Scientific Research Publishing

Journal of Transportation Technologies, 2025, 15(1), 179-202
https://www.scirp.org/journal/jtts
ISSN Online: 2160-0481
ISSN Print: 2160-0473

Reproducibility in Transportation Research: Importance, Best Practices, and Dealing with Protected and Sensitive Data

Jonathan S. Wood[1]*, **Ida van Schalkwyk**[2]

[1]Department of Civil, Construction & Environmental Engineering, Iowa State University, IA, USA
[2]Washington Department of Transportation, WA, USA
Email: *jwood2@iastate.edu; vanschi@wsdot.wa.gov

How to cite this paper: Wood, J.S. and van Schalkwyk, I. (2025) Reproducibility in Transportation Research: Importance, Best Practices, and Dealing with Protected and Sensitive Data. *Journal of Transportation Technologies*, **15**, 179-202.
https://doi.org/10.4236/jtts.2025.151010

Received: December 28, 2024
Accepted: January 23, 2025
Published: January 26, 2025

Abstract

Reproducibility is a key aspect of the scientific method as it provides evidence for research claims. It is essential to promote openness, accessibility, and collaboration within the scientific community. This article aims to provide an introduction to best practices in reproducibility that are relevant to the transportation research community, to discuss issues and barriers to reproducibility, and to describe methods for addressing these issues. This article starts by discussing openness and transparency, then discusses several key best practices for reproducibility in transportation engineering, highlighting common methods and techniques, as well as the associated benefits. The paper concludes with a discussion of the key barriers to implementing reproducibility practices in transportation research and potential solutions. The barriers include existing culture and attitudes, data sensitivity, insufficient methodological detail, lack of code sharing, limited validation, additional time and research burden, and skill and knowledge gaps. Discussing each of these items provides an opportunity for the transportation research community to evolve to become one that embraces the openness and transparency of reproducibility.

Keywords

Reproducibility, Openness, Transparency, Scientific Method, Responsible Research

1. Introduction

Scientific evidence is a cornerstone of engineering progression and improvement. As a standardized approach for establishing facts, the scientific method fails when supporting evidence is not provided or fully transparent. While interpretation of the supporting evidence does not guarantee that the scientific claims are correct, the scientific method fails without complete and transparent supporting evidence. Weak and opaque evidence provides little to no support for scientific claims, while strong reproducible and replicated evidence is required for strong claims. Furthermore, when new evidence that conflicts with earlier evidence is found, it is important that the previous evidence be available for scrutiny alongside the new evidence to enhance scientific knowledge and allow for direct comparison. Without this, evidence that has been widely accepted may be difficult to challenge (regardless of the accuracy of the accepted claims), creating barriers to growth in related scientific knowledge and progress [1].

Reproducibility is a fundamental scientific principle [2] that asserts the ability to consistently obtain the same results and interpretation using the original raw data and an identical process [3]. Its counterpart, replicability, also crucial to establishing scientific knowledge, involves achieving consistent results using a different dataset—and potentially different analysis methods—to address the same research question or hypothesis.

Replicability is vital as it helps minimize the likelihood of bias in results and establish whether independent research provides consistent evidence. Reproducibility, however, is key to building trust in the evidence generated by research studies and for establishing evidence that can be compared with future research findings and evidence (and thus, essential for true replication). Without the original evidence for comparison, it may be more difficult to publish conflicting research in the future and gain acceptance of the results by the scientific community since the conflicting results cannot be directly compared to the original evidence providing strong evidence for opposing findings more difficult (and providing strong evidence against what is considered to have scientific consensus difficult) [1]. This also makes published replication efforts suspect as it is not clear if the publications are systematically biased by being more publishable due to consistency with earlier publications [1] [4].

Several scientific fields have pushed for reproducibility as a standard [5]-[8] with one author referring to it as a "credibility revolution" [9]. Additionally, many initiatives by the National Academies [10] have been the focus of many initiatives to ensure that science and engineering advance as quickly as possible. Reproducibility implements the idea that the average researcher in the same field could reasonably use the same data source, apply the same data processing and analysis, and obtain the same results. When the raw data are not from readily available sources, part of reproducibility is providing the raw data. Reproducibility also requires full details on data processing and cleaning, analysis methods (including annotated code, when appropriate), and full disclosure of assumptions made. Results from the research itself are greatly influenced by decisions made during the processing, cleaning, linkage, and analysis of the data.

Although reproducibility does not guarantee that the research and evidence are correct (or unbiased), it provides transparency, improves trust, and allows other researchers and interested parties to evaluate and validate the evidence and results claimed [10] [11]. This is essential in the progression of science and for ethical dissemination of research [12]. Some have even opined that evidence must be reproducible to qualify as scientific evidence [13]. Others have argued that, before publication, the reviewers should be provided with adequate information and data to reproduce the results (in what is termed reproducibility)—or they shouldn't be asked to provide the reviews [14].

Reproducibility is starting to become a requirement for some transportation research funding and was included in the request for proposals for some projects (e.g., NCHRP 22-49: The Effect of Vehicle Mix on Crash Frequency and Crash Severity). Reproducibility in transportation research was the focus of a paper, which focused on computational analysis and computing technologies, and noted several benefits to reproducibility [15]. Some of the benefits listed include greater impacts of the research, increased reputations for the researchers, minimizing the chance of errors in the research, increased trust, productivity improvements, improved ease of other researchers performing extended work that is related, and improved stewardship of public resources [15].

Given the impacts of transportation engineering on people's lives, reproducibility has ethical implications. When misconduct occurs in research, it undermines the integrity and trust of the entire research community. Following best practices for reproducibility provides transparency that reduces the possibility of data fabrication and falsification. It also creates a research environment that promotes integrity and the responsible conduct of research [16]. Thus, the importance of reproducibility and reproducibility practices in transportation research cannot be understated.

A particularly problematic reproducibility-related issue with both scientific and ethical implications is the traditional "trust us" attitude that is sometimes encountered in scientific research. This attitude, where researchers may resist providing their data or code under the assumption that their results and interpretations should be trusted as is, or should be trusted because of their credentials or experience, fundamentally contradicts the principles of transparency and reproducibility central to scientific integrity. The "trust us" attitude hinders the ability of other researchers to validate and build upon the work and can also lead to the proliferation of erroneous findings [17]. This attitude can be viewed as unscientific, as it avoids the critical step of peer scrutiny that is a cornerstone of the scientific process [18], turning it instead into an act of faith. From an ethical perspective, this attitude is troubling for many reasons particularly if the research was supported with public funds (*i.e.*, an investment of public resources). Withholding crucial details of that research effectively prevents the public from fully benefiting or scrutinizing the investment [19]. Without transparency, the credibility of the evidence and claims is lost. Thus, one paper discussing the importance of reproducibility and open science stated that "transparency is superior to trust" [20].

In the context of transportation research, where research outcomes can have direct, tangible impacts on public safety, infrastructure planning, policy, and project budgets, the "trust us" attitude[1] is concerning. Inaccurate or unfounded research results could lead to incorrect design, policy, and regulatory decisions with far-reaching consequences [21]—underlining the necessity for reproducibility and transparency in research. Transportation research includes areas such as traffic safety, environmental sustainability, urban planning, and economic development. Misleading results in these areas can lead to inefficient use of resources, ineffective policies, and even endanger public safety. For instance, incorrect assumptions about traffic patterns may lead to inadequate infrastructure development, while unreliable environmental impact assessments can result in policies that fail to adequately protect ecosystems or mitigate pollution.

The role of reproducibility and transparency in transportation research is also critical in the context of emerging technologies and trends. For example, with the advent of autonomous vehicles, smart cities, and increasingly interconnected transportation networks, the complexity of transportation systems is escalating. Research in these areas requires rigorous validation to ensure that the systems are safe, reliable, and effective. Reproducibility ensures that the findings are robust and reliable, forming a sound basis for the development and implementation of these technologies.

