ANNUAL REVIEWS

Annual Review of Statistics and Its Application

The Role of Statistics in Promoting Data Reusability and Research Transparency

Sarah M. Nusser

Department of Statistics, Iowa State University, Ames, Iowa, USA; email: nusser@iastate.edu

Annu. Rev. Stat. Appl. 2023. 10:145–64

First published as a Review in Advance on November 18, 2022

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

https://doi.org/10.1146/annurev-statistics-033121-105114

Keywords

data reuse, data sharing, data quality, open science, reproducibility, replicability

Abstract

The value of research data has grown as the emphasis on research transparency and data-intensive research has increased. Data sharing is now required by funders and publishers and is becoming a disciplinary expectation in many fields. However, practices promoting data reusability and research transparency are poorly understood, making it difficult for statisticians and other researchers to reframe study methods to facilitate data sharing. This article reviews the larger landscape of open research and describes contextual information that data reusers need to understand, evaluate, and appropriately analyze shared data. The article connects data reusability to statistical thinking by considering the impact of the type and quality of shared research artifacts on the capacity to reproduce or replicate studies and examining quality evaluation frameworks to understand the nature of data errors and how they can be mitigated prior to sharing. Actions statisticians can take to update their research approaches for their own and collaborative investigations are suggested.

1. INTRODUCTION

Data generated from research studies are increasingly valued as a primary research product by funders, societies, publishers, and scholars. The elevated importance of publicly accessible research data is a natural outgrowth of multiple recent transformations in research culture and practice (R. Soc. 2012; NASEM 2018, 2019). First, the enormous quantity and variety of data accessible to researchers have led to an explosive array of new types of investigations and methodologies that leverage and increase the value of data sources. Second, in response to the reproducibility crisis in scientific research, scholarly communities are demanding more transparency, especially more thorough documentation of studies with shared data, code, and other artifacts so that others can reproduce or replicate study findings. Finally, these forces, combined with the internet's vastly increased capacity for seamless remote collaboration and sharing of digital objects, have profoundly changed the practice of research through the paradigm of open science (hereafter referred to as open research[1]), which views open communication and transparency in research as essential to evaluating research claims and advancing knowledge (R. Soc. 2012, NASEM 2018).

As a result, research funders, scholarly publishers, and disciplinary communities now commonly expect a peer-reviewed article to be paired with publicly accessible digital artifacts that enable studies to be reproduced and replicated and new studies to be explored with the shared data. Key among these artifacts are the data that generate research findings and the documentation necessary to understand and reuse the data. Funders and publishers are additionally requiring researchers to make accessible detailed descriptions of the methods used to collect, process, and analyze the data, often in the form of computer code for data processing and analyses. Some scholars view this evolution as ultimately leading to a time when the data themselves will be the key research output from studies to which papers are attached (Bourne 2005, Mons et al. 2011, Velterop & Schultes 2020).

Many disciplines are not well equipped to fully embrace these new forms of transparency. In particular, standard research practice for many scholarly fields does not include the necessary elements to ensure shared data are carefully prepared, well documented, and thus able to withstand public scrutiny. Doing so requires attention to study design, measurement, and data processing and analysis in a way that proactively anticipates the need to share data and other research outputs as artifacts of a scholarly paper. These topics all connect to statistical practice and will fundamentally change how the field of statistics intersects with other disciplines and how statisticians approach data sharing and transparency in their own research.

This article explores these topics through the concept of data reusability—the capacity for another user to assess the relevance of a data source for their study, evaluate its quality and limitations, and appropriately use the data in answering new questions. While some disciplines have robust data sharing practices, what makes data reusable in practice is not well understood by most researchers. In particular, more emphasis on statistical thinking is needed to construct a rigorous research process that will generate high-quality data and other artifacts.

The remainder of this review is organized as follows. Section 2 provides background on open research and research transparency and how they are shaping research policy and culture. Section 3 explores the concept of data reusability, reviews factors that promote reusability by data users, and describes practices for producing reusable data. Section 4 focuses on statistical elements of data generation by considering issues that arise in reproducibility and replicability studies and discussing frameworks for assessing the quality of data before they are shared and as they are

[1]Open research (or open scholarship) is an alternative term for open science that recognizes that research transparency and sharing of research artifacts apply to nonscientific fields.

used. Section 5 offers actions that statisticians and other researchers can take to support research transparency in their own investigations. The article concludes by encouraging statisticians to become more involved in learning about and applying transparent research methods in their teams and their own work, as well as to increase their attention to statistical aspects of data generation that affect data quality and inference.

2. CHANGES IN RESEARCH CULTURE AND POLICY

2.1. Open Research and Research Transparency

The importance of data sharing was presciently articulated in a 1985 National Research Council report (NRC 1985), but it was not until the early 2010s that a mass movement for data sharing emerged. In a report titled "Science as an Open Enterprise," the Royal Society outlined recommendations for updating the centuries-old concept of openness through peer-reviewed articles, expanding modes of exchange to capitalize on today's computing infrastructures and collaborative tools (R. Soc. 2012). This report called for "intelligent openness" of shared digital research products—in other words, it called for products to be accessible to others, intelligible through complete and understandable documentation, assessable for evaluating shared outputs and research findings, and usable for future purposes. The report spawned influential initiatives in the European Union and elsewhere to promote research data sharing (e.g., FAIR data, discussed below). In the USA, a White House policy memo (Holdren 2013) directed major government research funders to require public access to research data and publications from funded projects except when prevented by privacy and confidentiality, proprietary, or national security concerns. Many governments and funders responded in kind (NASEM 2018), and a new era of open research and data sharing commenced.

To improve and accelerate the practice of data sharing and reuse in open science, Wilkinson et al. (2016) proposed the now widely accepted FAIR Guiding Principles for sharing data and other artifacts (**https://www.go-fair.org/fair-principles**), reflecting the recommendations laid out in the Royal Society's report. Largely focused on the machine actionability of shared data, the FAIR principles call for data to be findable via internet search, accessible to a person wanting to explore and use a data source, interoperable with computer systems and other data through the use of common standards and vocabularies, and reusable by another user (Wilkinson et al. 2016). The reusability principle aims to facilitate data reuse by providing detailed metadata (i.e., documentation describing the data source) that enable data to be replicated or combined in novel settings. In particular, this principle calls for data and metadata to be "richly described with a plurality of accurate and relevant attributes" that include a clear and accessible data usage license and detailed provenance and that adhere to domain-relevant community standards (Wilkinson et al. 2016). The global community's understanding and implementation of these principles have continued to mature through more complete specifications of metadata requirements to facilitate FAIR implementation and data stewardship competencies required to effectively prepare and share data.

Open research hinges on research transparency as a long-standing and necessary part of the scientific process that enables evaluation, correction, and extension of research findings. Research transparency is a multidimensional concept that includes sharing of data and information on study design, methods, materials, and analyses used during an investigation. Dimensions include data transparency (or sharing), which provides the basic evidence that undergirds research findings (Moravcsik 2019); production transparency, which specifies the steps taken to review, edit, and manipulate data prepared for analysis (Moravcsik 2019); and analytic transparency, which refers to the statistical procedures and code that generate the research results described in an article

(Lupia & Elman 2014). All of these forms of transparency are important in enabling a new user to understand, assess, and appropriately reuse the data source.

2.2. Broader Ecosystem Influences on Open Research and Transparency

A complex ecosystem of organizations and initiatives influences research culture and practice and has been instrumental in promoting open research and transparency. Research institutions, disciplinary societies, scholarly publishers, and research funders promote research transparency in their policies and programs with a mix of mandates, rewards, and frameworks that guide system change in the research ecosystem. While this review focuses on choices statisticians and other researchers can make to foster transparency and data reuse, it is useful to understand the external context and activities that are shaping the move toward open research.

