2022

Exclusion Cycles: Reinforcing Disparities in Medicine

Ana Bracic
Michigan State University
Shawneequa L. Callier
George Washington University - School of Medicine and Health Sciences
Nicholson Price
University of Michigan Law School, wnp@umich.edu

Recommended Citation

This is the author's version of the work. It is posted here by permission of the AAAS for personal use, not for redistribution. The definitive version was published in Price, W. Nicholson II. Bracic, Ana, Shawneequa L. Callier, and W. Nicholson Price II. "Exclusion Cycles: Reinforcing Disparities in Medicine." Science 377, no. 6611 (2022): 1158-1160. DOI: https://doi.org/10.1126/science.abo2788

Exclusion Cycles in Medicine: Clinical Practice, Data Analytics, & Policy Implications

Authors: Ana Bracic[1]†, Shawneequa L. Callier[23]†, W. Nicholson Price II[45]†*

Affiliations:

[1] Department of Political Science, Michigan State University, East Lansing, MI, USA.

[2] Department of Clinical Research and Leadership, George Washington University School of Medicine and Health Sciences, Washington, DC, USA

[3] Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

[4] University of Michigan Law School, Ann Arbor, MI, USA

[5] Centre for Advanced Studies in Biomedical Innovation Law, University of Copenhagen, Copenhagen, Denmark

*Corresponding author. Email: wnp@umich.edu.

† These authors contributed equally to this work.

Abstract: Minoritized populations face exclusion across contexts from politics to welfare to medicine. In medicine, exclusion manifests in substantial disparities in practice and in outcome. While these disparities arise from many sources, the interaction between institutions, dominant-group behaviors, and minoritized responses shape the overall pattern and are key to improving it. We apply the theory of exclusion cycles to medical practice, the collection of medical big data, and the development of artificial intelligence in medicine. These cycles are both self-reinforcing and other-reinforcing, leading to dismayingly persistent exclusion. The interactions between such cycles offer lessons and prescriptions for effective policy.

One-Sentence Summary: Clinical practice, medical big data collection, and medical AI constitute self-reinforcing and interacting cycles of exclusion.

Social exclusion of minoritized populations is a pernicious and intractable problem across different domains, from politics to medical practice. In medicine, substantial disparities exist in both experiences and health outcomes for minoritized populations, with origins in systemic racism, implicit bias, historical practice, and social determinants of health (1, 2). We draw on a theory of "exclusion cycles" developed in the context of nonmedical social interactions (3) to link known dominant-group and minoritized-group behaviors and demonstrate their self-reinforcing interactions. Interlinked cycles help reveal why exclusion and racial disparities are so intractable in medicine, despite efforts to reduce them on the part of physicians and health systems through strategies focused on individual parts of the cycle such as diverse workforce recruitment or implicit bias training (1). This framework highlights particular dangers that may arise through expanding use of big data and artificial intelligence (AI)–based systems in medicine, making bias especially intractable unless tackled directly and early.

In the prototypical exclusion cycle, anti-minority culture gives rise to discrimination by some members of the dominant group. Navigating spaces that discrimination has constrained, some members of the minoritized group respond by developing strategies such as withdrawal and self-advocacy. The dominant group misunderstands and may resent many of these strategies, assuming that they are inherent to the minoritized group rather than an outcome of the dominant group's discriminatory behavior, which reinforces antiminority culture and feeds discrimination further.

The exclusion cycle theory, which emphasizes agency on the part of both dominant and minoritized group members, maps onto problematic dynamics in medical practice and research recruitment. We apply our discussion of exclusion cycles to minoritized patients—and we use the term "minoritized" deliberately. Rather than being a minority organically, people are often marginalized by others and consequently minoritized by those with greater social power; individuals may subsequently be treated as minoritized even if successful despite imbalanced power dynamics. Although our principal examples come from the extensive empirical work on the experience of Black patients and research participants in the United States, similar exclusion cycles can be found for many groups, whether minoritized on the basis of race, ethnicity, gender identity, disability, some other marker, or a combination of them. Exclusion cycles transcend specific identities and have the potential to cause pervasive problems throughout medical practice and research.

