AMA Journal of Ethics
Illuminating the Art of Medicine

Epidemiology and Clinical Practice

January 2025, Volume 27, Number 1: E1-65

AMA Journal of Ethics®
January 2025, Volume 27, Number 1: E3-5

FROM THE EDITOR

How Historical Legacies Inform Contemporary Epidemiology and Medicine

Emily L. Graul, MSc and Christopher K. Wong, MD, MSc

Essential to the practice of evidence-based medicine are concurrent and recurrent epidemiological analyses of health determinants and outcomes that are well-designed, high-quality, and transparent interventional and noninterventional studies.[1,2,3,4,5,6,7,8,9] The disciplines of medicine and epidemiology are closely intertwined and share the aim of improving health and well-being but are distinguished by their scope: medicine centers on personalized health care for individuals and families, whereas epidemiology encompasses and studies the health of populations.

Their intersection gives rise to notable ethics questions regarding marginalized communities that are typically excluded from or underrepresented in analyses, as well as people lost to follow-up in trials or misrepresented, mis-sampled, or mis-aggregated in data.[10,11,12] These communities' members are also typically ones health systems disproportionately and inequitably fail to reach: those who are uninsured or underinsured or have job, transportation, or food insecurity; those who live with chronic illnesses or disabilities; or those who experience language and cultural barriers when attempting to access care. Epidemiological data thus reflect wider social and structural inequity, biasing how evidence is applied in clinical practice: specifically, guidelines and formulas that draw upon epidemiological research can stem from and propagate social and institutional biases, exacerbating health inequity.[13,14,15] Some consequences of inequity include individuals' and communities' distrust of health care and of how their data are categorized in research and used.[16]

In this issue of the *AMA Journal of Ethics*, we asked contributors to focus on both the past and the present when addressing why interfaces between epidemiology and medicine are clinically and ethically significant. Epidemiology and health research have contributed to medicine's advancement, but such efforts have had ethical shortcomings that deserve attention. Contributors to this issue consider topics such as the inception of epidemiology's and medicine's integration as part of colonialism and industrialization and how these historical legacies undermine both fields today. They also outline the relationships between various institutions and organizations that play roles in bridging epidemiology and medicine for evidence-based health care, including the parties that provide health care; encode data on people's health information; store, process, and interpret data for research; and translate epidemiological findings into clinical practice guidelines. Furthermore, contributors cover the role of institutions that educate health

practitioners and researchers, including fostering awareness of epidemiological bias. Finally, our issue's experts outline the origins of today's health data coding and classification systems and canvass the myriad complexities of thoughtfully handling data. At the heart of all of these endeavors is building and maintaining trust and respect among persons whose experiences are represented by data, clinicians caring for patients, and clinician-scientists who use patients' data.

By examining historical and present convergences of medicine and epidemiology and implications of those convergences, we aim to support clinicians taking a critical eye to evidence that should inform competent, compassionate practice while enhancing epidemiologists' consideration of peoples' lived experiences as they work to generate evidence from data. We hope this issue can serve as a lens and resource for clinicians and trainees to improve health care practice, health research, clinical outcomes, and equity.

References

1. Guyatt G; Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*. 1992;268(17):2420-2425.
2. Garg PK, Platt JM, Hirsch JA, et al. Association of neighborhood physical activity opportunities with incident cardiovascular disease in the Cardiovascular Health Study. *Health Place*. 2021;70:102596.
3. Jones WS, Mulder H, Wruck LM, et al; ADAPTABLE Team. Comparative effectiveness of aspirin dosing in cardiovascular disease. *N Engl J Med*. 2021;384(21):1981-1990.
4. Joseph J, Pajewski NM, Dolor RJ, et al; PREVENTABLE Trial Research Group. Pragmatic evaluation of events and benefits of lipid lowering in older adults (PREVENTABLE): trial design and rationale. *J Am Geriatr Soc*. 2023;71(6):1701-1713.
5. Mangione CM, Barry MJ, Nicholson WK, et al; US Preventive Services Task Force. Statin use for the primary prevention of cardiovascular disease in adults: US Preventive Services Task Force recommendation statement. *JAMA*. 2022;328(8):746-753.
6. Cholesterol Treatment Trialists' Collaboration. Effect of statin therapy on muscle symptoms: an individual participant data meta-analysis of large-scale, randomised, double-blind trials. *Lancet.* 2022;400(10355):832-845.
7. Gupta A, Madhavan MV, Poterucha TJ, et al. Association between antecedent statin use and decreased mortality in hospitalized patients with COVID-19. *Nat Commun.* 2021;12(1):1325.
8. Collins R, Reith C, Emberson J, et al. Interpretation of the evidence for the efficacy and safety of statin therapy. *Lancet.* 2016;388(10059):2532-2561.
9. Riestenberg RA, Furman A, Cowen A, et al. Differences in statin utilization and lipid lowering by race, ethnicity, and HIV status in a real-world cohort of persons with human immunodeficiency virus and uninfected persons. *Am Heart J.* 2019;209:79-87.
10. Mays VM, Echo-Hawk A, Cochran SD, Akee R. Data equity in American Indian/Alaska Native populations: respecting sovereign nations' right to meaningful and usable COVID-19 data. *Am J Public Health*. 2022;112(10):1416-1420.

11. Xiao H, Vaidya R, Liu F, Chang X, Xia X, Unger JM. Sex, racial, and ethnic representation in COVID-19 clinical trials: a systematic review and meta-analysis. *JAMA Intern Med.* 2023;183(1):50-60.
12. Michos ED, Van Spall HGC. Increasing representation and diversity in cardiovascular clinical trial populations. *Nat Rev Cardiol.* 2021;18(8):537-538.
13. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *N Engl J Med*. 2020;383(9):874-882.
14. Visweswaran S, Sadhu E, Morris M, Samayamuthu M. Clinical algorithms with race: an online database. *medRxiv*. 2023;2023.07.04.23292231.
15. Eneanya ND, Boulware LE, Tsai J, et al. Health inequities and the inappropriate use of race in nephrology. *Nat Rev Nephrol.* 2022;18(2):84-94.
16. Mathur R, Rentsch CT, Venkataraman K, et al. How do we collect good-quality data on race and ethnicity and address the trust gap? *Lancet*. 2022;400(10368):2028-2030.

Emily L. Graul, MSc is an MD/PhD student at Emory University in Atlanta, Georgia, in the Medical Scientist Training Program and a trainee associate editor for *Thorax*. She received her master's degree in epidemiology from the London School of Hygiene and Tropical Medicine. She previously worked at Imperial College London in England, where she conducted research in cardiopulmonary epidemiology and held teaching roles.

Christopher K. Wong, MD, MSc is an internal medicine and pediatrics resident at Baylor College of Medicine in Houston, Texas. He received his medical degree from Baylor and his master's degree in global health nutrition from the London School of Hygiene and Tropical Medicine. His professional interests include global health, transition medicine, and health policy.

Citation
AMA J Ethics. 2025;27(1):E3-5.

DOI
10.1001/amajethics.2025.3.

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

AMA Journal of Ethics®
January 2025, Volume 27, Number 1: E6-13

CASE AND COMMENTARY: PEER-REVIEWED ARTICLE

Whom Should We Regard as Responsible for Health Record Inaccuracies That Hinder Population-Based Fact Finding?

Kathleen M. Akgün, MD, MS and Shelli L. Feder, PhD, APRN

Abstract

Electronic health records (EHRs) have revolutionized the scale, speed, and granularity at which health data can be collated and summarized for epidemiologic purposes. However, population-level analyses of patient-level data are only as reliable as the accuracy or completeness of patient reporting, clinician data entry, and how systems are programmed. This commentary on a case argues that responsibility for the validity of EHR data should be shared among key stakeholders, including patients. This commentary also proposes models for EHR data inquiry, data entry, and review processes that incorporate roles of community partners, frontline clinicians, and health science experts.

Case

T is a 43-year-old woman who has well-controlled asthma and visits Dr A for an annual checkup from a rural part of the state. Dr A reviews T's electronic health record (EHR), noting no documentation of COVID-19 vaccination. Dr A remembers results from a national study that rurality was associated with decreased odds for vaccination and asks T about her reasons for not getting any shots. "I did get 2 shots last year, but I didn't get them here. That's probably why you're not seeing them. I plan to get a booster as soon it's available."

Dr A wonders why many other patients' EHRs contain incomplete or inaccurate information and why. "Not only does this affect how I plan my time with my patients, but poor-quality data hinders epidemiological surveillance and tracking of population-level vaccine uptake. Wasn't the harrowing transition we've all just made to EHRs supposed to eliminate problems like this?"

Commentary

EHRs have become ubiquitous in clinical care in economically developed settings, including much of the United States. We review the evolution of these tools from their use in clinical care and billing to population-level health studies and pragmatic clinical trials. We then identify sources of biases and inaccuracies in EHR data and consider the ethics and consequences of using EHR-based data in research. Finally, we discuss the responsibilities of maintaining EHR data accuracy and propose ways to promote

engagement among key stakeholders (eg, health care systems and payers, EHR developers, patients, clinicians, and researchers) in building an accurate, representative EHR. We illustrate these issues in a study of vaccination outcomes for patients enrolled in rural and urban health systems.

Brief History of the EHR

Clinical information systems were the predecessors of the modern-day EHR and were first utilized in single clinical sites as early as the 1960s.[1] Efforts to transform health record keeping with EHR technology were promulgated with the development of the Department of Veteran Affairs' VistA and Computerized Patient Record System in the 1970s and 1980s.[1] Since then, health care organizations have been incentivized—and eventually mandated—to transition from paper charts to EHRs: first with the passage of the Health Information Technology for Economic and Clinical Health Act (HITECH) as part of the 2009 American Recovery and Reinvestment Act. Through incentive payments, HITECH sought to maximize EHRs' potential to improve patient safety (including by minimizing illegible handwriting and standardizing health data collection, entry, and reporting) and build EHRs into the scaffolding of health care delivery.[2] With the passage of the 21st Century Cures Act in 2016, the HITECH regulations were expanded to require the use of EHRs.[3]

In addition to improving patient safety, the EHR has been a boon for researchers. The proliferation of the EHR has enhanced the analyzability of clinical encounters through typed clinical note documentation accompanied by structured billing codes. Patient demographics are routinely collected and grouped by researchers to estimate the prevalence of health behaviors, determine at-risk patient populations, identify health disparities, and screen potential participants for clinical trial enrollment.[4] Epidemiologic studies based on EHR data are critical for measuring population-level outcomes, including hospitalization or death. EHRs also allow for standardized data collection with the use and dissemination of templates for clinical notes, transforming unstructured text into structured data elements that can be easily extracted from the EHR.[5]

Sources of Error in EHR Data

Although the EHR has many benefits, limitations of data collection can affect data quality and bias research findings. Common domains of EHR data quality for research purposes include accuracy, completeness, consistency, credibility, and timeliness (see Table 1).[6]

Table 1. Quality Domains for Data in the Electronic Health Record

Quality domain	Definition	Example from vaccination scenario
Accuracy	Extent to which data in EHR is valid "representation of the real-world value"	Vaccine receipt would indicate patient was vaccinated
Completeness	Frequency of missing data and patterns	Vaccine status entered across rural, suburban settings independent of race or ethnicity or payer
Consistency	Predictability of data collected in different systems or databases	Vaccine data across systems (eg, Medicare, VHA) uses comparable variable definitions
Credibility	Plausibility, believability of data	Dates of vaccination begin after vaccine available (12/14/20) and do not exceed present day
Timeliness	Lapse of time between data entry and ability to measure variable of interest	Vaccine data collated by week following launch of vaccination campaign to report vaccination trends

Domains and definitions adapted from Feder.[6]
Abbreviations: EHR, electronic health record; VA, Veterans Health Administration.

Like any system data, EHR data are only as strong as their inputs.[4] As the case illustrates, barriers to accurate vaccination documentation begin with the patient-clinician interaction. Inaccurate data inputs may result from poor patient-clinician communication and a lack of patient understanding or opportunity to ask for clarification about the questions being asked.[1] When reviewing their EHR, patients not uncommonly perceive mistakes.[7] Among 22 889 US participants of the OpenNotes study who read their notes and completed error questions, 4830 (21%) identified an EHR mistake, 2043 (42%) of whom reported that the mistake was serious.[8] In addition to mistakes, time constraints could lead to inaccurate EHR data. Additional sources of inaccurate data, which are intrinsic to clinical care and not unique to the EHR, include patient preferences regarding disclosure of sensitive information, the receipt of out-of-network care or at care centers utilizing different EHR systems, asynchronous data entry, or clinician omission (see Table 2).[9,10,11,12,13]

Table 2. Sources of Electronic Health Record Errors and Examples of Mechanisms of Missingness

Errors in EHR data input	**Examples in clinical EHR data**
Data entry errors	Incorrect medications, laboratories, or vital signs
Cut and paste errors	Dated health information; lack of updated conditions, medications, or procedures
Chart management errors	Charting information in wrong EHR
Chart completion errors	Delayed or incomplete chart documentation
Incorrect order entry	"Sound-alike" medication prescribing entered in error but could be folded into other clinicians' documentation for the patient
Submechanisms of missingness	**Examples in clinical EHR data**
Data elements	Exposures, confounding variables, outcomes, relevant variables
Time points	Baseline, varying time for follow-up
Likelihood of measurement during clinical encounter	Blood draws not done at all primary care visits; telehealth visits
Outside care	Out-of-network subspecialty care
Changing clinical practice standards	Screening leads to more incidental findings

Text adapted from Haneuse et al[9] and Bowman.[10]
Abbreviation: EHR, electronic health record.

Data entry tools like drop-down menus, copy-paste features, and automatic laboratory value entries can enhance efficiency but can also contribute to system-level errors and omissions, perpetuating biases and inequities.[14,15,16] In the case example, and as reported in a recent study,[17] rurality was associated with decreased COVID-19 vaccination. However, data entry was incomplete, confounded by human factors that might be exacerbated in rural settings. Inaccurate or incomplete data entry can contribute to sweeping but biased generalizations about treatment disparities, which very well could exist, but are incompletely ascertained due to missing data.[4,5,10,18,19]

Effects of Regulations, Missingness, and Representativeness on Research

Individual patient privacy and agency could be at odds with the need for high-quality population-level health data. Governing bodies overseeing research activities provide one layer of protection for patients by minimizing privacy risks and ensuring data security and investigator integrity and compliance with established rules of behavior in

research. Deidentification of EHR data (in compliance with the Health Insurance Portability and Accountability Act of 1996 Privacy Rule) is standard practice when collating system-wide EHR data for research purposes.[20] However, as EHR data is increasingly stripped of identifiers for subsequent public sharing and analyses, designating research based on such data as "not human subjects research" could jeopardize this oversight, with incompletely understood consequences for population-level studies and inferences. In addition, patient opt-out features can perpetuate biases in the final data based on who is or is not choosing this option.[11,12,13]

Moreover, when epidemiologists analyze EHR data for population health impacts, they often encounter missingness, wherein data for variables of interest are unavailable for each included observation for various reasons. Relying solely on quantitative data inputted by clinical teams thus could limit conclusions, telling incomplete stories. Yet solely focusing on solutions like outreach does not improve rural vaccination rates in the setting of incomplete measurement. Qualitative and mixed methods studies could provide important context for EHR data capture and assist researchers in confirming and contrasting findings derived from the EHR. For example, patient interviews could identify barriers to and mechanisms of vaccine uptake and how and where patients are getting vaccinated. Studies assessing clinician perspectives of EHR data entry options and workflows could also uncover reasons for missing or erroneous patient vaccination history data.

Finally, a fundamental vulnerability of collated, population-level data is whether the included sample is indeed representative of the intended source population. Systemic biases, community engagement, and intersectionality across minoritized groups, sex or gender identity groups, and race or ethnicity groups could influence the visibility of specific populations in EHR data and the resultant output. Population health, health equity, and policy experts have begun to identify sources of and strategies for dealing with bias in the EHR.[21] Patient-reported outcomes and health and general literacy are key areas that can be targeted to reduce bias in EHR-based studies.[22]

Stakeholders' Responsibilities

Just as the causes of EHR inaccuracies are multifaceted, so are responsibilities for ensuring EHR data fidelity, which are shared among key stakeholders: health care systems, vendors, clinicians, patients, and researchers using the data.[23] Vendors and health care systems remain accountable to the general public for EHR functionality, usability, and accuracy. Clinicians must remain engaged with health care systems to ensure their data entry maximizes efficient use of EHR data for clinical documentation, review, and research purposes.[10] Reframing the roles of patients as partners in data generation rather than as study subjects or participants could motivate patients to contribute to solutions to high-quality health data collection.[4] Specifically, inviting patients to identify strategies for EHR data entry that most align with their preferences could enhance the completeness, credibility, and timeliness of their data. For example, because patient consent is not routinely obtained before using EHR data, patients may be unaware of how third parties could use their data, and reidentification of patients might be easier in certain types of studies, including genetics or rare disease studies or those using diagnostic imaging or clinical text notes.[24] "Opt-out" features could return some agency to patients over their EHR data use in research but is not routinely done across health systems. However, this patient-centered approach could, itself, contribute to biased data, as patients who participate in such efforts might not reflect the overall patient population of interest.