Given the importance of reproducibility in transportation research, the objective of this article is to provide an introduction to best practices in reproducibility that are relevant to the transportation engineering research community, to discuss issues and barriers to reproducibility, and to describe methods for addressing the issues and barriers. While the best practices represent an ideal and something that should be a goal, it is recognized that the current culture in transportation research is significantly different from other scientific fields that developed these best practices. It will take time to evolve the transportation research culture including changes in attitudes toward openness and transparency, willingness to share raw data, and full methodological details; new skill sets will need to be developed; new infrastructure for storing and disseminating reproducibility data and other materials; changes in the roles and responsibilities of research panels and organizations funding research; and changes in resources provided to support the research (which should result in greater returns on the research investments) this paper serves as a starting point to develop a long-term vision and plan, as well as start discussions aimed at creating a credibility revolution in transportation research.

2. Best Practices for Reproducibility

Science is progressing towards a culture that embraces transparency and openness, fueled partly by technological advancements that facilitate data sharing [22]. To establish effective reproducibility practices in transportation research, it is es-

[1] This is a form of a logical fallacy similar to "Appeal to Authority".

sential to follow established best practices that promote transparency, accessibility, and collaboration. This section starts by discussing openness and transparency (key concepts for reproducibility practices), then discusses several key best practices for reproducibility in transportation engineering research, highlighting common methods and techniques, as well as the associated benefits. Although many of these practices are often considered an added demand in terms of time and resources, they are essential for the progression of science and research. The need for additional resources is discussed later as one of the potential barriers to reproducibility, and potential solutions are provided.

Each of the best practices discussed in this and the following sections fits into an overall framework for reproducibility. A general process for a reproducibility framework is shown in **Figure 1**. As shown, there are considerations for data sharing (including the collection, cleaning, and processing of the data), the methods, and the papers and reports that are published on the research. Each of the topics and considerations in this flowchart is discussed in this paper starting with the discussion on openness and transparency (which is one of the fundamental goals of reproducibility).

The general concept for the overall framework shown in **Figure 1** recognizes that not all data or projects can share the raw data (or in some cases, any data) publicly. It also recognizes other potential barriers to full openness and transparency. The overall framework provides suggestions for providing the maximum amount of transparency possible while balancing potential limitations and legal requirements. This includes potentially providing modified data, using the project panel or other external researchers to validate the data, data processing, and results, providing full documentation and details of code and methods (including assumptions made, models estimated, etc.), publishing open access whenever possible, providing supplementary materials and appendices, publishing on permanent servers, etc.

2.1. Emphasis on Openness and Transparency

Openness and transparency in research activities lay the groundwork for improving reproducibility. Openness refers to the willingness to share all details and evidence. Transparency primarily refers to disclosing all elements involved in the research process [23]. Transparency aids in ensuring that all processes and decisions involved in the research are made evident, allowing others to scrutinize and validate the research work undertaken [24] and make inferences about the likelihood of applicability of research results to local conditions or contexts. For instance, the openness of methodology is an essential aspect of transparency that can significantly contribute to reproducibility. The complete disclosure of methods, including all preprocessing and analytical steps, is necessary for any researcher attempting to replicate the original study [25]. The disclosure should not be limited to mere descriptions but should include the minutiae of all experimental settings, analytical assumptions, operational definitions, control variables, etc. The algorithms and software used in the research process should also be fully detailed, specifying the software packages, versions, and even the parameters used [26]. Therefore, the exhaustive description of these elements can enable the reproduction of the results under similar conditions, augmenting the scientific integrity of the research.

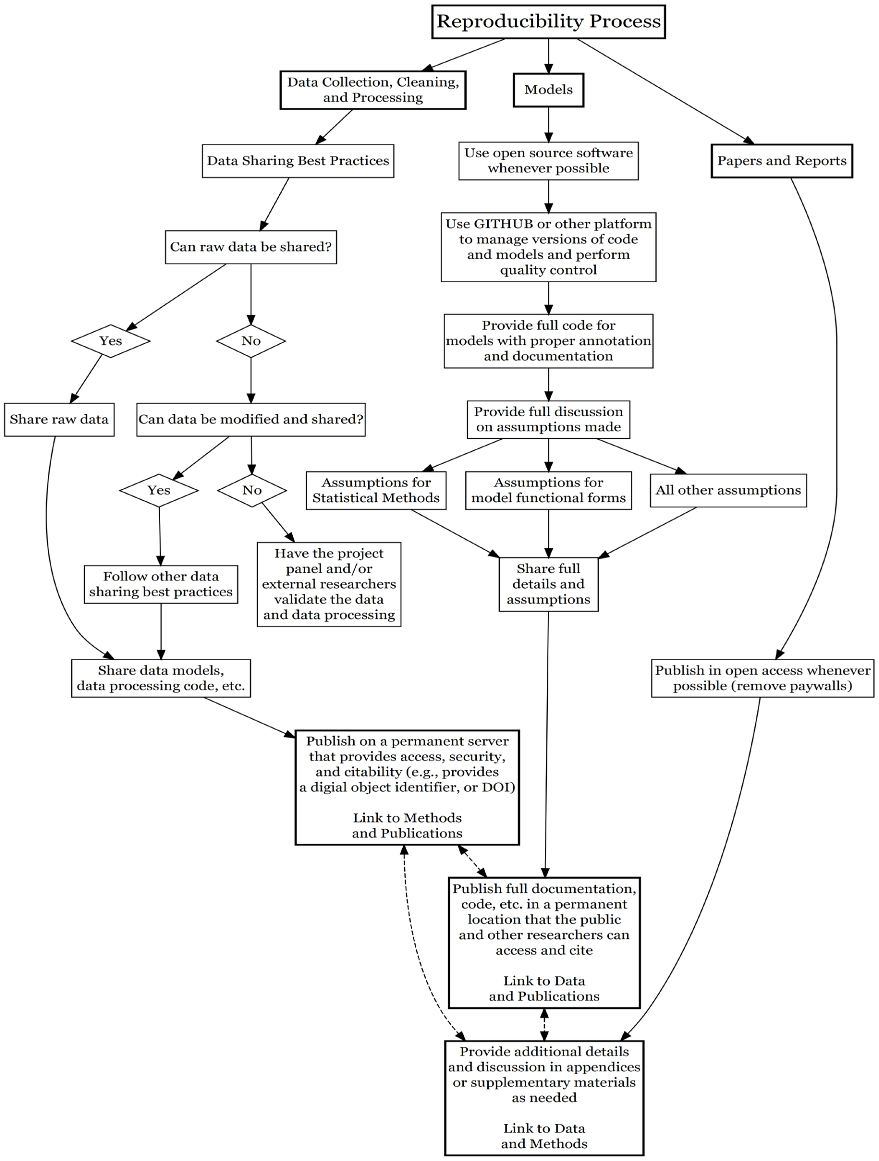
Reproducibility Process
Data Collection, Cleaning, and Processing
Models
Papers and Reports
Data Sharing Best Practices
Use open source software whenever possible
Can raw data be shared?
Use GITHUB or other platform to manage versions of code and models and perform quality control
Yes
No
Provide full code for models with proper annotation and documentation
Share raw data
Can data be modified and shared?
Provide full discussion on assumptions made
Yes
No
Assumptions for Statistical Methods
Assumptions for model functional forms
All other assumptions
Follow other data sharing best practices
Have the project panel and/or external researchers validate the data and data processing
Share full details and assumptions
Share data models, data processing code, etc.
Publish in open access whenever possible (remove paywalls)
Publish on a permanent server that provides access, security, and citability (e.g., provides a digial object identifier, or DOI)
Link to Methods and Publications
Publish full documentation, code, etc. in a permanent location that the public and other researchers can access and cite
Link to Data and Publications
Provide additional details and discussion in appendices or supplementary materials as needed
Link to Data and Methods

Figure 1. Overall process for reproducibility best practices.