Most funders have deepened their data sharing policies to require public access to research data as a condition of funding, with exceptions for privacy, proprietary, or national security concerns. For example, the US National Institutes of Health (NIH) significantly updated their data management and sharing policy with more-detailed expectations for documenting shared data and code and postgrant reviews by the program officer (NIH 2020). The European Research Council has similar policies and emphasizes adhering to FAIR principles (ERC 2022). Funders typically allow the grant budget to include support for robust data stewardship and appropriate computing and data storage infrastructure, and some sponsors also offer funding programs to establish services to ease the burden of sharing and reusing data and training programs for data sharing tools and research.

Research institutions have responded to these requirements by updating campus policies, computing infrastructure, and staffing to support public access to data. Higher education and scientific associations have fostered this transformation through initiatives to promote institutional and cultural change. For example, US higher education associations have convened academic institutions for discussions on creating campus teams, data policies, and information technology and training needed to support researchers in responsibly sharing their data, which resulted in the publication of a "Guide to Accelerate Public Access to Research Data" for stewarding change at research institutions (Smith et al. 2021). The US National Academies convened the Roundtable on Aligning Incentives for Open Science to catalyze changes in academic research practice and culture, and published a toolkit for fostering open science practices with helpful tips on research practice (NASEM 2021a). In the second phase of their work, Roundtable participants are organizing communities for research institutions, funders, and disciplinary societies to discuss potential strategies and challenges that arise in supporting open research (e.g., **https://www.heliosopen.org**).

Societies also represent a critical lever in changing culture and practice and have held their own convenings to explore how they can support research transparency and openness via scholarly publishing and disciplinary norms. For example, the American Geophysical Union collaborated with more than a dozen societies to host a yearlong data sharing seminar series (**https://wesharedata.org**), attended by numerous societies and a wide range of individuals and organizations in the USA and abroad. Some societies have developed initiatives to improve the understanding of research transparency and researcher responsibilities. Examples include the Data Access and Research Transparency initiative for political science research (Lupia & Elman 2014) and similar guidelines for artificial intelligence research (Gundersen et al. 2018). Societies have also established awards for data sharing and reuse. For example, the Federation of American Societies for Experimental Biology has collaborated with the NIH to support the DataWorks! Prize (**https://www.faseb.org/resources/data-science-and-informatics/dataworks**).

Scholarly publishing is another driver because of its key role in disseminating knowledge from research studies and documenting achievements of researchers. Many publishers have

implemented policies and operations to support research transparency using the Transparency and Openness Promotion (TOP) Guidelines for publishers established by the Center for Open Science (Open Sci. Collab. 2015). The TOP Guidelines have eight modules of standards for scholarly publishers that can be adopted incrementally. Several modules support data, process, and analytic transparency, including modules for sharing of design decisions, data, and code for analytic methods. Recognition of a researcher's contributions is fostered through citation guidelines for data, code, and materials. Publishers can advocate for reproducibility and reduce publication bias (i.e., bias in the published record resulting from submitting and accepting manuscripts on the basis of the direction or significance of the outcomes) by adopting modules for publishing replication studies and null results. Modules for preregistration of studies and of analysis plans can be implemented to reduce the potential for *p*-hacking and increase transparency of planned hypotheses.

In the USA, federal agencies are adopting open research and transparency practices for their internal research activities. For example, the National Aeronautics and Space Administration has initiated an agency-wide program to fully adopt the paradigm of open science in its research operations (NASA 2022). Even statistical offices are considering transparency and reusability, as exemplified by the National Academies of Sciences, Engineering, and Medicine (NASEM) and discussions in the European Union (e.g., Luhman et al. 2019, NASEM 2021b). More broadly, the Foundations for Evidence-Based Policy Making Act of 2018 (H.R. 4174) is leading to a government-wide transformation in archiving and reusing data for policy and program development and evaluation.

The movement toward increased research transparency is being embraced by all sectors of the research ecosystem. While much progress has been made, work remains across all sectors, particularly in how research culture and recognition are defined by scholarly fields and academia. With respect to data sharing, a key challenge to widespread adoption by researchers is a better understanding of what practices will enable data, production, and analytic transparency as a foundation to effective, efficient, and impactful reuse of publicly available research data.

3. DATA REUSABILITY AND RESEARCH PRACTICE

3.1. Data Reusability

Information scientists define data reuse as "the secondary use of data for a purpose other than the original intention of the data producer" (Faniel et al. 2016, p. 1404). In practice, the goal of data reuse ranges from reproducibility studies, which attempt to reproduce the findings and statistics cited in an article using the same data and procedures, to replication studies that use the same research methods to collect new data, possibly in a different context, to entirely new research endeavors that analyze shared data to address a novel question, sometimes in combination with other data sources (e.g., metanalysis).

Data reusability is the ability of a new user to reuse data for a new purpose. Data reusability in this context recognizes that in the process of data reuse, researchers seek to accomplish several tasks, including selecting the data source, understanding the data and their context, and appropriately analyzing the shared data for new purposes. This perspective is related to the R in FAIR but is more focused on researcher practice.

Fostering data reusability is obviously essential to reaping the benefits of shared data. This section explores the attributes of a shared data source that enable reusers to execute tasks in using the data for a new purpose. Much of what is needed is under the control of the researcher who produces the shared data source. Thus, this section also examines approaches to producing and sharing research data that facilitate research transparency and address data producer concerns.

3.2. Data Reuser Needs

Information scientists have examined data reuse behaviors in several disciplines to better understand the tasks researchers engage in and the context needed to select and confidently reuse shared data. The specific information and strategies used in evaluating data reusability vary by study purpose and disciplines, but some commonalities exist across fields of study.

Data users evaluate reusability of a potential data source by determining the relevance of the data to their study goals, understanding the content of the data, and evaluating whether the data are trustworthy (Faniel & Jacobsen 2010). The relevance of a data source is generally assessed at a high level via a small set of scoping parameters that express study requirements. For example, researchers may look for the presence of specific variables important to their study goals and assess whether the time period, location, and/or population represented by the data will support their investigation (Faniel & Jacobsen 2010).

Understanding the data and evaluating their credibility are more detailed processes that depend on granular information (Faniel & Yakel 2017, Faniel et al. 2019). While metadata describing the variables in the data are obviously critical, many other types of documentation are essential to understanding the context of the data generation process, data quality, how to analyze the data, and their appropriate uses. Data users need detailed descriptions of study materials and procedures to evaluate methods that generated the data, in addition to well-documented code for processing, analyzing, and creating statistical summaries from the data. Researchers also use this information to identify assumptions, flaws, and limitations in the data and evaluate how they affect their own investigation. Other information on data provenance and permissible uses is expressed in data use agreements or licenses (Willis & Stodden 2020). Data reusers rely on the quality and completeness of the documentation to understand the data, and consider the professionalism and completeness of the documentation, the use of standards, and their own perception of its correctness in assessing their trustworthiness (Faniel & Yakel 2017).

Information external to the shared digital objects also contributes to assessments of data reusability. Researchers review prior published articles based on the data to validate their quality, understand their scope of application, and gain insights into how to properly analyze the data (Faniel et al. 2019). In addition, reusers examine reputational information to evaluate both the quality of the data and their credibility, particularly for the repository where data reside and the data producer's institutional affiliation and graduate training.