Finally, data scientists and researchers are responsible for ensuring that high-quality EHR data are appropriately analyzed in population-level analyses. They should also remain vigilant in recognizing biased results of the analyses performed[25] by explicitly addressing the 5 domains of data quality—accuracy, completeness, consistency, credibility, and timeliness—throughout the research process. Research teams also must defend against data breaches and must routinely address limitations and biases of measured data in study manuscripts and other output.[6]

Accountable EHR Data Use

How can we improve and innovate EHRs to enhance the accuracy of vaccine documentation and other data? An integrated US health system is an attractive answer and could serve as the platform for clinical data integration, but it is unlikely to gain favor in the current politically polarized environment. The Affordable Care Act (ACA) of 2010 was the first reform to the US health care system in a generation and offered a mechanism to build towards integration by leveraging Medicaid infrastructure for beneficiary data collection,[26] but state-level single-payer policies have failed to gain political traction.[27] The ACA remains a lightning rod for fiscally conservative policy makers, and, although public support appears to have grown over time, an overhaul to the US health system is unlikely to be successful in the coming decade.

Given this landscape, individuals and organizations should undertake to improve EHR data quality.

- Health information vendors should continually reassess EHR data collection tools and interfaces with patients, clinicians, and data scientists to optimize their product's usability.
- Clinicians must advocate for meaningful approaches to EHR use to support clinical care, rather than simply plodding through required data fields for billing purposes.
- Organizations should consider expanding, standardizing, and integrating the data collected. Although patient-entered EHR data can be an attractive option to increase accuracy of patient data while also empowering patients to control their health care narrative firsthand, it could still result in data skewed towards patients who are computer literate and have access to broadband internet services. Vaccination and medication use data collection could be standardized by automating linkages between pharmacy manufacturing lots or similar measures and EHRs, thereby improving care tracking. Government- and private, nonprofit-supported applications, such as the Immunization Information System and health information exchanges, also hold promise for integrating and harmonizing health information—from vaccinations to medications, subspecialist evaluations, and clinical testing results—across participating regions.[6,7,8,9]

Limitations

Despite attention to data quality, there exist no standard methods for assessing EHR data quality.[28] We acknowledge that the aforementioned proposed solutions for improving the accuracy of EHR data are practical only for measurable processes and outcomes of care. Other important aspects of care, such as patient-clinician demographic concordance or general communication styles, are not currently captured in structured data available in large health care system EHRs. Alternate care delivery models (minute clinics, telehealth, concierge medicine) might also have uncertain impacts on EHR completeness going forward. As conventions shift in health care delivery

and its documentation, key stakeholders are necessary to determine how best to tackle pressing health priorities of the population.

Conclusion

EHR use has revolutionized health information collection and analysis. This growth has led to opportunities to generate important reports about the health of hundreds of millions of people practically in real time. Steadfast commitment to high-quality data collection and reporting is necessary for all parties along the pathway of data generation: from EHR developers, programmers, and vendors to patients, clinicians, and epidemiologists. Pulling back the curtain on how each of these groups generate and interact with EHR data is imperative to assure measurement of accurate population-level health outcomes.

References

1. Atherton J. Development of the electronic health record. *Virtual Mentor.* 2011;13(3):186-189.
2. Trout KE, Chen LW, Wilson FA, Tak HJ, Palm D. The impact of meaningful use and electronic health records on hospital patient safety. *Int J Environ Res Public Health*. 2022;19(19):12525.
3. Phelan D, Gottlieb D, Mandel JC, et al. Beyond compliance with the 21st Century Cures Act Rule: a patient controlled electronic health information export application programming interface. *J Am Med Inform Assoc.* 2024;31(4):901-909.
4. Hollister B, Bonham VL. Should electronic health record-derived social and behavioral data be used in precision medicine research? *AMA J Ethics*. 2018;20(9):E873-E880.
5. Vuong K, Ivers R, Hall Dykgraaf S, Nixon M, Roberts G, Liaw ST. Ethical considerations regarding the use of pooled data from electronic health records in general practice. *Aust J Gen Pract*. 2022;51(7):537-540.
6. Feder SL. Data quality in electronic health records research: quality domains and assessment methods. *West J Nurs Res*. 2018;40(5):753-766.
7. Klinger EV, Carlini SV, Gonzalez I, et al. Accuracy of race, ethnicity, and language preference in an electronic health record. *J Gen Intern Med*. 2015;30(6):719-723.
8. Bell SK, Delbanco T, Elmore JG, et al. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. *JAMA Netw Open*. 2020;3(6):e205867.
9. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. *JAMA Netw Open*. 2021;4(2):e210184.
10. Bowman S. Impact of electronic health record systems on information integrity: quality and safety implications. *Perspect Health Inf Manag*. 2013;10(fall):1c.
11. El Emam K, Jonker E, Moher E, Arbuckle L. A review of evidence on consent bias in research. *Am J Bioeth*. 2013;13(4):42-44.
12. Kho ME, Duffett M, Willison DJ, Cook DJ, Brouwers MC. Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ*. 2009;338:b866.
13. de Man Y, Wieland-Jorna Y, Torensma B, et al. Opt-in and opt-out consent procedures for the reuse of routinely recorded health data in scientific research and their consequences for consent rate and consent bias: systematic review. *J Med Internet Res.* 2023;25:e42131.

14. Boyd AD, Gonzalez-Guarda R, Lawrence K, et al. Equity and bias in electronic health records data. *Contemp Clin Trials*. 2023;130:107238.
15. Rozier MD, Patel KK, Cross DA. Electronic health records as biased tools or tools against bias: a conceptual model. *Milbank Q*. 2022;100(1):134-150.
16. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. *J Med Internet Res*. 2018;20(5):e185.
17. Bernstein E, DeRycke EC, Han L, et al. Racial, ethnic, and rural disparities in US veteran COVID-19 vaccine rates. *AJPM Focus*. 2023;2(3):100094.
18. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*. 2018;178(11):1544-1547.
19. Getzen E, Ungar L, Mowery D, Jiang X, Long Q. Mining for equitable health: assessing the impact of missing data in electronic health records. *J Biomed Inform*. 2023;139:104269.
20. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. US Department of Health and Human Services. Reviewed October 25, 2022. Accessed February 22, 2024. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
21. Sun M, Oliwa T, Peek ME, Tung EL. Negative patient descriptors: documenting racial bias in the electronic health record. *Health Aff (Millwood)*. 2022;41(2):203-211.
22. 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.
23. Sittig DF, Singh H. Rights and responsibilities of users of electronic health records. *CMAJ*. 2012;184(13):1479-1483.
24. Holmes JH, Beinlich J, Boland MR, et al. Why is the electronic health record so challenging for research and clinical care? *Methods Inf Med*. 2021;60(1/2):32-48.
25. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol*. 2021;21(1):234.
26. Stulberg D. The Patient Protection and Affordable Care Act and reproductive health: harnessing data to improve care. *J Health Polit Policy Law*. 2013;38(2):441-456.
27. Sparer MS. States as policy laboratories: the politics of state-based single-payer proposals. *Am J Public Health*. 2019;109(11):1511-1514.
28. Lewis AE, Weiskopf N, Abrams ZB, et al. Electronic health record data quality assessment and tools: a systematic review. *J Am Med Inform Assoc*. 2023;30(10):1730-1740.

Kathleen M. Akgün, MD, MS is an associate professor of medicine at VA-Connecticut Health Care System (VACHS) and the Yale School of Medicine in New Haven, Connecticut. A health services researcher focused on health care delivery and outcomes and advocacy, she also serves as co-director of the West Haven Network of Dedicated Enrollment Sites and as the director of the medical intensive care unit and co-chair of the Clinical Ethics Committee at VACHS.

Shelli L. Feder, PhD, APRN is an assistant professor at the Yale School of Nursing in New Haven, Connecticut. She is also an associate program director for the Yale National Clinician Scholars Program, an affiliate investigator at the PRIME Center, and an organizational health services researcher. She has over a decade of clinical experience as an advanced practice nurse in hospice, palliative care, and cardiovascular settings. Dr Feder's research program aims to create innovative models of care that improve access to high-quality, timely palliative care for people with cardiopulmonary conditions.

Editor's Note

The case to which this commentary is a response was developed by the editorial staff.

Citation

AMA J Ethics. 2025;27(1):E6-13.

DOI

10.1001/amajethics.2025.6.

Acknowledgements

Dr Feder and Dr Akgün acknowledge support from grant 1R56HL16652301A1 from the National Heart, Lung, and Blood Institute (primary investigator: Dr Feder) and from grant 13407 from the Center of Innovation of the Veterans Health Administration.

Conflict of Interest Disclosure

Dr Akgün has a contract with the Department of Veterans Affairs Central Office. Dr Feder disclosed no conflicts of interest.

The people and events in this case are fictional. Resemblance to real events or to names of people, living or dead, is entirely coincidental. 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

AMA Journal of Ethics®
January 2025, Volume 27, Number 1: E14-20

MEDICAL EDUCATION: PEER-REVIEWED ARTICLE

What Should Health Professions Students Learn About Data Bias?

Douglas Shenson, MD, MPH, MA, MS, Beverley J. Sheares, MD, MS, and Chelesa Fearce

Abstract

In epidemiology, bias is defined as systematic deviation from the truth, and it can arise at different stages of scientific investigation (eg, data collection, methodological application, and outcomes analysis). Epidemiological bias can appear as a consequence of data bias (usually categorized as selection bias or information bias) or social bias (prejudice). Such forms of bias may occur separately or together. This article explores what health professions students should learn about the relationship between data bias and social bias—generated by racial, ethnic, gender, or other kinds of prejudice, singly or in combination—as a source of ethical and clinical concern in health care practices and policies that influence patient care and community health.

The American Medical Association designates this journal-based CME activity for a maximum of 1 AMA PRA Category 1 Credit™ *available through the* AMA Ed Hub™. *Physicians should claim only the credit commensurate with the extent of their participation in the activity.*

Thinking Critically About Evidence

Bias is defined as the presence of systematic error in a study, and its adverse impact has significant ethical importance: a cascade of diagnoses or broader interventions that are erroneous, leading to treatment plans that harm patients and populations. The recognition of data bias is a foundational clinical medicine skill because evidence-based practice depends on accurate information. Students tend to consider the answer to the question, “Are the numbers in this table correct?” to be straightforward. But accuracy cannot be assessed by scrutiny of numerical data alone; only a close look at the methods that produced those values can reveal bias. In short, data bias emerges from a *process*. A reader in the health professions must understand the steps that generate the numbers and the assumptions made by investigators about their data sources. It is important to know whether bias stems from the availability of information, its collection, its methodological manipulation, or the analysis of findings.

Broader judgment comes into play because epidemiological bias, which includes data bias, does not arise from methodological errors alone. It can also result from socially discriminatory choices that inform data selection, classification, and analysis. Health professions students are often surprised that, as used in epidemiology, the term *bias* is

putatively unconnected from its everyday meaning: prejudice or devaluation of a particular social group. Here, we argue that the everyday and scientific meanings of the term are at times closely interrelated. A focus on implicit bias in the clinical realm enhances this awareness,[1] and recent attention to the need for more equity in public health data[2] reinforces the importance of the issue. This article canvasses concepts pertaining to epidemiological bias that health professions students should understand, regardless of whether bias is generated by epidemiological errors or by racial, ethnic, gender, or other prejudices. Such sources of inaccuracy are of ethical and clinical concern because they can influence patient care and community health.

Importance of Accurate Data

The goal of epidemiological and clinical research is to produce accurate data that are useful. The presence of bias in a study implies that there are systematic errors in the data. Nonetheless, bias is not a dichotomous concept: it can exist to a greater or lesser degree and may distort a true association in one direction or the other.[3] Where it is present, bias will influence the validity of data for the population under study (internal validity) and for populations for whom results are assumed to be relevant (external validity). For example, in a drug trial, internal validity represents the extent to which observed outcomes can be ascribed to the treatment regimen, allowing for causal inference. There can be no external validity (broader effectiveness) without internal validity, although the presence of the second does not guarantee the first.[4]

Sources of Data Bias

Data bias is a capacious concept. Its largest categories are selection bias and information bias, which in turn encompass numerous subvarieties.[5] Notably, certain study designs are structurally vulnerable to data bias. For example, retrospective cohort studies are particularly prone to selection bias. In such chart-based studies, the investigator identifies a cohort that has been assembled in the past, identifies potential predictor variables from measurements made in the past, and evaluates outcome variables. Since data will likely not have been collected for research, some charts might be excluded due to missing but crucial information.[6] Interviewer bias can occur in case-control studies if investigators question patients who are "cases" more intensively about exposures that are already known to be associated with the disease.[3,7] Even randomized controlled studies are vulnerable to bias resulting from misallocation of participants, insufficient data blinding, or loss of subjects to follow-up.[8]

Selection bias. In most studies, only a sample of the target population is chosen for observation or intervention. Consequently, studies are susceptible to selection bias, that is, to the recruitment of a nonrepresentative assemblage of subjects.[9] Individuals within the sample may systematically differ with respect to social and economic status, educational level, age, or other consequential characteristics. Such errors can obscure causal associations between an exposure, such as a treatment, and a health-related outcome.[10] Biases in which errors of inclusion or exclusion play a role often have their own designation or eponym. This inventory of biases includes nonresponse bias, volunteer bias, Berksonian bias, attrition bias, incidence-prevalence bias, confounding by indication, surveillance bias, and other named biases.[5,9]

Information bias. Information bias can arise during or after data collection and refers to systematic errors in the measurement of variables or classification of subjects. Errors of measurement can occur because of faulty instrumentation or discernment, the latter of which includes recall bias, interview bias, observer bias, or confirmation bias.[9,11] As a

rule, understanding the relationship between an exposure and an outcome requires subjects to be classified into categories, such as “exposed” or “non-exposed,” and to isolate variables responsible for differing outcomes.[12,13] Misclassifications commonly arise in observational studies but can be present in randomized controlled studies.[12,13] Nondifferential and differential misclassification bias refer to whether measurement error due to misclassification of subjects is symmetrically or asymmetrically distributed between the intervention and comparison groups. For example, in a study of the impact of a drug on obesity, the scales used to weigh patients may not all be accurate. Depending on whether those inaccuracies are similar (say, 5% higher for all patients) or dissimilar, this error will have a divergent bearing on the results.

Bias should be differentiated from other problems of accuracy, particularly from confounding. Confounding describes an association between 2 variables—an exposure and outcome—that appears causal but that surfaces only due to influence from a hidden yet consequential variable. A well-cited example is the association between heavy coffee drinking and cancer of the pancreas.[7] This mirage is present only because heavy coffee drinkers are more likely than light or non-coffee drinkers to smoke cigarettes, an action responsible for the elevated risk of pancreatic cancer. Confounding represents a distortion of the relationship between exposure and outcome due to the presence of one or more extraneous variables, and, like data bias, it can lead to incorrect inferences about causality. Typically, it is not possible to correct for data bias, whereas if a confounding variable is known and measured, the real effect of the exposure on the outcome can be obtained by adjustment for this factor. In sum, confounding produces errors of interpretation despite the accuracy of the measurement.[4]

Random error, in contrast to bias, is nonsystematic and affects the precision rather than the validity of research findings. This lack of exactness results from sampling variability, producing errors that are unsystematic. Data can be both biased and imprecise, but, unlike bias, lack of precision is best addressed by increasing a study’s sample size.

Social Bias and Data Bias

In the epidemiological literature, data biases are implicitly considered oversights, mistakes, or unavoidable failures in research protocols. Indeed, epidemiologists distinguish sharply between data bias and social bias: “Bias undermines the internal validity of research. Unlike the conventional meaning of bias—i.e., prejudice—bias in research denotes deviation from the truth.”[14] In short, data bias is an operational error and social bias (prejudice) is a disposition of judgment. This distinction is not always clear, however. The consequences of social bias can lead researchers to deviate from the truth, and clinicians can collect biased data by using measurement tools that have social biases structured into them.