2.2. Ensuring Access to Raw Data and Related Information

Making the raw data freely accessible to the wider research community, as appropriate, based on data ownership and other restrictions, serves as a cornerstone of reproducible research. This includes all raw data whenever possible, including data obtained from public agencies, data collected by the research team, and other data sources used. The data should be easily accessible and stored in a permanent location. Simply mentioning where or how the raw data were obtained is inadequate in all cases, except the rare case. However, some research may be able to reference the original source, provided that it is easily obtained by other researchers and the public, and the specific queries and specifications used to process the data obtained are stored and easily reapplied to obtain the same original dataset to the researchers, and doing so does not create a barrier to reasonable access of the raw data. Additionally, stating that the authors are willing to share the data with anyone who requests them directly is also an inadequate practice. As computers are replaced, data files are moved, jobs are changed, students and researchers leave, or any number of potential events occur, data are lost, corrupted, and forgotten.

Anyone who works in research learns quickly that it is not always possible to provide all raw data. While sharing the raw data should be considered standard practice, there are specific reasons that the raw data should not or cannot be shared. Thus, the practice of open sharing of the raw data should only be deviated from when ethically, contractually (e.g., some proprietary data) sources, or legally required. When any of these is the case, alternative approaches for providing data may be used. Relevant methods and approaches for this will be discussed later. If the researchers do not own the data (e.g., the data collected and processed is often owned by the funding agency), the researchers should work with the funding agency to ensure that the data are accessible to the broader research community. In cases where this is not possible, an alternative is to provide the data and code to an independent research group which then reviews, processes, and validates that the study was reproduced and the evidence supports the claims.

Accessibility of the raw data allows other researchers to validate the original results, reducing the potential for biases or errors [27]. Data sharing has the added benefit of promoting novel investigations and can provide opportunities for researchers to derive new insights through reanalysis [28]. It also supports the development of future researchers in that it can serve as a baseline or reference for typical approaches to dataset development and linkage.

However, data accessibility should be accompanied by appropriate metadata and data dictionaries to enable accurate interpretation. This includes comprehensive descriptions of how the data was collected, what each data point represents, and the measurement units [29]. It is also important to provide metadata, data models (which show how various datasets and tables are related and linked, as well as other potential information), annotated processing code, and summaries of the data. These aspects can significantly reduce ambiguity, increase comprehension, and enhance the ease of both reproducibility and replicability. In addition, it enables and facilitates the implementation of the research in that it provides the necessary technical details for data elements to be collected or derived by the implementing agency.

It should be noted that these best practices go beyond simply providing the final dataset used in the reported results or in a publication such as the Data in Brief journal, although this would often be more open and transparent than is currently common practice within the transportation research community.

One of the key requirements for ensuring reproducibility is the safe storage, wide dissemination, and long-term preservation of raw data, along with associated metadata, data models, processing codes, and data summaries. Central to this practice is using data repositories—dedicated storage locations where data is maintained and made accessible to other researchers, as appropriate.

There are several public repositories available to researchers across different disciplines. For instance, there are numerous existing options such as figshare [30], Mendeley Data [31], Dryad Digital Repository [32], Harvard Dataverse [33], Open Science Framework [34], Zenodo [35], and Science Data Bank [36]. McMaster University has an extensive list of open data repositories on their webpage [37]. When choosing a repository, researchers should consider factors like the reputation and longevity of the repository, whether it provides a digital object identifier (DOI) for datasets, and its data licensing policies [38]. An established institutional database that meets the standards of longevity, providing a DOI, data licensing, and security could also be used.

Data repositories typically employ redundancy measures to prevent data loss and corruption over time, such as multiple copies stored in different geographical locations. In addition, many repositories follow the FAIR principles (Findability, Accessibility, Interoperability, and Reusability), ensuring that datasets are safely stored and easily discoverable and accessible for reuse [39].

Adherence to open data formats (such as comma-separated values (CSV) or tab-separated values (TSV) files for tabular data), and avoidance of proprietary data formats, is a common practice to ensure data remains interpretable and usable in the long term as these formats are less likely to become obsolete. Researchers should strive to provide comprehensive metadata with the data, which enables a clear understanding of the data by others and increases the chances of its reuse [40].

2.3. Detailed Disclosure of Research Methodologies

Reproducibility in transportation research is heavily reliant on detailed and comprehensive disclosure of research methodologies. A thoroughly elucidated methodology goes beyond a cursory overview or summary; it should serve as a blueprint that guides other researchers, enabling them to follow the exact steps taken and thus replicate the study [41].

This entails providing exhaustive descriptions and justifications of the experimental designs and statistical methodologies employed, step-by-step narratives of data collection protocols, and clear instructions on how the collected data were processed and cleaned. Methodological transparency is indispensable, especially in cases where sophisticated or complex analyses were carried out. Often, without proper disclosure of these intricate methods, the risk of introducing errors or bias significantly increases, thus undermining the reliability of the study results [42].

Furthermore, the importance of disclosing "failed" experiments, unexpected results, and negative findings cannot be overstated. Such transparency provides a more holistic picture of the research process, aids in identifying potential pitfalls, and ensures that the same mistakes are not perpetuated in future research endeavors [43].

For enhancing comprehension, using flow diagrams, pseudocode, graphical tools, or detailed algorithmic descriptions can be particularly beneficial. These visual representations are instrumental in elucidating complex procedures or data flows and make it easier for other researchers to understand, reproduce, and extend the study methodologies [44].

When computation or scripting plays a role in the research process, it's crucial to provide not only the raw scripts but also detailed annotations for the code. Annotated code is a key element of reproducible computational research as it allows other researchers to understand, replicate, and extend the computational methods. This can be achieved through effective commenting within the code, documenting changes and rationales, and explaining the role and function of each step and variable [45]. While some may argue that annotating code will add to research project costs, it is widely recognized in the programming industry that annotating code is the best practice and expectation for any coder. Often, when writing a paper for publication, providing fully detailed methodology processes, assumptions, flow diagrams, etc., distracts from the key messages and story that the paper is attempting to convey. Thus, the full and detailed disclosure of the research methodologies will often need to be provided as supplementary material, appendices, links to repositories, or other methods that provide the full details in a reasonably accessible manner. This may also be the case for research reports submitted to funding agencies. However, this should never be used as an excuse for not providing full and detailed disclosure of the methodologies. This is even more important when using public funding as this information forms part of the basic research product because it is key to understanding potential implementation strategies and the limitations of the research.

2.4. Utilization of Open-Source Software

Utilizing open-source software is a fundamental practice in reproducibility efforts. Open-source software, due to its transparent nature, fosters a culture of collaboration, innovation, and reproducibility. The open-source ecosystem is characterized by a unique blend of attributes, such as cost-free availability, extensive user-driven documentation, dynamic communities, and frequent updates that address security and performance issues, making it an ideal choice for researchers seeking reproducible and robust results [46].

Open-source software platforms like R and Python have been widely adopted across scientific communities, including transportation research, due to their extensive libraries, powerful analytical capabilities, and the reproducibility they foster. These platforms allow researchers to share their code freely, encouraging review, validation, and improvement [47].

Although proprietary software might sometimes be necessary for specific tasks, it introduces challenges to reproducibility due to associated costs, accessibility issues, and closed-source nature. In these cases, researchers are recommended to share detailed algorithms or pseudocodes that delineate the proprietary software's procedures. This would help other researchers recreate the same procedures using open-source alternatives [48].