Exclusion Cycles in Medical Practice

The clinical encounter

Patient-provider interactions show self-reinforcing exclusion cycles in action (4) (see the figure). Medicine has pervasive aspects of antiminority culture. Among the best studied is perception and treatment of pain in Black patients (5). Black patients are frequently believed to feel pain less severely, often based on the belief that Black patients are bio-logically different from white patients (anti-minority culture). Accordingly, physicians are more likely to prescribe inadequate doses of pain treatments (discrimination) (5). Such poor and discriminatory treatment may lead Black patients to withdraw from treatment relationships or engage in self-advocacy in future encounters with physicians (response strategies) (4). And physicians may readily attribute such responses to characteristics of the patients themselves, such as minoritized patients being distrustful or noncom-pliant (4), rather than interpreting them as justifiable responses to discrimination (attribution error). This error leads to a stronger antiminority culture and a refreshed cycle of exclusion. Although each separate stage of the cycle has been well

characterized, even sophisticated recent treatments of clinical bias do not link the stages into the self-rein-forcing cycle (6).

Research participation and recruitment

The dynamics of research participation, including those related to big data and AI, can display their own exclusion cycles. Antiminority culture appears in entrenched perceptions of minoritized patients as "inherently mistrustful" and thus less interested in re-search participation (7). This helps enable discriminatory recruitment practices, such as lessened efforts to recruit minoritized patients directly or reliance on convenient but demographically biased samples (8). In one big-data study, a putatively race-neutral convenience sample of perioperative patients was in fact 86.7% white, substantially less diverse than the overall hospital system's 69.8% white patient population (8). Minoritized patients, recognizing the discriminatory nature of lessened research engagement, may be justifiably reluctant to participate in research studies, not least because of the demonstrated lack of trust-worthiness on the part of institutions (9); in the same big-data study, Black patients declined research participation at nearly twice the rate of white patients (8). Finally, big-data researchers may erroneously attribute lessened patient participation not as a justifiable reaction to discriminatory practices, but rather as an inherent attribute of minoritized patients—contributing to and strengthening antiminority culture.

The Impact of Big Data and AI

The use of large-scale datasets and sophisticated machine learning techniques promises positive transformation for the health system, including by improving diagnoses, monitoring patients, and improving quality. Nevertheless, big data and AI show disturbingly similar dynamics of exclusion (see the figure)—or surreptitious and problematic inclusion in undesirable practices (9)—but with the added challenges of automation, scale, and the deceptive appearance of objectivity. Indeed, AI exclusion cycles are especially pernicious because of prevalent views that AI systems can decrease bias over time as they learn from patient experience or decrease physician bias by providing a more objective view (10). We explain how AI systems might instead entrench bias through exclusion cycles.

First, antiminority culture. AI systems themselves cannot have negative views of minoritized groups. But the humans who write, validate, and deploy AI may be racist or biased, especially given coders' lack of diversity, leading to systems that incorporate antiminority culture (11). Even if AI systems are designed by unbiased coders striving for neutrality, those systems derive data from and exist within a medical sys-tem that has its own antiminority culture; those views are embedded in the patterns that AI learns. Not only do training data used to develop AI reflect a biased system, but they also typically inadequately rep-resent minoritized patients, based in part on insufficient outreach and the inclusion challenges highlighted above. Perniciously, antiminority culture embedded in data hides beneath a veneer of mechanical objectivity, which may lead to erroneous conclusions that special efforts at inclusion and equity are unwarranted, or that merely increasing dataset size will deal with problems of bias (10).

Second, discrimination. AI systems can readily embody bias and discrimination. AI based on biased and incomplete data is unfortunately likely to be less accurate about understudied patients, providing lower-quality recommendations and analyses for those patients (8, 12). For instance, algorithms trained to detect cancerous skin lesions may perform worse on patients with darker skin, because the training datasets principally comprise data from light-skinned patients (13). When minoritized patients are included in underlying data-sets, AI systems are likely to reflect

bias in their recommendations. For instance, an AI system used to recommend follow-up coordination for patients at high risk of medical complications reflected precisely this bias: Because it was trained on underlying medical practice that provided less care to Black patients, the algorithm predicted lower risk for Black patients than similarly situated white patients—and consequently was less likely to recommend interventions (14). Notably, these biased recommendations do not stay relegated to computer systems; instead, they feed back into the behavior of physicians, affecting the cycle of exclusion in human-based medical practice. That is, the feedback of the exclusion cycle in data strengthens the exclusion cycle in care.

