Rigor and Reproducibility in Federally-Funded Scientific Research: What are the Right Questions to Ask?

2017

Didi Cross, PhD, PMP
Director, Research Data Science & Evaluation (RDS&E), Clarivate Analytics
Executive Summary

Science managers face a range of challenges when it comes to managing the rigor and reproducibility of work in their program areas. Federal government personnel in this function have a particular challenge given the size of their portfolios and the breadth of stakeholders to which they are accountable. Deciding how to use funding judiciously and allocating funds to studies that rigorously apply scientific methods, are reproducible and transparent, and provide an appropriate level of duplication is hampered by inconsistent definitions and use of these terms (e.g., reproducibility, rigor, robustness, replicability), and misaligned incentives for academic scientists. However, increasing attention to and understanding of the issues related to rigor and reproducibility of federally-funded scientific research is helping to mitigate some of the risks associated with uncertainty in the current scientific climate.

This paper distinguishes between rigor, reproducibility, and duplication (or verification or corroboration). Rigor addresses how studies are designed and conducted. Reproducibility (specifically, methodological reproducibility) addresses whether the information that is provided in the study is sufficient to re-run the experiment and analysis, and thereby produce the same finding given the same data and tools. The amount and nature of duplication, either repeating a study exactly or conceptually verifying prior findings, is also a distinct and important consideration.

The purpose of this paper is to empower those who manage scientific research in the federal government by providing insight, guidance and highlighting the key questions to ask regarding rigor and reproducibility. This paper aims to provide:

- An introduction to the challenges
- Definitions of terms in the academic and trade literature
- Important considerations
- Guidance on questions to ask
- Recommended resources for more information
Science is inherently complicated. The scientific method involves systematically formulating and testing hypotheses, thereby increasing understanding of the natural world. Crucial to understanding the veracity of the findings from completed studies is the ability to assess the design, selection, application, and reporting of the knowledge, methods, and materials used for research—together, comprising what is termed “rigor and reproducibility” in science. The heart of this matter is an issue of judicious management of limited resources.

The National Institutes of Health (NIH) website\(^1\) emphasizes the crucial importance:

> “Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results. When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research.”

However, in recent years, studies investigating the reporting of scientific research have called into question the rigor and reproducibility of published findings. John Ioannidis of the Stanford University School of Medicine\(^2\) points out that this ought, in fact, to be expected: “Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true” (emphasis added). He hypothesizes that this is due to a combination of reasons, including flexibility in study design and data analysis, personal biases and prejudices, and existing incentive structures for scientists.

---

Glenn Begley, formerly of Amgen, and Lee Ellis of the University of Texas M.D. Anderson Cancer Center, examined a number of landmark, preclinical papers relevant to industry in hematology and oncology. They found they could not confirm published findings for many of these influential pieces of research:

“Fifty-three papers were deemed ‘landmark’ studies. It was acknowledged from the outset that some of the data might not hold up, because papers were deliberately selected that described something completely new... Nevertheless, scientific findings were confirmed in only 6 (11%) of cases. Even knowing the limitations of preclinical research, this was a shocking result.” (emphasis added).

Further complicating the issue is inconsistent use of the relevant terminology—rigor, reproducibility, and duplication (or verification or corroboration). For clarity, the following are useful definitions of these terms:

- **Rigor** addresses study design and execution. Scientific experiments should be designed so that the findings are more likely to be valid. Although the validity of a particular finding cannot always be immediately determined, many aspects of study design and analysis can be assessed at an early stage. These include the rationale for undertaking the experiment in a certain way, selection and characterization of the materials used, and the existing knowledge on which these choices are based. Those conducting and managing science can determine whether the results are likely to be reliable based on these and other characteristics. Results that are unreliable (i.e., those obtained from questionable choices of experimental or analytical methodology, and/or uncharacterized materials) are, of course, less likely to be valid.

**Key Question:**
Given that this study was designed and conducted this way, to what extent should the findings be believed? Are there reasons to doubt these findings?

---

• **Reproducibility** deals more closely with reporting the features of scientific research that describe the experiment and sharing the data underlying the analysis. This reporting and data sharing can facilitate repetition of the analysis. Data sharing does not, of course, readily facilitate the re-running of an experiment, however, the structure of data may be helpful in understanding particular details of the experimental design. Ideally, these details should be documented explicitly elsewhere.

**Key Question:**
Is the information that is provided sufficient enough to re-run the experiment and analysis, and thereby produce the same finding?