Sometimes, open-source software access to public agencies may be restricted due to security concerns or protocols. In those instances, investments in proprietary software products that are widely used and supported in academia and research should be considered as it increases the likelihood of usability for reproducibility.

2.5. Version Control and Documentation

Employing version control systems (VCS) is critical to ensure reproducibility. VCS, like Git, offers a mechanism to track and record changes to a project's data, scripts, and documentation. This allows researchers to compare and contrast different project versions, isolate the impact of specific changes, and revert to previous versions if necessary. This systematic recording of changes enhances the reliability and reproducibility of the research [49].

Documentation plays a pivotal role in VCS. Every change, addition, or deletion to the project should be accompanied by thorough documentation clearly stating what was modified, why the change was necessary, and who was responsible for it. This improves transparency, promotes a greater understanding of the evolution of the project, and facilitates collaboration between team members [50].

The use of platforms like GitHub and Bitbucket is highly recommended as these can integrate VCS with cloud-based storage, which further enhances the accessibility of the project's files and facilitates collaboration, both internally and externally. Detailed documentation combined with version control enables other researchers to follow the entire research process, ensuring a seamless replication of the study [51].

2.6. Institutional Support for Reproducibility Standards

Institutional support is crucial in encouraging and facilitating adherence to reproducibility standards. The institutions can assist in providing researchers with access to tools and resources essential for reproducible research, including data repositories, open-source software, and version control systems. Moreover, institutions can also help by providing training and workshops to foster a culture of reproducibility. The institutional endorsement, expectation, and promotion of reproducibility standards underscore the importance of these practices, encouraging researchers to abide by them [23].

In addition to resources and training, institutions can play an active role in promoting open science policies and advocating for the wide dissemination and accessibility of research findings. They can also establish clear guidelines and protocols to ensure the integrity and reliability of research, thus fostering a research environment that is conducive to reproducibility. Encouraging researchers to publish their data, methodology, and code in public repositories can be part of such guidelines [20]. In supporting open science, institutions may also choose to create open repositories that meet established open data standards and the institution's needs as a whole.

The institutional policies and practices can also incorporate measures to recognize and reward researchers who follow reproducibility practices. This can take the form of incentives, awards, or even consideration in promotion and tenure decisions, which can motivate researchers to adopt reproducibility standards in their work [52].

2.7. Building Collaborations and Partnerships

Collaborations and partnerships significantly bolster the culture of reproducibility. Engaging with diverse teams of researchers facilitates a broader range of skills and perspectives and introduces an additional layer of review and verification of the research process. Collaborations can occur at various levels—within a department, across departments, between institutions, or even internationally. These collaborative relationships provide an ideal platform for knowledge exchange, innovation, and mutual learning, thus enhancing the reliability and reproducibility of the research [53].

Such partnerships also extend to collaborations with the wider community, including stakeholders, policymakers, industry partners, and the public. These partnerships can provide valuable input into the research process, foster a sense of ownership and engagement, and contribute to the practical applicability and societal impact of the research. Engaging with non-academic partners can also help establish data-sharing agreements or access proprietary data while still adhering to reproducibility standards [54].

When collaborations involve data sharing or co-creation, it is crucial to establish clear agreements regarding data management, ownership, and access. These agreements, often formalized as data management plans, ensure that all parties clearly understand the responsibilities and expectations related to the data, which is critical for maintaining the integrity and reproducibility of the research [55].

3. Barriers to Reproducibility in Transportation Research and Relevant Solutions

The previous section provided an overview of best practices for reproducibility from other scientific fields that are appropriate for transportation research. However, it is well recognized that numerous barriers exist to the widespread adoption and implementation of these practices. This section discussed some of these key issues and barriers as well as potential solutions. Many, if not all, of these barriers have been faced by other scientific fields. While the solutions were not derived and accepted overnight, if fields including medical, public health, psychology, etc., can find approaches to overcome their barriers to reproducibility, then barriers to reproducibility in transportation research should be considered solvable.

3.1. Cultural and Attitudinal Barriers

A key barrier to accepting and implementing reproducibility best practices is based on the existing research culture, attitudes surrounding open science and transparency, and misperceptions related to reproducibility, data sharing, etc. While the "trust us" attitude is one of these problematic attitudes, there are several other attitudes and misperceptions that can lead to significant barriers to reproducibility. Some of these are simply a lack of understanding of reproducibility, reproducibility methods, and the benefits of openness and transparency. Some best practices and benefits related to them have already been discussed. However, some researchers may raise concerns that providing full details, code, and data could potentially act as evidence against their particular research and claims.

The potential that providing full methodological details, raw data, annotated code for data processing and analysis, and any other relevant details for reproducibility provides opportunities for others to find errors or other mistakes, as well as opportunities to challenge the results and claims, is a key concern for many researchers. Thus, this is a concern that openness and transparency may act as a form of self-incrimination. It is important to remember and accept that no one is perfect and that everyone makes mistakes. Thus, when someone finds an error, alternative result, or interpretation, this is not a negative reflection on the original researchers. Instead, this should be expected to occur and be accepted (although it may take time for the research culture to accept and adapt to this paradigm). The openness and sharing are evidence that the original researchers were not hiding anything and were not involved in academic misconduct. It could be argued that following reproducibility best practices, such as utilizing established methods to overcome practical, legal, and ethical barriers, is a form of academic misconduct. This is based on failing to provide supporting evidence for the scientific claims. This failure is, in essence, a refusal to follow scientific standards based on fear or the "trust us" attitude.

Institutional and policy reforms might be required to address the "trust us" attitude and promote reproducibility. Researchers should be encouraged and incentivized to share their raw data, analysis code, and detailed methodologies. Journals and funding agencies can play a crucial role by making data sharing and openness a requirement for publication or funding [20]. Moreover, education and training in data management and reproducible research practices should be incorporated into the curriculum of researchers at all levels [45].

Another concern is that some researchers may not collect and use the optimal data or approaches for a research project if they know they must share the raw data, full methodological details, etc. In some cases, this could be related to proprietary data. In other cases, it may be due to the time it would take to prepare the full reproducibility files and information (*i.e.*, wanting to minimize the time and effort required). This may also be due to a lack of comfort with methods for sharing sensitive data. Solutions to these issues will be discussed in the next section.

A further concern that is rarely stated is hesitancy to share data and fully detailed methods, as that results in not having a competitive advantage in future related research. When the data are not shared (or full details/code for new methods), the original researchers may use the data from the earlier project as an advantage over other researchers when submitting proposals on the related project for funding. In today's competitive research environment, such an approach for research using public funding raises ethical concerns as public funding was used to process the data and carry out the analysis.

When the research culture has evolved to be strongly opposed to openness and transparency with reproducibility best practices, this is a symptom that a credibility revolution is necessary in that field. This is possible through education, incentives, requirements by funding agencies and publishers, etc. Further discussion on solutions is provided below.

3.2. Sensitive Data

In some research, sensitive data are used. This may include protected data, such as experiments or other data collection, that requires Institutional Review Board (IRB) approval. It may also involve data that have potentially sensitive or personally identifying information (PII). When sensitive data are used, this presents a barrier to data sharing. However, this is not a barrier that is unique to transportation research. Other fields, such as medicine, psychology, and others, have at least as many sensitive data issues and have developed methods and approaches to overcome this as a barrier to reproducibility. Some common methods for handling sensitive data include:

1) Data Masking or Pseudonymization
2) Data Swapping or Permutation
3) Noise Addition
4) K-Anonymity
5) L-Diversity
6) Generalization and Suppression
7) Differential Privacy

These methods do not provide the complete raw data with the full details, yet often provide it in a form as close to the raw data as possible while overcoming the issues with the sensitive data elements. Details for each of these methods are discussed below.

