REVIEW ARTICLE

AI IN MEDICINE
Jeffrey M. Drazen, M.D., *Editor*

Considering Biased Data as Informative Artifacts in AI-Assisted Health Care

Kadija Ferryman, Ph.D., Maxine Mackintosh, Ph.D., and Marzyeh Ghassemi, Ph.D.

Artificial intelligence (AI) tools used in medicine, like AI used in other fields, work by detecting patterns in large volumes of data. AI tools are able to detect these patterns because they can "learn," or be trained to recognize, certain features in the data. However, medical AI tools trained with data that are skewed in some way can exhibit bias, and when that bias matches patterns of injustice, the use of the tools can lead to inequity and discrimination. Technical solutions such as attempting to fix biased clinical data used for AI training are well intentioned, but what undergirds all these initiatives is the notion that skewed clinical data are "garbage," as in the computer science adage "garbage in, garbage out." Instead, we propose thinking of clinical data as artifacts that, when examined, can be informative of societies and institutions in which they are found.

Viewing biased clinical data as artifacts can identify values, practices, and patterns of inequity in medicine and health care. Examining clinical data as artifacts can also provide alternatives to current methods of medical AI development. Moreover, this framing of data as artifacts expands the approach to fixing biased AI from a narrowly technical view to a sociotechnical perspective that considers historical and current social contexts as key factors in addressing bias. This broader approach contributes to the public health goal of understanding population inequities and also provides novel ways to use AI as a means of detecting patterns of racial and ethnic correction, missing data, and population inequities that are relevant to health equity.

From the Johns Hopkins Berman Institute of Bioethics and the Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore (K.F.); Genomics England and the Alan Turing Institute, London (M.M.); and the Department of Electrical Engineering and Computer Science and the Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA (M.G.). Dr. Ferryman can be contacted at kadija.ferryman@jhu.edu or at the Johns Hopkins Berman Institute of Bioethics, 1809 Ashland Ave., Baltimore, MD 21205.

This article was updated on August 31, 2023, at NEJM.org.

N Engl J Med 2023;389:833-8.
DOI: 10.1056/NEJMra2214964

MEDICAL AI AND BIAS

We are witnessing the ascendance of AI. AI tools such as ChatGPT and DALL-E can appear to mimic human intelligence, but they are computer programs that categorize, classify, learn, and filter data to solve problems, make predictions, and perform other seemingly intelligent tasks. AI tools used in medicine, like AI used in other domains, work by detecting patterns in large volumes of data. For example, AI can learn to detect anomalies in medical images after being trained on large numbers of images of anomalies. Medical AI has displayed impressive capabilities, especially in the field of radiology. Some AI tools are at least as accurate as highly experienced radiologists in identifying disorders in medical images.[1,2]

However, if medical AI tools are trained with data that are skewed in some way, these tools can exhibit bias. For example, an AI tool could be developed to detect disease in chest films. The tool would be trained with the use of a data set composed of thousands of images of chest films with or without disease. The AI would learn to identify diseases from these images. Then, when shown a new image, the AI tool would be able to determine whether evidence of disease was present on the

chest radiograph. Ideally, this tool would be highly accurate at identifying disease, and it would also be accurate for everyone. However, the tool would become biased if the training data included a preponderance of images with particular characteristics, such as chests of a certain size or shape or a pattern of difference in the way that the images were marked as showing or not showing disease.

This example is rooted in reality. Medical AI tools, like other AI applications, can become biased because of biases, both known and unknown, in the training data, and the bias may reflect societal inequities. A recent article exploring the use of AI to diagnose disease on the basis of chest images noted that even when trained with data sets of thousands of images, the AI model exhibited a pattern of underdiagnosis in underserved and racial and ethnic minority groups.[3] This pattern was especially glaring in intersectional groups such as Black and Hispanic women. A medical AI tool like this one is not only biased but is also a source of health inequity because there are already unjust health disparities in underserved and racial and ethnic minority groups (Fig. 1). For example, in the United States, Black persons are less likely than White persons to receive an early diagnosis of lung cancer.[4]