The implication of researcher practices in assessing data reusability is that much more than the metadata describing the data variables must be shared if data are to be reusable by others. Faniel et al. (2019) summarize context reusers' needs by grouping them into three types: (*a*) data production information, including research objectives, data collection methods, and materials used, as well as data issues such as missingness, analysis methods, and producer information; (*b*) repository information, including data provenance, repository curation practices, and repository reputation and history; and (*c*) data reuse information, including information on prior reuse, advice on reuse (from papers or workshops), and terms of use. Stodden (2015) offers a complementary list of items that facilitate reproducibility reuses of data. She emphasizes the details needed on the statistical and computational methods used to generate research findings, such as data processing and analysis steps and information on software and algorithms used in computations.

3.3. Producer Approaches to Improving Data Reusability

The requirements associated with generating and sharing reusable data sources often represent a profound change in a researcher's practice. Data producers face many unknowns, such as how to adapt their research process to create more comprehensive documentation and address

concerns about reputational exposure. They are also frequently unaware of the benefits of transparent research practices to the quality and impact of their research.

While creating effective documentation can be difficult due to the complexity of the research process and the implicit knowledge researchers have about specific tasks in the process (Borgman 2012, Faniel et al. 2019), approaches are emerging that improve the process and minimize effort of developing documentation. Alter & Gonzalez (2018) note that creating clear and transparent documentation of the research process and data is most easily accomplished as an integrated part of the standard research pipeline. To facilitate sharing and reuse of data, Dempsey et al. (2022) recommend adopting a perspective that views data as always being updated and benefiting from the application of FAIR practices throughout the research project. By continuously applying FAIR practices, their approach avoids ineffective and burdensome practices of post hoc data formatting and documentation development that often limit the reusability of shared data.

Incorporating transparency into a research pipeline calls for a much stronger emphasis on prestudy planning (Alter & Gonzalez 2018, Dempsey et al. 2022). Proactively identifying standards for protocols and data, specifying the workflows needed for each task and automating where possible, and determining what documentation will be collected and how it will be captured and maintained before collecting any data all significantly reduce time and cost burdens and improve the consistency of protocol execution and of data quality and reusability (Mikytuck et al. 2022).

Tools for managing and documenting workflows and for dynamically updating data and documentation are increasingly available to assist in this effort, although more training and education are needed. Alter & Gonzalez (2018) offer a list of resources for understanding research workflows and tools for documentation that support production and analytic transparency for social science data, and NASEM (2019) offer a review of workflow tools and approaches. Many researchers already use interactive collaboration tools for managing and tracking versions of workflows, software, and data. Common platforms include GitHub, Jupyter notebooks, and the Open Science Framework. Dempsey et al. (2022) recommend a number of lightweight applications that can facilitate capture of data and workflows so that data are born FAIR. Recognizing that not all researchers have the skills to employ these tools, they discuss the benefits of incorporating a computer scientist into research teams to support automation in the research pipeline, which notably increases efficiency for the researchers by improving error detection and correction, algorithms for data analysis, and preparation of data and code to accompany peer-reviewed manuscript submissions.

The care taken in standardizing, automating, and documenting the research process has additional benefits. Producers are often concerned about risking their reputations by exposing inadvertent errors in their research output. Although unintended errors are an inevitable part of the research process, the work put into creating workflows and clear protocol documentation minimizes the risk of data generation and processing errors and increases the ability of a new user to understand the data source and how to reuse it. Transparency in the discovery and resolution of errors is also a well-accepted practice and is essential to efficient scientific progress.

Another benefit is reducing the potential for data to be misused by others, a concern expressed by data producers (Fecher et al. 2015). Quality documentation about the research process, especially about how data are analyzed, is important to prevent misuse. In addition, creating a data use agreement or license for the data is critical because it describes the data, their ownership, terms for their use and attribution, and requirements for protecting them (Alter & Gonzalez 2018). Data licenses are especially useful for data that need to be restricted due to privacy and confidentiality, proprietary, or national security concerns.

Many researchers do not share data because of confidentiality or proprietary restrictions. However, even if data are restricted, information about the restricted data can still be made publicly accessible, including the existence of the data, their content, and the terms of access

(Alter & Gonzalez 2018). In some fields, repositories have been specially developed to standardize and reduce burden in addressing confidentiality. For example, TalkBank (**https://talkbank.org**) supports both producers and users of video data of children for language development research by offering a secure repository with standardized protocols and required data access and use agreements. The Inter-university Consortium for Political and Social Research (ICPSR; **https://www.icpsr.umich.edu**) has an extensive history of secure access for social science data and offers multiple mechanisms for restricted data access.

More broadly, repositories are a key resource in making data or their existence publicly accessible and ensuring long-term preservation of research artifacts. Much of the work on reusability has aimed to define repository practices to increase data reusability (e.g., Faniel & Yakel 2017). Repository curation practices increasingly place a priority on evaluating data and documentation prior to deposit, sometimes involving additional curation to improve the accessibility and quality of shared objects (e.g., Peer & Green 2012). Repositories often support FAIR principles to the extent possible, which includes ensuring that all digital objects are assigned unique, resolvable persistent identifiers (PIDs) for future users and creating automated processes for data curation and use (Cousijn et al. 2022). Note that PIDs are critical for automation and documentation in data sharing. In addition to identifying the data and metadata, they provide links to context such as the researcher, their institution, their funder, and publications that use the data. PIDs are instrumental in tracking future uses of the data and giving credit to the researchers who produced the shared resource, and they are integral to services such as DATACITE and CHORUS that automatically interlink articles, data, and other shared artifacts with their authors, institutions, and funding sources.

Data reusability can be affected by the availability of computational and software resources required to use shared data. To the degree possible, researchers should use open source software and tools to improve access to the resources needed to reuse data sources (NASEM 2019). An emerging practice for facilitating reuse, especially for restricted or complex data sources, is the use of "containers," or software packages that contain all of the elements needed to run an application in any computing environment. These technologies (e.g., Docker) enable the data and accompanying software code to be shared as an application that does not depend on future data users' computing environment. Whole Tale (Brinckman et al. 2019) is another platform to package data, software, and computing environments for reproducibility studies and future reuse.

Other practices that data producers may be unaware of include data embargo periods and data availability statements. Data producers concerned about completing their research prior to others reusing the data can make the existence of the data publicly known and establish a limited embargo on the data source's use before it will be shared, although this option may be less available given the recent US White House policy mandate (Nelson 2022) to make publications and their supporting research data from federally funded research immediately publicly accessible. Also, journals are now requiring manuscripts to include data availability statements, which describe where and how supporting data can be accessed, facilitate credit for the producer, and create more opportunity for future use and downstream impact of the shared data.

4. STATISTICAL CONSIDERATIONS FOR IMPROVING DATA REUSABILITY

Statistical science is at the heart of ensuring the quality and effectiveness of research investigations. Statistical thinking promotes scholarly rigor and is essential for generating reusable data and ensuring the quality of data and other products generated from a study. Most germane for data reusability are the rigor of the study design for acquiring and processing data and the quality

of the data themselves. This section discusses reproducibility and replicability as special forms of data reuse that serve as a check on different elements of the research process and which can highlight flaws that arise in preparing data for sharing. Approaches to assessing data quality are also summarized and identified as an area needing additional attention by statisticians.

4.1. Data Reusability in Reproducibility and Replicability Studies

Reusable research data are essential for evaluating the reproducibility or replicability of research findings (Hardwicke et al. 2018). Conversely, the broader benefits of reusable and extensible research can be realized when research artifacts are created to support reproducibility (Goeva et al. 2020).