3.2.1. Data Masking or Pseudonymization

Data masking, known as pseudonymization, is a method for handling data with PII. In essence, it changes or hides the values of variables in the dataset, replaces the values, or masks part or all of the sensitive data to maintain the statistical properties of the data while protecting individual identities. It also improves data security on data servers [56].

Data masking can be accomplished using many approaches. Some common methods include [57] [58]:

1) Shuffling
2) Scrambling
3) Substitution
4) Variance
5) Masking Out
6) Nullifying
7) Encryption

Details on these methods and examples of how each is applied are readily available online. Thus, the details for implementing these methods are not provided here.

3.2.2. Noise Addition

Noise addition is a method that adds random noise to the data, which helps obscure individual data points [59]. This method is often used in differential Privacy, a mathematical technique that ensures the Privacy of individuals in a dataset while allowing for statistical analysis. Multiple methods are available for applying noise addition. A discussion on methods for optimal noise addition for data privacy was developed in an IEEE paper [60]. A method for using decision trees to ensure data privacy while maintaining the ability to apply data mining and analysis is also available [61].

3.2.3. K-Anonymity

K-Anonymity is a method that ensures that each person (or other entity, as relevant) in the dataset is indistinguishable from at least k-1 individuals, even if someone knows all of the quasi-identifiers. Quasi-identifiers are the pieces of information that could potentially identify an individual when linked with other data. This is considered the most common method for handling sensitive data for sharing in reproducibility efforts [62].

Several established K-Anonymity methods exist, including [62] [63]: microaggregation, bucketization, suppression, and generalization. Within these methods, there are multiple approaches each of which has been implemented in various software packages, including R [64] and Python [65].

3.2.4. L-Diversity

L-diversity is a privacy principle for protecting sensitive data similar to k-anonymity but with additional security. It was introduced to deal with the shortcomings of k-anonymity, which could still leak information when sensitive attributes within the k-anonymous groups are homogeneous [66]. The l-diversity principle stipulates that there are at least "l" distinct values for each sensitive attribute in each group of records sharing a set of quasi-identifiers. This ensures that even if an attacker knows all quasi-identifiers of an individual, they cannot determine the individual's sensitive attributes with certainty [66].

3.2.5. Generalization and Suppression

The concept for generalization and suppression is to reduce the granularity of data to protect individuals' Privacy. For example, group them into brackets instead of providing exact ages (e.g., 20 - 30, 30 - 40). If some data is too identifiable, it may be suppressed. This leads to data loss compared to the raw data. However, if other methods that do not result in loss of the information that directly impacts the statistical properties of the data cannot adequately address the sensitivity of the data, this can be used. While this is an accepted method, it limits the data's openness and transparency. Thus, when researchers apply generalization and suppression, they should be careful not to aggregate more than is necessary. Additionally, they should provide full details for how they determined the aggregation values used, how sensitive the results are to different ranges or values used, etc.

3.2.6. Differential Privacy

Differential Privacy is a mathematical framework that quantifies privacy leakage in a dataset. It provides a guaranteed measure of Privacy by injecting noise into the results of data queries to protect the Privacy of individuals in the dataset [67]. Differential Privacy is particularly relevant in scenarios where multiple queries on a dataset may lead to an increased risk of re-identification. This concept has gained popularity recently and is used by tech companies like Apple and Google for data analysis while preserving user privacy [68].

3.3. Lack of Access to Raw Data

Often the researchers themselves either do not have access to the raw data or are required to agree not to share the data provided to them in its raw form. For example, the Highway Safety Information System (HSIS) is a source of crash and related data funded by the Federal Highway Administration (FHWA) that is used in many transportation safety research projects. To track the use of the HSIS data, instructions not to share the data are given when requests for HSIS data are fulfilled. This becomes an easy argument against the practicality of providing raw data and data processing information, data models, etc.

When researchers are not allowed to share the raw data, there are several options to ensure openness and transparency. They can provide the detailed request they made (so that others can make the same request). Then, using the details for the data processing, data models, etc., other researchers could, in theory, reproduce the same results. It is worth noting that failure to reproduce the results in these cases could be due to differences in the data delivered from the requests, even if they were identical. In the case of HSIS, the codes used to process many of the previous requests are stored, so this should not be an issue in the particular case of HSIS data.

When the raw data are not available, the researchers do not have permission to share, and there are no reasonable solutions, the researchers should still provide the detailed data models[2], full methods, and annotated data processing code as

[2] A good practice, with or without providing the data, is to provide Entity-Relationship Diagrams using open-source software, such as R [69].

well as well-defined data as close to the raw form as possible. While this is not ideal regarding reproducibility and providing supporting evidence, it shows openness on the part of the researchers and allows for a review of the work in as much detail as possible.

3.4. Data Contracts, Proprietary Data, and Data Ownership

The issues of data contracts, proprietary data, and data ownership are often intertwined. Some datasets used in transportation research may be subject to data contracts or licensing agreements restricting data sharing. This is common with proprietary data collected by private entities that consider their data as a competitive asset [70].

Furthermore, the question of who owns the data may create barriers to reproducibility. In some cases, the data are collected and owned by public entities; in others, the data may be owned by private entities or individuals. Determining who has the right to share or restrict access to the data can be a complex legal issue [71] [72].

Researchers can address these barriers by using data use agreements and non-disclosure agreements, where appropriate, to provide access to data for other researchers in a way that complies with contractual and legal obligations [70] [73]. It is also critical for researchers to provide a comprehensive description of the datasets used, including any restrictions on their use or sharing, to ensure that other researchers and the funding agency understand the limitations of reproducing the research.

Given that limitations and barriers to sharing raw data are an inherent part of research practice, several methods for addressing them have been discussed. **Figure 2** provides a flowchart for determining appropriate methods for data sharing, accounting for potential limitations and barriers to data sharing. As shown, the raw data should be shared if possible. If this is not possible, there are methods for anonymizing the data and dealing with other sensitivity issues. This may sometimes allow the data to be shared in a form close to the raw data. If this is the case, the methods used to change the data should be clearly specified and discussed. If the raw data can be queried or requested by other researchers or the public, simply providing the details of the query may be adequate. For each of these options, data models, methods, and processing codes should also be provided. When none of these are possible, having the research panel or an independent group review and validate the data and results can be used as an approach to providing evidence that the research is reproducible.

3.5. Insufficient Methodological Detail

Insufficient methodological detail is a key barrier to reproducibility in many scientific fields, not just transportation research. If researchers do not provide enough detail about their methodology, including the design of their experiments, the statistical methods used, and the analytical software used. In that case, it can be nearly impossible for other researchers to reproduce their work [20].

To address this issue, researchers should aim to provide as much detail as possible about their methods. This includes describing the techniques and algorithms used and providing information about any parameters or settings that could affect the results [74]. Researchers should also consider publishing their protocols and codes. Suppose in repositories, which can facilitate the sharing of detailed methodological information [74]. In the cases of Machine Learning (ML) and Deep Learning (DL), the details of hyperparameters and methods used (e.g. learning rate, loss function, etc.), the random seed, software versions, hardware specifications, and all other relevant values used should also be provided to ensure full reproducibility [75] [76].

3.6. Lack of Code Sharing and Use of Proprietary Software

Lack of code sharing is a significant barrier to reproducibility in transportation research. When researchers do not share the code they used for their analyses, it makes it more difficult for others to reproduce their results [26]. This problem is compounded when researchers use proprietary software that may not be accessible to all other researchers [26].