In this example, AI bias, which is also referred to as algorithmic bias, is consequential because it can lead to algorithmic discrimination. The White House Office of Science and Technology Policy recently identified algorithmic discrimination as a key issue in its Blueprint for an AI Bill of Rights.[5] Research and scholarly communities also recognize the potential for AI bias to become algorithmic discrimination. Some have offered technical solutions such as attempting to fix biased clinical data used for AI training. One way to fix training data is to include demographically representative data sets by bringing together, or "federating," data from various clinical institutions.[6] Other solutions include artificially creating demographic diversity by imputing data that are missing from some demographic categories or by creating new synthetic data where data do not exist.[7] Efforts are also under way to create new, diverse, and representative data sets for AI by including in the data sets a broad diversity of people rather than artificially creating diverse data or patching different data sets together. The National Institutes of Health recently launched the Bridge2AI program, a $130 million effort to build, from the ground up, diverse data sets that can be used to train and build new medical AI tools.[8]

NOT JUST A DATA BIAS PROBLEM

Although each of these efforts is well intentioned and can achieve some progress toward minimizing AI bias and downstream discrimination, what undergirds all these initiatives is the notion that skewed clinical data are "garbage," as in the aforementioned computer science adage "garbage in, garbage out," meaning that bad or faulty data lead to bad or faulty analytic outputs. Although we recognize that skewed or missing data can lead to algorithmic bias and discrimination, we propose an alternative approach to AI bias. We think of these data as artifacts. In the archeological and historical sense, artifacts are objects that, when examined, can provide information about societies, including institutions, activities, and values. Artifacts are important because of what they can reveal about earlier societies, even if they reveal beliefs and practices that may be at odds with those in contemporary societies.

In a similar way, we can think of clinical data used for AI as artifacts that can reveal what may be uncomfortable truths. For example, the widely cited research by Obermeyer and colleagues on algorithmic bias in medicine reveals that health care expenditures are lower for sicker Black persons than for healthier White persons, resulting in an algorithm that distributes health care resources inequitably.[9] However, just as we would not view artifacts that show harm as garbage or as objects that should be fixed, so too we should not ignore current clinical artifacts. When viewed as an artifact that can illuminate social values, the biased clinical data identified by Obermeyer and colleagues show, as sociologist Ruha Benjamin writes, that "Black patients do not 'cost less'…they are valued less."[10] Thus, when skewed clinical data are considered as informative artifacts, not garbage, we can harness the power of pattern recognition in AI to help us understand what these patterns mean in historical and contemporary social contexts. Below are three examples of how viewing biased clinical data as artifacts can identify values, practices, and patterns of inequity in health care. Examining clinical data as artifacts can also provide alternatives to current methods of medical AI development.

HEALTH DATA ARTIFACTS AND VALUES

There has been growing attention to the application of racial and ethnic correction factors in clinical data. For example, in 2021, the Chronic Kidney Disease Epidemiology Collaboration reported on a new equation to estimate a measure of kidney function (the glomerular filtration rate), without the use of a racial correction.[11] This equation previously "corrected" for the supposedly higher muscle mass of Black persons. Research has shown that racial corrections in medicine can be traced back to the practice of using White male bodies as the reference, or norm, against which other bodies and physiological functions are measured. Although genetic ancestry may provide some clinically relevant information, such as genetic variants that confer protection against disease,[12] there is a growing recognition that some racial and ethnic corrections in medicine need to be reevaluated, since the evidence supporting them may be dated, and the use of these corrections may deepen health inequities.[13]

An awareness of the history of racial correction of clinical data is important because clinical prediction models may build on the embedded logic that there is a biologically determinative relationship between race and aspects of physiology, such as lung function.[14,15] These data and assumptions can then be imported into the development of medical AI tools. Seemingly invisible biases such as racially "corrected" clinical data can be hard to fix with purely technical means if the history of racial correction is unrecognized. Here we emphasize that racist values such as White normality or supremacy, though disavowed in contemporary medicine, can affect practice in the present as well as the development of future medical AI tools if these data are used as training sets. Upstream examination of clinical data as artifacts by interdisciplinary groups comprising clinical staff, patients, engineers or developers, and social science and humanities scholars can reveal important, yet implicit histories and other factors shaping the data. This kind of intervention can help identify data that would result in discriminatory AI tools downstream and suggest interventions for addressing the deep causes of these skewed data, such as reevaluating racial correction in clinical practice.

