AMA Journal of Ethics®
January 2025, Volume 27, Number 1: E27-33

STATE OF THE ART AND SCIENCE: PEER-REVIEWED ARTICLE

How Should Meaningful Evidence Be Generated From Datasets?

Caroline E. Morton, MRCGP and Christopher T. Rentsch, PhD

Abstract

Datasets are often considered "ideal" when they are large and contain longitudinal and representative data. But even research that uses ideal datasets might not generate high-quality evidence. This article emphasizes the roles that transparency plays in enhancing observational epidemiological findings' credibility and relevance and argues that epidemiological research can produce high-quality evidence even when datasets are not ideal. This article also summarizes strategies for bolstering transparency in key phases of research planning and application.

Dataset Size and Scope

In epidemiology research, the quality and believability of findings often hinge on the size and scope of datasets. Datasets are considered to be "ideal" if they are large and the data they contain are longitudinal and largely representative of the underlying population. However, the assumption that large datasets inherently produce high-quality epidemiological research is misleading and should be challenged, as there is more to a high-quality observational study than just the size of the data input. This article summarizes the roles of transparency and systematic reporting of methodologies, data sources, and analytic code in enhancing the credibility of observational studies. This article also posits that high-quality research is possible with datasets that are not ideal, provided that there is a high level of transparency throughout the research process. By examining the different stages of a research study—from conception to execution—this article outlines practical strategies that can be used to increase transparency.

Well-Formulated Research Questions

A critical aspect of epidemiological research is formulating a research question. A well-crafted question guides the entire research process. It helps in defining the scope and design of the study, in identifying an appropriate data source to answer the question, and in selecting appropriate statistical methods to answer the question.

There are excellent resources for guidance on developing health research questions.[1,2,3,4] Briefly, the question should be relevant, answerable (through the collection and analysis of data), and specific. It should address a gap in knowledge or a

pressing public health issue. The question should also have practical implications for health care, policy, or further research.

A commonly used framework to specify a clinical research question is referred to as PICOT, which contains 5 elements: population to be studied, intervention or exposure used in the study, comparator or reference group for treatment group comparisons, outcome or result to be measured, and timeframe or duration of data collection. As an example of use of this framework, one study of mRNA COVID-19 boosters aimed "to compare the effectiveness [outcome] of a third dose of either the BNT162b2 [exposure] or the mRNA-1273 [comparator] vaccine among US veterans who had completed an mRNA vaccine primary series [population] and received a third dose between 20 October 2021 and 8 February 2022" or between "1 January and 1 March 2022 [timeframe]."[5]

Data Provenance

Key to transparency is understanding the context in which and the intent for which data were collected. Although there are many different types of bias,[6,7,8,9,10] it can be helpful to think about them in terms of *selection bias* and *information bias*. Selection bias is bias that arises when the study sample systematically differs from the population (for example, self-selection into a study). Information bias is bias that arises when key variables are not measured accurately (for example, self-reported disease status). Each of these biases encompasses a number of more specific biases that should be considered further, depending on study design and data source.[11]

Another way to consider potential biases is to break down the stages of data collection into steps. There are 4 key steps at which bias can occur that involve choices about: (1) location, (2) participant, (3) research team, and (4) software used, each of which are described below. Knowledge of the location in which the participant's data were collected is critical for identifying any potential differential bias introduced by the setting (eg, the need to pay or have health insurance). For example, was a patient being seen in routine health care, in an emergency setting, or at a private health clinic? Next, researchers should consider the participants and whether their health behavior introduces any biases. For example, participants who are captured in a dataset, either through routine care or in a specific research study, can often request that their information be removed at a later stage (ie, "opt-out"), potentially introducing bias after initial data collection. The research team is also an important but underestimated source of potential bias. Recognizing this source of bias requires consideration of team members' background, training, and unconscious biases, which may affect the types of data that are recorded. Finally, the software used to create the dataset may also introduce biases by prompting researchers or clinicians to enter data in a specific way or to ask particular questions.

It is not always possible to completely remove biases arising from data provenance, but it is important to acknowledge and account for them when possible in the interests of transparency and accuracy, as these biases might affect both the results and the generalizability of findings to a different population. The questions researchers ask may differ between countries—particularly between those with and without nationalized health care systems—so it is important to understand the context in which and the purpose for which data were collected.