An important example of overlap between data bias and social bias is found in research on risk factors for cardiovascular disease in young men identified as Black. The available data are fraught with selection bias. Given the enormity of the population with a history of incarceration and the disproportionate incarceration of Black men,[15,16] the exclusion of incarcerated persons from household-based surveys poses a large obstacle to obtaining unbiased samples. Examples of surveys that exclude people who are currently incarcerated include the National Health and Nutrition Examination Survey and the National Health Interview Survey.[16]

Information bias arising from a legacy of medical racism continues to affect diagnostic and eligibility criteria. Indeed, race is embedded in clinical algorithms and decision-making tools across many medical specialties.[17,18] For example, in 2019 researchers revealed algorithmic bias in a widely used medical artificial intelligence tool that incorporates health care costs into the prediction of clinical risk, with deleterious consequences for Black patients. Since the health care system spends more money, on average, on White patients than on Black patients, the tool returns higher risk scores for White patients than for Black patients. Use of this tool might have led to more referrals for White patients to specialty services, perpetuating both spending discrepancies and race bias in health care.[18] Moreover, in pulmonary medicine, the observation of differences in lung capacity between a population characterized in the 1800s as "Full Blacks" and White soldiers was attributed to a biological difference associated with race rather than the effect of enslavement and environmental exposures now known to alter lung function. Subsequently, a "race correction" was built into equations used in spirometry.[19] Whether in the assessment of occupational lung diseases such as asbestosis or "objective" eligibility for lung transplantations, the incorporation of biased reference standards for lung function can lead to worse outcomes for Black patients.

Links between social and data biases are also evident in biomedical research. For example, the evolving field of precision medicine is driven by lab-based sequencing of the genetic code, creating large databases that are curated and organized to extract clinically relevant information. The underrepresentation of non-European populations in genomic databases, like all selection bias, is problematic for clinical care because the exclusion of such data limits their generalizability.[20,21] Moreover, while at the cellular level racial identity is nonexistent, once a pathophysiological process is understood and given a label, the resulting diagnostic category can take on racialized associations, leading to information bias. There are many examples of such racialized diagnostic categories, including sickle cell disease, sarcoidosis, gallstones, and cystic fibrosis. In the clinical setting, this racial "essentialism" leads to assumed or missed diagnoses, misclassification through confirmation bias, and harmful consequences.[21]

Social biases that contribute to data bias are not limited to race. A body of public health research documents gaps in national survey data of sexual orientation and gender identity.[22,23,24] Although data collection procedures have evolved, prior to 2016, biological sex in Behavioral Risk Factor Surveillance System telephone surveys could be assigned based on a respondent's "vocal timbre," a practice vulnerable to confirmation bias.[25] Research documents that this approach has resulted in substantial misclassification of answers, especially those of persons who identify as transgender or gender diverse. [25,26,27] Moreover, the collection of data on sex and gender, where explicitly sought, does not by itself guarantee validity, as there is widespread misunderstanding of the meaning of sex and gender.[24]

Conclusion

Data and social biases are oblique to one another: they are separate frames, but at times they interlock; together, they contribute to epidemiological bias. Data bias refers to systematic errors in a sequence of tasks that produces data; social bias refers to actions and attitudes that can shape those operations. And when these frames coincide, it is not always clear which is a subset of the other. The exclusion of a group from a survey or study can reflect selection bias, but this exclusion may more accurately be ascribed to prejudice.

While there is a path to identifying data bias, there are no shortcuts. Some degree of bias is always present in a published study, so the challenge of bias recognition is ongoing.[3] Awareness of the nature and types of bias in research studies allows for a more meaningful scrutiny of results and conclusions. As researchers, careful planning is needed in each step of research design, and, when presenting results, a full acknowledgment of any sources of bias is essential.[28] The health professional's commitment to a close examination of evidence must remain steadfast, as the presence of bias—whether of epidemiological or social origin—undermines the provision of effective and acceptable clinical care.

References

1. Gopal DP, Chetty U, O'Donnell P, Gajria C, Blackadder-Weinstein J. Implicit bias in healthcare: clinical practice, research and decision making. *Future Healthc J*. 2021;8(1):40-48.
2. Ponce NA, Lau DT. Toward more equitable public health data: an AJPH special section. *Am J Public Health*. 2023;113(12):1276-1277.
3. Pannucci CJ, Wilkins EG. Identifying and avoiding bias in research. *Plast Reconstr Surg*. 2010;126(2):619-662.
4. Rothman KJ. *Modern Epidemiology*. Little Brown & Co; 1986.
5. Delgado-Rodríguez M, Llorca J. Bias. *J Epidemiol Community Health*. 2004;58(8):635-641.
6. Talari K, Goyal M. Retrospective studies—utility and caveats. *J R Coll Physicians Edinb*. 2020;50(4):398-402.
7. Gordis L. *Epidemiology.* 5th ed. Elsevier/Saunders; 2014.
8. Higgins JPT, Savović J, Page MJ, Elbers RG, Sterne JAC. Assessing risk of bias in a randomized trial. In: Higgins J, Thomas J, Chandler J, et al, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane Training; 2024:chap 8. Accessed June 24, 2024. https://training.cochrane.org/handbook/current/chapter-08
9. Jager KJ, Tripepi G, Chesnaye NC, Dekker FW, Zoccali C, Stel VS. Where to look for the most frequent biases? *Nephrology (Carlton)*. 2020;25(6):435-441.
10. Hsu JL, Banerjee D, Kuschner WG. Understanding and identifying bias and confounding in the medical literature. *South Med J*. 2008;101(12):1240-1245.
11. Glick M. Believing is seeing: confirmation bias. *J Am Dent Assoc*. 2017;148(3):131-132.
12. Moseley AM, Pinheiro MB. Research note: evaluating risk of bias in randomised controlled trials. *J Physiother*. 2022;68(2):148-150.
13. Bosdriesz JR, Stel VS, van Diepen M, et al. Evidence-based medicine—when observational studies are better than randomized controlled trials. *Nephrology (Carlton)*. 2020;25(10):737-743.
14. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet*. 2002;359(9302):248-252.
15. Robey JP, Massoglia M, Light MT. A generational shift: race and the declining lifetime risk of imprisonment. *Demography*. 2023;60(4):977-1003.
16. Wang EA, Redmond N, Dennison Himmelfarb CR, et al. Cardiovascular disease in incarcerated populations. *J Am Coll Cardiol*. 2017;69(24):2967-2976.
17. Cerdeña JP, Plaisime MV, Tsai J. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. *Lancet*. 2020;396(10257):1125-1128.
18. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *N Engl J Med*. 2020;383(9):874-882.

19. Braun L. Race correction and spirometry: why history matters. *Chest*. 2021;159(4):1670-1675.
20. Landry LG, Ali N, Williams DR, Rehm HL, Bonham VL. Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Aff (Millwood)*. 2018;37(5):780-785.
21. Tsai J. How should educators and publishers eliminate racial essentialism? *AMA J Ethics*. 2022;24(3):E201-E211.
22. Baker KE, Streed CG Jr, Durso LE. Ensuring that LGBTQI+ people count—collecting data on sexual orientation, gender identity, and intersex status. *N Engl J Med*. 2021;384(13):1184-1186.
23. Patterson CJ, Sepúlveda MJ, White J, eds. *Understanding the Well-Being of LGBTQI+ Populations*. National Academies Press; 2020.
24. Jacobs JW, Bibb LA, Shelton KM, Booth GS. Assessment of the use of sex and gender terminology in US federal, state, and local databases. *JAMA Intern Med*. 2022;182(8):878-879.
25. Riley NC, Blosnich JR, Bear TM, Reisner SL. Vocal timbre and the classification of respondent sex in US phone-based surveys. *Am J Public Health*. 2017;107(8):1290-1294.
26. Tordoff D, Andrasik M, Hajat A. Misclassification of sex assigned at birth in the behavioral risk factor surveillance system and transgender reproductive health: a quantitative bias analysis. *Epidemiology*. 2019;30(5):669-678.
27. Gonzales G, Tran NM, Bennett MA. State policies and health disparities between transgender and cisgender adults: considerations and challenges using population-based survey data. *J Health Polit Policy Law*. 2022;47(5):555-581.
28. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. *J Multidiscip Healthc*. 2016;9:211-217.

Douglas Shenson, MD, MPH, MA, MS is an adjunct associate professor in the Section of General Internal Medicine at the Yale School of Medicine (YSM) in New Haven, Connecticut, where he is deputy leader of YSM's Health Equity Thread. Dr Shenson is also director of YSM's required preclinical course: "Populations & Methods: The Application of Epidemiology and Biostatistics to Public Health."

Beverley J. Sheares, MD, MS is the inaugural leader of the Health Equity Thread at the Yale School of Medicine in New Haven, Connecticut, where she is an associate professor of pediatrics in the Pulmonary, Allergy, Immunology, and Sleep Medicine Section. Dr Sheares' clinical, teaching, and research experiences all coalesce around reducing health disparities and promoting equity.

Chelesa Fearce is a MD/PhD student studying chemistry at Yale University in New Haven, Connecticut, who is interested in drug development for psychiatric disorders. After graduating from Spelman College in 2017, she spent 2 years at the National Institutes of Health studying dopamine receptor signaling. Chelesa plans to make health equity an integral part of her career as a physician-scientist.

Citation
AMA J Ethics. 2025;27(1):E14-20.

DOI
10.1001/amajethics.2025.14.

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

AMA Journal of Ethics®
January 2025, Volume 27, Number 1: E21-26

AMA CODE SAYS: PEER-REVIEWED ARTICLE

Which Values Should Guide Evidence-Based Practice?

Amber R. Comer, PhD, JD

Abstract
This article draws on opinions in the AMA *Code of Medical Ethics* and applies them to evidence-based practice.

The American Medical Association designates this journal-based CME activity for a maximum of 1 AMA PRA Category 1 Credit™ *available through the* AMA Ed Hub™. *Physicians should claim only the credit commensurate with the extent of their participation in the activity.*

Evidence in Clinical Practice

Prior to the emergence and availability of evidence-based reviews, physicians and patients made decisions based on anecdotal data, opinion, experience, judgment, conjecture, and conventional wisdom.[1,2] In 1982, the first textbook describing the methodology of translating biomedical science into clinical practice, *Clinical Epidemiology: The Essentials*,[3] set the stage for what would eventually become what we now call evidence-based medicine (EBM).[1] EBM incorporates the best available scientific evidence when making decisions about an individual patient's care.[4] In the years since the adoption of EBM, it has become not only the clinical standard of care, but also an ethical expectation.[5]

The age-old adage that medicine is both a science and an art has been strengthened by the emergence of EBM; however, questions remain regarding how to elevate the science without sacrificing the art of medicine, the latter of which includes the clinician's compounding of clinical experience, intuition, knowledge of the patient and their preferences and goals, and even the social landscape through which the patient presents. This article explores the ethical issues clinicians face in clinical practice when combining EBM and the art of medicine during medical decision-making. Additionally, this article offers practical clinical recommendations for how to overcome these common ethical dilemmas.

Applying EBM to Patients

Practicing EBM raises several ethical challenges. The first pertains to balancing the science and art of medicine when making evidence-based decisions about patients' care or key communications to patients. Incorporation of clinical expertise with science is important because using only science to make medical decisions fails to take the patient's preferences and values into consideration. Indeed, the art of medicine refers to a patient-centered approach that includes observing and listening to patients and

respecting patients' values, culture, and opinions rather than seeing patients solely as diseased persons in need of a cure.[6] While seemingly straightforward, the caveat that practicing medicine requires an established relationship between clinicians and patients highlights the imbalance in the science-focused approach and that it is the "art" aspect of medicine that resolves it.

The idea of medical practice as a balance of science and art can be better understood through the works and influence of Sir William Osler, a Canadian physician whose legacy on the teaching and practice of medicine continues to influence modern practices, including evidence-based technique. In the context of medicine's growing "biologized view of the sick person,"[7] the quotation attributed to Osler, "The good physician treats the disease, the great physician treats the patient with the disease," can be interpreted as a statement recognizing the need to holistically evaluate a patient and encouraging the continued practice of the ancient Greek-inspired art of observation within medicine.[7,8]

One challenge of balancing scientifically promising or evidence-based care options with a patient's values and opinions has to do with how to manage care of patients who ask for treatments or interventions that are not evidence based or who refuse evidence-based treatments or interventions. Clinicians have a duty to respect patient autonomy, which entails that patients or their surrogates should consent to care they receive. To express respect for a patient's autonomy, though, is not to blindly agree with a patient's decisions, as clinicians have additional ethical responsibilities to balance autonomy and evidence-based care and, in some cases, must adhere to political and legal boundaries. Clinicians are then faced with the challenge of deciding when and if it is ethically acceptable to offer or withhold an evidence-based treatment or procedure to support patient autonomy.

What Does the *Code* Say About Evidence?

The American Medical Association (AMA) *Code of Medical Ethics* recognizes that high-quality medical decisions require physicians to practice both the science and the art of medicine. Opinion 5.5, "Medically ineffective interventions," states: "physicians should only recommend and provide interventions that are medically appropriate—i.e., scientifically grounded—and that reflect the physician's considered medical judgment about the risks and likely benefits of available options in light of the patient's goals for care."[9] When providing recommendations, a physician has a "primary ethical obligation … to promote the well-being of individual patients."[10] However, this obligation can conflict with a physician's ethical duty to use "best available evidence" in instances when the patient or their surrogate requests a treatment or intervention that is not evidence based or when an evidenced-based treatment or intervention is refused.[11] In these instances, the AMA *Code* offers the guidance that "[p]hysicians are not required to offer or to provide interventions that, in their best medical judgment, cannot reasonably be expected to yield the intended clinical benefit or achieve agreed-on goals for care."[9] Importantly, the AMA *Code* recognizes that "respecting patient autonomy does not mean that patients should receive specific interventions simply because they (or their surrogates) request them."[9] Conversely, the AMA *Code* explicitly states that "a patient who has decision-making capacity may accept or refuse any recommended medical intervention."[12]

Applying the AMA *Code* in Practice

How to balance the science and art of medicine when making evidence-based decisions. The AMA *Code* uses the phrases "the physician's considered medical judgment" and "best medical judgment"[9] to describe the standard for making medical recommendations in clinical practice. Although it is the standard of care and an ethical expectation to use the best available evidence— including by referencing up-to-date, evidence-based clinical practice guidelines—evidence alone is not definitive because the results of research studies are interpretations of aggregate data that may change based upon study design, methodology, participant sample, and analysis methods. Therefore, reasonable and intelligent people may disagree on interpretations of the data and the generalizability of these interpretations for use with their individual patients.[2] Thus, medical judgment goes beyond merely identifying the best available evidence and requires that the clinician understand the patient's preferences, values, and goals of medical care.[2,13] Put differently, the process by which clinicians synthesize generalized knowledge garnered from EBM with clinical experience and skills and with individual patients' preferences, values, and medical care goals makes both science and art inherent to medical judgment.

How to manage care of patients who ask for treatments or interventions that are not evidence based. The patient-clinician relationship is a mutual relationship founded upon trust, and while the goal of the relationship is to provide beneficial care to the patient, both parties have their own obligations and rights. Thus, when a patient asks for a treatment or intervention that is not evidence based, approval or denial of the request requires balancing patient and clinician autonomy. First, while clinicians have the ethical obligation to respect patient autonomy, they are not ethically obligated to deliver care that will not have a reasonable chance of benefiting their patient. Furthermore, in the event that acquiescing to requests for treatments or interventions that are not evidence based might place the patient's or the general publics' health at risk, the ethical obligation to prevent harm warrants a clinician's decision to deny these requests. To preserve the medical judgment of physicians with the intention of supporting the safety and well-being of patients and the public, AMA policy recommends that physicians maintain their autonomy and have final say regarding the delivery of high-quality patient care, including by determining which diagnostic tests to run, whether a patient should be hospitalized, when interventions become extraordinary, what treatment methodology to apply, and when it is appropriate to terminate the patient-physician relationship.[13]

This recommendation is not in lieu of respecting patient autonomy and does not ignore the art of medicine, as developing and agreeing upon a care plan is a collaborative effort between clinicians and patients or surrogates with the prioritization of their consent. Rather, this recommendation balances the art and science of medicine via the physician's using science and evidence to safeguard a patient based upon holistic assessment of the patient and their needs. The AMA *Code* also recommends that physicians explain their rationale for not offering the requested intervention or treatment to the patient and offer an alternative if appropriate. Moreover, the AMA *Code* addresses the importance of transparency in maintaining trust, which is essential to the patient-physician relationship.[14] Therefore, if a patient suggests a treatment or intervention that a physician disapproves of using their medical judgment, then the physician should provide information about all other appropriate treatment options, including potential risks and benefits.[14]

How to manage care of patients who refuse evidence-based treatments or interventions. Although clinicians have decision-making authority for the care they choose to deliver, this charge must be balanced with the ethical obligation to obtain informed consent for medical treatment from the patient or their surrogate when the patient lacks decision-making capacity. Obtaining informed consent for treatment requires that the clinician inform the patient about the best available evidence, including treatment options' limits and benefits, so that the patient can determine if they are willing to assume the risk of harm in exchange for the potential benefit of treatment. If the patient has capacity and has been appropriately informed, they have the legal and ethical right to refuse all medical treatments or interventions, even those that may preserve or prolong their life. Although a patient has the right to refuse treatments and interventions, it is important to take the time to identify if there are any underlying reasons for the refusal—for example, fear, a prior bad experience, or a misunderstanding about the nature of the disease or treatment—that can be addressed through further conversation and support.