NASEM (2019, p. 46) define reproducibility as "obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis." This definition is synonymous with "computational reproducibility." NASEM also define replicability as "obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data." Reproducibility and replicability are important elements of the process of self-correction in scholarship (McNutt 2020), but they contribute in different ways.

A reproducibility study explores whether another user can obtain the same statistical summaries cited in an article based on the data and code used to produce the results. To determine reproducibility, one attempts to generate the same or sufficiently similar outcomes as those obtained in the original study. Nonreproducibility may arise from statistical issues associated with the conduct of the study (Stodden 2015) or difficulties in the accessibility and reusability of data, code, and study documentation (NASEM 2019). Goodman et al. (2016) note that reproducibility analyses address primarily the user's trust that the findings and shared outputs are indeed what have been represented in the article.

Replicability studies generate new data using the same methods to assess whether outcomes are consistent with evidence offered by the original study or collection of studies. In contrast to reproducibility, replicability studies focus on adding evidence that could confirm or provide insights into the veracity of conclusions from prior studies (Goodman et al. 2016). Inconsistent outcomes can be useful and suggest new areas to pursue or novel insights into our understanding. Alternatively, differences between replication and original study outcomes may emanate from shortcomings in a study's design, measurement process, analysis, or documentation needed to perform the study with the same methods and newly collected data (NASEM 2019). In the following subsections, we briefly examine statistical issues that that arise in reproducibility and replication studies that may affect data reusability.

4.1.1. Examples of data reuse in reproducibility and replicability studies. Studies from a wide range of fields have evaluated the reproducibility and replicability of published outcomes (e.g., NASEM 2019, tables 4-1 and 4-2). Most of these studies involve selecting a set of journal articles and attempting to reproduce or replicate original study findings using information provided with the article or post hoc by the authors. These studies' findings illustrate several statistical issues that hamper data reusability, many of which are entangled in problems with accessibility and documentation of research outputs.

Reproducibility studies require access to the data used to generate findings; the computational steps taken, ideally as executable code; and information on the computing environment (NASEM 2019). Efforts to reproduce computational results have failed due to reusability issues related to access and the quality of information provided. For example, Hardwicke et al. (2018) evaluated data reusability for papers published in *Cognition* after its data sharing policy was implemented. These authors found that 38% of data shared with articles following the implementation of a

data sharing policy were not "in-principle reusable," that is, accessible, complete, and understandable. The primary hurdles were the understandability of the data source and the completeness of documentation, rather than accessibility. In evaluating computational reproducibility for a sample of articles with in-principle-reusable data, Hardwicke et al. (2018) found that 31% required author assistance to reproduce the outcome and 37% were not reproducible even with author assistance. This finding mirrors a broader evaluation by NASEM (2019, p. 9), which concluded that "a number of systematic efforts to reproduce computational results across a variety of studies have failed in more than one-half of the attempts made, mainly due to insufficient detail on digital artifacts, such as data, code and computational workflow." Hardwicke et al. (2018) describe numerous preventable statistical errors hindering their reproducibility study, even when data were in-principle reusable. These include the calculation, use, and reporting of standard deviations or standard errors; *p*-values; test-statistic values; effect sizes; means or medians; degrees of freedom; and confidence intervals.

As with reproducibility studies, replication studies require the use of data, code, workflows, and other study documentation. They depend on shared data and summary statistics in designing a replicate study and assessing whether outcomes of the replications are consistent. For example, Errington et al. (2021) evaluate the replicability of preclinical cancer biology studies, reporting on their attempts to replicate 193 experiments described in 53 high-impact articles. The authors found that none of the papers was documented sufficiently for them to independently design a replication protocol. Similar to the findings of Hardwicke et al. (2018), statistical reporting issues included failing to provide key descriptive and inferential statistics (27% of experiments) or the test associated with a test outcome (21% of experiments). Accessibility was a larger issue than reported by Hardwicke et al. (2018). Only 2% of experiments were associated with publicly accessible data, and when requests were made, authors provided raw data for only 16% of the experiments, summary statistics for 15%, and nothing for 68%. When they had enough information to calculate sample sizes for their replications, Errington et al. (2021) found that many required larger sample sizes for the replication experiment than used in the original study, for instance, 25% larger for animal experiments.

4.1.2. Statistical considerations for data reuse in the context of reproducibility and replicability. The state of reproducibility and replicability studies clearly indicates a need for more involvement by statisticians in improving data reusability and minimizing errors in basic statistical operations. Stodden (2015) discusses these issues through the lens of "statistical reproducibility," providing guidance on areas statisticians should highlight in their own work and collaborations with researchers. She calls for increasing research transparency, especially by giving attention to the quality and completeness of shared data, providing detail on steps taken during the research process, and evaluating the sensitivity of results to the underlying data and models used to generate the findings. NASEM (2019) also highlight the need for increased transparency and outline similar statistical design and reporting issues. Their recommendations apply to data generated from designed studies and studies that incorporate external data sources such as administrative data or forms of "found" data (e.g., transactional data, sensor streams).

Design of the study and data generation mechanism clearly affects the quality and reusability of the data. For experimental and observational studies, a common issue affecting replicability is the use of inadequate sample sizes for drawing meaningful conclusions (Stodden 2015). Ideally, sample sizes are determined in response to a design goal, but frequently resource constraints are used to establish sample sizes. In these instances, it is especially important to evaluate whether meaningful differences can be detected via a power analysis or whether the sample size is such that effect sizes or other parameters can be estimated with precision. NASEM (2019) mention other practices

that negatively influence the usability of data and are familiar to statisticians, such as failure to randomize or blind studies or to account for variables in the design that reduce uncertainties when making inferences on study goals.

Other statistical issues are more prominent when reusing found or large data sources. These data sources often lack detailed provenance information describing how the data were generated. NASEM (2019) and Stodden (2015) underscore the need to characterize uncertainties in these data sources but note the difficulty of this task when the data generating mechanism is hidden from the user. They encourage assessing the data's coverage with respect to the inference population, determining whether important variables have been omitted, and looking for other error sources that may limit the inferences made from found or administrative data. Given a lack of information on the data generation process and the possibility that assumptions for statistical analyses may not be met, Stodden (2015) recommends that the sensitivity of research findings be evaluated from the perspective of how results change if the underlying data or the models used to produce the findings are perturbed to a reasonable degree.

An increasingly common approach to promoting rigor in design and analysis is to preregister (i.e., make publicly accessible) study designs and analysis plans prior to conducting studies. While not statistical per se, preregistration is a form of transparency that documents the statistical design decisions for the research project prior to initiating the study. An especially impactful form of preregistration is the use of registered reports (NASEM 2021a). Registered reports are articles that describe study objectives and methodologies and are peer-reviewed and published prior to knowing study outcomes, with a commitment to publish results regardless of the outcome. Decisions to accept a peer-reviewed registered report are based on the significance of the research question and the quality of the proposed methods (Errington et al. 2021). Registered reports ensure a refereed paper for the authors even if the results are null and represent a form of research transparency that combats analysis flaws such as *p*-hacking, cherry-picking results, hypothesizing after results are known (HARKing), and publication bias toward studies that demonstrate statistically significant findings (NASEM 2019).