HEALTH DATA ARTIFACTS AND PRACTICES

Viewing skewed health data as artifacts worthy of close examination can also identify health care practices, which can point the way to sociotechnical solutions to problems with data and data-centric tools such as AI. For example, gender identity is often missing in clinical data. Instead of thinking only of ways to fix these data or abandon the reams of data we already have, we can examine them for the rich information they present and consider what the missingness of data suggests about clinical and social practices, such as a lack of uniformity in terms referring to sex and gender in clinical parlance and the continued use, in medical intake forms, of outdated gender identity terms that may not apply to everyone.[16] The missing data could also suggest that some persons may not feel comfortable and supported in disclosing this information and that medical staff may lack the training or authority to collect it.[17]

An artifact approach to health data also facilitates novel applications of the capabilities of AI. Because AI can quickly identify patterns, it can spot missingness in clinical data, such as the absence of certain racial groups, which can serve as a hypothesis-generating tool that can catalyze new, interdisciplinary research on clinical care and health inequities.[18] If we approach these data as artifacts, we move away from the predominant framing of bias in AI as an issue that can be solved through technical means, such as by imputing missing data or creating new data sets.

HEALTH DATA ARTIFACTS AND PATTERNS OF INEQUITY

Examining health data as artifacts rather than as garbage can also help reveal patterns of inequity across populations in health care. Unfortunately, there are numerous examples of unjust health disparities, or health inequities, specifically among racial and ethnic minority groups in the United States. Health data reflect these disparities. As mentioned above, lung cancer is more likely to be diagnosed at an advanced stage of disease in Black patients than in White patients. If used to train a cancer prediction algorithm, this bias in the data might predict lower survival among Black patients. The lower predicted sur-