In practice, it is imperative to first determine whether a patient retains the capacity to make decisions regarding their health. Such determinations will rely heavily on the patient's ability to understand—and to communicate their understanding of—the risks and benefits associated with treatment or interventions.[15] To that same end, assessing a patient's decision-making capacity is critical to initiating the informed consent conversation that will outline the best available evidence, again including treatment options' potential limitations and benefits. Should patient capacity be determined to be limited, then health decisions, including those requiring informed consent, should be made by the appointed surrogate. However, if the patient maintains capacity but refuses evidenced-based treatment, then the clinician may ask questions to ascertain whether the reason for refusal could be addressed in other ways, such as through a goals-of-care conversation or by providing additional support. It is nevertheless important to remember that patients with capacity, or surrogates representing patients with limited capacity, have the legal and ethical right to refuse any treatment or intervention. In cases in which refusal of a treatment or an intervention would result in patient suffering or even death, physicians are encouraged to consult with a palliative care specialist to assist with the goals-of-care discussion or to provide support to the patient and family in their decision to refuse.

Conclusion

Application of EBM in clinical practice raises several ethical challenges, including how to balance the science and art of medicine when making evidence-based decisions for patients, how to manage patients who ask for treatments or interventions that are not evidence based, and how to manage patients who refuse treatments or interventions that are based on evidence. To balance the science and art of medicine, clinicians should synthesize the generalized knowledge garnered from EBM with both their clinical knowledge and skills and the preferences, values, and goals of the individual patient so that they can offer medically appropriate and scientifically grounded treatments that reflect their best medical judgment. Clinicians are not ethically obligated to deliver care that in their medical judgment will not benefit the patient, and because clinicians have the ultimate decision-making authority regarding how care is delivered, patients should not be given treatments simply because they demand them. Although clinicians have autonomy regarding the care they choose to deliver, this charge must be balanced with the ethical and legal right of patients to refuse any medical treatment or intervention, even if it will prolong or preserve their life.

References

1. Zimerman AL. Evidence-based medicine: a short history of a modern medical movement. *Virtual Mentor*. 2013;15(1):71-76.
2. Jones R, Nieto Y, Rizzo JD, et al; Steering Committee for Evidence-Based Reviews of the American Society for Blood and Marrow Transplantation. The evolution of the evidence-based review: evaluating the science enhances the art of medicine—statement of the Steering Committee for Evidence-Based Reviews of the American Society for Blood and Marrow Transplantation. *Biol Blood Marrow Transplant*. 2005;11(11):819-822.
3. Fletcher RH, Fletcher SW, Wagner EH. *Clinical Epidemiology: The Essentials.* Williams & Wilkins; 1982.
4. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ*. 1996;312(7023):71-72.
5. Ahuja RB. Ethical practice of evidence-based medicine: a review for plastic surgeons. *Indian J Plast Surg*. 2013;46(1):11-17.
6. da Mota Gomes M, Haynes RB. William Osler (1849-1919) at the roots of evidence-based medicine. *Can J Gen Intern Med*. 2019;14(4):23-27.
7. Craxì L, Giardina S, Spagnolo AG. A return to humane medicine: Osler's legacy. *Infez Med*. 2017;25(3):292-297.
8. Osler W. *The Evolution of Modern Medicine*. Yale University Press; 1921.
9. American Medical Association. Opinion 5.5 Medically ineffective interventions. *Code of Medical Ethics*. Accessed February 26, 2024. https://code-medical-ethics.ama-assn.org/ethics-opinions/medically-ineffective-interventions
10. American Medical Association. Opinion 11.1.2 Physician stewardship of health care resources. *Code of Medical Ethics*. Accessed July 19, 2024. https://code-medical-ethics.ama-assn.org/ethics-opinions/physician-stewardship-health-care-resources
11. American Medical Association. Opinion 11.2.1 Professionalism in health care systems. *Code of Medical Ethics*. Accessed July 19, 2024. https://code-medical-ethics.ama-assn.org/ethics-opinions/professionalism-health-care-systems
12. American Medical Association. Opinion 1.1.3 Patient rights. *Code of Medical Ethics*. Accessed February 26, 2024. https://code-medical-ethics.ama-assn.org/index.php/ethics-opinions/patient-rights
13. House of Delegates. Physician decision-making in health care systems H-285.954. American Medical Association. Updated 2017. Accessed July 19, 2024. https://policysearch.ama-assn.org/policyfinder/detail/285.954?uri=%2FAMADoc%2FHOD.xml-0-2078.xml
14. American Medical Association. Opinion 11.2.4 Transparency in health care. *Code of Medical Ethics*. Accessed August 7, 2024. https://code-medical-ethics.ama-assn.org/ethics-opinions/transparency-health-care
15. Barstow C, Shahan B, Roberts M. Evaluating medical decision-making capacity in practice. *Am Fam Physician*. 2018;98(1):40-46.

Amber R. Comer, PhD, JD is the director of ethics policy and the secretary of the Council on Ethical and Judicial Affairs at the American Medical Association in Chicago, Illinois. She is also an associate professor of health sciences and medicine at Indiana University. Dr Comer is an expert in medical decision-making for patients with critical illness.

Citation
AMA J Ethics. 2025;27(1):E21-26.

DOI
10.1001/amajethics.2025.21.

Conflict of Interest Disclosure
Author 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

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

AMA Journal of Ethics®
January 2025, Volume 27, Number 1: E34-43

POLICY FORUM: PEER-REVIEWED ARTICLE

What Are High-Quality Race and Ethnicity Data and How Are They Used in Health Equity Research?

Christopher T. Rentsch, PhD, Moneeza K. Siddiqui, PhD, MPH, and Rohini Mathur, PhD, MS

Abstract

The COVID-19 pandemic changed public awareness of the importance of high-quality race and ethnicity data for identifying and redressing widely documented racial and ethnic health inequity. This article emphasizes the importance of high-quality race and ethnicity data in health equity research, as highlighted by the COVID-19 pandemic. The article defines what constitutes high-quality race and ethnicity data, discusses challenges in using these data, and provides 2 cases that illustrate the role of these data in identifying and redressing health inequity. Finally, this article advocates for the use of accurate, standardized, and granular data and highlights the need for community engagement and trust building to improve data quality and research outcomes.

What Are Race and Ethnicity Data?

Race and ethnicity classifications reflect how particular groups of people have been racialized— that is, how their racial or ethnic identity has been shaped by historical and political forces. In particular, the ways racial and ethnic groups are defined depend on social, cultural, political, and geographical context. Although the terms *race* and *ethnicity* have evolved over time, race has historically referred to broad categories of people that are divided arbitrarily based on ancestral origin and physical characteristics.[1] The United States (US) Census Bureau acknowledges that race is "a social definition ... and not an attempt to define race biologically, anthropologically, or genetically."[2] In the US, ethnicity has historically referred to a person's cultural identity (eg, language, customs, religion)—namely, as Hispanic or Latino, Latina, or Latinx.[1] In the United Kingdom (UK), however, the term *ethnicity* encompasses both of the above-mentioned concepts and is defined as the "various ways in which a person may choose to define their ethnic group ... include[ing] common ancestry, elements of culture, identity, religion, language and physical appearance."[3] While the concepts of race and ethnicity are broad social constructs, they do not preclude the existence of biological or genetic variation that may affect health outcomes.[4] In this article, we use both terms—*race* and *ethnicity*—to refer to these social constructs, in line with recent proposals to use unified race and ethnicity categories.[5]

The COVID-19 pandemic changed public awareness of the importance of high-quality race and ethnicity data for identifying and redressing widely documented racial and ethnic health inequity.

In health equity research, concepts of race and ethnicity can be thought of as proxies for structural and individual racism and discrimination.[6,7] In turn, research findings on racial or ethnic health differences, typically reported at a group or community level, are often a proxy for a range of health determinants, including—but not limited to—education, income, employment, housing, beliefs and behaviors, language and culture, and embodied experiences of racism and discrimination.[8] Thus, collecting high-quality data on race and ethnicity can be a key first step to quantifying health inequity, which is needed as a basis for making policies that aim to redress inequity. In this article, we define what constitutes high-quality race and ethnicity data, discuss the challenges in using these data, and provide 2 case studies that illustrate the role of these data in identifying and redressing health inequity.

Characteristics of High-Quality Race and Ethnicity Data

Accurate and comprehensive data on race and ethnicity are critical for conducting effective health equity research to guide policy development. Essential characteristics of high-quality race and ethnicity data include high levels of completeness, self-reported collection, consistency, and granularity, as described below.

As with any data captured in routine health care settings, the completeness of data is related to access and health care usage, even in countries where health care is free at the point of access. Despite universal primary health care in the UK, certain population groups, such as migrants, attend primary care less frequently.[9] These important differences in access can greatly affect the completeness of race and ethnicity data, limiting our ability to redress inequity in populations often with the greatest health care need. The self-report of an individual's own racial or ethnic identity (as opposed to data recorded by an observer based on visual assessment or other indirect methods) is essential for accuracy.[10,11] While an individual's identity might not fit into categories listed, use of consistent and standardized categories during collection and in published research minimizes discrepancies, enhances comparability, and allows for monitoring patterns over time. Greater granularity in racial and ethnic categories allows for better representation of racial and ethnic identities, provided analyses avoid combining relatively smaller groups into an "other" category that potentially obscures inequity. The quantity and validity of standard ethnic categories may evolve over time to reflect the changing ethnic makeup of a population. For example, the "mixed" ethnicity group is the largest growing ethnic group in the UK[12] and in the US,[13] and more granular breakdowns of this high-level, catchall group will be essential for identifying the needs of the population over the long-term.

Pandemic-Prompted Change

The COVID-19 pandemic has highlighted and exacerbated racial and ethnic inequity in health care and health outcomes.[14] Our understanding of this inequity was made possible by research leveraging routinely collected race and ethnicity data available in health care records and insurance claims databases. While several countries[15,16,17] recognize the importance of collecting race and ethnicity data, others consider the collection of such data illegal, making it impossible to directly quantify and redress inequity in these settings.[18,19] Collection of race and ethnicity data is an imperfect system, and current practices often suffer from inconsistencies in self-reported

collection, standardization, and granularity of categories. However, these shortcomings should not preclude the use of existing race and ethnicity data to examine patterns in the health needs of minoritized populations.

The pandemic was a catalyst for change in research culture. The urgent need for responsive research led to widespread changes in how we use, share, and communicate about data. First, the pandemic resulted in initiatives (as demonstrated in the cases below) that improved the speed, safety, and transparency of research. Second—and also related to use—it placed health inequity research in a global spotlight. Early in the pandemic, press reports suggested that racially and ethnically minoritized groups were disproportionately affected by COVID-19 relative to their White counterparts.[20,21,22] Hypotheses included excess occupational exposure to the SARS-CoV-2 virus, greater barriers in accessing health care, and lack of culturally and linguistically appropriate public health communications.[23,24] There was a clear and urgent need to formally evaluate the potential for racial and ethnic inequity associated with the pandemic. Third, the pandemic led to novel collaborations across sectors and disciplines, including community partnerships and engagement. For example, Latino communities in California engaged in community-academic partnerships to develop culturally appropriate health interventions addressing testing barriers.[25] Fourth, the pandemic required researchers to facilitate public understanding to help narrow the "trust gap" between themselves and the public concerning how people's health and administrative data are used for research.[26] These changes in research practice hold promise for more rapidly translating scientific research into policy aimed at redressing health inequity.

Cases

Below, we provide 2 use cases that demonstrate the benefits and challenges of using race and ethnicity data to identify and redress inequity in health care utilization and outcomes. The cases we selected represent health care systems in the US and UK that offer care largely free of charge, thereby minimizing significant cost barriers to health care utilization. However, inequity in access to health care remains in both systems.[27,28] Disentangling the impact of health care access from observed inequity in health outcomes remains a challenge, as any underrepresentation of marginalized groups in the data can compromise the ability to accurately assess and redress health inequity.

Case 1: racial and ethnic disparities in COVID-19 pandemic in the US and UK. In the US, we highlight research leveraging longitudinal electronic health record data from the Department of Veterans Affairs (VA). The VA is the largest integrated health care system in the US and provides comprehensive health care to more than 9 million veterans annually nationwide at over 1300 points of care.[29] Since 2003, the VA has routinely collected self-reported race and ethnicity data during intake and at outpatient and inpatient visits.[11] In the UK, we highlight research conducted using OpenSAFELY,[30] a novel software platform developed on behalf of NHS England to support rapid, responsive research on COVID-19. At its inception in 2020, OpenSAFELY included electronic health records that contained self-reported ethnicity[31] for 25 million people, covering 40% of the English population.[32]

Within VA data, researchers identified stark disparities among racial and ethnic minoritized groups in the risk or prevalence of testing positive for COVID-19[33,34,35] and in COVID-19 hospitalizations.[36] However, among those who tested positive, there were no observed disparities in subsequent mortality,[33] which has been attributed to the care

received in the VA health care system, as health disparities in the VA tend to be smaller than in the private sector.[37]

Nevertheless, at a population level, the substantial excess burden of SARS-CoV-2 infection among racially and ethnically minoritized groups inevitably translated to excess mortality in these communities in the US[38] and UK.[39] In the US, American Indian and Alaska Native (AI/AN) patients "experienced the largest absolute and relative increases in mortality during the pandemic," although they represented only 1% of the VA population.[40] The OpenSAFELY studies found similar ethnic disparities in testing positive, hospitalization, and mortality.[41] In the UK, these data were used to additionally identify factors—such as living in deprived areas[42] and residing in large, multigenerational households[43]—associated with SARS-CoV-2 infection and mortality. Thanks to large sample sizes, researchers were able to undertake comparisons among more granular ethnicity groups, which identified widening inequity in COVID-19 mortality among South Asian groups, especially the Bangladeshi community, in the second wave of the pandemic. These findings led to further work in which the crude household size variable was redefined as a measure of multigenerational living. This work showed that 66% of people of South Asian ethnicity live in multigenerational households compared to 23% of White groups and 49% of Black groups and that multigenerational living and living alone were both associated with increased risk of COVID-19.[43] In both countries, however, the lack of data on wider social determinants of health, such as employment and contact patterns, in large-scale electronic health record systems limited investigating these factors further.

Despite these limitations, the rapid, responsive way of working during the pandemic meant that researchers in both countries were collaborating in large, multidisciplinary teams, enabling rapid transformation of research findings into responsive policy recommendations, including for tailored, culturally responsive public health messaging concerning prevention and, eventually, vaccination. For example, the VA created a COVID-19 Equity Dashboard to track and visualize infection and vaccination rates by race and ethnicity and other demographic factors, enabling targeted outreach and intervention.[44] Additionally, the VA conducted virtual listening sessions between veterans of color and demographic-matched professionals to increase vaccination rates, which were crucial for building trust and for addressing vaccine hesitancy and historical injustices in medicine.[44] In the UK, targeted communication and engagement strategies, such as leveraging local influencers through the Community Champions scheme and utilizing flexible deployment models that support vaccinations during religious events and in places of worship, were essential to improving vaccine uptake among ethnic minorities and combatting misinformation.[45]

To maximize transparency and trust in its research, each study conducted using the OpenSAFELY platform is required to preregister a complete study protocol and publicly share all code that extracts and analyzes data.[46,47] This transparency aims to assure all stakeholders—including patients, professionals, and policy makers—that data were used as intended and handled and interpreted appropriately.

Case 2: using ethnicity data to develop targeted public health interventions. For over 30 years, the Clinical Effectiveness Group (CEG) at Queen Mary, University of London, has utilized electronic health record data to generate valuable insights and innovations, thereby facilitating health and social care improvements. The CEG enhances learning health systems in one of London's most diverse and deprived areas, the borough of

Tower Hamlets. By employing a cycle of analysis, feedback, and interaction, the CEG effectively bridges research, policy, and practice, driving public health advancements and reducing inequity.

The learning health system at work is demonstrated in redressing ethnic inequity in measles mumps and rubella (MMR) vaccination. It was found that "between 2006 and 2008 … Tower Hamlets had the highest rates of confirmed measles [in the UK], with 24 cases per 100 000 … compared with a national figure of 2 per 100 000."[48] Using routinely collected primary care data, the CEG was able to demonstrate significant ethnic inequity in MMR uptake. In Tower Hamlets, focus group work with Somali parents suggested that MMR vaccine uptake was low partly on account of safety concerns related to autism. Thanks to high-quality ethnicity recording (which was over 97% complete for children under 5), the researchers were able to analyze data for the Somali group separately from the broader ethnic category of Black African/Caribbean.

