Improving Reproducibility in the Face of Data Access Restrictions and other Data Complexities

ISSUE BRIEF

Concerns about the reproducibility and replicability of research results have been expressed in both scientific and popular media. As these concerns came to light, Congress requested that the National Academies of Sciences, Engineering, and Medicine assess the extent of issues related to reproducibility and replicability and offer recommendations for improving rigor and transparency in scientific research.

The National Academies' report *Reproducibility and Replicability in Science* (2019) defines reproducibility and replicability and examines the factors that may lead to non-reproducibility and non-replicability in research. This report provides recommendations to researchers, academic institutions, journals, professional societies, and funders on steps they can take to improve reproducibility and replicability in science.

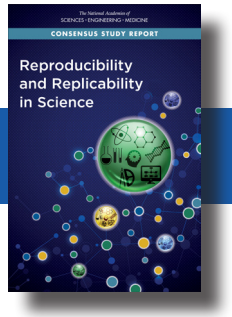
This brief offers highlights from the report, focusing on challenges for reproducibility and replicability presented by research employing proprietary or restricted-access datasets, datasets that continually add data, and datasets with other complexities that may impact research reproducibility and replicability.

DEFINING REPRODUCIBILITY AND REPLICABILITY

The terms "reproducibility" and "replicability" are often used interchangeably, but the report proposes that each term be used to refer to a separate concept. Reproducibility means computational reproducibility—obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis. Replicability means obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data. In short, reproducing research involves using the original data and code, while replicating research involves new data collection and similar methods used by previous studies.

These two processes also differ in the expected outcome of a comparison between two results. In general, when a researcher transparently reports a study and makes available the underlying digital artifacts, such as data and code, the results should be computationally reproducible. In contrast, even when a study was rigorously conducted according to best practices, correctly analyzed, and transparently reported, it may fail to be replicated.

The report emphasizes that any determination of replication between two results needs to take account of both proximity (the closeness of one result to the other, such as the closeness of the mean



Reproducibility and Replicability in Science

www.nationalacademies.org/reproducibilityinscience

The National Academies of SCIENCES ENGINEERING MEDICINE

values) and uncertainty (variability in the measures of the results). A full list of principles and characteristics to consider in assessing replication can be found in Chapter 5 of the report.

REPRODUCIBILITY

The committee's definition of reproducibility focuses on computation because most scientific and engineering research disciplines use computation as a tool. However, access to data, how scientists use software, and how the combination of the two can be verified is neither uniform nor robust. These shortfalls have implications for reproducibility.

The remainder of this document highlights the committee's findings on how to improve computational reproducibility in the presence of complex data. To ensure the reproducibility of computational results, researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results to enable other researchers to repeat the analysis.

IMPROVING REPRODUCIBILITY WHEN DATA OR ACCESS ARE COMPLEX

The report identifies several situations that make reproducibility challenging, if not impossible. A frequent cause is the presence of ***nonpublic data***. Data or code may not be publicly releasable for licensing, privacy, or commercial reasons. For example, commercial data may be proprietary; privacy laws (such as the Health Insurance Portability and Accountability Act, HIPAA, or Title 13 U.S.C., Census law) may restrict sharing of personal information. The report's **Recommendation 4-1** that researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results suggests that researchers should endeavor to describe such data as precisely and transparently as possible, while complying with any legal or ethical restrictions, thus avoiding the issue of "inadequate record keeping" highlighted by the report. For instance, researchers may be able to provide the database schema for nonpublic data, even if they cannot provide the data itself, and they may be able to provide detailed descriptions about the access protocols, even when that access is complex and costly.