Third, response strategies to negotiate space constrained by discrimination. Members of minoritized communities, facing a demonstrably exclusionary medical system, may be understandably less interested in participating in big-data research (8, 9). Underrepresentation of groups in big datasets has been most thoroughly documented in the context of genomic databases and genome-wide association studies (12). But underrepresentation in other big datasets is also pervasive; indeed, a key goal of the US National Institutes of Health "All of Us" initiative is to develop a more representative dataset for precision health (with some notable success). How minoritized patients will respond to the use of AI remains uncertain (10); these systems are still being deployed, and it is too early to know whether differences in the quality of recommendations will lead to justifiable differences in trust and adherence. However, as long as discriminatory or poorly performing recommendations exist and are coupled with ensuing withdrawal or resistance strategies, AI's potential to democratize excellence in care is impeded.

Fourth and finally, the attribution error. Minoritized group status is an easy proxy for harder-to-observe attributes. Many out-comes vary substantially by race (2). Much of this variance is due to a system with substantial embedded racism and racial bias (2), and some of it may be due to justifiable minoritized community response strategies adopted in response to that sys-tem. The outcomes that AI "sees" in data will likely vary based on these underlying realities—and be augmented by biases and response strategies in data collection. In a learning AI system, if poor recommendations derived from poor initial data result in justifiably poor adherence to AI recommendations as its own response strategy, future iterations would be even more biased and perform even worse. Putting all this together, it would be surprising if AI did not attribute differences in patient out-comes, and correspondingly in diagnoses or recommendations, to minoritized status (writ broadly). AI systems are thus likely to commit the attribution error themselves, at least implicitly, and use race as a problematic proxy, directly or indirectly—leading to the next bout of antiminority culture embedded within the data and system and strengthening the cycle anew.

AI systems will not always reinforce systemic and data biases. Indeed, AI systems can be used to diagnose bias embedded within apprenticeship-like medical training and other existing medical practices (15). Potentially, AI systems using unsupervised learning could even identify and account for underlying disparities, improving equity in care (10). However, given the self-reinforcing nature of AI-involved exclusion cycles, a long history of bias in medical care (1, 2), and disparities in data collected, the potential negative impacts demand focus even against the positive possibilities of AI.

Policy Implications

Although the pieces of exclusion cycles in clinical encounters and AI and big data may already be known, connecting the pieces explicitly into exclusion cycles has three principal implications:

First, exclusion can be self-reinforcing, including in AI; second, exclusion cycles between medical practice and AI can interweave and reinforce one another; and third, effective policy interventions will need to take these self- and other-reinforcing dynamics into account.

At the most basic level, the dynamics of exclusion in medicine are self-reinforcing. Racial bias embedded in medical practice and research data collection self-perpetuates by leading to discrimination, which may lead to strategies such as resistance and withdrawal by minoritized patients, and, consequently, to physicians' erroneous attribution of problems to those patients' inherent characteristics. In AI, this cycle also exists, with the deceptive shield of algorithmic objectivity. Machines cannot be racist, but they can and do participate in self-reinforcing exclusion cycles, based both on the involvement of imperfect humans and training on data that reflect and embody systemic racism. Given this initial involvement, systemic bias will be perpetuated and reinforced through the dynamics of big data and AI themselves. There is no reason to expect improvement over time (10), absent active intervention.

Crucially, exclusion cycles in medicine are also other-reinforcing. Even putatively unbiased AI systems learn bias from biased training data and return biased recommendations—but similarly, human physicians, nurses, and other care providers, even if unbiased themselves, will receive biased recommendations and analyses from such AI systems. Disturbingly, we should then expect that the addition of bias from AI systems would itself then be reinforced within care-based exclusion cycles.

Imagine a perfectly unbiased physician in an unbiased care system that imports a highly touted AI system trained in a real-world biased health care system to improve patient pain care. The AI system has learned from biased data to recommend inadequate doses of medication to Black patients and makes those recommendations to the unbiased physician. If clinicians dutifully take the advice from the AI system, prescribing inadequate doses, the system has essentially replicated an antiminority culture, driven discriminatory treatment, and set the stage for understandable patient response strategies and physician attribution errors. After enough cycles of this, the care exclusion cycle can continue even without the AI intervention. Skill fade and automation bias—in which providers rely on automation, including AI, and gradually lose expertise in, e.g., racially equitable provision of care—may augment this process. Though individual-stage dynamics have been noted—biased AI can obviously lead to biased care (13)—the ability of entire self-reinforcing cycles to re-inforce one another has gone unrecognized.