By 2011, Tower Hamlets had virtually achieved herd immunity and had the highest rates of MMR vaccination in London, thanks to efforts that were responsive to the local context.[48] The CEG demonstrated that achieving herd immunity for childhood vaccinations was an achievable goal in an ethnically and socially diverse population. The high-quality ethnicity data available to researchers allowed them "to identify characteristics of the difficult to reach groups, including significant differences in uptake across different ethnicities."[48]

Changes in management and the withdrawal of financial incentives meant that the gains were not sustained long-term. Ten years later, MMR immunization rates in London dropped to levels disproportionately lower than the rest of the UK, partly due to the pandemic.[49] Inequity widened, prompting renewed efforts to reach herd immunity for MMR. In February 2022, the CEG launched a quality improvement program to redress falling rates of childhood immunizations. Research is now underway to fully evaluate the program, which will generate the evidence base to inform practice and policy going forward.[50] One suggested policy action is to include national measures to tackle these inequities by financially incentivizing general practitioners to deliver timely routine childhood vaccinations in primary care.[49,50,51]

Current Key Challenges

Achieving representative data collection presents significant challenges, especially in diverse populations in which socioeconomic inequity, access to health care, and geographic location can influence data quality and availability. It is further complicated in systems where race and ethnicity data collection can be skewed by the nature of health care provision. Although health care systems like the VA or the UK's National Health Service are largely free at the point of contact, those who are marginalized might be less likely to interact with health care systems and be represented in the data.

While the above cases constitute positive examples of using existing large-scale race and ethnicity data, data injustices remain. For example, the term *data genocide* has been used to describe the lack of AI/AN data available in the US during the pandemic.[52] As a result, AI/AN communities exercised communal ownership of health data to drive public health responses tailored to their specific needs.[53] Greater community engagement is crucial in redressing health inequity and building trust between researchers and marginalized communities.

To overcome these challenges, as a start, we point to recent guidance on the reporting of race and ethnicity in scientific research.[54] We also note a call for action to bring about data justice "regarding the reporting and analysis of publicly-funded work involving racialized groups."[6]

Conclusion

Ensuring high-quality race and ethnicity data through collection of self-reported, standardized, and granular data is crucial for meaningful analysis aimed at identifying health inequity. Provided that researchers discuss limitations in the collection and classification of data, analyzing data by race and ethnicity can yield crucial insights into health patterns and serve as a critical basis for redressing health inequity.

References

1. Stamper K. Why we confuse race and ethnicity: a lexicographer's perspective. Conscious Style Guide. February 13, 2019. Accessed May 8, 2021. https://consciousstyleguide.com/why-we-confuse-race-ethnicity-lexicographers-perspective/
2. About the topic of race. US Census Bureau. Revised March 1, 2022. Accessed June 27, 2024. https://www.census.gov/topics/population/race/about.html
3. Ethnic group, England and Wales: Census 2021. Office for National Statistics. November 29, 2022. Accessed June 27, 2024. https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/bulletins/ethnicgroupenglandandwales/census2021
4. Borrell LN, Elhawary JR, Fuentes-Afflick E, et al. Race and genetic ancestry in medicine—a time for reckoning with racism. *N Engl J Med*. 2021;384(5):474-480.
5. Flores G. Language barriers and hospitalized children: are we overlooking the most important risk factor for adverse events? *JAMA Pediatr*. 2020;174(12):e203238.
6. Krieger N. Structural racism, health inequities, and the two-edged sword of data: structural problems require structural solutions. *Front Public Health*. 2021;9:655447.
7. Lett E, Asabor E, Beltrán S, Cannon AM, Arah OA. Conceptualizing, contextualizing, and operationalizing race in quantitative health sciences research. *Ann Fam Med.* 2022;20(2):157-163.
8. Social determinants of health. World Health Organization. Accessed June 27, 2024. https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1
9. Zhang CX, Boukari Y, Pathak N, et al. Migrants' primary care utilisation before and during the COVID-19 pandemic in England: an interrupted time series analysis. *Lancet Reg Health Eur*. 2022;20:100455.
10. Bastos JL, Peres MA, Peres KG, Dumith SC, Gigante DP. Socioeconomic differences between self- and interviewer-classification of color/race. Article in Portuguese. *Rev Saude Publica*. 2008;42(2):324-334.
11. Sohn MW, Zhang H, Arnold N, et al. Transition to the new race/ethnicity data collection standards in the Department of Veterans Affairs. *Popul Health Metr*. 2006;4(1):7.
12. Fothergill L. Census reveals new chapter in story of mixed-race Britain. Migration Museum. December 7, 2022. Accessed August 15, 2024. https://www.migrationmuseum.org/census-reveals-new-chapter-in-story-of-mixed-race-

britain/#:~:text=Source%3A%20Office%20for%20National%20Statistics%20–%20Census%202021&text=The%20number%20of%20people%20identifying%20as%20'White%20and%20Asian'%20rose,%3A%20up%2061%25%20to%20467%2C113
13. Parker K, Horowitz JM, Morin R, Lopez MH. Multiracial in America: proud, diverse and growing in numbers. Pew Research Center. June 11, 2015. Accessed August 15, 2024. https://www.pewresearch.org/social-trends/2015/06/11/multiracial-in-america/
14. Katikireddi SV, Lal S, Carrol ED, et al. Unequal impact of the COVID-19 crisis on minority ethnic groups: a framework for understanding and addressing inequalities. *J Epidemiol Community Health*. 2021;75(10):970-974.
15. Improving how we report ethnicity. New Zealand Ministry of Social Development. Accessed June 27, 2024. https://www.msd.govt.nz/about-msd-and-our-work/tools/how-we-report-ethnicity.html
16. Deb S, Sud M, Coburn N, et al. Race and ethnicity research in cardiovascular disease in Canada: challenges and opportunities. *Can J Cardiol*. 2024;40(6):1172-1175.
17. Anjana RM, Unnikrishnan R, Deepa M, et al; ICMR-INDIAB Collaborative Study Group. Metabolic non-communicable disease health report of India: the ICMR-INDIAB national cross-sectional study (ICMR-INDIAB-17). *Lancet Diabetes Endocrinol*. 2023;11(7):474-489.
18. Chopin I, Niessen J, eds. *Combating Racial and Ethnic Discrimination: Taking the European Legislative Agenda Further*. Commission for Racial Equality; Migration Policy Group; 2002. Accessed June 27, 2024. https://www.migpolgroup.com/_old/wp-content/uploads/2016/10/81.CombatingRacialandEthnicDiscrimination-TakingtheEuropeanLegislativeAgendaFurther_03.02.pdf
19. Al-Zubaidi Y. Racial and ethnic statistics in Sweden: has the socialization process started yet? In: Carlson L, ed. *Equality*. Stockholm Institute for Scandinavian Law; 2022:425-450. *Scandinavian Studies in Law*. Vol 68. Accessed June 27, 2024. https://scandinavianlaw.se/pdf/68-18.pdf
20. Aldridge RW, Lewer D, Katikireddi SV, et al. Black, Asian and minority ethnic groups in England are at increased risk of death from COVID-19: indirect standardisation of NHS mortality data. *Wellcome Open Res*. 2020;5:88.
21. Reyes C, Husain N, Gutowski C, St Clair S, Pratt G. Chicago's coronavirus disparity: Black Chicagoans are dying at nearly six times the rate of white residents, data show. *Chicago Tribune*. April 7, 2020. Updated April 8, 2020. Accessed June 28, 2024. https://www.chicagotribune.com/2020/04/07/chicagos-coronavirus-disparity-black-chicagoans-are-dying-at-nearly-six-times-the-rate-of-white-residents-data-show/
22. Thebault R, Tran AB, Williams V. The coronavirus is infecting and killing black Americans at an alarmingly high rate. *Washington Post*. April 7, 2020. Accessed June 28, 2024. https://www.washingtonpost.com/nation/2020/04/07/coronavirus-is-infecting-killing-black-americans-an-alarmingly-high-rate-post-analysis-shows/
23. Fothergill A, Maestas EG, Darlington JD. Race, ethnicity and disasters in the United States: a review of the literature. *Disasters*. 1999;23(2):156-173.
24. Webb Hooper M, Nápoles AM, Pérez-Stable EJ. COVID-19 and racial/ethnic disparities. *JAMA*. 2020;323(24):2466-2467.

25. Garibay KK, Durazo A, Vizcaíno T, et al. Lessons from two Latino communities working with academic partners to increase access to COVID-19 testing. *Prog Community Health Partnersh*. 2024;18(1):1-9.
26. Mathur R, Rentsch CT, Venkataraman K, et al. How do we collect good-quality data on race and ethnicity and address the trust gap? *Lancet*. 2022;400(10368):2028-2030.
27. Ajayi Sotubo O. A perspective on health inequalities in BAME communities and how to improve access to primary care. *Future Healthc J*. 2021;8(1):36-39.
28. Ward RE, Nguyen XT, Li Y, et al; VA Million Veteran Program. Racial and ethnic disparities in US veteran health characteristics. *Int J Environ Res Public Health*. 2021;18(5):2411.
29. Veterans Health Administration: about VHA. US Department of Veterans Affairs. Updated September 12, 2024. Accessed October 1, 2024. https://www.va.gov/health/aboutvha.asp
30. Secure analytics platform for NHS electronic health records. OpenSAFELY. Accessed June 27, 2024. https://www.opensafely.org/
31. Hull SA, Mathur R, Badrick E, Robson J, Boomla K. Recording ethnicity in primary care: assessing the methods and impact. *Br J Gen Pract*. 2011;61(586):e290-e294.
32. Andrews CD, Mathur R, Massey J, et al; OpenSAFELY Collaborative. Consistency, completeness and external validity of ethnicity recording in NHS primary care records: a cohort study in 25 million patients' records at source using OpenSAFELY. *BMC Med*. 2024;22(1):288.
33. Rentsch CT, Kidwai-Khan F, Tate JP, et al. Patterns of COVID-19 testing and mortality by race and ethnicity among United States veterans: a nationwide cohort study. *PLoS Med*. 2020;17(9):e1003379.
34. Ferguson JM, Abdel Magid HS, Purnell AL, Kiang MV, Osborne TF. Differences in COVID-19 testing and test positivity among veterans, United States, 2020. *Public Health Rep*. 2021;136(4):483-492.
35. Ferguson JM, Justice AC, Osborne TF, Magid HSA, Purnell AL, Rentsch CT. Geographic and temporal variation in racial and ethnic disparities in SARS-CoV-2 positivity between February 2020 and August 2021 in the United States. *Sci Rep*. 2022;12(1):273.
36. Razjouyan J, Helmer DA, Li A, et al. Differences in COVID-19-related testing and healthcare utilization by race and ethnicity in the Veterans Health Administration. *J Racial Ethn Health Disparities*. 2022;9(2):519-526.
37. Peterson K, Anderson J, Boundy E, Ferguson L, McCleery E, Waldrip K. Mortality disparities in racial/ethnic minority groups in the Veterans Health Administration: an evidence review and map. *Am J Public Health*. 2018;108(3):e1-e11.
38. Weinberger DM, Rose L, Rentsch C, et al. Excess mortality among patients in the Veterans Affairs health system compared with the overall US population during the first year of the COVID-19 pandemic. *JAMA Netw Open*. 2023;6(5):e2312140.
39. Strongman H, Carreira H, De Stavola BL, Bhaskaran K, Leon DA. Factors associated with excess all-cause mortality in the first wave of the COVID-19 pandemic in the UK: a time series analysis using the Clinical Practice Research Datalink. *PLoS Med*. 2022;19(1):e1003870.
40. Weinberger DM, Bhaskaran K, Korves C, et al. Excess mortality in US veterans during the COVID-19 pandemic: an individual-level cohort study. *Int J Epidemiol*. 2023;52(6):1725-1734.

41. Mathur R, Rentsch CT, Morton CE, et al; OpenSAFELY Collaborative. Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: an observational cohort study using the OpenSAFELY platform. *Lancet*. 2021;397(10286):1711-1724.
42. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. 2020;584(7821):430-436.
43. Wing K, Grint DJ, Mathur R, et al. Association between household composition and severe COVID-19 outcomes in older people by ethnicity: an observational cohort study using the OpenSAFELY platform. *Int J Epidemiol*. 2022;51(6):1745-1760.
44. Leder SC, List JM, Chandra R, Jones KT, Moy E. VA research and operations uniting to combat COVID-19 inequities. *Health Equity*. 2023;7(1):296-302.
45. Third quarterly report on progress to address COVID-19 health inequalities. Gov.UK. May 2021. Updated September 3, 2021. Accessed June 27, 2024. https://www.gov.uk/government/publications/third-quarterly-report-on-progress-to-address-covid-19-health-inequalities/third-quarterly-report-on-progress-to-address-covid-19-health-inequalities
46. OpenSAFELY. GitHub. Accessed February 27, 2024. https://github.com/OpenSAFELY
47. OpenSAFELY jobs. OpenSAFELY. Accessed February 27, 2024. https://jobs.opensafely.org/
48. Cockman P, Dawson L, Mathur R, Hull S. Improving MMR vaccination rates: herd immunity is a realistic goal. *BMJ*. 2011;343:d5703.
49. Firman N, Marszalek M, Gutierrez A, et al. Impact of the COVID-19 pandemic on timeliness and equity of measles, mumps and rubella vaccinations in North East London: a longitudinal study using electronic health records. *BMJ Open*. 2022;12(12):e066288.
50. Marszalek M, Hawking MKD, Gutierrez A, et al. Implementation of a quality improvement programme using the Active Patient Link call and recall system to improve timeliness and equity of childhood vaccinations: protocol for a mixed-methods evaluation. *BMJ Open*. 2023;13(1):e064364.
51. Primary Care Strategy and NHS Contracts Group. *Update to the GP Contract Agreement 2020/21-2023/24*. British Medical Association; NHS England; 2020. Accessed June 27, 2024. https://www.england.nhs.uk/wp-content/uploads/2020/03/update-to-the-gp-contract-agreement-v2-updated.pdf
52. Data genocide of American Indians and Alaska Natives in COVID-19 data. Urban Indian Health Institute. Accessed June 27, 2024. https://www.uihi.org/projects/data-genocide-of-american-indians-and-alaska-natives-in-covid-19-data/
53. Huyser KR, Horse AJY, Kuhlemeier AA, Huyser MR. COVID-19 pandemic and Indigenous representation in public health data. *Am J Public Health*. 2021;111(suppl 3):S208-S214.
54. Flanagin A, Frey T, Christiansen SL; AMA Manual of Style Committee. Updated guidance on the reporting of race and ethnicity in medical and science journals. *JAMA*. 2021;326(7):621-627.

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.

Moneeza K. Siddiqui, PhD, MPH is a lecturer in genetic epidemiology at Queen Mary University of London in England who previously served as a principal investigator in precision medicine at the University of Dundee in Scotland, where she led research in pharmacogenetics and type 2 diabetes. She obtained an MPH from Columbia University and a PhD from the University of Dundee. Her research focuses on comparisons across ancestries using genetics and multi-omics methods to understand the heterogenous presentation of type 2 diabetes in South Asians.

Rohini Mathur, PhD, MS is a professor and the chair of health data science at Queen Mary University of London in England. She is also the academic lead of the Clinical Effectiveness Group, a clinically driven and academically supported quality improvement and research center, where she leads research with local and international partners. She obtained an MS and PhD from the London School of Hygiene & Tropical Medicine. Specializing in health inequalities research, she aims to generate an evidence base to inform tailored approaches to the management of cardiometabolic disease by harnessing global data on pathophysiology, clinical outcomes, and treatment response in ethnically and geographically diverse populations.

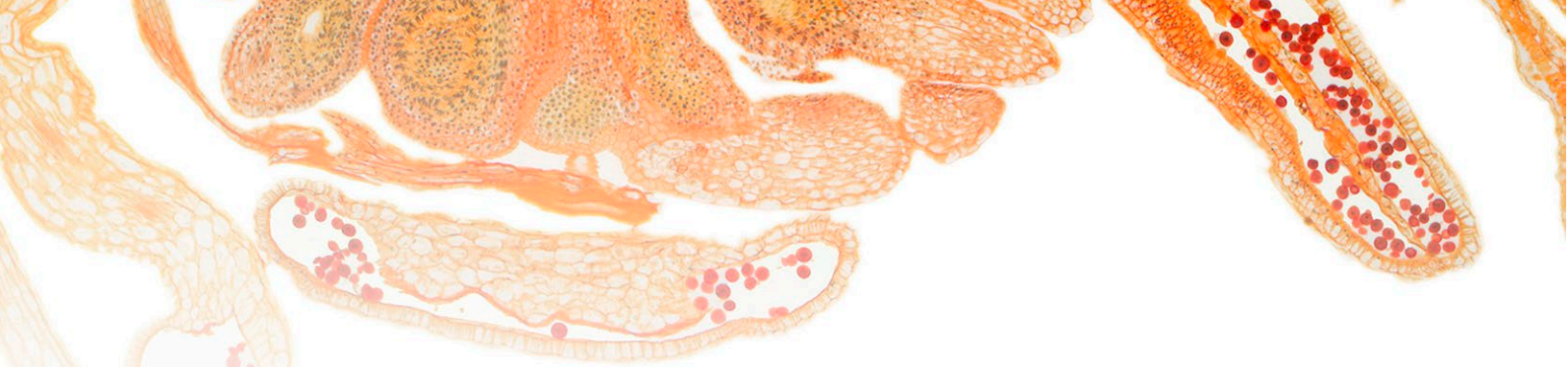
Citation
AMA J Ethics. 2025;27(1):E34-43.