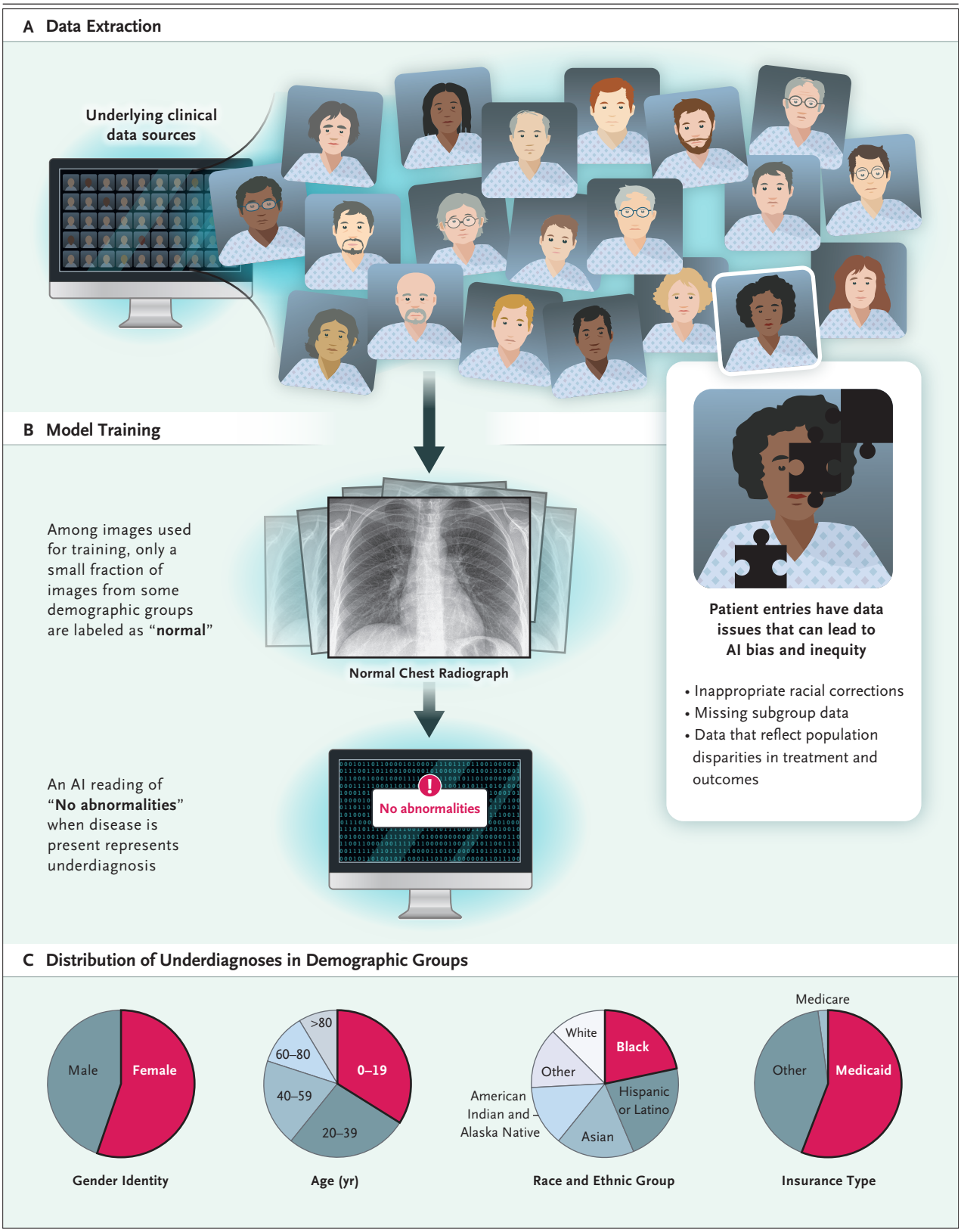

Figure 1 (facing page). Bias in Medical Artificial Intelligence (AI).

The use of AI in a health-related risk or outcome prediction task (in this case, detection of disease in chest radiographs in underserved patient populations) described by Seyyed-Kalantari et al.[3] is shown. As shown in Panel A, data are first extracted from clinical sources that reflect the contexts in which the data were acquired and recorded. Human biases, device-related biases (e.g., pulse oximetry showing incorrect blood oxygenation in patients with dark skin), and systemic biases from these sources are reflected in the data. As shown in Panel B, models are trained to maximize overall performance, which may result in benefit to one group at the expense of others. Models may also be unable to capture necessary interaction effects between clinical features and group attributes. As shown in Panel C, model audits are performed after training to ensure that important metrics, such as the incidence of false positive "underdiagnoses," are not markedly lower in one subgroup than in others. In the pie charts, red indicates the greatest incidence of false positive underdiagnoses. Subgroup performance audits are a key first step in revealing underlying issues that should be addressed before model integration.

vival, in turn, could affect the treatment options offered to these patients, especially in the case of treatment triage or rationing systems that favor patients who are expected to have better outcomes.

A purely technical response to this biased algorithm would be to use alternative data or to exclude the disease stage at diagnosis as an input. However, viewing these data as an artifact can help reveal patterns of inequity that bring these differences at diagnosis to the foreground. The history of these data shows that just 2 years ago, the lung-cancer screening guidelines were changed because they had been disproportionately classifying Black persons as ineligible for early cancer screening.[19,20] Examining health data as artifacts helps illuminate a pattern of population-level exclusion from preventive medical care. Without an awareness of this history, the data show a population that is predisposed to poor medical outcomes, and this kind of interpretation could undergird the development of new AI prediction tools, which could, in turn, lead to new instances of undertreatment and exclusion (Table 1).

CONCLUSIONS

The growing attention to bias within the AI and health care communities is a welcome development, especially as we continue to experience the ebbs and flows of the coronavirus disease 2019 pandemic. However, the harms of AI have often been imprecisely and narrowly considered as a data bias problem. Although there is value in innovating computational ways of altering data sets and engaging diverse participants in biomedical research, these cannot be the only solutions, and they should not rely on the implicit notion that past and current health data have little to offer AI research and development today.

We propose shifting from a focus on the deficits in health data to a consideration of these data as artifacts of human activities and values. We recognize the irony that artifact analysis in fields such as archeology is linked to a history of colo-

Table 1. Technical and Artifact-Based Approaches to Data Issues in Medical Artificial Intelligence (AI).