From a policy intervention standpoint, tackling bias in either system independently is thus woefully insufficient. Bias in either the human care cycle or the AI–data overlay can reintroduce bias at the other level, even positing the absence of bias in human actors themselves. Bias is a systemic infection that cannot be treated in only one place; it must be treated systemically. Accordingly, efforts to decrease bias in medical care at the physician level must be coordinated with (i) efforts to ensure that algorithms trained on existing data do not themselves incorporate biases, and (ii) efforts to guard against the introduction of self-reinforcing biases in algorithms that are deployed into the care process—in contrast to existing proposals that tend to treat problems of care and AI bias largely separately (6, 8, 13). Research into best practices for coordination and de-biasing interventions could help, as could placing minoritized physicians and data scientists on care and research teams.

Other interventions may best be targeted at different exclusion cycle stages. Interventions focused on algorithms, whether during regulatory review or other governance efforts, are likely best focused on discriminatory outputs—which can be most readily measured—and on attribution errors, especially for explainable AI systems. That is, breaking the exclusion cycle for algorithms is likely easiest by identify-ing discriminatory outcomes and ensuring that system learning about performance does not problematically incorporate new biases about patient characteristics. Carefully evaluating performance data in the process of retraining could help break the propagation of exclusion cycles—for in-stance, by distinguishing attribution errors (new learnings about minoritized patients) from decreased adherence to poor (or even accurate) predictions. Antiminority culture is tougher to address, and patient-focused strategies are ethically inappropriate to tackle without first addressing other parts of the cycle. Attempting to promote minoritized group trust in current biased systems is more likely to understandably backfire than to help fix the cycle. Rather, minoritized community strategies to combat systemic constraints will hopefully evolve as the system improves, amid necessary efforts to demonstrate institutional trustworthiness to minoritized communities.

By contrast, interventions for physicians can more readily focus on antiminority culture (through education on systemic biases, including in AI) and discrimination (through monitoring and evaluation of disparate treatment). A rich literature describes such potential interventions (1, 2, 4).

The interactions that we describe here occur within the context of broader systemic racism in health care (1, 2, 4, 5). We focus here on big data and AI, relatively new systemic elements, but patient and provider experiences—and interventions to improve them—are also profoundly influenced by barriers to access, proximity to care, language barriers, the hidden curriculum of medicine and its depictions of minoritized groups, and other factors. Increasing equity must include interventions across the health care system (1, 2). Although we can-not address the full scope of systemic racism here, we suspect the exclusion cycle framework will provide a useful tool in analyzing and addressing other aspects of systemic racism. Future empirical research could analyze the details of exclusion cycles in medicine and elsewhere.

The particular interventions we mention here are known. But the combination is important; these issues cannot be addressed seriatim but must be faced in coordination. Though the addition of big data and AI to medicine promises substantial gains, they complicate the picture for reducing bias and require careful efforts to ensure that progress on one front is not rapidly lost on another.

References and Notes

1. M. A. Fair, S. B. Johnson, Addressing racial inequities in medicine. *Science*. **372**, 348–349 (2021).

2. I. of M. of the N. Academies, *Unequal Treatment; Confronting Racial and Ethnic Disparities in Healthcare* (National Academies Press, 2003).

3. A. Bracic, *Breaking the exclusion cycle: How to promote cooperation between majority and minority ethnic groups* (Oxford University Press, 2020).

4. D. B. Matthew, *Just Medicine: A Cure for Racial Inequality in American Health Care* (NYU Press, 2018).

5. K. M. Hoffman, S. Trawalter, J. R. Axt, M. N. Oliver, Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *PNAS*. **113**, 4296–4301 (2016).

6. M. Sun, T. Oliwa, M. E. Peek, E. L. Tung, Negative Patient Descriptors: Documenting racial bias in the electronic health record. *Health Affairs*. **41**, 203–211 (2022).

7. L. M. Crawley, African-American participation in clinical trials: situating trust and trustworthiness. *Journal of the National Medical Association*. **93**, 14S (2001).

8. K. Spector-Bagdady, S. Tang, S. Jabbour, W. N. Price, A. Bracic, M. S. Creary, S. Kheterpal, C. M. Brummett, J. Wiens, Respecting Autonomy And Enabling Diversity: The Effect Of Eligibility And Enrollment On Research Data Demographics. *Health Affairs*. **40**, 1892–1899 (2021).

9. S. Callier, S. M. Fullerton, Diversity and Inclusion in Unregulated mHealth Research: Addressing the Risks. *The Journal of Law, Medicine & Ethics*. **48**, 115–121 (2020).

10. B. Babic, S. Gerke, T. Evgeniou, I. G. Cohen, Algorithms on regulatory lockdown in medicine. *Science*. **366**, 1202–1204 (2019).