To improve research transparency and enable assessments of reproducibility and replicability, many disciplines have developed guidelines specifying the detail with which statistical designs (and other elements of the study) should be described. For example, in the brain and behavioral sciences, Prager et al. (2019) provide practical recommendations for the information needed to demonstrate statistical rigor in a publication, such as specifying the experimental design and units, sample size determinations and power calculations, descriptions of the raw data and transformations of variables, data quality considerations such as outliers and how they were treated, and clear indications of deviations from plans and how they were addressed. Similarly, author instructions for animal research (ARRIVE 2.0) require inclusion of study design details such as treatment and groups, experimental units, sample size and its determination, inclusion and exclusion criteria, randomization, and blinding (Percie du Sert et al. 2020). To fulfill this requirement and support rigor in study design and analysis, recommendations in a report published by the NIH repeatedly center on the need for statistical input at the beginning of a study as a necessary step in addressing reproducibility and replicability of research studies (ACD 2021).

4.2. Evaluating Data Quality

Data reusability depends on the data being of good quality. Many researchers do not know how to assess the quality of a data source or examine what factors may promote its quality. Quality evaluation is an underdeveloped component of data sharing practice, in part because approaches vary in relation to the types of measurements and study designs used by a discipline. Fortunately,

frameworks that could be useful in establishing a broad approach to data review are starting to emerge from statistics and information science.

4.2.1. Statistical frameworks for data quality. Evaluating data quality from a statistical perspective is a multifaceted endeavor (Keller et al. 2017). It is illustrative to consider the context of survey statistics and methodology, which has a long history of examining data quality and its impacts on analysis through error frameworks (Groves & Lyberg 2010). The concept of total survey error (TSE) articulates uncertainties that arise in the survey process from errors in representing the target population and obtaining survey responses (Groves et al. 2009). For example, consider errors of representation. Coverage error occurs when a sampling frame used to sample units for measurement either omits part of or includes more than the scope of the target population or its subpopulations. During data collection, nonresponse for a reporting unit occurs when no data are collected for the unit, which can be problematic if response rates differ across meaningfully different parts of the population. In order to adjust for these representation issues, researchers calculate survey weights for each responding unit to reflect the number of population units represented by the responding unit, using covariates to mitigate potential bias. Similarly, TSE acknowledges that errors occur in responses. Specification error describes the mismatch between the often unobservable target concept and the measurement method used to obtain data. Systematic or random measurement error can arise from errors in the measurement process (e.g., recall bias in dietary intake often skews toward more desirable behaviors; respondents have difficulty recalling what they consumed). Processing error occurs when errors are introduced in editing data, handling missing data, or other computational steps. Evaluating these error sources not only helps a researcher decide what adjustments can be made in acquiring and processing the data but also provides information for articulating potential issues and limitations in the data source.

Because surveys have become more difficult to conduct, the TSE paradigm has been adapted to incorporate big data sources such as administrative data, transactional data, social media data, and other found sources (NASEM 2017). Zhang (2012) presents a two-phase model for multiple data sources that starts with understanding the error in each data source, followed by evaluating errors generated from integrating data sources into a single source. Using this framework, Amaya et al. (2020) outline the process of generating analysis data from big data and survey data sources, noting that, in contrast to the traditional survey data response process, steps in the big data process are iterative and nonlinear in nature. They introduce the Total Error Framework, with error components arising in both big data and survey data. Sources of error include coverage, sampling, specification, nonresponse/missing data, processing, modeling/estimation, and analytic error. The nature of some error sources is similar for survey response and big data. For example, coverage error due to undercoverage, overcoverage, and duplication applies to both types of data, even if the process of representation is different. In contrast, processing error may be larger for big data due to difficulties in data linkage. As noted above, the lack of information on big data generation mechanisms creates extra difficulties in adjusting for missingness in modeling and estimation. The Total Error Framework offers a broader platform than TSE for considering a more formalized approach to quality evaluation of shared research data.

Other data quality assessments account for the purpose associated with the reuse (Keller et al. 2017). In a separate extension of TSE, the Total Survey Quality Framework was developed to incorporate components related to the purpose of using the data (Biemer 2010). The accuracy dimension reflects TSE components and the degree to which TSE is minimized. Other components reflect earlier discussions from information science research on data reusability (Section 3.2) and FAIR principles (Section 2.1). For example, accessibility (data are easy to access), relevance (data satisfy users' needs), usability/interpretability (data are clearly documented and metadata are

well managed), completeness (data have the elements needed to satisfy objectives), and credibility (data are considered trustworthy by the community) are all part of the process data users employ to assess the reusability of data. Remaining framework components include coherence (estimates from different sources can be reliably combined), comparability (comparisons across demographic, spatial, and temporal dimensions are valid), and timeliness (data are delivered on schedule).

These statistical error frameworks indicate that data quality is affected by both the errors that arise in the data generating process and the context associated with the use of the data. While data producers should consider future uses in preparing their shared data, they will not have full knowledge of future reuses of their shared data and thus are in a better position to evaluate and minimize error components related to constructing the analysis data. Reusers, on the other hand, will want to consider both data generation and purpose-driven error components to assess quality and data reusability.

4.2.2. Information science, publishing, and repository perspectives on data quality. Information scientists define data quality in the context of curating digital objects for future use, often expressed as "the ability of a data collection to meet user requirements" (Faniel et al. 2016, p. 1405). They focus on attributes of shared digital objects that enable data reuse, mirroring statistical frameworks that incorporate the purpose of the data. For example, Faniel et al. (2016) found that user satisfaction with data reuse was associated with data completeness (in terms of coverage and missing data), accessibility, ease of use (well managed, easily manipulated), and credibility (accurate, reliable, well documented to avoid misuse and increase confidence in use) and with the quality of the documentation.

Repositories assess data quality as part of the data curation and archival process. Some approaches directly target reproducibility and replicability on the basis of a "replication standard" that requires "sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author" (King 1995, p. 444). Peer et al. (2014) summarize data review approaches used by selected domain-specific and generalist repositories to meet this standard. Data quality review is considered a curatorial activity prior to making data accessible, with the purpose of maintaining and protecting the value of research data. Curatorial tasks support nonstatistical requirements to facilitate reusability, such as resolving format issues to ensure access and reusability, reviewing data to ensure they adhere to confidentiality restrictions, creating documentation and metadata to promote interpretability, and assigning digital object identifiers to facilitate appropriate attribution. The ICPSR offers an oft-copied example of this approach, with protocols for checking submissions of data, study descriptions and methodology documentation for confidentiality issues and completeness, formatting files to be accessible, and ensuring consistency between data and documentation (see the Ingest chapter in ICPSR 2022). The Yale Institution for Social and Policy Studies (ISPS) has gone a step further by establishing a "replication" repository (sensu King 1995) to support reproducible research (Peer & Green 2012). It employs statistical experts and embeds review of data content in the curatorial process, extending protocols that mirror the ICPSR approach to include reproduction of original research results using data and submitted code.

Approaches to curating for data quality are evolving. In the context of earth science data reuse, Peng et al. (2022) report on emerging efforts to create guidelines for documenting, sharing, and reusing quality information for individual data sets that align with FAIR principles. Their four-part framework conceptualizes quality from the perspectives of science (the data generation process), product (elements that relate to creating a shared data source), stewardship (curation-related steps), and services (data uses). While it is too early to tell how data quality documentation will be defined and assessed, these authors' efforts offer a more integrated platform for information scientists to holistically curate for data quality.