Data Issues	Technical Approach	Alternative or Complementary Artifact-Based Approach
Racial corrections	Attempt to correct model performance after development in order to approximate differences in performance observed between groups	Convene interdisciplinary group to examine history of data and current clinical use; adjust problem formulation (e.g., design model to diagnose inequities[18]), adjust model assumptions, or both
Missing data	Collect additional data on groups; impute missing samples with the use of individual or group data; remove populations that are likely to have data missing from datasets	Convene interdisciplinary group to examine reasons why data are missing (e.g., lack of access or earned mistrust); increase education on structural barriers to medical care
Population disparities (e.g., disparities in diagnosis, treatment, or expenditures)	Use alternative data from diverse sources; exclude data points or variables with population differences as inputs for an AI model; disclose overall diagnostic accuracy and robustness checks	Examine population-level differences in undertreatment and exclusion; allow persons with limited social power or capital to influence the development of AI[21] (e.g., conduct community participatory research to understand health care needs), and create new AI tools if necessary

nial exploitation and extraction. However, we draw on the tradition of historical artifact examination practiced by anthropologists such as Zora Neale Hurston, who aimed to illuminate undervalued histories and practices, as well as the work of current scholars who argue for the importance of using an archival approach as an alternative to algorithmic fairness, and we apply these insights to health care.[21,22] Examining health care data as artifacts expands the technical approach to data bias in AI development, offering a sociotechnical approach that considers historical and current social contexts as important factors. This expanded approach serves the public health goal of understanding population inequities and suggests novel uses of AI to detect health equity–relevant data patterns. We propose this reframing so that the development of AI in health care can reflect our commitment and responsibility to ensure equitable health care now and in the future.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

REFERENCES

1. Rudolph J, Huemmer C, Ghesu F-C, et al. Artificial intelligence in chest radiography reporting accuracy: added clinical value in the emergency unit setting without 24/7 radiology coverage. Invest Radiol 2022;57:90-8.
2. Rajpurkar P, Lungren MP. The current and future state of AI interpretation of medical images. N Engl J Med 2023;388:1981-90.
3. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27:2176-82.
4. American Lung Association. State of lung cancer: racial and ethnic disparities. October 28, 2022 (https://www.lung.org/research/state-of-lung-cancer/racial-and-ethnic-disparities).
5. White House Office of Science and Technology Policy. Blueprint for an AI bill of rights: making automated systems work for the American people. October 2022 (https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf).
6. Kennedy S. Researchers to propose framework to address federated learning challenges. HealthITAnalytics, February 16, 2023 (https://healthitanalytics.com/news/researchers-to-propose-framework-to-address-federated-learning-challenges).
7. Burlina P, Joshi N, Paul W, Pacheco KD, Bressler NM. Addressing artificial intelligence bias in retinal diagnostics. Transl Vis Sci Technol 2021;10:13.
8. National Institutes of Health. Bridge to Artificial Intelligence Highlights. August 31, 2023 (https://commonfund.nih.gov/bridge2ai/highlights).
9. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447-53.
10. Benjamin R. Assessing risk, automating racism. Science 2019;366:421-2.
11. Inker LA, Eneanya ND, Coresh J, et al. New creatinine- and cystatin C-based equations to estimate GFR without race. N Engl J Med 2021;385:1737-49.
12. 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;384:474-80.
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;383:874-82.
14. Braun L. Breathing race into the machine: the surprising career of the spirometer from plantation to genetics. Minneapolis: University of Minnesota Press, 2014.
15. Brems JH, Ferryman K, McCormack MC, Sugarman J. Ethical considerations regarding the use of race in pulmonary function testing. Chest 2022;162:878-81.
16. Kronk CA, Everhart AR, Ashley F, et al. Transgender data collection in the electronic health record: current concepts and issues. J Am Med Inform Assoc 2022;29:271-84.
17. Cruz TM, Smith SA. Health equity beyond data: health care worker perceptions of race, ethnicity, and language data collection in electronic health records. Med Care 2021;59:379-85.
18. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. Nat Med 2020;26:16-7.
19. Haddad DN, Sandler KL, Henderson LM, Rivera MP, Aldrich MC. Disparities in lung cancer screening: a review. Ann Am Thorac Soc 2020;17:399-405.
20. Rivera MP, Katki HA, Tanner NT, et al. Addressing disparities in lung cancer screening eligibility and healthcare access: an official American Thoracic Society statement. Am J Respir Crit Care Med 2020;202(7):e95-e112.
21. Davis JL, Williams A, Yang MW. Algorithmic reparation. Big Data Soc 2021;8(2) (https://journals.sagepub.com/doi/10.1177/20539517211044808).
22. Hurston ZN, Plant DG, Walker A. Barracoon: the story of the last "Black cargo." New York: Amistad, 2018.