DOI
10.1001/amajethics.2025.34.

Conflict of Interest Disclosure
Dr Mathur's salary is partly funded through a life sciences consortium, which includes numerous pharmaceutical companies. The other 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



AMA Journal of Ethics®

January 2025, Volume 27, Number 1: E44-50

POLICY FORUM: PEER-REVIEWED ARTICLE

How Should Epidemiologists Respond to Data Genocide?

Abigail Echo-Hawk, MA, Sofia Locklear, PhD, Sarah McNally, MPH, Lannesse Baker, MPH, and Sacena Gurule, MPA

Abstract

Data quality for and about American Indian/Alaska Native (AI/AN) people is undermined by deeply entrenched, colonial practices that have become standard in US federal data systems. This article draws on cases of maternal mortality and COVID-19 to demonstrate the ethical and clinical need for inclusive, diverse, and accurate data when researching AI/AN health trends. This article further argues that epidemiologists specifically must challenge implicit bias, question methods and practices, and recognize colonial, racist reporting practices about AI/AN people that have long undermined data collection, analytical, and dissemination practices that are fundamental to epidemiological research.

The American Medical Association designates this journal-based CME activity for a maximum of 1 AMA PRA Category 1 Credit™ *available through the* AMA Ed Hub™. *Physicians should claim only the credit commensurate with the extent of their participation in the activity.*

First Data Gatherers

American Indian and Alaska Native (AI/AN) people were the first data gatherers in what is now called the United States.[1] Indigenous communities have consistently been empirically rigorous, collecting both quantitative and qualitative data for health and well-being purposes.[2] Prior to colonization, AI/AN individuals and communities had robust health, and AI/AN health care practices have long been utilized in Western medicine.[3] Indigenous knowledge is still passed down through generations, despite settler colonialism's initiation in the late 1400s as one of the most influential social determinants of AI/AN health.[4,5,6] Settler colonial logic is a "logic of elimination,"[4] whereby settler colonizers purposefully try to deplete and eliminate original people and their cultures through genocide. Less widely known is how settler colonial genocidal practices have influenced data.[7]

Data Sovereignty

Prior scholars have deemed the exclusion of AI/AN people from federal data, such as the US census, to be "statistical genocide."[7] The Urban Indian Health Institute (UIHI)—a division of the Seattle Indian Health Board and the only national Tribal Epidemiology Center that serves urban dwelling AI/AN populations by providing public health support

through data, research, and evaluation—considers statistical genocide to be a form of data erasure contributing directly to a larger colonial project of data genocide. Data genocide, as defined by UIHI, is "the elimination of Indigenous people in data resulting in the non-fulfillment of treaty and trust responsibilities due to 'lack' of data on urban and rural tribal communities."[8] Data genocide also includes the erasure of Indigenous people through aggregating data and misclassifying Indigenous people within datasets. Even when collected, any data about nation-based Indigenous people in the United States must respect federal treaty rights, a tenet of which is Indigenous data sovereignty. Indigenous data sovereignty is the right of each Tribe to exercise sovereignty over the collection, ownership, and application of data that aligns Indigenous customs, values, and ways of knowing.[9] Data sovereignty extends to any health information collected about Indigenous people and must be respected to ensure that collection and use of the data align with Indigenous principles and is guaranteed by the United Nations Declaration of the Rights of Indigenous People, which the United States announced support for in 2010.[10]

Data Invisibility

One striking example of data genocide is the invisibility of AI/AN people in maternal mortality rates. AI/AN women, along with Black women, have some of the highest rates of pregnancy-related mortality deaths, with a significant increase seen in 2021 associated with the COVID-19 pandemic.[11] Yet this fact often goes ignored in most analyses of maternal mortality rates, with AI/AN people being lumped into an "other" category, thereby erasing their racial and political identity as Indigenous and eliminating the ability to disaggregate the data and identify disparate outcomes for this group. Collapsing racial and ethnic data into an other category is often rationalized by small sample sizes. Yet data genocide—through individuals being racially misclassified within federal data sets—contributes to shrinking the sample size.[7,8] Through racial misclassification, Indigenous people are made invisible while simultaneously being labeled as "other." Consequently, calculation of maternal mortality deaths, which are linked to the social determinants of health,[12] now lies in the hands of a system that determined that AI/AN birthing people were too small of a population to separate out—or to do so precisely—within statistical analyses,[13,14] making invisible the reality of maternal mortality for AI/AN women. These practices are racist because they reify settler colonial power's embeddedness in data systems, data analysis, and data dissemination by not collecting and reporting data on Indigenous people's race and ethnicity.

This problem is avoidable. Yet it is further exacerbated by common data practices spanning collection to dissemination. The use of a single-race AI/AN category illustrates how these data practices are rooted in data genocide.[8] Despite AI/AN being one of "the largest growing multi-racial groups in the United States,"[15] it is common practice for government, academic, and other agencies to use only a single-race AI/AN category in their analyses, effectively shrinking the sample size of specific groups through dilution, potentially overlooking statistically significant differences, and upholding a former colonial practice by the US government to determine who was AI/AN based on blood quantum.[16] There is no scientifically valid reason to use only a single-race AI/AN category in data analysis and dissemination, and, as a result of the advocacy of tribal nations, only Tribes, not the US government, can determine who is a tribal member.[17] Yet statistical and other agencies continue to use this outdated, nonscientific, and colonial data practice. The authors recognize this practice as structural racism in data. To uproot this structural racism, the field of epidemiology must challenge implicit bias, question what has become standard methodological practice, and recognize the

unintended and very real consequences of this practice and other colonial data practices on AI/AN and other populations, such as Pacific Islanders, impacted by ongoing colonialism.

The UIHI's report, "Data Genocide of American Indians and Alaska Natives in COVID-19 Data,"[8] which discussed AI/AN COVID-19 data reporting for all US states, illustrated the detrimental effect of the elimination of AI/AN in data, as it resulted in misallocation of federal funds meant to address the pandemic,[18] despite AI/AN being one of the groups most detrimentally affected by the virus.[19] In one of the first studies published on COVID-19 infection rates in AI/AN, the authors were only able to include data reported to the Centers for Disease Control and Prevention for 23 states, as the rest of the nation was not reporting a minimum of 70% complete race and ethnicity data, effectively limiting the understanding of the virus in a paper that was intended to inform public health and clinical practice.[19] Colonial data practices effectively prohibit researchers and clinicians from accessing the information they need to make data-driven decisions in research, policy, programming, and practice. In order to attain health equity, these practices must be challenged. For example, an individual can be an enrolled member of a federally recognized Tribe and categorized as American Indian, yet at the same time be racialized on a phenotypical level as Black or White, resulting in racial misclassification in medical records. These differences in racial reporting are crucial to capture within the data, as AI/AN patients can have significantly worse COVID health outcomes, for example, than White or Black patients.[20] Disaggregated data on Indigenous peoples' Tribal and community belonging, race, and ethnicity is vital in order for researchers to fully understand the diverse and complex picture of Indigenous health.[21]

Consequences of Data Genocide

Ongoing data genocide contributes to social theories of health inequalities like "deaths of despair" to explain why non-Hispanic White mortalities due to suicide, drug overdose, and alcoholic liver disease exceed the death rates of other racial groups, while ignoring the extreme health inequity Indigenous people experience.[22] In fact, the validity of such theories is challenged when data about AI/AN people are appropriately included in analyses.[22] Excluding AI/AN people from the data or subsuming them (thereby rendering them invisible) under an other category harms not only Indigenous people themselves—as inaccurate pictures of their colonially imposed health inequity due to data genocide are presented—but also those in other racial groups, as data genocide of Indigenous people misrepresents the data and promotes misunderstanding of health inequity among persons and communities designated as other. While Friedman et al demonstrate that, indeed, Indigenous people are experiencing much higher rates of deaths of despair than their non-Hispanic White counterparts,[23] we strongly stand against the language of "despair" when analyzing deaths of any type for any racial group. This phrasing places blame on an individual's emotional states and emotional points of intolerance instead of framing these deaths within the uninhabitable structures that settler colonialism and capitalism created. Data genocide has implications for the lived experiences of today's AI/AN people and communities. Genocide happening within data collection, analysis, and dissemination hides lived realities of poor health outcomes, such as the alarmingly high rates of maternal child mortality for Indigenous women, and masks the contemporary ways in which settler colonialism affects AI/AN persons' and communities' health.

Rethinking Data Practices

To address data genocide at a basic level within clinical data collection and analysis, several small changes can be made. UIHI's "Best Practices for American Indian and Alaska Native Data Collection"[15] recommends a myriad of best practices that are grounded in and stem from Indigenous values. This framework specifies ensuring that any data collected about AI/AN people include a multiracial category and that those people are counted in the AI/AN category during analysis. Or, in other words, "if the AI/AN individual identifies as another race, include the individuals who are AI/AN in any combination with any other race and include those who identify as Latinx/Hispanic. In the event the definition cannot be as inclusive as stated above, the next less inclusive definition should be used, i.e. AI/AN alone."[15] International efforts led by the Māori Indigenous Sovereignty Network in New Zealand include creating a platform for Māori Tribal information managers to access existing government datasets, to which they then can add their own Tribal data and analysis; the platform is an efficient tool at merging governmental data with supplementary Tribally collected and owned data.[24]

The UIHI's "Best Practices" also identifies opportunities to train staff, doctors, and data analysts on proper race data collection.[15] Such training includes an understanding of race as a social construction and not as biological essentialism,[25] learning about the political status of AI/AN individuals and Tribes, and understanding the impacts of racialization on health and the various ways in which these impacts must be captured in our ever-growing multiracial society. Additionally, epidemiologists must be trained on small population methodologies and Indigenous statistics[26] for quantitative data analyses. Yet it isn't just quantitative data about AI/AN people that must be meaningfully included; qualitative data must also be collected that can add rich nuance to our understandings of Indigenous health. Last, and most important, those who collect data should engage in conversations with local Tribes and urban Native communities on Indigenous data sovereignty and what data collection practices work best for their communities and geographies.

References

1. Rodriguez-Lonebear D. Building a data revolution in Indian country. In: Kukutai T, Taylor J, eds. *Indigenous Data Sovereignty: Toward an Agenda*. Australian National University Press; 2016:253-274. *Centre for Aboriginal Economic Policy Research Monographs*; vol 38.
2. Decolonize data: accurate data tells accurate stories. Urban Indian Health Institute. Accessed February 28, 2024. https://www.uihi.org/projects/decolonizing-data-toolkit/
3. Redvers N, Blondin B. Traditional Indigenous medicine in North America: a scoping review. *PLoS One*. 2020;15(8):e0237531.
4. Wolfe P. Settler colonialism and the elimination of the native. *J Genocide Res*. 2006;8(4):387-409.
5. Wispelwey B, Tanous O, Asi Y, Hammoudeh W, Mills D. Because its power remains naturalized: introducing the settler colonial determinants of health. *Front Public Health*. 2023;11:1137428.
6. McKay DL, Vinyeta K, Norgaard KM. Theorizing race and settler colonialism within US sociology. *Sociol Compass*. 2020;14(9):e12821.
7. Anner J. To the US Census Bureau, Native Americans are practically invisible. *Minor Trendsetter*. 1990;4(1):15-21.
8. Data genocide of American Indians and Alaska Natives in COVID-19 data. Urban Indian Health Institute. February 15, 2021. Accessed August 21, 2024.

https://www.uihi.org/projects/data-genocide-of-american-indians-and-alaska-natives-in-covid-19-data/
9. Kukutai T, Taylor J, eds. *Indigenous Data Sovereignty: Toward an Agenda*. Australian National University Press; 2016. *Centre for Aboriginal Economic Policy Research Monographs*; vol 38.
10. United Nations. United Nations Declaration on the Rights of Indigenous Peoples. *Hum Rights Q*. 2007;33(3):909-921.
11. Thoma ME, Declercq ER. Changes in pregnancy-related mortality associated with the coronavirus disease 2019 (COVID-19) pandemic in the United States. *Obstet Gynecol*. 2023;141(5):911-917.
12. Wang E, Glazer KB, Howell EA, Janevic TM. Social determinants of pregnancy-related mortality and morbidity in the United States: a systematic review. *Obstet Gynecol*. 2020;135(4):896-915.
13. Trost SL, Beauregard J, Njie F, et al. Pregnancy-related deaths among American Indian or Alaska Native persons: data from maternal mortality review committees in 36 US states, 2017-2019. Centers for Disease Control and Prevention. May 28, 2024. Accessed July 3, 2024. https://www.cdc.gov/maternal-mortality/php/data-research/2017-2019-aian.html
14. Hoyert DL. Maternal mortality rates in the United States, 2021. Centers for Disease Control and Prevention. Reviewed March 16, 2023. Accessed July 3, 2024. https://www.cdc.gov/nchs/data/hestat/maternal-mortality/2021/maternal-mortality-rates-2021.htm
15. Best practices for American Indian and Alaska Native data collection. Urban Indian Health Institute. Updated August 26, 2020. Accessed February 28, 2024. https://www.uihi.org/download/best-practices-for-american-indian-and-alaska-native-data-collection/
16. Haozous EA, Strickland CJ, Palacios JF, Solomon TGA. Blood politics, ethnic identity, and racial misclassification among American Indians and Alaska Natives. *J Environ Public Health*. 2014;2014:321604.
17. Tribal enrollment process. US Department of the Interior. Accessed July 3, 2024. https://www.doi.gov/Tribes/enrollment#:~:text=Rarely%20is%20the%20BIA%20involved
18. Skinner A, Raifman J, Ferrara E, Raderman W, Quandelacy TM. Disparities made invisible: gaps in COVID-19 data for American Indian and Alaska Native populations. *Health Equity*. 2022;6(1):226-229.
19. Hatcher SM, Agnew-Brune C, Anderson M, et al. COVID-19 among American Indian and Alaska Native persons—23 states, January 31-July 3, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(34):1166-1169.
20. Musshafen LA, El-Sadek L, Lirette ST, Summers RL, Compretta C, Dobbs TE 3rd. In-hospital mortality disparities among American Indian and Alaska Native, Black, and White patients with COVID-19. *JAMA Netw Open*. 2022;5(3):e224822.
21. Huyser KR, Locklear S. Reversing statistical erasure of Indigenous peoples: the social construction of American Indians and Alaska Natives in the United States using national data sets. In: Walter M, Kukutai T, Gonzales AA, Henry R, eds. *Handbook of Indigenous Sociology*. Oxford University Press; 2021:247-262.
22. Case A, Deaton A. Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. *Proc Natl Acad Sci U S A*. 2015;112(49):15078-15083.

23. Friedman J, Hansen H, Gone JP. Deaths of despair and Indigenous data genocide. *Lancet*. 2023;401(10379):874-876.
24. Kukutai T. How Indigenous communities in New Zealand are protecting their data. *Science*. 2024;384(6691):eado9298.
25. Smedley A, Smedley BD. *Race in North America: Origin and Evolution of a Worldview*. Westview Press; 2012.
26. Walter M, Andersen C. *Indigenous Statistics: A Quantitative Research Methodology*. Routledge; 2016.

Abigail Echo-Hawk, MA (Pawnee) is the executive vice president of the Seattle Indian Health Board and the director of the Urban Indian Health Institute. She works to support the health and well-being of urban Indian communities and tribal nations across the United States by leading public health professional teams' development of culturally competent and culturally relevant health and policy interventions with tribal and urban Indian communities across the country.

Sofia Locklear, PhD (Lumbee) is currently an assistant professor of sociology at the University of Toronto Mississauga in Ontario, Canada. Previously, Sofia worked at the Urban Indian Health Institute. Her research investigates the racialization of Indigenous people, whiteness, and racial inequities more broadly. Her work currently focuses on housing outcomes for urban Indigenous people living across the United States.

Sarah McNally, MPH (mixed-Tongan and Rotuman) currently works as the Decolonizing Data fellow at the Urban Indian Health Institute. She holds a BS in psychology and in public health from Wayne State University and an MPH in health behavior and health education from the University of Michigan in Ann Arbor. Her professional interests include advancing health equity for Native Hawaiian and Pacific Islanders living in the United States by improving the quality and utilization of federal health data resources.

Lannesse Baker, MPH (Anishinaabe/Turtle Mountain Chippewa) is the public health officer at the Urban Indian Health Institute. She earned an MPH at the University of Minnesota in Minneapolis and is currently a PhD candidate in Indigenous health at the University of North Dakota studying American Indian/Alaska Native maternal child health. She is dedicated to improving the health of AI/AN people though community-driven, solutions-focused work that leads to structural change.