• The degrees of **duplication** on a detailed level, and the verification or corroboration on a more conceptual level, are both valuable contributions. Duplicating or repeating an experiment with the same materials, protocol, and analysis can serve to confirm that the experiment was conducted correctly. Running a new experiment or study using different experimental techniques, but studying the same phenomena, serves to verify or corroborate the existence of those phenomena. The appropriate amount and level of granularity of duplication, verification, and corroboration varies across research areas.

Another challenge arises as individual scientists find they must “publish or perish.” Meaning, the metrics on which individual scientists’ careers are measured—which directly influence their incentives—focus strongly on the number of papers published, particularly the number published in high-profile journals. These metrics are not strictly aligned with the broader goal of science—to advance knowledge. Marc Edwards and Siddhartha Roy of Virginia Tech explain further:

“Over the last 50 years, we argue that incentives for academic scientists have become increasingly perverse in terms of competition for research funding, development of quantitative metrics to measure performance, and a changing business model for higher education itself. Furthermore, decreased discretionary funding at the federal and state level is creating a hypercompetitive environment between government agencies... for scientists in these agencies, and for academics seeking funding from all sources—the combination of perverse incentives and decreased funding increases pressures that can lead to unethical behavior.”

Even in the absence of unethical behavior, there must be a willingness to ask questions about study design, analysis, reporting, and conscious and unconscious personal bias at all stages of the evaluation and management of scientific research. To aid this process, this white paper presents questions that program directors, evaluation officers, and others can ask of themselves and their grantees. By asking these questions, the scientific community, and in particular the funding agencies that support them, can develop a set of leading practices to enable the judicious management of scientific resources.

---

Rigor, Reproducibility, and Duplication

This paper examines the challenges above by asking three separate but important questions (see Figure 1):

- **Question 1: Is the scientific study rigorous?**
  Is the proposed science likely to be reliable and valid? Unreliable findings may be produced due to multiple causes, ranging from outright fraud to unconscious bias and everything in between. Biological variability should be considered in designing experiments as well as considering whether findings are reliable.

- **Question 2: Is the scientific study reproducible (or replicable or transparent)?**
  To what degree are the details and procedures of the study reported, and can another investigator replicate the findings? This has been the focus of many “open data” initiatives and much of the discussion around reproducibility and replicability. However, differential skills at applying experimentation and analysis techniques should be considered as well.

- **Question 3: To what extent does the scientific study duplicate or verify other studies?**
  A related question is: How much overlap and what kind of verification are appropriate? A certain amount of duplication is necessary, beneficial, and intended in science, but judicious management means not wasting resources on duplication that is inappropriate or unnecessary.

Continue reading to examine these questions in more depth, and to discover resources for more information on each.

Figure 1. The Three Key Questions to Ask about Scientific Research
Question 1:

Is the Scientific Study Rigorous?

The first of the three key questions addresses the rigor of the proposed scientific study. NIH has issued guidelines for reviewers to assess rigor in grant applications, defining scientific rigor as “the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results. This includes full transparency in reporting experimental details so that others may reproduce and extend the findings.” This definition incorporates rigor in reporting (which is covered in Question 2), as well as rigor in designing studies. To aid preparers and reviewers of grant applications, NIH provides a helpful infographic, resource chart, and reviewer guidance.

One potential way to evaluate study rigor is to evaluate it against five pillars of rigor that Arturo Casadevall of mBio and Ferric Feng of Infection and Immunity propose, and a sixth pillar that former National Institute of Mental Health (NIMH) Director Thomas Insel adds. These include:

- Redundant experimental design
- Sound statistical analysis
- Recognition of error
- Avoidance of logical fallacies
- Intellectual honesty
- Biological variability

With regard to redundant experimental design, Thomas Insel adds that “[m]any of the common standards for reducing bias in clinical research, such as randomization and keeping raters ‘blind’ to the experimental condition, are not always adhered to in preclinical studies. As a result, bias can creep into the experiment, leading the investigator to find what she or he is looking for and ignore results that are not consistent.”

With regard to the second of these pillars—sound statistical analysis—Insel explains that “Statistics are critical for defining group differences. But sometimes what we see in practice is the use of statistics to fish for a difference that is not present.” Expanding on this pillar, Harvey Motulsky of GraphPad Software indicates that studies may suffer from P-hacking (re-analyzing data until the desired result is obtained), overemphasis on P values, excess use of statistical hypothesis testing, and excessive reliance on standard errors. Goodman et al. of Stanford University define multiplicity to include “testing many hypotheses in one experiment, testing one hypothesis many times or in multiple ways in one or more studies, and other maneuvers that virtually guarantee a chance observation that will appear to strongly support some hypotheses.” Goodman explains the importance of multiplicity:

“Multiplicity, combined with incomplete reporting, might be the single largest contributor to the phenomenon of non-reproducibility, or falsity, of published claims.”

