter. Both questionable research practices and outright fraud erode public trust in biomedical science.

What action items do we recommend, then, to preclinical investigators as well as to laboratory directors, research mentors, medical educators, and others? Administrative solutions (e.g., a clinicaltrials.gov for preclinical research, where crucial elements of such studies would be registered) seem premature, in part because of the diversity of preclinical research, which makes a one-size-fits-all solution unlikely. Instead, we recommend encouraging a change in the culture of preclinical research toward greater transparency and rigor. A crucial element of such rigor is specification of protocols ahead of time (e.g., prespecification of data analysis plans) to reduce selective reporting. Establishing clear data management and analysis protocols and sufficiently detailed experimental protocols, which allow others to repeat experiments if desired, are steps in the right direction.

Another step in the right direction is for senior investigators to take greater ownership of the details of their research. This step, however, is admittedly more difficult given the current level of overcommitment of many senior investigators. Such a step can include implementing active laboratory management practices, such as random audits of raw data (or at least reduced reliance on data summaries), more hands-on time overseeing experiments, and encouraging a healthy skepticism among all concerned. Part of the NIH’s efforts to improve rigor and reproducibility is directed toward empowering less senior members of the research team to speak up if they suspect that something is amiss.

In closing, the guide we present here need not be limited to clinical trials and preclinical studies. Indeed, it also applies to the field of medical education research. Various thoughtful critiques call for improved reproducibility within this discipline, and others recommend larger sample sizes, more soundly designed experiments, more consistent application of clear conceptual frameworks, and better balancing of direct and conceptual replication. These critiques primarily address issues related to the rigor of research, and in doing so, they also address reproducibility, as rigor + transparency = reproducibility.

Funding/Support: None reported.
I'm an old and broken man, my face chipped away by life's decisions. Nurses come and go. Medical students, young doctors, seasoned doctors. Coming into my room to do things to me. They sigh. They shake their heads. I know I'm a mess. I lay in my bed alone day after day enduring procedure after procedure. The pain is unbearable, and yet they keep coming back. Now they tell me my treatment is failing. Where does it all end?

I can only imagine the musings of the patient I see as I walk into a room dimly lit only by the sunshine slipping through the blinds. There he lies, his face and shoulders covered in specks of dried blood. I've come for the next procedure. He is resigned to more pain. It is just one more item on my seemingly endless to-do list. This task is going to take forever. He's withdrawn. I'm exhausted. I don't have time to think or feel anything about this patient or his pain. Or do I? Has anybody?

My phone won't stop ringing. My pager won't stop beeping. More requests for medications. Orders for morning labs. When are you going to come talk to this worried patient and her family? Is that another trauma alert? I don't even have time for the patient in front me. I'm overwhelmed, even irritated. Is this really medicine?

I spot a King James Bible on his bedside table. I take a deep breath and start the procedure. I should make small talk. Good doctors make small talk.

“Sir, what's your favorite passage?”

“I don't know,” he replies. “I really like the New Testament.”

“Oh! What have you been reading?”

“Nothing, I can't find my glasses.”

“One of my favorite passages is Psalm 139. Are you familiar with it?”

“No.”

“Would you like me to read it to you?”

He nods. So much for small talk. I silence my phone and sit down by his bedside. I don't have time for this. But I will make time.

“How precious to me are your thoughts, O God! How vast is the sum of them! Were I to count them, they would outnumber the grains of sand—when I awake, I am still with you.”

As I begin to read, he breathes a heavy sigh. I look up from the page in time to see tears rolling down his emaciated cheeks. Rolling through the dried blood. Rolling over his sutures. Dripping from his disfigured chin. I take his hand. He squeezes mine. I keep reading. His tears keep falling. And suddenly I am keenly aware that he doesn't need my medical degree at all.

I stop perseverating over my to-do list—all the patients yet to be seen, all the orders yet to be written. Not even the ache of exhaustion can dissuade me. I grab a washcloth and wet it with warm water. I begin to wipe away the tears and the blood. He moans.

“I'm sorry. Did I hurt you?” He shakes his head and I realize he hasn't been bathed since his surgery. He asks me to continue.

I came into his room on a mission to complete one more routine task, and I left knowing that sometimes the most powerful medicine comes in the form of a tangible expression of humanity, the careful attention to the soul. It is so much more than tending to the frail body of a fellow human being. From this day forward, I am determined to resist the temptation to treat patients and their ailments as a list of tasks to be accomplished. Instead I resolve to pursue a practice of medicine that is characterized by choosing compassion above all else.

Cheralyn J. Hendrix, MD

C.J. Hendrix is a resident, Department of Surgery, George Washington University Hospital, Washington, DC; e-mail: cjhendrix@gwu.edu; Twitter: @CHendrix_MD.

An Academic Medicine Podcast episode featuring this article is available through iTunes, the Apple Podcast app, and SoundCloud.