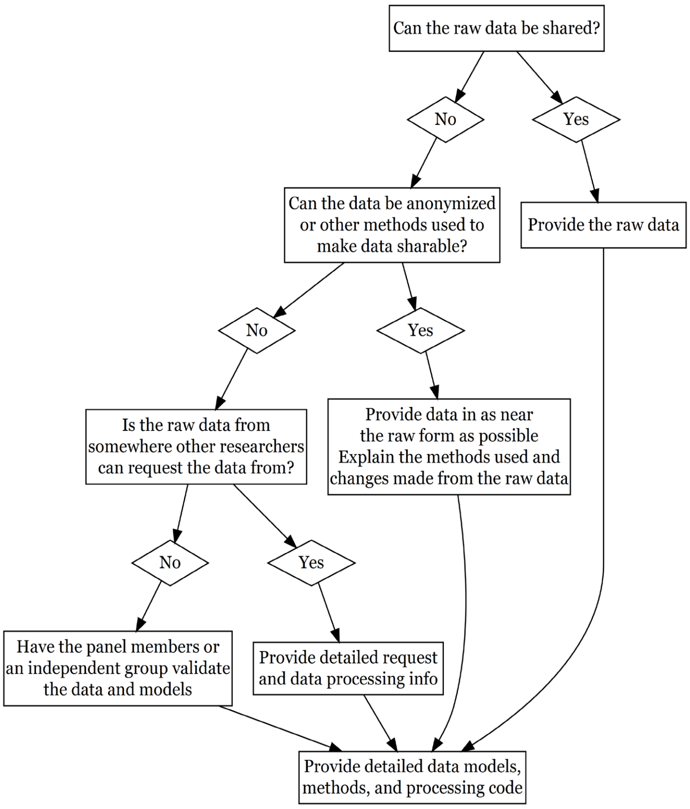
Can the raw data be shared?
No
Yes
Can the data be anonymized or other methods used to make data sharable?
Provide the raw data
No
Yes
Is the raw data from somewhere other researchers can request the data from?
Provide data in as near the raw form as possible Explain the methods used and changes made from the raw data
No
Yes
Have the panel members or an independent group validate the data and models
Provide detailed request and data processing info
Provide detailed data models, methods, and processing code

Figure 2. Flowchart for determining the best methods for data sharing, considering potential barriers and limitations.

To overcome this barrier, researchers should aim to share their code as much as possible. This can be achieved using open-source platforms like GitHub, which allows code versioning and collaborative work [77]. Furthermore, to increase the accessibility and reproducibility of their work, researchers should consider using open-source software instead of, or in addition to, proprietary software when possible [78] or software used widely in academia and by research institutions.

3.7. Limited Validation

Limited validation refers to the inadequacy of data or model validation in a study. Researchers might focus heavily on developing their model or analysis but may not spend enough time validating the results. This is a significant issue because, without adequate validation, it's impossible to know whether the results are reliable or if they merely represent a spurious correlation [79].

To ensure reproducibility, researchers should always include rigorous validation as part of their research process. This might include cross-validation, out-of-sample testing, or other validation techniques. The validation process and its results should be thoroughly documented and reported in the research output [23].

3.8. Added Time and Resource Burdens

Ensuring that research is reproducible may add significant time and resource burdens to the research process. Preparing data and code for sharing, writing detailed methodological descriptions, and performing extensive validation can all take considerable time and effort [20]. It is, however, debatable whether this step is optional because staff overturn is accelerating and because it is unreasonable to expect any researcher to remember all the details associated with a research project after the fact.

While this added burden may seem daunting, it's important to remember that reproducibility is a fundamental aspect of scientific research. To manage this challenge, researchers should consider reproducibility as an integral part of the research process rather than an added task. They could also use tools and services that facilitate reproducible research, such as open-source software, code repositories, and data-sharing platforms [23]. Implementing best practices for documenting the full research process, including annotated code, data models, etc., also improves the ability to quickly and easily check the work of graduate students and junior engineers/researchers. Thus, in some cases, it may require some initial training to be provided, but it can potentially reduce the time requirements for the senior researchers to check and review the work.

Funding agencies can address this concern of added time and resource requirements by considering the added time and resources required for proper reproducibility practices when allocating funding for research—which may require a shift in expectations based on past experiences. However, this can and should be tied to requirements for reproducibility, ownership of data, and data sharing, which should be explicitly agreed on in the contract. Independent validation of the data and analysis should be included in the project when the data or other portions of the project cannot or will not be made fully available. Additionally, the panel members supervising the research should be trained on the minimum standards expected for reproducibility and should work as partners with the research team to ensure that reproducibility best practices are followed.

3.9. Skill and Knowledge Gap

Finally, a significant barrier to reproducibility is the skill and knowledge gap in both data analysis and reproducible research practices. Not all researchers have the skills to perform sophisticated data analyses, prepare data and code for sharing, or understand the intricacies of validation [80].

Addressing this gap requires both training and cultural change within the research community. This could include formal training in data analysis and reproducibility and promoting a culture that values and rewards reproducible research. Furthermore, senior researchers and supervisors should lead in implementing and promoting good practices in data management, analysis, and sharing [81].

4. Conclusions and Recommendations

The issues affecting reproducibility in transportation research are multifaceted but not impossible. Other scientific fields have shown that, while perceived and real barriers to reproducibility may be daunting, they can be overcome. To enhance reproducibility in transportation research, it is crucial to address existing culture and attitudes as well as data sensitivity, lack of access to raw data, contracts, and data ownership issues, insufficient methodological detail, lack of code sharing, limited validation, additional time and research burden, and skill and knowledge gaps.

Promoting a culture of openness and transparency in transportation research is key to addressing many of these issues. This includes sharing datasets, methodologies, and code, with necessary safeguards for sensitive data. Encouraging independent validation of research findings, maintaining detailed documentation of research processes, and investing in reproducibility from the beginning of the research process can also enhance reproducibility.

Addressing the skill and knowledge gap requires concerted efforts from educational and research institutions. Introducing training programs and courses on reproducible research can equip researchers with the necessary skills. Training for researchers and faculty may also be necessary. Training for oversight panels and funding agencies may also be required.

Furthermore, funding agencies, institutions, and academic journals must support and promote reproducibility. They can do this by introducing policies that encourage or require data and code sharing, providing guidelines for detailed methodological reporting, creating repositories and other resources that support reproducibility best practices, and acknowledging the value of replication studies. Reproducibility could be a criterion for new awards; journals could encourage articles specifically devoted to the reproducibility of other studies; funding agencies could use a record of reproducibility or follow reproducibility best practices as proposal evaluation criteria and ratings of research quality (such as the star quality ratings for Crash Modification Factors [82]) could include reproducibility criteria.

Overall, while the path toward a research community that fully embraces and practices reproducibility may be challenging, advancing the science of transportation research is necessary. The change will not be immediate, nor should it be expected to be. The evolution towards this goal will be challenging and will undoubtedly face many challenges, yet the many brilliant minds in the transportation research community have the potential to solve these issues as they arise. By working towards this goal, researchers can enhance the reliability and credibility of their findings, fostering trust and collaboration within the scientific community. This will benefit funding agencies while supporting and, perhaps, accelerating the implementation of research findings.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Socol, Y., Shaki, Y.Y. and Yanovskiy, M. (2019) Interests, Bias, and Consensus in Science and Regulation. *Dose-Response*, **17**. https://doi.org/10.1177/1559325819853669

[2] Open Science Collaboration (2015) Estimating the Reproducibility of Psychological Science. *Science*, **349**, aac4716. https://doi.org/10.1126/science.aac4716

[3] Gundersen, O.E. (2021) The Fundamental Principles of Reproducibility. *Philosophical Transactions of the Royal Society A*: *Mathematical*, *Physical and Engineering Sciences*, **379**, Article ID: 20200210. https://doi.org/10.1098/rsta.2020.0210

[4] Nissen, S.B., Magidson, T., Gross, K. and Bergstrom, C.T. (2016) Publication Bias and the Canonization of False Facts. *eLife*, **5**, e21451. https://doi.org/10.7554/elife.21451

[5] Landis, S.C., Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., *et al*. (2012) A Call for Transparent Reporting to Optimize the Predictive Value of Preclinical Research. *Nature*, **490**, 187-191. https://doi.org/10.1038/nature11556

[6] Pusztai, L., Hatzis, C. and Andre, F. (2013) Reproducibility of Research and Preclinical Validation: Problems and Solutions. *Nature Reviews Clinical Oncology*, **10**, 720-724. https://doi.org/10.1038/nrclinonc.2013.171

[7] Rooney, A.A., Cooper, G.S., Jahnke, G.D., Lam, J., Morgan, R.L., Boyles, A.L., *et al*. (2016) How Credible Are the Study Results? Evaluating and Applying Internal Validity Tools to Literature-Based Assessments of Environmental Health Hazards. *Environment International*, **92**, 617-629. https://doi.org/10.1016/j.envint.2016.01.005

[8] McNutt, M. (2014) Journals Unite for Reproducibility. *Science*, **346**, 679-679. https://doi.org/10.1126/science.aaa1724

[9] Vazire, S. (2018) Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science*, **13**, 411-417. https://doi.org/10.1177/1745691617751884

[10] Engineering National Academies of Sciences (2019) Reproducibility and Replicability in Science. National Academies Press.