Regarding biological variability, Insel asserts that this pillar is particularly difficult to address. He describes a situation in which three labs found large differences in results when following the same experimental protocols on inbred mouse strains. Other researchers (Clayton & Collins, Clayton, and Lorsh et al.) also point out the importance of this pillar to be able to determine if the study is rigorous.

11 http://jpet.aspetjournals.org/content/jpet/351/1/200.full.pdf.
Further complicating the issue is inconsistent use of the relevant terminology—rigor, reproducibility, and duplication (or verification or corroboration).

NIH website
Question 2:
Is the Scientific Study Reproducible (or Replicable or Transparent)?

The second of the three key questions addresses the data and information that researchers provide to facilitate re-running the same or similar studies. Although these elements are to be reported at the time of publishing scientific results, they would ideally be monitored throughout the course of the research.

There are varying degrees to which one scientist might seek to reproduce the findings of another. Steven Goodman et al. propose one helpful way to consider this (see Table 1):

Many initiatives (e.g., the CONSORT statement and the Transparency and Openness Promotion Guidelines described on the next page and some elements of the Open Science and Open Data movements) focus on what is described as methods reproducibility above.

Other frameworks for considering types or degrees of reproducibility exist as well. Based on definitions from Prasad Patil et al.:

- “A study is deemed reproducible if it encompasses a population, hypothesis, experimental design, experimenter, data, analysis plan, and code. These parameters must produce the same results in a new analysis for the study to be considered truly reproducible.”

- A study is “strongly replicable” if given a population, hypothesis, experimental design, analysis plan, and code, the researcher obtains consistent estimates when recollecting data and performing the analysis using the original study design and methodology.

Even though some of the definitions in Prasad Patil et al. may differ from the literature and policy discussions, the process is instructive. However, use of Goodman’s terminology is recommended, as it aligns with a linear process of science.

David Crotty of the Oxford University Press raises an important related consideration: Is the research accurate but not easily reproduced? He points out that considering reproducible research valid and irreproducible research invalid is too simplistic, and reasons that some research may be completely valid, reliable, and accurate but not easily reproducible. This can be attributed to differences in technical ability in conducting particular experimental techniques.

Three labs doing behavioral experiments with the same inbred mouse strains discovered huge differences in results even when they intentionally tried to follow precisely the same experimental protocols.”

– Former NIMH Director, Thomas Insel

<table>
<thead>
<tr>
<th>Table 1.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Methods Reproducibility</strong></td>
</tr>
<tr>
<td>Is meant to capture the original meaning of reproducibility, that is, the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results.</td>
</tr>
<tr>
<td><strong>Results Reproducibility</strong></td>
</tr>
<tr>
<td>Refers to what was previously described as ‘replication,’ that is, the production of corroborating results in a new study, having followed the same experimental methods.</td>
</tr>
<tr>
<td><strong>Inferential Reproducibility</strong></td>
</tr>
<tr>
<td>Not often recognized as a separate concept, is the making of knowledge claims of similar strength from a study replication or reanalysis.</td>
</tr>
</tbody>
</table>

Question 3: To What Extent Does the Scientific Study Duplicate or Corroborate Other Studies?

The third question addresses whether a given study does or should corroborate or duplicate other studies. David Crotty\(^{22}\) asserts that “I suspect that we may just have to accept the notion that research requires some level of redundancy. If we create a repository of failed experiments and no one will risk performing (or funding) a new attempt where others have failed in the past, then we may block important discoveries from happening.” He provides the following example to further explain this: “Suppose some researcher proposes that compound X will cure disease Y, but unknowingly uses a contaminated sample of X in the experiments and the cure fails. Do we prevent anyone else from testing that hypothesis and miss out on the potential cure?”

According to Goodman et al.\(^{23}\), inferential reproducibility (defined in the previous section) yields insights relevant to verification. Goodman et al. assert that inferential reproducibility is closely related to corroboration:

“This dimension of reproducibility…refers to the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study.”

If a study theorizes the existence of a biological or natural phenomenon, and the presence of that phenomenon can indeed be detected using different experimental and analytical approaches, then that phenomenon is more likely to have been accurately characterized.

…Do we prevent anyone else from testing that hypothesis and miss out on the potential cure?”