11. R. Benjamin, Race after technology: Abolitionist tools for the new Jim Code. *Social Forces* (2019).

12. C. N. Rotimi, S. L. Callier, A. R. Bentley, Lack of diversity hinders the promise of genome science. *Science*. **371**, 565 (2021).

13. A. S. Adamson, A. Smith, Machine learning and health care disparities in dermatology. *JAMA dermatology*. **154**, 1247–1248 (2018).

14. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. **366**, 447–453 (2019).

15. E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, Z. Obermeyer, An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*. **27**, 136-140 (2021).

Acknowledgements

Funding: Provide complete funding information, including grant numbers, complete funding agency names, and recipient's initials. Each funding source should be listed in a separate paragraph.

Novo Nordisk Foundation grant NNF17SA0027784 (WNP).

Author contributions:

Conceptualization: AB, WNP

Analysis and writing: AB, SLC, WNP

Competing interests: Authors declare that they have no competing interests.

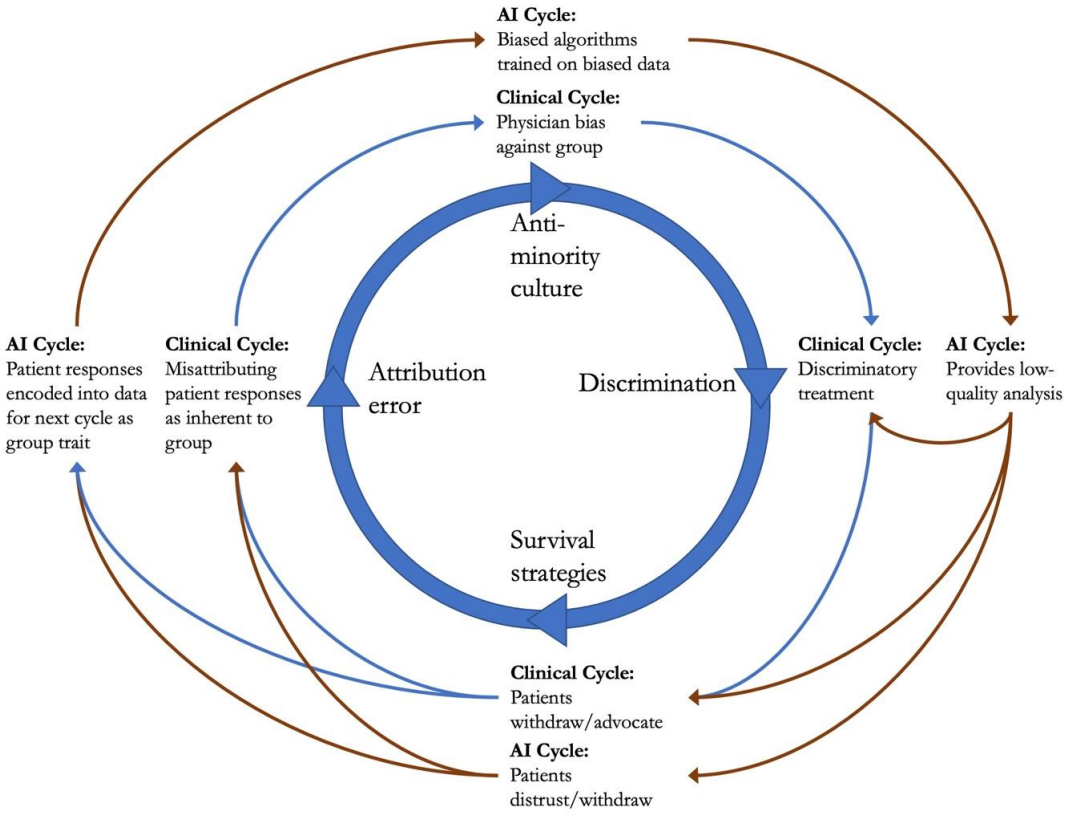

Fig. 1. Medical practice and AI create reinforcing exclusion cycles. Overlapping exclusion cycles in medicine. Around a generalized structure of the four-step cycle, we depict two interacting exclusion cycles: clinical encounters and AI products that result from, and influence, data surrounding those clinical encounters. Each cycle self-perpetuates, but the cycles also interact at various points. We highlight four: Low-quality AI analysis can result in clinical discrimination and patient withdrawal; clinical patient withdrawal or advocacy can influence what information is encoded in future AI data cycles about patient groups, and patient distrust of AI cycles (whether knowing the AI is poor-quality or simply in response to poor quality AI-driven care) can result in physician beliefs about those patients.