Preregistration

A key approach to improving transparency and trust in research is the preregistration of study protocols.[12,13,14,15] By preregistering study protocols, researchers establish a clear blueprint of their research objectives, methods, and analytical plans before data analysis begins. This proactive approach mitigates risk of publication bias,[16] which refers to the suppression of whole studies—for example, those without statistically significant results.[17] Preregistration also reduces the potential for researchers to disseminate misleading or erroneous results by holding them accountable to their stated methodologies and hypotheses. Transparent documentation of study protocols enables stakeholders, including peer reviewers and readers, to evaluate the integrity and robustness of the study, thereby bolstering confidence in the validity of its findings.

One area of epidemiology in which preregistration is increasingly common is real-world evidence (ie, data derived from sources such as electronic health records, registries, medical claims, and patient self-monitoring) of drug safety and effectiveness.[18] As the US Food and Drug Administration expands the use of real-world evidence in its drug approval decision-making processes,[19] the use of preregistered protocols is critical to improving transparency. Although sponsors and researchers are required to preregister certain clinical trials and report results to ClinicalTrials.gov,[20] a similar system for real-world studies did not exist until recently. The Open Science Framework developed by the Center for Open Science aims to fill that gap by offering a free, open-source platform to preregister protocols in addition to other study materials.[21]

A common misconception of preregistering study plans is that they are fixed. On the contrary, protocols are flexible and can be edited as the study progresses. However, transparency is achieved by having all deviations from the original protocol documented across the lifetime of a study. Beyond improving transparency, preregistration cultivates a culture of collaboration and open science,[22,23] encouraging reproducibility within the research community.

Code Sharing

Once data have been made available to researchers, it is usually necessary to "clean" the dataset to get it into a format conducive to analysis. This usually means, at a minimum, applying the prespecified criteria to obtain a narrower dataset. Prespecified criteria might be a particular population (eg, males 65 years or older) or patients with a particular duration of follow-up (eg, at least 1 year of follow-up after baseline. Data cleaning also includes the creation of variables of interest, including exposures and outcomes. Covariates and other variables of interest may be important for adjustment of models, stratification, or identifying subpopulations. Dataset preparation is typically carried out using scripted code, such as Python, SAS, Stata, or R.

Since defining these variables and running analyses are fundamental to understanding how the research protocol was applied to the raw data, researchers should strongly consider publicly sharing code. GitHub is a free service widely used for code sharing.[24] A key advantage of GitHub is that every update to code is time stamped, allowing proper version control. Licences can be applied to the code to allow (if permitted) reuse and adaptation. Whenever possible, efforts should be made to add good documentation to any code—including, but not limited to, in-line code comments, a README file, and software versions.

One prominent example of good preregistration practices and code sharing is OpenSAFELY.[25] OpenSAFELY is a highly secure, transparent, trusted research environment for analysis of electronic health records data arising from primary and secondary care. All platform activity is publicly logged so that anyone at any time can review what code is being run against the data through the OpenSAFELY jobs-server.[26] Before any code is submitted, researchers must preregister a study protocol, which is posted publicly. All software for data management and analysis is shared on GitHub, automatically and openly, for scientific review and efficient reuse.[27]

Interpretation

Studies are carried out to generate answers to important research questions. It is therefore crucial to appropriately interpret results of a given analysis in a way that generates meaningful, clear, and believable evidence. Accordingly, researchers should clearly explain how conclusions were drawn from the data and how analyses were carried out. The findings should also be presented within the wider context of previously published literature. Do the results fit in with what is already known about the topic? If not, more questions should be asked to interrogate what potential biases might be at play.

When interpreting results, special consideration should be given to issues that are known to cause confusion, such as the differences between absolute and relative risk[28,29,30] and whether the results can be generalized to a wider population.[31,32] A lay summary can clarify potentially confusing issues while helping to explain some of the findings without the statistical jargon. Additionally, a clear, comprehensive figure that conveys a study's key findings is often what many readers look to first,[33] so authors should be mindful of this preference when developing and selecting the results to put in figures. Finally, infographics and expert opinion pieces can aid understanding if placed alongside a particularly controversial or difficult-to-understand analysis.