[11] Grant, S., Wendt, K.E., Leadbeater, B.J., Supplee, L.H., Mayo-Wilson, E., Gardner, F., *et al*. (2022) Transparent, Open, and Reproducible Prevention Science. *Prevention Science*, **23**, 701-722. https://doi.org/10.1007/s11121-022-01336-w

[12] Resnik, D.B. and Shamoo, A.E. (2016) Reproducibility and Research Integrity. *Accountability in Research*, **24**, 116-123.

https://doi.org/10.1080/08989621.2016.1257387

[13] Loeb, A. (2021) To Qualify as "Scientific", Evidence Has to Be Reproducible. https://www.scientificamerican.com/article/to-qualify-as-scientific-evidence-has-to-be-reproducible/

[14] Stark, P.B. (2018) Before Reproducibility Must Come Preproducibility. *Nature*, **557**, 613-613. https://doi.org/10.1038/d41586-018-05256-0

[15] Zheng, Z. (2021) Reasons, Challenges, and Some Tools for Doing Reproducible Transportation Research. *Communications in Transportation Research*, **1**, Article ID: 100004. https://doi.org/10.1016/j.commtr.2021.100004

[16] Diaba-Nuhoho, P. and Amponsah-Offeh, M. (2021) Reproducibility and Research Integrity: The Role of Scientists and Institutions. *BMC Research Notes*, **14**, Article No. 451. https://doi.org/10.1186/s13104-021-05875-3

[17] Begley, C.G. and Ellis, L.M. (2012) Raise Standards for Preclinical Cancer Research. *Nature*, **483**, 531-533. https://doi.org/10.1038/483531a

[18] Popper, K.R. (2002) Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge.

[19] Nielsen, M.A. (2014) Reinventing Discovery: The New Era of Networked Science. Princeton University Press。

[20] Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., *et al.* (2017) A Manifesto for Reproducible Science. *Nature Human Behaviour*, **1**, Article No. 21. https://doi.org/10.1038/s41562-016-0021

[21] Saltelli, A. and Giampietro, M. (2017) What Is Wrong with Evidence Based Policy, and How Can It Be Improved? *Futures*, **91**, 62-71. https://doi.org/10.1016/j.futures.2016.11.012

[22] Woelfle, M., Olliaro, P. and Todd, M.H. (2011) Open Science Is a Research Accelerator. *Nature Chemistry*, **3**, 745-748. https://doi.org/10.1038/nchem.1149

[23] Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., *et al.* (2015) Promoting an Open Research Culture. *Science*, **348**, 1422-1425. https://doi.org/10.1126/science.aab2374

[24] Kidwell, M.C., Lazarević, L.B., Baranski, E., Hardwicke, T.E., Piechowski, S., Falkenberg, L., *et al.* (2016) Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, **14**, e1002456. https://doi.org/10.1371/journal.pbio.1002456

[25] McKiernan, E.C., Bourne, P.E., Brown, C.T., Buck, S., Kenall, A., Lin, J., *et al.* (2016) How Open Science Helps Researchers Succeed. *eLife*, **5**, e16800. https://doi.org/10.7554/elife.16800

[26] Stodden, V., Seiler, J. and Ma, Z. (2018) An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility. *Proceedings of the National Academy of Sciences*, **115**, 2584-2589. https://doi.org/10.1073/pnas.1708290115

[27] Leonelli, S. (2016) Data-Centric Biology. University of Chicago Press. https://doi.org/10.7208/chicago/9780226416502.001.0001

[28] Hardwicke, T.E. and Ioannidis, J.P.A. (2018) Populating the Data Ark: An Attempt to Retrieve, Preserve, and Liberate Data from the Most Highly-Cited Psychology and Psychiatry Articles. *PLOS ONE*, **13**, e0201856. https://doi.org/10.1371/journal.pone.0201856

[29] Peng, R.D. (2011) Reproducible Research in Computational Science. *Science*, **334**, 1226-1227. https://doi.org/10.1126/science.1213847

[30] Figshare (2023) About Figshare. https://knowledge.figshare.com/about

[31] (2023) Mendeley Data. https://data.mendeley.com/

[32] Dryad (2023) Home-Publish and Preserve Your Data. https://datadryad.org/stash

[33] Harvard Dataverse (2020) For Researchers—Harvard Dataverse Support. https://support.dataverse.harvard.edu/researchers

[34] (2023) OSF: Open Science Framework. https://osf.io/

[35] Zenodo (2013) European Organization for Nuclear Research and Open AIRE. https://www.zenodo.org/

[36] ScienceDB (2023) Science Data Bank. https://www.scidb.cn/en/introduction

[37] McMaster University (2023) Open Access Data Repositories. https://mira.mcmaster.ca/research/open-access-data-repositories

[38] Piwowar, H.A., Day, R.S. and Fridsma, D.B. (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE*, **2**, e308. https://doi.org/10.1371/journal.pone.0000308

[39] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., *et al.* (2016) The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, **3**, Article No. 160018. https://doi.org/10.1038/sdata.2016.18

[40] Michener, W.K. (2015) Ten Simple Rules for Creating a Good Data Management Plan. *PLOS Computational Biology*, **11**, e1004525. https://doi.org/10.1371/journal.pcbi.1004525

[41] Simons, D.J. (2014) The Value of Direct Replication. *Perspectives on Psychological Science*, **9**, 76-80. https://doi.org/10.1177/1745691613514755

[42] Baker, M. (2016) 1, 500 Scientists Lift the Lid on Reproducibility. *Nature*, **533**, 452-454. https://doi.org/10.1038/533452a

[43] Matosin, N., Frank, E., Engel, M., Lum, J.S. and Newell, K.A. (2014) Negativity Towards Negative Results: A Discussion of the Disconnect between Scientific Worth and Scientific Culture. *Disease Models & Mechanisms*, **7**, 171-173. https://doi.org/10.1242/dmm.015123

[44] Stodden, V., Guo, P. and Ma, Z. (2013) Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. *PLOS ONE*, **8**, e67111. https://doi.org/10.1371/journal.pone.0067111

[45] Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., *et al.* (2016) Enhancing Reproducibility for Computational Methods. *Science*, **354**, 1240-1241. https://doi.org/10.1126/science.aah6168

[46] Scacchi, W. (2010) The Future of Research in Free/Open Source Software Development. *Proceedings of the FSE/SDP Workshop on Future of Software Engineering Research*, Santa Fe New, 7-8 November 2010, 315-319.

[47] Ince, D.C., Hatton, L. and Graham-Cumming, J. (2012) The Case for Open Computer Programs. *Nature*, **482**, 485-488. https://doi.org/10.1038/nature10836

[48] Sonnenburg, S., Braun, M.L., Ong, C.S., *et al.* (2007) The Need for Open Source Software in Machine Learning. *Journal of Machine Learning Research*, **8**, 2443-2466.