Scholarly journals requiring data accessibility may also support a form of data quality review through reproducibility evaluations. Christian et al. (2018) describe efforts by the *American Journal of Political Science* (AJPS) to evaluate compliance with AJPS data sharing standards, including reproducibility assessments. AJPS uses a dedicated repository service hosted by the University of North Carolina's Odum Institute for Research in Social Science, which, like ISPS, has staff with statistical expertise. The approach is similar to that described by Peer et al. (2014) for evaluating King's (1995) replication standard. An issue faced by AJPS is the scalability of such reviews. Christian et al. (2018) report that approximately 8 hours are devoted to fully reviewing a data source, with only 10% of data sources shared with an article passing checks on the first run. Willis & Stodden (2020) describe approaches taken by journals in their computational reproducibility reviews (including the *Journal of the American Statistical Association*), noting issues of scalability and variability in journal initiatives as well as researcher challenges. They propose the concept of "assessable computational research artifacts," which are digital objects containing information required to reproduce findings and outcomes of the journal's reproducibility evaluation, thereby offering a persistent and linked record of the reproducibility review.

Researchers have attempted to address the scalability of quality checking through automation of repository processes. For example, Stonebraker et al. (2013) describe an end-to-end data curation system designed to operate at scale and minimize human intervention. They define data curation as "the act of discovering a data source(s) of interest, cleaning and transforming the new data, semantically integrating it with other local data sources, and deduplicating the resulting composite" (p. 1), addressing several issues that arise in nondesigned data. They suggest that features of any end-to-end curation system should rely on automation to achieve scalability, incorporate a process for data cleaning, build an interface that accommodates use by nonprogrammers, and allow for a data integration process that is adapted as new sources become available. They also note that the ability to rely on automation is greater when complete or at least considerable information is available about the content and format of a data source. A second example is given by Pezoulas et al. (2019), who outline an automated data quality assessment framework for curating medical data that incorporates some statistical review. The approach hinges on data cleaning as the foundation of data quality and on developing an automated system to scale the cleaning process for large and complex medical data sources. Key elements of their process include detection of missing values, outliers, and similarity in records (rows) or highly correlated variables (columns) for further treatment, reflecting some of the error components discussed in quality frameworks. In many of these attempts, the treatment of data is simplistic, but the concepts of automation and large-scale quality assessment could be useful.

Finally, disciplinary repositories play an important role in establishing norms that promote data reusability by ensuring that data are well documented and preserved for future access and reuse. Some of the oldest and most effective examples are repositories developed for specific data types that are used in a wide range of studies, such as genomic sequences, brain images, and astronomic observations. Through community-based efforts, many of these specialist applications have developed standards and tools to evaluate the quality and reusability of data to be deposited in the repository. These archives generally use controlled vocabularies and formats for data content and file specifications, with algorithms that validate quality and compliance with standards before data are ingested into the repository. By focusing on a specific and common form of data that is driving advances in a field, these specialist repositories overcome hurdles associated with scaling data quality checks and create services that reduce producer burden in preparing data. A long-standing example is the Protein Data Bank, which is the "global archive of experimentally determined three-dimensional structure data for biological macromolecules" (wwPDB Consort.

2019, p. D520). More recently, neuroscientists have developed a standard for sharing brain images (Gorgolewski et al. 2016), which has been adopted by community repositories such as the OpenfMRI repository (Poldrack et al. 2013).

5. ACTIONS FOR STATISTICIANS AND OTHER RESEARCHERS

Research practice in all fields, including statistics, is evolving to place a greater emphasis on rigor and research transparency. As responsible scholars, it behooves statisticians to increase their awareness of these transformations and update their practices to align with the new ways the scientific method is being implemented. At a high level, this means taking a more holistic view of how their own or their collaborators' data are prepared for sharing and future reuse.

A good starting point to maximize the quality, reusability, and impact of shared data is to spend more time in the planning phase on how reusability can be fostered. While future uses for shared data may not be known, some researchers have found it helpful to consider their own reuse of data at a later date (Nusser et al. 2021, Martone & Nakamura 2022, Mikytuck et al. 2022). It is also useful to consider what it would take for someone to reproduce or replicate a research study without assistance from the producer. Below is a list of topics to consider in planning to share reusable data (they are interrelated but listed separately for clarity):

1. Plan for sharing. At the beginning of the study, extend planning beyond the design of the research study itself to define what products will be shared to foster reusability and how they will be made accessible. Products should include data, code, methods, materials, and other artifacts needed to enable a future user to effectively use the data without assistance from the data producer. Restricted data require additional attention for permission for future use (e.g., through institutional review boards), terms of use expressed in the data license, and how the data or their existence are made accessible (Sections 3.1 and 3.2).
2. Plan for error prevention. Consider error evaluation frameworks to identify potential sources of error in data generation and with respect to the study purpose, and identify methods to avoid or minimize the impact of errors in data sources and manipulations. Consider and minimize statistical issues commonly arising in reproducibility and replication studies (Section 4).
3. Plan for methods and documentation capture. Identify a collaborative tracking and versioning platform to assist in managing the application and documentation of study methods. Develop processes and conventions that can be consistently applied as workflows and automated to the degree possible for data acquisition, checking, and updating and for processing and analyzing data. Consider future reusability by evaluating the quality and reasonableness of the data as they are acquired and creating commented and versioned code. Use existing disciplinary standards and open source resources, as well as approaches that allow data to be born FAIR (Section 3.3).
4. Find a trusted repository that prioritizes reuse and transparency. When products are to be preserved and made accessible, consider repository venues that are easily visible to others who might benefit from future use of the data source, especially domain-specific repositories. When a generalist repository is used, consider what kind of curation they support and the features that demonstrate that the repository is trustworthy (e.g., data quality review, assigning PIDs) (Section 3.3).
5. Choose publishers that require sharing of research artifacts, data availability statements, and quality checks on artifacts. Evaluate journal policies and practices in relation to the TOP Guidelines to assess their support for research transparency. Look for checklists that specify expectations for transparency in manuscripts and research products (Section 2.2).

6. Stay informed and contribute to advancing research culture. Stay aware of and get involved with disciplinary and pan-disciplinary transformations related to research transparency, open research, and frameworks such as FAIR. Campus research offices and libraries typically offer the resources and training needed to understand what is involved in making data publicly accessible and adopting new tools to reduce burden and enable sharing among collaborators. International initiatives such as RDMkit (**https://rdmkit.elixir-europe.org**) and GO-FAIR offer additional resources. Explore initiatives in your own and your collaborators' fields to learn about current and emerging practices. Journals such as *PLOS*, *Science*, and *Nature* regularly publish articles on new developments in open research practice (Section 2).

6. CONCLUSIONS

Data sharing has emerged as an essential element of scientific practice. Data are a primary source of evidence supporting study conclusions and thus play a strong role in research transparency. They fuel the potential to expand our knowledge base as researchers pose new questions with shared data. Making data publicly accessible is now expected by funders and publishers unless there are reasons to restrict data access. While many elements of the research ecosystem are evolving, the basic requirement to share data has arrived.

Because data sharing is necessary but not sufficient for their effective reuse, the field of statistics should consider the broader context of open science, research transparency, and other policy changes that are affecting nearly all researchers in producing and sharing data. This review has introduced the key concept of data reusability, which connects directly to the field of statistics through an emphasis on rigor and error prevention and is informed by information science practices for curating research artifacts for future use.

As researchers with expertise in study design, data handling, and inference, statisticians play an important role in fostering statistically sound research studies that generate quality data. Many of the statistical mishaps in planning a study, collecting data, and preparing a data source are familiar to statisticians. What is less familiar is how they are expressed in the paradigm of open science and the downstream impact of errors on future use and the reputation of those who publish data. The plethora of statistical errors demonstrated through reproducibility and replication studies emphasize the need for statistical expertise at all phases of research. This echoes the call by Brownstein et al. (2019) for more involvement by statisticians in scientific studies, emphasizing that nearly all elements of research studies have statistical elements and almost no elements are purely statistical. Disciplines also recognize that they need this expertise, as is evident in the recommendations in the NIH report on enhancing rigor and transparency in animal research (ACD 2021).