Need for Institutional Resources

Transparency and open science can support reproducibility and the responsible conduct of research, but they are not a guarantee of scientific rigor or equitable science.[34] An epidemiological study can be transparently reported but still come to the wrong conclusions, as was the case for a seminal study on the protective effect of high-density lipoprotein cholesterol on the risk of coronary heart disease.[35] Moreover, data sharing and other open science practices can pose a resource burden on scientists without institutional support; these barriers can be particularly marked for scientists in lower-resource settings.[36,37,38] Even when resources are available, concerns have been raised that the movement towards open data risks "perpetuating a neocolonial dynamic," wherein it is necessary for researchers to pay for access or purchase costly software or training in order to use data effectively.[39] Finally, data sharing, in particular, requires careful consideration to ensure patient privacy and respect the original consent processes.[40,41]

Conclusion

High-quality epidemiological research is not always achieved even with an ideal dataset. Transparency is often underutilized as a way to increase the believability—and therefore the meaningfulness—of epidemiological findings from observational research. Transparency can be achieved through specifying a well-formulated research question, acknowledging limitations arising from data provenance, preregistering analysis plans, code sharing, and making measured interpretations. Preregistration and code sharing

are pivotal practices for fostering not only transparency and credibility but also accountability for and reproducibility of research designs and analyses. These practices also combat biases and promote a culture of collaboration and open science. Ultimately, increasing the adoption of these modern practices in epidemiology could serve as a cornerstone for building trust among researchers, patients, and the broader public.

References

1. Kloda L, Bartlett JC. Formulating answerable questions: question negotiation in evidence-based practice. *J Can Health Libr Assoc*. 2014;34(2):55-60.
2. Lipowski EE. Developing great research questions. *Am J Health Syst Pharm*. 2008;65(17):1667-1670.
3. Thabane L, Thomas T, Ye C, Paul J. Posing the research question: not so simple. *Can J Anaesth*. 2009;56(1):71-79.
4. Tully MP. Research: articulating questions, generating hypotheses, and choosing study designs. *Can J Hosp Pharm*. 2014;67(1):31-34.
5. Dickerman BA, Gerlovin H, Madenci AL, et al. Comparative effectiveness of third doses of mRNA-based COVID-19 vaccines in US veterans. *Nat Microbiol*. 2023;8(1):55-63.
6. May T. How America's health data infrastructure is being used to fight COVID-19. Datavant blog. May 26, 2020. Accessed July 11, 2024. https://www.datavant.com/blog/how-americas-health-data-infrastructure-is-being-used-to-fight-covid-19
7. Miksad RA, Abernethy AP. Harnessing the power of real-world evidence (RWE): a checklist to ensure regulatory-grade data quality. *Clin Pharmacol Ther*. 2018;103(2):202-205.
8. Boyd AD, Gonzalez-Guarda R, Lawrence K, et al. Potential bias and lack of generalizability in electronic health record data: reflections on health equity from the National Institutes of Health Pragmatic Trials Collaboratory. *J Am Med Inform Assoc*. 2023;30(9):1561-1566.
9. Sauer CM, Chen LC, Hyland SL, Girbes A, Elbers P, Celi LA. Leveraging electronic health records for data science: common pitfalls and how to avoid them. *Lancet Digit Health*. 2022;4(12):e893-e898.
10. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*. 2018;361:k1479.
11. Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ, eds. *Modern Epidemiology*. 4th ed. Wolters Kluwer; 2021.
12. Berger ML, Sox H, Willke RJ, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE special task force on real-world evidence in health care decision making. *Pharmacoepidemiol Drug Saf*. 2017;26(9):1033-1039.
13. Wang SV, Schneeweiss S, Berger ML, et al; joint ISPE-ISPOR Special Task Force on Real World Evidence in Health Care Decision Making. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0. *Pharmacoepidemiol Drug Saf*. 2017;26(9):1018-1032.
14. Simmons JP, Nelson LD, Simonsohn U. Pre-registration: why and how. *J Consum Psychol*. 2021;31(1):151-162.
15. Logg JM, Dorison CA. Pre-registration: weighing costs and benefits for researchers. *Organ Behav Hum Decis Process*. 2021;167:18-27.