[49] Blischak, J.D., Davenport, E.R. and Wilson, G. (2016) A Quick Introduction to Version Control with Git and Github. *PLOS Computational Biology*, **12**, e1004668. https://doi.org/10.1371/journal.pcbi.1004668

[50] Ram, K. (2013) Git Can Facilitate Greater Reproducibility and Increased Transparency in Science. *Source Code for Biology and Medicine*, **8**, Article No. 7. https://doi.org/10.1186/1751-0473-8-7

[51] Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L. and Teal, T.K. (2017) Good Enough Practices in Scientific Computing. *PLOS Computational Biology*, **13**, e1005510. https://doi.org/10.1371/journal.pcbi.1005510

[52] (2017) Challenges in Irreproducible Research. *Nature*, **546**, 583.

[53] Teixeira da Silva, J.A. (2015) Negative Results: Negative Perceptions Limit Their Potential for Increasing Reproducibility. *Journal of Negative Results in BioMedicine*, **14**, Article No. 12. https://doi.org/10.1186/s12952-015-0033-9

[54] Bos, N., Lodi, S., Meyer, M., *et al.* (2012) The Importance of Systems Thinking to Address Obesity. *Nutrition Reviews*, **75**, 94-106.

[55] Whitmire, A.L., Baldwin, M.K., Desselle, B.C., *et al.* (2016) Neuroscience Data Integration through the Brokering of Standards. *Neuroinformatics*, **14**, 273-287.

[56] Cuzzocrea, A. and Shahriar, H. (2017) Data Masking Techniques for Nosql Database Security: A Systematic Review. 2017 *IEEE International Conference on Big Data* (*Big Data*), Boston, 11-14 December 2017, 4467-4473. https://doi.org/10.1109/bigdata.2017.8258486

[57] Sarada, G., Abitha, N., Manikandan, G. and Sairam, N. (2015) A Few New Approaches for Data Masking. 2015 *International Conference on Circuits*, *Power and Computing Technologies* [*ICCPCT*-2015], Nagercoil, 19-20 March 2015, 1-4. https://doi.org/10.1109/iccpct.2015.7159301

[58] Cobb, M. (2022) What Is Data Masking? Techniques, Types and Best Practices. https://www.techtarget.com/searchsecurity/definition/data-masking

[59] Mivule, K. (2013) Utilizing Noise Addition for Data Privacy, an Overview. https://arxiv.org/pdf/1309.3958

[60] He, J., Cai, L. and Guan, X. (2018) Preserving Data-Privacy with Added Noises: Optimal Estimation and Privacy Analysis. *IEEE Transactions on Information Theory*, **64**, 5677-5690. https://doi.org/10.1109/tit.2018.2842221

[61] Kadampur, M.A. and Somayajulu, D. (2010) A Noise Addition Scheme in Decision Tree for Privacy Preserving Data Mining. http://arxiv.org/abs/1001.3504.

[62] Slijepčević, D., Henzl, M., Klausner, L.D., Dam, T., Kieseberg, P. and Zeppelzauer, M. (2021) *k*-Anonymity in Practice: How Generalisation and Suppression Affect Machine Learning Classifiers. *Computers & Security*, **111**, Article ID: 102488. https://doi.org/10.1016/j.cose.2021.102488

[63] De Capitani di Vimercati, S., Foresti, S., Livraga, G. and Samarati. P. (2023) *k*-Anonymity: From Theory to Applications. *Transactions on Data Privacy*, **16**, 25-49.

[64] Templ, M., Meindl, B., Kowarik, A. and Gussenbauer, J. (2023) SdcMicro: Statistical Disclosure Control Methods for Anonymization of Data and Risk Estimation. R Package Version 5.7.5.

[65] Fujita, T. (2022) Anonypy: Anonymization Library for Python. Protect the Privacy of Individuals. https://github.com/glassonion1/anonypy

[66] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. (2007) L-Diversity: Privacy beyond *k*-Anonymity. *ACM Transactions on Knowledge Discovery from Data*, **1**, 3. https://doi.org/10.1145/1217299.1217302

[67] Dwork, C. (2008) Differential Privacy: A Survey of Results. In: Agrawal, M., Du, D., Duan, Z. and Li, A., Eds., *Theory and Applications of Models of Computation*.

TAMC 2008, Springer, 1-19. https://doi.org/10.1007/978-3-540-79228-4_1

[68] Apple Differential Privacy Team (2017) Learning with Privacy at Scale. https://machinelearning.apple.com/research/learning-with-privacy-at-scale

[69] Wood, J. and Basulto-Elias, G. (2024) ERD-Builder: Entity Relationship Diagrams Builder, 2024. R Package Version 1.0.0. https://CRAN.R-project.org/package=ERDbuilder

[70] Birch, K., Cochrane, D. and Ward, C. (2021) Data as Asset? The Measurement, Governance, and Valuation of Digital Personal Data by Big Tech. *Big Data & Society*, **8**. https://doi.org/10.1177/20539517211017308

[71] Asswad, J. and Marx Gómez, J. (2021) Data Ownership: A Survey. *Information*, **12**, Article 465. https://doi.org/10.3390/info12110465

[72] Padova, Y. (2021) Data Ownership versus Data Sharing: And What about Privacy? *Lex Electronica*, **26**, 38-73. https://heinonline.org/HOL/P?h=hein.journals/lexel26i=33

[73] Bezuidenhout, L.M., Leonelli, S., Kelly, A.H. and Rappert, B. (2017) Beyond the Digital Divide: Towards a Situated Approach to Open Data. *Science and Public Policy*, **44**, 464-475. https://doi.org/10.1093/scipol/scw036

[74] Leek, J.T. and Peng, R.D. (2015) Reproducible Research Can Still Be Wrong: Adopting a Prevention Approach. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 1645-1646. https://doi.org/10.1073/pnas.1421412111

[75] Hartley, M. and Olsson, T.S.G. (2020) Dtoolai: Reproducibility for Deep Learning. *Patterns*, **1**, Article ID: 100073. https://doi.org/10.1016/j.patter.2020.100073

[76] Alahmari, S.S., Goldgof, D.B., Mouton, P.R. and Hall, L.O. (2020) Challenges for the Repeatability of Deep Learning Models. *IEEE Access*, **8**, 211860-211868. https://doi.org/10.1109/access.2020.3039833

[77] Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F.d.V., *et al.* (2016) Ten Simple Rules for Taking Advantage of Git and Github. *PLOS Computational Biology*, **12**, e1004947. https://doi.org/10.1371/journal.pcbi.1004947

[78] Prlić, A. and Procter, J.B. (2012) Ten Simple Rules for the Open Development of Scientific Software. *PLOS Computational Biology*, **8**, e1002802. https://doi.org/10.1371/journal.pcbi.1002802

[79] Bokhari, E. and Hubert, L. (2018) The Lack of Cross-Validation Can Lead to Inflated Results and Spurious Conclusions: A Re-Analysis of the Macarthur Violence Risk Assessment Study. *Journal of Classification*, **35**, 147-171. https://doi.org/10.1007/s00357-018-9252-3

[80] Teixeira da Silva, J.A. (2013) The Need for Post-Publication Peer Review in Plant Science Publishing. *Frontiers in Plant Science*, **4**, Article 485. https://doi.org/10.3389/fpls.2013.00485

[81] Michener, W.K. (2015) Ecological data sharing. *Ecological Informatics*, **29**, 33-44. https://doi.org/10.1016/j.ecoinf.2015.06.010

[82] U.S. Department of Transportation (2022) CMF Clearinghouse. https://www.cmfclearinghouse.org/sqr.php