An underdeveloped area in data sharing culture is giving more explicit attention to fostering data quality. While it is difficult to avoid unintended errors entirely, the potential for error can be more thoughtfully considered. Ideally, the potential for error at various stages in a research workflow is evaluated in the planning process. Approaches and tools that can reduce or eliminate errors should be considered and adopted whenever possible. Additionally, it is important to evaluate error in the data as they are produced, processed, and analyzed. In characterizing potential errors, a good start for nonsurvey data sources is the Total Quality Framework proposed by Amaya et al. (2020). This kind of framework can help statisticians and other researchers consider potential quality issues in more detail and help them bring error assessment to the fore in collaborations.

Statisticians also need to focus beyond the data per se. The reusability of data requires other contextual information, particularly well-documented and clean code for processing and analyzing

the data and detailed descriptions of methods and materials involved in generating, processing, and analyzing data. Overall, the field of statistics does not have a strong tradition of or training in the curation of code or data, although some subfields like survey statistics have more experience. Nor does statistical training involve efficient approaches for detailed record keeping, as is the case for lab scientists. It behooves statisticians to join other disciplines in learning more about how to capture information about the research process to support data reusability and research transparency more broadly.

DISCLOSURE STATEMENT

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

ACKNOWLEDGMENTS

I thank David C. Walker for his assistance in reviewing literature, Alyssa M. Mikytuck and Gizem Korkmaz for their collaborations in our funded project to understand reusability from researcher perspectives, and the reviewers and editors for their insightful comments. The writing of this review was supported in part by the National Science Foundation under EAGER award 2039677; any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Additional support was provided by the Biocomplexity Institute and Initiative and by the University of Virginia Strategic Initiative Fund Project 160.

LITERATURE CITED

ACD (Advis. Comm. Dir.). 2021. *ACD Working Group on Enhancing Rigor, Transparency and Translatability in Animal Research: final report*. Rep., NIH, Bethesda, MD. **https://acd.od.nih.gov/documents/presentations/06112021_ACD_WorkingGroup_FinalReport.pdf**

Alter G, Gonzalez R. 2018. Responsible practices for data sharing. *Am. Psychol.* 73(2):146–56. **https://doi.org/10.1037/amp0000258**

Amaya A, Biemer PP, Kinyon D. 2020. Total error in a big data world: adapting the TSE framework to big data. *J. Surv. Stat. Methodol.* 8:89–119. **https://doi.org/10.1093/jssam/smz056**

Biemer P. 2010. Total survey error: design, implementation, and evaluation. *Public Opin. Q.* 74(5):817–48. **https://doi.org/10.1093/poq/nfq058**

Borgman C. 2012. The conundrum of sharing research data. *J. Am. Soc. Inf. Sci. Technol.* 63(6):1059–78. **https://doi.org/10.1002/asi.22634**

Bourne P. 2005. Will a biological database be different from a biological journal? *PLOS Comput. Biol.* 1(3):e34. **https://doi.org/10.1371/journal.pcbi.0010034**

Brinckman A, Chard K, Gaffney N, Hategan M, Jones MB, et al. 2019. Computing environments for reproducibility: capturing the "Whole Tale." *Future Gener. Comput. Syst.* 94:854–67. **https://doi.org/10.1016/j.future.2017.12.029**

Brownstein NC, Louis TA, O'Hagan A, Pendergast J. 2019. The role of expert judgment in statistical inference and evidence-based decision-making. *Am. Stat.* 73(Suppl. 1):56–68. **https://doi.org/10.1080/00031305.2018.1529623**

Christian T-ML, Lafferty-Hess S, Jacoby WG, Carsey T. 2018. Operationalizing the replication standard. *Int. J. Digit. Curation* 13(1):114–24. **https://doi.org/10.2218/ijdc.v13i1.555**

Cousijn H, Habermann T, Krznarich E, Meadows A. 2022. Beyond data: sharing related research outputs to make data reusable. *Learn. Publ.* 35:75–80. **https://doi.org/10.1002/leap.1429**

Dempsey WP, Foster I, Fraser S, Kesselman C. 2022. Sharing begins at home: how continuous and ubiquitous FAIRness can enhance research productivity and data reuse. *Harvard Data Sci. Rev.* 4(3). **https://doi.org/10.1162/99608f92.44d21b86**

ERC (Eur. Res. Counc.). 2022. *Open research data and data management plans, version 4.1*. Grant Inf., ERC, Brussels. **https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf**

Errington TM, Denis A, Perfito N, Iorns E, Nosek BA. 2021. Reproducibility in cancer biology: challenges for assessing replicability in preclinical cancer biology. *eLife* 10:e67995. **https://doi.org/10.7554/eLife.67995**

Faniel IM, Frank R, Yakel E. 2019. Context from the data reuser's point of view. *J. Doc.* 75(6):1274–97. **https://doi.org/10.1108/JD-08-2018-0133**

Faniel IM, Jacobsen TE. 2010. Reusing scientific data: how earthquake engineering researchers assess the reusability of colleagues' data. *Comput. Support. Coop. Work* 19:355–75

Faniel IM, Kriesberg A, Yakel E. 2016. Social scientists' satisfaction with data reuse. *J. Assoc. Inf. Sci. Technol.* 67(6):1404–16. **https://doi.org/10.1002/asi.23480**

Faniel IM, Yakel E. 2017. Practices do not make perfect: disciplinary data sharing and reuse practices and their implications for repository data curation. In *Curating Research Data*, Vol. 1: *Practical Strategies for Your Digital Repository*, ed. LR Johnston, pp. 103–26. Chicago: Assoc. Coll. Res. Libr.

Fecher B, Friesike S, Hebing M. 2015. What drives academic data sharing. *PLOS ONE* 10(2):e0118053. **https://doi.org/10.1371/journal.pone.0118053**

Goeva A, Stoudt S, Trisovic A. 2020. Toward reproducible and extensible research: from values to action. *Harvard Data Sci. Rev.* 2(4). **https://doi.org/10.1162/99608f92.1cc3d72a**

Goodman SN, Fanelli D, Ioannidis JPA. 2016. What does research reproducibility mean? *Sci. Transl. Med.* 8:341. **https://doi.org/10.1126/scitranslmed.aaf5027**

Gorgolewski K, Auer T, Calhoun V, Craddock RC, Das S, et al. 2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3:160044. **https://doi.org/10.1038/sdata.2016.44**

Groves RM, Fowler FJ Jr., Couper MP, Lepkowski JM, Singer E, Tourangeau R. 2009. *Survey Methodology*. Hoboken, NJ: Wiley. 2nd ed.