Sacena Gurule, MPA (Bishop Paiute Tribe) is a senior program manager on the special projects team of Urban Indian Health Institute. She completed a bachelor of arts degree in American Indian studies and in sociology-criminology at Fort Lewis College and a master's degree in public administration at California Baptist University.

Citation
AMA J Ethics. 2025;27(1):E44-50.

DOI
10.1001/amajethics.2025.44.

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

AMA Journal of Ethics®
January 2025, Volume 27, Number 1: E51-57

MEDICINE AND SOCIETY: PEER-REVIEWED ARTICLE

What Should Clinicians Know About How Coding Influences Epidemiological Research?

Jennifer Quint, PhD and Alex Brownrigg, PhD, MA

Abstract

Coded health care data from patients' health records are used in epidemiological research, especially on incidence or prevalence of disease; for drug safety monitoring or long-term cohort tracking; and to inform policy making. This article briefly summarizes the evolution of internationally recognized coding ontologies and nomenclature and describes applications of coded electronic health record (EHR) data in day-to-day health care operations, research, auditing, and policy development. This article also illuminates how errors can occur when EHR information is coded, considers errors' consequences, and suggests strategies for mitigating errors and improving overall use of coded EHR data.

A History of Health Care Data Coding

The classification or "coding" of diseases dates back to 17th-century England.[1] At that time, codes were collected as part of the London Bills of Mortality to enable frequent causes of death to be recorded. While "Found dead in the Fields at St. Mary Islington"[1] no longer has a code, a desire to capture such granularity in our health care systems remains today.

What would become known as the *International Classification of Diseases* (*ICD*) coding system was adopted by the International Statistical Institute in 1893, and diagnostic terms were introduced in the sixth revision of the *ICD* in 1948.[2,3] Health care communities had recognized the *ICD* system officially before publication of the first volume of the ninth revision in 1977, at which point the *ICD* was expanded to include additional detail at the subcategory level. With each edition of the *ICD*, the number of codes increases, which facilitates billing and administration and the use of these data for audit and research purposes.

This article briefly summarizes the evolution of internationally recognized coding ontologies and nomenclature and describes applications of coded electronic health record (EHR) data in day-to-day health care operations, research, auditing, and policy development. This article also illuminates how errors can occur when EHR information is

coded, considers errors' consequences, and suggests strategies for mitigating errors and improving overall use of coded EHR data.

Types and Complexity of Codes

In addition to the *ICD*, other coding systems have evolved, the most commonly used of which is the SNOMED CT system, a consistent vocabulary for recording clinical information that is considered to be "the most comprehensive, multilingual clinical healthcare terminology" in existence.[4] SNOMED CT was released in its current format in 2002 as a combination of reference terminology and clinical terms.[5] The currently used coding systems in health care are summarized in the Table. It should be noted that individual *ICD* or SNOMED CT codes are added and retired over time, with the result that multiple codes exist to code for the same condition.[6]

Table. Summary of Coding Systems Currently Used in Health Care

System	Type of coding	Use	Where used
ICD-10	Classification	Statistics, billing	Globally
OPCS-4	Classification	Statistics, billing	UK
Read system	Terminology	Clinical	UK, to be retired
SNOMED CT	Terminology	Clinical	Globally
Dm+d	Terminology	Medicines	UK

Abbreviations: Dm+d, Dictionary of Medicines and Devices; *ICD*, *International Statistical Classification of Diseases and Related Health Problems*; *OPCS, Office of Population, Census and Surveys Classification of Interventions and Procedures*; *SNOMED CT*, *Systemized Nomenclature of Medicine—Clinical Terms*; UK, United Kingdom.

The complexity of coding is likely to increase, given that health care is increasingly reliant on technology and digital medical records. More data sources are becoming available (eg, patient-facing apps and wearable devices), which are linkable to other health care data sources that are accessible, both to patients and for research and policy making. This interconnectivity and accessibility make understanding of the use and accuracy of health care data all the more important. In addition, tools for using the data are becoming more complex, with artificial intelligence (AI) and machine learning algorithms that automate coding being used more frequently.[7] Regardless of the methodology used, however, the accuracy of the coding underpinning EHR data is paramount to the data's usefulness. There is a certain degree of false hope that AI will solve problems that current data strategies cannot (such as identifying individuals at high risk of disease), but the bottom line is that if the coding is not right to begin with, no amount of AI will make data analysis any better.

Beyond the importance of using data for day-to-day health care decisions for an individual, data are used for other reasons, ranging from monitoring quality of care and benchmarking services to measuring public health trends and disease epidemiology. Published papers using these data for research cover a wide variety of topics.[8]

Training Clinicians About Coding

In the United Kingdom (UK), medical coders undertake hospital coding, translating what is written in the medical records into *ICD-10* codes, which are ultimately entered into hospital episode statistics (HES) and Office for National Statistics mortality data. HES are used by national bodies and regulators, including the Department of Health and Social Care and NHS England, for the purpose of health care analytics. The data are also available for research in deidentified format with appropriate permissions. There are

strict rules concerning hospital coding and data entry, and, in the UK, medical coders are trained, as they are in other countries.[9,10] Coders follow algorithms, which include instructions, such as coding a disease in place of symptoms in most cases; if a diagnosis is only possible, it cannot be coded, whereas if it is probable (not an impression or suspected), it can be coded. While there is clear guidance concerning what can be coded and how, there is too often little or no coordination between medical coders and medical staff, with coders having to interpret and decipher what has been written and medical staff not being aware of the nuances of coding rules.[11,12] This compartmentalization can lead to inaccuracies in the data. One example is discrepancies in national respiratory audit data entry by clinicians and therefore spurious case ascertainment results. These discrepancies arise because data that do not meet inclusion criteria for the audit based on coding rules might be entered into the audit anyway by health care professionals.[13] Ultimately, clinical staff are vital in ensuring accurate data acquisition and, ultimately, data quality.

In primary care in the UK, data entry is usually undertaken by health care professionals at the point of inputting the data during a consultation. Codes are often assigned via dropdown menus or attached to keywords in the background of the system. Even in this setting, however, as well as globally,[14,15] health care professionals have minimal training as to the importance of choices of codes used or how they inform policy and contribute to audit and research. There is no formal requirement to teach UK doctors about coding classifications and terminologies, and a recent survey of UK medical schools found huge variation in the importance given to the area.[16]

Consequences of Coding Errors

At an individual level, inaccuracy in a person's medical record can have significant consequences, and, in the UK, data once entered generally cannot be removed, although codes do exist to indicate a disease has resolved. For example, a patient's record could contain a code for a disease they do not have, or there could be ambiguous granularity in diagnostic criteria that makes it difficult for new physicians seeing the patient to make decisions. Moreover, important aspects of care, such as identifying unpaid carers, is often not coded, thereby limiting offers of carer support. Errors can also be problematic at a system and population level.[17]

Knowledge and understanding of systems are essential for accurate use of health care data beyond clinical practice. Data may be missing from the EHR for a variety of reasons (eg, something is unknown or an individual declined to answer), which can introduce bias. Less obviously, health care professionals might be reluctant to code information related to wider determinants of health due to stigma or stereotyping and worries about how it might affect patients' insurance coverage and job prospects. For example, health care professionals might be reluctant to code for a diagnosis, such as HIV, that the patient does not want to disclose if there is concern that insurers or employers could somehow find out about the diagnosis. Moreover, the variety of disease code sets used for clinical or billing purposes can result in different estimates of prevalence. Use of less accurate estimates for resource allocation planning can have a knock-on effect in terms of financial distributions that can ultimately be detrimental to patient care.[18,19] Likewise, use of different disease code sets in research has resulted in mixed findings, such that associations between exposure and outcome variables are found to be present or not,[20] and in the inability to make comparisons due to heterogeneity between coding systems. Inconsistency in results and, ultimately, variability of conclusions can undermine the value of these data for research.

Coding errors not only contribute to biased outcomes but have ethical implications if used by insurance or pharmaceutical companies for personal gain.[21] Companies' primary purpose, however, in using data from EHRs, pharmacy records, and billing and reimbursement documentation, is "to monitor medicine consumption and pharmaceutical spending, and to assess safety and providers' compliance with guidelines."[22] Accurate and objective information is essential to guide policy making and spending and to avoid exacerbating health inequalities, lengthening waiting lists, and inappropriately prioritizing services. The earlier that data—and more complete data—can be made available, the more robust will be estimates and forecasts. However, politicization of epidemiological data can lead to misalignment of incentives and evaluations.[23]

Improving EHR Data Use

Ultimately, there needs to be trust in those using the data. Closer working relationships between health care professionals and medical coders and clinical ownership of codes and data are essential for mitigating errors and improving use of EHR data. Beyond individual efforts, there needs to be regulation and accreditation of health care data professionals and clearly defined roles for health care professionals in supplying context when inputting data. In research studies, reporting of codelists and of algorithms and methodology needs to be transparent so that analyses are reproducible. Audit programs are helpful for improving coding standards and could be undertaken as part of national audit programs for quality improvement. As with any research, integrity is key, and auditors need to be as transparent as possible. As a society, we also need to guard against people exploiting any uncertainty that arises from miscoding (or poor data quality) to advance their own agendas, which leads to a politicization (and mistrust) of health data.

In the same way that researchers would never undertake a clinical trial without clear definitions of endpoints, we should encourage consensus on and standardization of important disease endpoints for observational work using EHR data. Work has been undertaken to harmonize various coding ontologies by mapping to a common data model (eg, Observational Medical Outcomes Partnership), thereby allowing federated data analytics. While these efforts at standardization can speed up research and make cross-country or system comparisons easier to undertake, there is still potential for biased outcomes as the risk of cumulative errors and the complexity of the systems grows.

We must also accept that, in the UK, it will never be appropriate to remove information that has been entered in the EHR. In the same way that if we write something in error in a medical record, we cross it out and date and sign it, there are resolved codes that can be used in the EHR, but it would be inappropriate to ever delete something that has been included.

Conclusion

In the UK, we have moved from paper records to secure data environments in less than 15 years, which is relatively high speed, considering the complexity of health care. Most patients and the public are keen for their data to be used in health management so that it can be based on robust estimates of risk calculated from accurate, standardized data, although they may have questions about how the data will be used in research and by whom.[24] Given that the data are imperfect, it is important for health care professionals

to communicate any limitations, biases, and caveats that can originate from miscoding and that are relevant to day-to-day decision-making. From a public perspective, it is important that policy makers be provided with the highest-quality information to develop policy that prioritizes the right services for people who need them and reduces growing health inequalities.

Few doubt that clinical coding systems have led to improvements in health care research and provided benefits to patients and the public. They have allowed data to be linked at a personal level, enabled more detailed studies and standardized analytics, allowed for real-time analytics, and will provide training data for next-generation AI. Yet further improvements are needed. Standardization is becoming even more important, as once disparate data sources are being linked for federated analyses as part of national and international collaborations. Study findings can be influenced by lack of standardized coding and definitions, as well as by inclusion and exclusion criteria, missing data, and the like, and the effects of these factors are likely to be exacerbated if people train AI and use other new technologies without thorough testing, validation, and understanding of the algorithms. Accordingly, regulation, accreditation, and accountability will be important to maintain the integrity of health data and research.

References

1. Boyce N. Bills of Mortality: tracking disease in early modern London. *Lancet*. 2020;395(10231):1186-1187.
2. Classifications and Terminologies Team. History of the development of the ICD. World Health Organization; 2021. Accessed August 20, 2024. https://cdn.who.int/media/docs/default-source/classification/icd/historyoficd.pdf
3. Hirsch JA, Nicola G, McGinty G, et al. ICD-10: history and context. *AJNR Am J Neuroradiol*. 2016;37(4):596-599.
4. What is SNOMED CT? SNOMED International. Accessed October 3, 2024. https://www.snomed.org/what-is-snomed-ct
5. Overview of SNOMED CT. National Library of Medicine. Reviewed October 14, 2016. Accessed July 3, 2024. https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html
6. MacRae C, Whittaker H, Mukherjee M, et al. Deriving a standardised recommended respiratory disease codelist repository for future research. *Pragmat Obs Res*. 2022;13:1-8.
7. Dong H, Falis M, Whiteley W, et al. Automated clinical coding: what, why, and where we are? *NPJ Digit Med*. 2022;5(1):159.
8. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol*. 2017;106(1):1-9.
9. NCCQ. Institute of Health Records and Information Management. Accessed July 3, 2024. https://www.ihrim.co.uk/education-and-cpd/overseas-students/registration-forms/nccq
10. Building healthcare software—clinical coding, classifications and terminology. NHS England. Updated August 24, 2023. Accessed July 3, 2024. https://digital.nhs.uk/developer/guides-and-documentation/building-healthcare-software/clinical-coding-classifications-and-terminology
11. Terminology and Classifications Delivery Service. *National Clinical Coding Standards ICD-10*. 5th ed. NHS England; 2023. Accessed October 3, 2024. https://classbrowser.nhs.uk/ref_books/ICD-10_2023_5th_Ed_NCCS.pdf

12. *ICD-10-CM Official Guidelines for Coding and Reporting*. Centers for Medicare and Medicaid Services; 2021. Accessed July 3, 2024. https://www.cms.gov/files/document/2021-coding-guidelines-updated-12162020.pdf
13. Singh S, Legg M, Garnavos N, et al. National Asthma and Chronic Obstructive Pulmonary Disease Audit Programme (NACAP): pulmonary rehabilitation clinical audit 2019: clinical audit interim report. Royal College of Physicians; 2020. Accessed July 3, 2024. https://www.rcp.ac.uk/media/1tzfeeqi/nacap_prplusclinical_audit_report_julyplus2020_0.pdf
14. Alyahya MS, Khader YS. Health care professionals' knowledge and awareness of the ICD-10 coding system for assigning the cause of perinatal deaths in Jordanian hospitals. *J Multidiscip Healthc*. 2019;12:149-157.
15. Asadi F, Afkhami S, Asadi F. Promotion of training course on ICD-10 poisoning coding: necessity to adopt preventive strategies. *BMC Med Educ*. 2023;23(1):903.
16. Health Data Research UK; Medical Schools Council; NHS England; NHS Education Scotland. Survey of data science in UK medical school curricula. Health Data Research UK; Medical Schools Council; 2023. Accessed August 27, 2024. https://www.hdruk.ac.uk/wp-content/uploads/2023/11/Survey-of-Data-Science-in-UK-Medical-School-Curricula-Report-August-2023.pdf
17. 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.
18. Stone PW, Osen M, Ellis A, Coaker R, Quint JK. Prevalence of chronic obstructive pulmonary disease in England from 2000 to 2019. *Int J Chron Obstruct Pulmon Dis*. 2023;18(18):1565-1574.
19. Morgan A, Gupta RS, George PM, Quint JK. Validation of the recording of idiopathic pulmonary fibrosis in routinely collected electronic healthcare records in England. *BMC Pulm Med*. 2023;23(1):256.
20. Whittaker H, Rothnie KJ, Quint JK. Exploring the impact of varying definitions of exacerbations of chronic obstructive pulmonary disease in routinely collected electronic medical records. *PLoS One*. 2023;18(11):e0292876.
21. Neubert A, Brito Fernandes Ó, Lucevic A, et al. Understanding the use of patient-reported data by health care insurers: a scoping review. *PLoS One*. 2020;15(12):e0244546.
22. Using routinely collected data to inform pharmaceutical policies. OECD Web Archive. April 19, 2023. Accessed July 3, 2024. https://web-archive.oecd.org/2023-04-20/503807-routinely-collected-data-to-inform-pharmaceutical-policies.htm
23. Wells CR, Galvani AP. Tackling the politicisation of COVID-19 data reporting through open access data sharing. *Lancet Infect Dis*. 2022;22(12):1660-1661.
24. Health data research explained. Health Data Research UK. Accessed July 3, 2024. https://www.hdruk.ac.uk/about-us/what-we-do/health-data-research-explained/

Jennifer Quint, PhD is a professor of respiratory epidemiology in the School of Public Health at Imperial College London in England. She is also an honorary consultant physician in respiratory medicine at the Royal Brompton Hospital and Imperial College London NHS Foundation Trust. She leads the Respiratory Electronic Health Record

group, a clinical epidemiology research group that focuses on maximizing the quality, linkage, and usage of electronic health record data for clinical and research purposes.

Alex Brownrigg, PhD, MA is a data scientist at the National Health Service in England. He was diagnosed with a potentially life-threatening rare disease and witnessed firsthand the importance of health data and clinical coding in diagnosing and treating disease. He has had several roles focused on patient and public engagement with Health Data Research UK and Genomics England and has a special interest in ethical use of data for public benefit.