16. Kicinski M, Springate DA, Kontopantelis E. Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Stat Med*. 2015;34(20):2781-2793.
17. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA*. 1990;263(10):1385-1389.
18. Wang SV, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ*. 2021;372:m4856.
19. US Food and Drug Administration. Framework for FDA's Real-World Evidence Program. US Food and Drug Administration; 2018. Accessed February 27, 2024. https://www.fda.gov/media/120060/download?attachment
20. Clinical trial reporting requirements. ClinicalTrials.gov. Updated September 17, 2024. Accessed September 30, 2024. https://clinicaltrials.gov/policy/reporting-requirements
21. OSF home. Accessed August 9, 2024. https://osf.io/
22. Kuhn TS. Historical structure of scientific discovery. *Science*. 1962;136(3518):760-764.
23. Mathur MB, Fox MP. Toward open and reproducible epidemiology. *Am J Epidemiol*. 2023;192(4):658-664.
24. GitHub. Accessed August 9, 2024. https://github.com/
25. OpenSAFELY. Accessed August 9, 2024. https://www.opensafely.org/
26. OpenSAFELY jobs. OpenSAFELY. Accessed February 27, 2024. https://jobs.opensafely.org/
27. OpenSAFELY. GitHub. Accessed February 27, 2024. https://github.com/OpenSAFELY
28. Tenny S, Hoffman M. Relative risk. In: *StatPearls*. StatPearls Publishing; 2024. Accessed November 21, 2024. https://www.ncbi.nlm.nih.gov/books/NBK430824/
29. Dupont WD, Plummer WD Jr. Understanding the relationship between relative and absolute risk. *Cancer*. 1996;77(11):2193-2199.
30. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: absolute risk reduction, relative risk reduction, and number needed to treat. *Perspect Clin Res*. 2016;7(1):51-53.
31. Rothman KJ, Greenland S. Validity and generalizability in epidemiologic studies. In: Balakrishnan N, Colton T, Everitt B, et al, eds. *Wiley StatsRef: Statistics Reference Online*. Wiley; 2014.
32. St Sauver JL, Grossardt BR, Leibson CL, Yawn BP, Melton LJ 3rd, Rocca WA. Generalizability of epidemiological findings and public health decisions: an illustration from the Rochester Epidemiology Project. *Mayo Clin Proc*. 2012;87(2):151-160.
33. Divecha CA, Tullu MS, Karande S. Utilizing tables, figures, charts and graphs to enhance the readability of a research paper. *J Postgrad Med*. 2023;69(3):125-131.
34. Knottnerus JA, Tugwell P. Promoting transparency of research and data needs much more attention. *J Clin Epidemiol*. 2016;70:1-3.
35. Davey Smith G, Phillips AN. Correlation without a cause: an epidemiological odyssey. *Int J Epidemiol*. 2020;49(1):4-14.
36. Gomes DGE, Pottier P, Crystal-Ornelas R, et al. Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proc R Soc B Biol Sci*. 2022;289(1987):20221113.

37. Bezuidenhout L, Chakauya E. Hidden concerns of sharing research data by low/middle-income country scientists. *Glob Bioeth*. 2018;29(1):39-54.
38. Quiroga Gutierrez AC, Lindegger DJ, Taji Heravi A, et al. Reproducibility and scientific integrity of big data research in urban public health and digital epidemiology: a call to action. *Int J Environ Res Public Health*. 2023;20(2):1473.
39. Evertsz N, Bull S, Pratt B. What constitutes equitable data sharing in global health research? A scoping review of the literature on low-income and middle-income country stakeholders' perspectives. *BMJ Glob Health*. 2023;8(3):e010157.
40. Meyer MN. Practical tips for ethical data sharing. *Adv Methods Pract Psychol Sci*. 2018;1(1):131-144.
41. Mostert M, Bredenoord AL, Biesaart MCIH, van Delden JJM. Big data in medical research and EU data protection law: challenges to the consent or anonymise approach. *Eur J Hum Genet*. 2016;24(7):956-960.

Caroline E. Morton, MRCGP is a senior clinical research fellow in health data engineering at Queen Mary University of London in England, where she works on building reproducible data pipelines for electronic health care research. Previously, she worked as a software engineer and epidemiologist for OpenSAFELY and for OpenCodelists at the Bennett Institute for Applied Data Science. Her work to date has been in using electronic health records to investigate chronic disease and building reusable systems for better research.

Christopher T. Rentsch, PhD is an associate professor at the London School of Hygiene & Tropical Medicine (LSHTM) in England and an adjunct assistant professor at the Yale School of Medicine in New Haven, Connecticut. He obtained an MPH from Emory University and a PhD from LSHTM. Dr Rentsch specializes in the use of electronic health records to generate real-world evidence of the safety and effectiveness of medications, with a focus on quantifying inequity in medication receipt and outcomes.

Citation
AMA J Ethics. 2025;27(1):E27-33.

DOI
10.1001/amajethics.2025.27.

Conflict of Interest Disclosure
Authors disclosed no conflicts of interest.

The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980