Groves RM, Lyberg L. 2010. Total survey error: past, present, and future. *Public Opin. Q.* 74(5):817–48. **https://doi.org/10.1093/poq/nfq065**

Gundersen OE, Gil Y, Aha DW. 2018. On reproducible AI: towards reproducible research, open science, and digital scholarship in AI publications. *AI Mag.* 39(3):56–68. **https://doi.org/10.1609/aimag.v39i3.2816**

Hardwicke TE, Mathur MB, MacDonald K, Nilsonne G, Banks GC, et al. 2018. Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*. *R. Soc. Open Sci.* 5:180448. **https://doi.org/10.1098/rsos.180448**

Holdren JP. 2013. *Increasing access to the results of federally funded scientific research*. Memo., Off. Sci. Technol. Policy, White House, Washington, DC, Febr. 22. **https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf**

ICPSR (Inter-Univ. Consort. Political Soc. Res.). 2022. *ICPSR: a case study in repository management*. Online Resour., ICPSR, Ann Arbor, MI. **https://www.icpsr.umich.edu/web/pages/datamanagement/lifecycle/index.html**

Keller SA, Korkmaz G, Orr M, Schroeder A, Shipp SS. 2017. The evolution of data quality: understanding the transdisciplinary origins of data quality concepts and approaches. *Annu. Rev. Stat. Appl.* 4:85–108. **https://doi.org/10.1146/annurev-statistics-060116-054114**

King G. 1995. Replication, replication. *PS Political Sci. Politics* 28:444–52

Luhman S, Grazzini J, Ricciato F, Meszaros M, Giannakouris K, et al. 2019. *Promoting reproducibility-by-design in statistical offices*. Paper presented at Conference on New Techniques and Technologies for Statistics, Brussels, March 14. **https://doi.org/10.5281/zenodo.3240198**

Lupia A, Elman C. 2014. Openness in political science: data access and research transparency. Introduction. *PS Political Sci. Politics* 47(1):19–42

Martone ME, Nakamura R. 2022. Changing the culture on data management and sharing: overview and highlights from a workshop held by the National Academies of Sciences, Engineering, and Medicine. *Harvard Data Sci. Rev.* 4(3). **https://doi.org/10.1162/99608f92.44975b62**

McNutt M. 2020. Self-correction by design. *Harvard Data Sci. Rev.* 2(4). **https://doi.org/10.1162/99608f92.32432837**

Mikytuck AM, Nusser SM, Korkmaz G. 2022. The interdependence of data producers and data users: how researchers' behaviors can support or hinder each other. MetaArXiv yp3ct. **https://doi.org/10.31222/osf.io/yp3ct**

Mons B, van Haagen H, Chichester C, Hoen PB, den Dunnen JT, et al. 2011. The value of data. *Nat. Genet.* 43:281–83. **https://doi.org/10.1038/ng0411-281**

Moravcsik A. 2019. *Transparency in Qualitative Research*. London: SAGE

NASA (Natl. Aeronaut. Space Adm.). 2022. *Transform to Open Science* (*TOPS*). Open Sci. Initiat., NASA, Washington, DC. **https://science.nasa.gov/open-science/transform-to-open-science**

NASEM (Natl. Acad. Sci. Eng. Med.). 2017. *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: Natl. Acad. **https://doi.org/10.17226/24893**

NASEM (Natl. Acad. Sci. Eng. Med.). 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, DC: Natl. Acad. **https://doi.org/10.17226/25116**

NASEM (Natl. Acad. Sci. Eng. Med.). 2019. *Reproducibility and Replicability in Science*. Washington, DC: Natl. Acad. **https://doi.org/10.17226/25303**

NASEM (Natl. Acad. Sci. Eng. Med.). 2021a. *Developing a Toolkit for Fostering Open Science Practices: Proceedings of a Workshop*. Washington, DC: Natl. Acad. **https://doi.org/10.17226/26308**

NASEM (Natl. Acad. Sci. Eng. Med.). 2021b. *Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies*. Washington, DC: Natl. Acad. **https://doi.org/10.17226/26360**

Nelson A. 2022. *Ensuring free, immediate, and equitable access to federally funded research*. Memo., Off. Sci. Technol. Policy, White House, Washington, DC, Aug. 25. **https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf**

NIH (Natl. Inst. Health). 2020. *Final NIH policy for data management and sharing*. Policy, NIH, Bethesda, MD. **https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html**

NRC (Natl. Res. Counc.). 1985. *Sharing Research Data*. Washington, DC: Natl. Acad. **https://doi.org/10.17226/2033**

Nusser SM, Cutcher-Gershenfeld JE, Mikytuck AM, Korkmaz G. 2021. *Fostering data reusability: increasing impact and ease in sharing and reusing research data*. Workshop Rep., Ia. State Univ./Univ. Va./Natl. Sci. Found., Washington, DC. **https://doi.org/10.5281/zenodo.5390123**

Open Sci. Collab. 2015. Promoting an open research culture. *Science* 348(6242):1422–25. **https://doi.org/10.1126/science.aab2374**

Peer L, Green A. 2012. Building an open data repository for a specialized research community: process, challenges and lessons. *Int. J. Digit. Curation* 7(1). **https://doi.org/10.2218/ijdc.v7i1.222**

Peer L, Green A, Stephenson E. 2014. Committing to data quality review. *Int. J. Digit. Curation* 9(1). **https://doi.org/10.2218/ijdc.v9i1.317**

Peng G, Lacagnina C, Downs RR, Ganske A, Ramapriyan HK, et al. 2022. Global community guidelines for documenting, sharing, and reusing quality information of individual digital datasets. *Data Sci. J.* 21:8. **http://doi.org/10.5334/dsj-2022-008**

Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, et al. 2020. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. *PLOS Biol.* 18(7):e3000410. **https://doi.org/10.1371/journal.pbio.3000410**

Pezoulas VC, Kourou KD, Kalatzis F, Exarchos TP, Venetsanopoulou A, et al. 2019. Medical data quality assessment: on the development of an automated framework for medical data curation. *Comput. Biol. Med.* 107:270–83

Poldrack R, Barch D, Mitchell J, Wager T, Wagner A, et al. 2013. Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front. Neuroinformatics* 7:12. **https://doi.org/10.3389/fninf.2013.00012**

Prager EM, Changers KE, Plotkin JL, McArthur DL, Bandrowski AE, et al. 2019. Improving transparency and scientific rigor in academic publishing. *Brain Behav.* 9:e01141. **https://doi.org/10.1002/brb3.1141**

R. Soc. 2012. *Science as an open enterprise*. Final Rep., R. Soc., London. **https://royalsociety.org/topics-policy/projects/science-public-enterprise/report**

Smith TL, Redd K, Nusser S, Samors R, Miller ER. 2021. *Guide to accelerate public access to research data*. Rep., Assoc. Am. Univ. Assoc. Public Land Grant Univ., Washington, DC. **https://doi.org/10.31219/osf.io/tjybn**

Stodden V. 2015. Reproducing statistical results. *Annu. Rev. Stat. Appl.* 2:1–19. **https://doi.org/10.1146/annurev-statistics-010814-020127**

Stonebraker M, Bruckner D, Ilyas IF, Beskales G, Cherniack M, et al. 2013. *Data curation at scale: the Data Tamer System*. Paper presented at 6th Biennial Conference on Innovative Data Systems Research (CIDR '13), Asilomar, CA, Jan. 6–9. **https://cs.uwaterloo.ca/~ilyas/papers/StonebrakerCIDR2013.pdf**

Velterop J, Schultes E. 2020. An academic publishers' GO FAIR implementation network (APIN). *Inf. Serv. Use* 40(4):333–41. **https://doi.org/10.3233/ISU-200102**

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018. **https://doi.org/10.1038/sdata.2016.18**

Willis C, Stodden V. 2020. Trust but verify: how to leverage policies, workflows, and infrastructure to ensure computational reproducibility in publication. *Harvard Data Sci. Rev.* 2(4). **https://doi.org/10.1162/99608f92.25982dcf**

wwPDB Consort. 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* 47:D520–28. **https://doi.org/10.1093/nar/gky949**

Zhang LC. 2012. Topics of statistical theory for register-based statistics and data integration. *Stat. Neerl.* 66(1):41–63. **https://doi.org/10.1111/j.1467-9574.2011.00508.x**