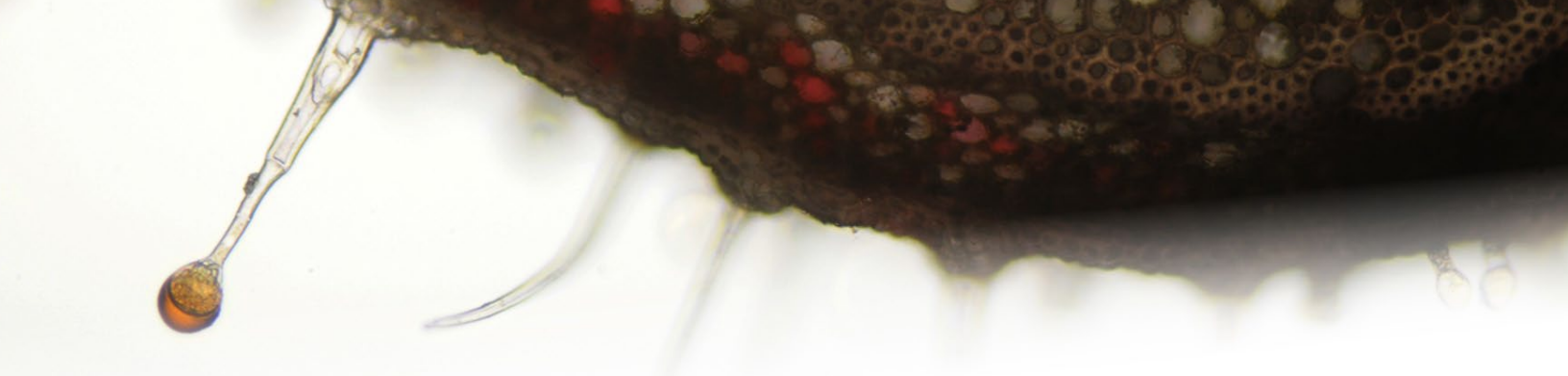
Citation
AMA J Ethics. 2025;27(1):E51-57.

DOI
10.1001/amajethics.2025.51.

Conflict of Interest Disclosure
Dr Quint reported receiving institutional research grants from the Medical Research Council, the National Institute for Health and Care Research, Health Data Research, GlaxoSmithKline, Boehringer Ingelheim, AstraZeneca, Insmed, and Sanofi and was paid for advisory board participation, consulting, or speaking by GlaxoSmithKline, Chiesi, and AstraZeneca. Dr Brownrigg 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



AMA Journal of Ethics®

January 2025, Volume 27, Number 1: E58-63

HISTORY OF MEDICINE: PEER-REVIEWED ARTICLE

Lessons From the Political History of Epidemiology for Divisive Times

H. K. Quinn Valier, PhD

Abstract

Historical precursors of the field we now call epidemiology date back to Hippocrates. Modern epidemiological science, however, developed as domestic and international infectious disease transmission accompanied industrialization, some nations' economic growth, and colonial powers' military expansion and dominance. This article canvasses ways in which modern epidemiology influenced public health innovation from the late 18th century through the mid-19th century. Specifically, this article suggests which lessons can be gleaned from political dimensions of epidemiology's history and applied to orientations to medicine and public health today.

The American Medical Association designates this journal-based CME activity for a maximum of 1 AMA PRA Category 1 Credit™ *available through the* AMA Ed Hub™. *Physicians should claim only the credit commensurate with the extent of their participation in the activity.*

Start With the Greeks

There are numerous ways to date the history of epidemiology. The term *epidemios* (from *epi* [on] and *demos* [people]) appears in the ancient texts of Hippocrates (c 460-c 370 BCE)—notably, in *Epidemics* and *Airs, Waters, Places*.[1] Both texts emphasized the natural—rather than supernatural—nature of disease and described health as a matter of a body in balance with its daily routines and external environment. The texts circulated at a momentous time, as Greek city states were seeking fresh territories through colonial conquest. More than just manuals for individual practitioners, *Epidemics* and *Airs, Waters, Places* were works intended to be of practical, political, and military use to Greek leaders weighing factors for and against the placement of new communities and military outposts.[2] From ancient times, then, the stimulus to collect and collate information about groups of people has transcended the concerns of individual clinicians and their patients; then, as now, politics matter.

In a sense, "politics matter" is an unsurprising and obvious claim to make about disease and society: epidemics threaten much beyond individual health. Overwhelming outbreaks of disease have destabilized militaries, markets, and governing powers, a feature as familiar to the Greeks as to the administrations that ruled through the Age of Imperialism in the 19th century. Recent discussions of "decolonizing" academic research have included this aspect of the history of epidemiology in their remit, and the

colonial roots and legacies of the discipline are now broadly known and acknowledged.[3] In focusing on how 19th- and 20th-century international public health practices systematically marginalized Indigenous voices and pathologized racial and cultural difference, recent scholars of decolonialization have offered numerous strategies for a more inclusive, equitable, and robust science.[4,5] Such discussion of the political dimensions of public health and epidemiology, while welcome, have, however, tended to focus on the fields of global health and social epidemiology. A possible unintended consequence of that focus risks obscuring a fundamental character of the discipline: that politics matter always and everywhere when interpreting patterns of health and disease. To engage with distributions of health and disease across any population is to engage with public policy, and, to that extent, the practice of any branch of epidemiology is therefore inherently "political." While there is justifiable concern that public engagement in politics can threaten scientific objectivity,[6] the practice of epidemiology has also been criticized for being too focused on empirical method—possibly to demonstrate its scientific bona fides and create some distance from contentious political issues—which itself raises questions of epidemiology's purpose and disciplinary responsibility to the public.[7] These are issues that epidemiologists have wrestled with for decades and are unlikely to be resolved anytime soon.

This article takes a long historical view of epidemiology to briefly revisit 3 famous contributors to public health from the late 18th through the mid-19th century. As will be shown, there was no one political vision for effecting social progress that was the norm and no "pure" science advocacy free from the presumption of or need for some form of political engagement. The pioneers discussed—Johann Peter Frank (1745-1821), Rudolf Virchow (1821-1902), and John Snow (1813-1858)—all showed a deep commitment to advocating for their science, but, importantly, each also demonstrated how their advocacy was shaped by beliefs and values absorbed from their respective social and political worlds.

Origins of Social Medicine

The application of statistical methods and probabilistic reasoning to problems of public health has roots in the 18th century and the work of Johann Peter Frank, who popularized the notion of "medical police." Frank's mammoth work, *System Einer Vollständigen Medicinischen Polizey* (*A Complete System of Medical Policy*), appeared in 6 volumes, the first of which was published in 1779.[8] A German physician and hygienist, Frank admired contemporary European Enlightenment philosophers and their emphasis on the primacy of human reason over superstition and dogma. The use of census tools and the central collection of vital statistics would, in Frank's view, drive top-down transformation by way of public health and social reform. Like other Enlightenment era thinkers, Frank believed that the inequalities that impeded health also impeded social progress. In a 1790 public lecture titled "The People's Misery: Mother of Diseases,"[9] he laid out the connections between disease, social conditions, and the need for policy-based action on the part of the physician. While later scholars have taken up this lecture title as something of a rallying cry for health equity,[10,11] it is worth noting that the means through which Frank believed social progress would be achieved were far from democratic. A committed mercantilist, Frank was a great believer in a zero-sum game of economic policy through which European nations sought to acquire colonial wealth to enrich themselves abroad while beggaring their neighbors and competitors at home. In Frank's authoritarian and paternalistic vision, populations foreign and domestic represented resources to be protected and made productive for the benefit of the state.[12] Medical police science directed state regulation of food, air, and water and

accompanied other measures intended to promote population growth and military preparedness, such as better education of midwives and improved childhood nutrition.

Along with widespread European political unrest in the 1840s, however, came a particularly brutal test of the medical police system when a devastating plague of typhus broke out in Prussian-administered Upper Silesia (now part of Poland). The acclaimed physician-scientist Rudolf Virchow was dispatched to investigate, and another chapter in the history of epidemiology was begun. Virchow employed extensive field-based observations, interviews, and statistical methods to compile his lengthy report on the epidemic, which he published in 1848.[13] Like Frank's work, Virchow's writing has since become foundational for origin stories about the rise of social medicine, histories that explicitly or implicitly align themselves with left-of-center advocacy regarding equity and the social and economic etiologies of illness in postindustrial society.[14]

It is a modern take and political perspective that arguably would have been quite alien to Virchow himself. Along with much-cited lines in his 1848 report, such as "Medical statistics will be our standard of measurement: we will weigh life for life and see where the dead lie thicker among the workers or among the privileged,"[15] are numerous other comments placing blame on local populations for their own misery.[13] For Virchow, the inferior culture and poor physical constitutions of Polish-speaking people showed just how in need they were of a strong Germanic hand. His firebrand writings and activities advocating for a liberal democracy and against the state machinery were not so much about "the people" as about who should be in control—old regime bureaucrats or new technocrats like himself?[16]

Epidemiology and Industrial Sanitation

Virchow's microscopic work and theories of cellular pathology place him in the vanguard of a scientific approach to medicine that would finally replace millennia of Hippocratic theories. Nonetheless, so committed was Virchow to the environmental regulation of public health that to the end of his life he remained an avowed "contingent contagionist" and germ theory skeptic.[17] In other words, Virchow argued that some diseases—like typhus—which were thought to be contagious were, in fact, generated by environmental filth—especially bad airs or "miasma"—leaving individuals weakened by poverty and poor social and living conditions the most susceptible to illness and death. By way of contrast, the British physician John Snow argued against miasmatic disease spread and aimed to show that some contagious illnesses did have a single cause and one that would produce illness regardless of the constitutional health of individuals in the affected population.

As a founding member of the London Epidemiological Society, Snow likely would have inevitably been drawn to the study of cholera during the outbreak of 1854.[18] How to control the disease was *the* urgent administrative and public health question of the day. An earlier outbreak of the disease in 1848-1849 had provoked Parliament to action, including by ordering private water companies supplying water from the city's River Thames to shift their intake locations upstream of the location where much of the city's sewage was dumped.[18] In his brilliant South London water study, Snow compiled statistical tables that looked at case counts in different London districts, along with the names of private companies and from where they drew their water supply.[18] While Snow went to great lengths to control for variables other than water source in his comparisons, recent historians have shown that contemporary critics of Snow had

reasonable grounds to question how well his districts and subdistricts did, in fact, compare.[19]

Snow's other major work on the Broad Street pump (located within Soho, an area of London close to his own neighborhood) was similarly purposed to find persuasive evidence for his belief that cholera was a waterborne disease. In that instance, he used not statistical tables but a dot map showing the incidence of cholera cases by their proximity to the pump.[20] Once again, the stiff opposition Snow faced is not adequately accounted for by the notion that his critics were ignorant or reactionary. It was a time of great peril and uncertainty, and too many aspects of Snow's work seemed (in the view of his critics) not to disallow miasmatic transmission alongside or instead of waterborne transmission. Snow's successful use of his map in getting the local authorities in Soho to remove the handle of the Broad Street pump was then as much an act of skillful negotiation and political persuasion as it was a self-evident scientific sweep of his doubters.[20]

Snow was not a politically active physician in the manner called for by Frank or Virchow, but his work shows his deep engagement with the same stark political realities of industrializing cities and the role that public agencies and policy might have in managing population health. For its part, the British government commissioned a wide-ranging report on the potential links between poor sanitation and epidemic disease and placed the prominent lawyer and social reformer Edwin Chadwick (1800-1890) in charge of it. His 1842 report, *The Sanitary Conditions of the Labouring Population*, profoundly influenced the passage of the Public Health Act 6 years later.[21] Chadwick's belief in miasmatic theory was deeply bound with his perceived need for wide-reaching legislation to transform the sanitary infrastructure of Britain and, with it, the health of a nation.[22] Although more suspicious of large-scale government bureaucracy than the Germans, British adherents of miasmatic theory did nonetheless worry about the potential for Snow's work to undermine new public health infrastructure and hard-won sanitary reforms. Snow, for his part, went some way to ease concerns (as evidenced by his success in getting the Broad Street pump shut down) as he continued to pursue the science. In the end, the great transformations of public health in Britain were owing to the work of both sanitarian miasmatists like Chadwick and proto-germ theorists like Snow.

Epidemiology as Political Science?

In a world of intense political polarization that is still reeling from the COVID-19 pandemic, history offers a reminder that we have endured calamitous times before. An awareness of political context and a willingness to engage with political influence would seem desirable for the ethical, professional conduct of epidemiological practice. Political partisanship on the part of the practitioner is neither necessary nor sufficient as a replacement for political awareness, but, I argue, neither is denial that there exists a political dimension to the science of epidemiology at all.

References

1. Martin PMV, Martin-Granel E. 2,500-year evolution of the term epidemic. *Emerg Infect Dis*. 2006;12(6):976-980.
2. Rosen G. *A History of Public Health*. Rev ed. Johns Hopkins University Press; 1993.
3. Adebisi YA. Decolonizing epidemiological research: a critical perspective. *Avicenna J Med*. 2023;13(2):68-76.

4. Petteway R, Mujahid M, Allen A, Morello-Frosch R. Towards a people's social epidemiology: envisioning a more inclusive and equitable future for social epi research and practice in the 21st century. *Int J Environ Res Public Health*. 2019;16(20):3983.
5. Feo Istúriz O, Basile G, Maizlish N. Rethinking and decolonizing theories, policies, and practice of health from the Global South. *Int J Soc Determinants Health Health Serv*. 2023;53(4):392-402.
6. Bracken M. Advocacy in epidemiology. *Am J Epidemiol*. 2006;163(suppl 11):S167.
7. Olshan AF, Diez Roux AV, Hatch M, Klebanoff MA. Epidemiology: back to the future. *Am J Epidemiol*. 2019;188(5):814-817.
8. Frank JP. Lesky E, ed. *A System of Complete Medical Police*. Johns Hopkins University Press; 1976.
9. Rosen G. Disease and social criticism: a contribution to a theory of medical history. *Bull Hist Med*. 1941;10(1):5-15.
10. Poverty and health. Editorial. *Am J Public Health*. 1969;59(4):587.
11. Geiger HJ. The political future of social medicine: reflections on physicians as activists. *Acad Med*. 2017;92(3):282-284.
12. Rosen G. Cameralism and the concept of medical police. *Bull Hist Med*. 1953;27(1):21-42.
13. Virchow RC. Report on the typhus epidemic in Upper Silesia. *Am J Public Health*. 2006;96(12):2102-2105. Excerpted from: *Virchows Arch Pathol Anat Physiol Klin Med*. 1848;2:143-332.
14. Porter D. How did social medicine evolve, and where is it heading? *PLoS Med*. 2006;3(10):e399.
15. Taylor R, Rieger A. Medicine as social science: Rudolf Virchow on the typhus epidemic in Upper Silesia. *Int J Health Serv*. 1985;15(4):547-559.
16. Figlio K, Weindling P. Was social medicine revolutionary? Rudolf Virchow and the revolutions of 1848. *Soc Soc Hist Med Bull (Lond)*. 1984;34:10-18.
17. Ackerknecht E. *Rudolf Virchow: Doctor, Statesman, Anthropologist.* University of Wisconsin Press; 1953.
18. Shapin S. Sick city: maps and mortality in a time of cholera. *New Yorker*. October 28, 2006. Accessed June 24, 2024. https://www.newyorker.com/magazine/2006/11/06/sick-city
19. Tulodziecki D. How (not) to think about theory-change in epidemiology. *Synthese*. 2021;198(suppl 10):2569-2588.
20. McLeod KS. Our sense of Snow: the myth of John Snow in medical geography. *Soc Sci Med*. 2000;50(7-8):923-935.
21. Poor Law Commissioners. *Sanitary Condition of the Labouring Population of Great Britain*. House of Commons, United Kingdom; 1842. Reports From Commissioners 26.
22. Worboys M. *Spreading Germs: Disease Theories and Practice in Britain, 1865-1900*. Cambridge University Press; 2000.

H. K. Quinn Valier, PhD is a research associate professor in the Tilman J. Fertitta Family College of Medicine and the director of student engagement for population health at the University of Houston in Texas.

Citation
AMA J Ethics. 2025;27(1):E58-63.

DOI
10.1001/amajethics.2025.58.

Conflict of Interest Disclosure
Author 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

AMA Journal of Ethics®
January 2025, Volume 27, Number 1: E64-65

ART OF MEDICINE

Right in the Eye

Kayla Mackenzie McCormick

Abstract

This illustration represents how a patient's view of themselves can be altered while going through iatrogenic trauma.

Figure. *Inner Perspective*

Media
Digital watercolor.

Caption
Empathic patient care involves a delicate balance between going about the business of diagnosing and treating patients and preserving their dignity and self-image. This balance can be upset in instances in which procedures that are routine for clinicians turn out to be sources of iatrogenic trauma for patients or in which a patient's perspective is lost or undervalued.

Kayla Mackenzie McCormick is a student at the School of the Art Institute of Chicago in Illinois.

Citation
AMA J Ethics. 2025;27(1):E64-65.

DOI
10.1001/amajethics.2025.64.

Conflict of Interest Disclosure
Author disclosed no conflicts of interest.

ISSN 2376-6980