Each study and dataset will differ in how it collects and manages data, but there are general steps to consider: data definition, collection, review and culling, and curation. Each step includes decisions that can affect reproducibility and replicability of results. Researchers may use or generate data where the data generating process is imprecisely described, in particular when the data generating process is not under the researcher's control. Examples include confidential data provided by national statistical agencies and private providers of data used in the social sciences, but also numerous "organic data" sources such as administrative data or incidentally collected data. See *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps* (National Academies of Sciences, Engineering and Medicine, 2019) for a discussion of data sources that are generated as by-products of everyday electronic activities, such as retail purchases, electronic health records, or cell phone location records. In cases where researchers employ such data sets and are privileged observers of the data providers and the generating process, they should endeavor to describe this process as fully as possible, in line with **Recommendation 4-1**. In addition, they should encourage the providers of confidential and proprietary data to furnish such information to respond to **Recommendation 4-1** (see also National Academies of Sciences, Engineering, and Medicine, *Methods to Foster Transparency and Reproducibility of Federal Statistics: Proceedings of a Workshop*, 2019). Similarly, data providers and researchers should endeavor to report uncertainties pertaining to the data generating process (**Recommendation 5-1**), an often challenging task when combining data or capturing flow data involve complex operations.

The report identified ***failure to archive*** the relevant digital artifacts necessary for reproducibility as an important issue. When the data generating process is continuous (in particular for organic data sources), archiving and accurately referencing the particular extract or version used by the researcher

remains a challenge. Query interfaces or application programming interfaces may change rapidly, posing additional challenges to computational reproducibility. When using nonpublic data, researchers often have only imperfect control over their computing and data storage environment. The data they use are subject to the same risk of obsolescence as data from any other data provider. The report's **Recommendation 6-5** encourages the National Science Foundation to create "code and data repositories for long-term archiving and preservation of digital artifacts that support claims made in the scholarly record." These repositories can be implemented within nonpublic areas, if necessary, or can themselves accommodate nonpublic code and data access, as some already do. Such nonpublic repositories would nevertheless provide ***public*** metadata, in line with the FAIR (findable, accessible, interoperable, re-usable) data principles referenced in the report, and attenuate the problem of data and code storage obsolescence when properly managed. Providers of restricted-access data may want to follow the same guidelines as open repositories. The minimum requirements for an archival repository are that it is searchable by providing a unique global identifier for the deposited artifact, has a stated guarantee of long-term preservation, and is aligned with a standard set of data access and curation principles. Most commonly, to meet these requirements, a digital object identifier is used as a unique global identifier, and long-term preservation guarantees are at least 10 years.

Restricted-access or very large datasets pose particular challenges for publication reproducibility audits, when journals assess the reproducibility of a manuscript's results prior to publication. Journals are encouraged to make every effort to conduct such checks (**Recommendation 6-4**), but gaining access to restricted-access data may take time, effort, and money that may be scarce in the publication process. Researchers may want to deposit as many related artifacts as possible, such as code or transformed, non-restricted data, in open repositories, in order to facilitate compliance with **Recommendation 6-4**. They may also want to negotiate access for journals at the start of the project, when signing data-use agreements with restricted-access and proprietary data providers. The report also emphasizes that communities that use data subject to some of the above sources of non-reproducibility should endeavor to develop alternative mechanisms for demonstrating reproducibility, possibly with funding by the National Science Foundation (**Recommendation 6-5**).

When computational reproducibility is not achievable due to data restrictions or other complexities, scientists may instead turn to a new study that attempts to replicate the original study of interest. In other words, an attempt to replicate may be the only recourse in the face of conditions that make reproducibility unachievable. Scientists have also developed tools to overcome challenges inherent in particular disciplines; examples include probabilistic forecasting in the field of geoscience, and genome-wide association studies, or GWAS, in the field of genetics. Transparency in methods, data collection, and analysis is paramount in fostering confidence in science when data access is restricted or is complicated.

To see the National Academies' body of work on scientific reproducibility, visit https://www.nap.edu/collection/89/reproducibility.

Division of Behavioral and Social Sciences and Education

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The nation turns to the National Academies of Sciences, Engineering, and Medicine for independent, objective advice on issues that affect people's lives worldwide.
www.national-academies.org