— David Crotty (2014)\(^{24}\)

A certain level of overlap is required to achieve progress. Whether this is called duplication, replication, verification, validation, corroboration, or something else, the degree of desirability and the nature of the overlap are likely to differ in and across research areas.

---


See how the National Institute on Aging is tackling rigor and reproducibility in research:

**Alzheimer’s Disease Preclinical Efficacy Database**

Dr. Refolo is the Director for Alzheimer’s Disease Drug Discovery and Development at the National Institute on Aging (NIA) where he has developed and manages a diverse portfolio of translational research programs and infrastructure.

Why did you create the AlzPED database?

We found that preclinical efficacy studies testing AD therapeutics were not rigorously conducted, contributing to failure of the same therapies in the clinic. Thus, the database was created to develop and promote the implementation of reproducibility strategies, including guidelines and best practices. It was also created to raise awareness about the elements of rigorous design for transparent reporting.

What data is included in the database? And what challenge does it address?

We have curated 300 research articles published between 1996 and 2016, extracted data on therapy type, therapeutic agent, therapeutic target, animal model, PI name, funding source, detailed outcome measures and perhaps most importantly 23 elements of study design. These elements of design are important because we and others believe that they are key indicators of the rigor of a study and drivers of the potential for reproducibility and translatability (i.e., if the therapeutic agent successfully enters the clinic) of a study.

For Alzheimer’s researchers, this does several things:

- provides researchers quick access to integrated data or preclinical therapy development for Alzheimer’s disease
- provides more detailed data on study design and data analysis
- draws attention to a standard set of “best practices” and thereby promotes practices which are likely to increase reuse and reproducibility of published data
- and highlights the effect of publication bias in favor of reporting positive findings.

Tell me more about the analysis

We examined a subset of 7 study design elements critical for human trials (sample size calculation, randomization, blinding for treatment, blinding for outcome measures, eligibility criteria, balance for gender, age matching) across 321 articles and saw that the majority of the studies (44% or 141 studies) did not include any of the 7 critical elements. About a third (34% or 109 articles) included only 1 critical element, the greatest number of critical elements in a study was 4 and only 1.5% or 5 studies included this many.

Who can benefit from this database?

Academic and industry researchers—provides a facile way to survey existing preclinical therapy development efforts, and raise their awareness about the elements of rigorous study design and requirements for transparent reporting.

Information scientists will be enabled to conduct systematic analyses of preclinical efficacy testing studies.

Funding agencies can use this new database as a tool for enforcement of requirements for transparent reporting and rigorous study design.

Looking forward - how can efforts and initiatives like AlzPED and its focus on best practices help scientists/researchers?

At this point, we do not know the impact on the research community. Any meaningful impact on the researchers will require coordinated advocacy by funding agencies. In the meantime, NIA will continue to curate additional published and soon unpublished studies into AlzPED. These data will be available for analysis by the research community, including academic, industry and government scientists enabling scientists to more effectively develop research strategies and plan experiments.

There is certainly a need for additional polices and guidelines aimed at improving reproducibility, rigor and translation of research across biomedicine and other areas. The NIH has issued several policies already.

Access the AlzPED database here.
Conclusion

To judiciously manage limited scientific resources, the evaluation of proposed or completed scientific research can be categorized into three distinct categories:

- **Is the scientific study rigorously designed?**
  - Is the proposed study design valid and will the findings likely be reliable?

- **Is the scientific study reproducible (or replicable or transparent)?**
  - To what degree are the details and procedures of the study reported, and can another investigator produce the same results?

- **To what extent does the scientific study duplicate or verify other studies?**
  - How much overlap is appropriate, and what kind of verification is needed?

This paper synthesizes some of the most relevant literature on this subject, poses these questions, and raises key considerations. The references in this paper provide significant additional information on rigor, reproducibility, and verification/duplication in scientific research. For readers seeking more in-depth insights, Clarivate Analytics encourages comment on this paper.

About Clarivate Analytics

Clarivate Analytics accelerates the pace of innovation by providing trusted insights and analytics to customers around the world, enabling them to discover, protect and commercialize new ideas faster. We own and operate a collection of leading subscription-based services focused on scientific and academic research, patent analytics and regulatory standards, pharmaceutical and biotech intelligence, trademark protection, domain brand protection and intellectual property management. Clarivate Analytics is now an independent company with over 4,000 employees, operating in more than 100 countries and owns well–known brands that include Web of Science, Cortellis, Derwent, CompuMark, MarkMonitor and Techstreet, among others. For more information, visit clarivate.com.

Funders, visit our website clarivate.com/speciality/funding