RESEARCH ARTICLE

Introducing the Team Card: Enhancing governance for medical Artificial Intelligence (AI) systems in the age of complexity

Lesedi Mamodise Modise[1‡], Mahsa Alborzi Avanaki[2], Saleem Ameen[3,4,5], Leo A. Celi[5,6,7*], Victor Xin Yuan Chen[1,8], Ashley Cordes[9], Matthew Elmore[10], Amelia Fiske[11], Jack Gallifant[5,12], Megan Hayes[13], Alvin Marcelo[14], Joao Matos[5,15,16], Luis Nakayama[5,17], Ezinwanne Ozoani[18], Benjamin C. Silverman[1,19,20], Donnella S. Comeau[2,19]
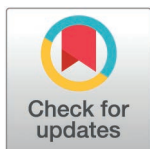
1 Center for Bioethics, Harvard Medical School, Boston, Massachusetts, United States of America, 2 Department of Radiology, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America, 3 Department of Biomedical Informatics, Harvard Medical School, Harvard University, Boston, Massachusetts, United States of America, 4 Tasmanian School of Medicine, College of Health and Medicine, University of Tasmania, Hobart, Tasmania, Australia, 5 Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, 6 Division of Pulmonary, Critical Care, and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America, 7 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America, 8 Faculty of Medicine, The Chinese University of Hong Kong, New Territories, Hong Kong SAR, 9 Indigenous Media in Environmental Studies Program and the Department of Data Science, University of Oregon, Eugene, Oregon, United States of America, 10 Duke Health, AI Evaluation and Governance, Duke University, Durham, North Carolina, United States of America, 11 Department of Preclinical Medicine, Institute of History and Ethics in Medicine, TUM School of Medicine and Health, Technical University of Munich, Bavaria, Germany, 12 Department of Critical Care, Guy's and St. Thomas' NHS Trust, London, United Kingdom, 13 Department of Environmental Studies, University of Oregon, Eugene, Oregon, United States of America, 14 Medical Informatics Unit, College of Medicine, University of the Philippines Manila, Philippines, 15 Faculty of Engineering, University of Porto, Portugal, 16 Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal, 17 Department of Ophthalmology, Sao Paulo Federal University, Sao Paulo, Brazil, 18 Machine Learning and Ethics Research Engineer, Innovation n Ethics, Dublin, Ireland, 19 Department of Human Research Affairs, Mass General Brigham, Somerville, Massachusetts, United States of America, 20 Institute for Technology in Psychiatry, McLean Hospital, Belmont, Massachusetts, United States of America

‡ Authors are in alphabetical order between the first and senior authors.
* leoanthonyceli@yahoo.com

Check for updates

Abstract

This paper introduces the Team Card (TC) as a protocol to address harmful biases in the development of clinical artificial intelligence (AI) systems by emphasizing the often-overlooked role of researchers' positionality. While harmful bias in medical AI, particularly in Clinical Decision Support (CDS) tools, is frequently attributed to issues of data quality, this limited framing neglects how researchers' worldviews—shaped by their training, backgrounds, and experiences—can influence AI design and deployment. These unexamined subjectivities can create epistemic limitations, amplifying biases and increasing the risk of inequitable applications in clinical settings. The TC emphasizes reflexivity—critical self-reflection—as an ethical strategy to identify and address biases stemming from the subjectivity of research teams. By systematically documenting team composition, positionality, and the steps taken to monitor and address unconscious bias, TCs establish a framework

for assessing how diversity within teams impacts AI development. Studies across business, science, and organizational contexts demonstrate that diversity improves outcomes, including innovation, decision-making quality, and overall performance. However, epistemic diversity—diverse ways of thinking and problem-solving—must be actively cultivated through intentional, collaborative processes to mitigate bias effectively. By embedding epistemic diversity into research practices, TCs may enhance model performance, improve fairness and offer an empirical basis for evaluating how diversity influences bias mitigation efforts over time. This represents a critical step toward developing inclusive, ethical, and effective AI systems in clinical care. A publicly available prototype presenting our TC is accessible at https://www.teamcard.io/team/demo.

Author summary

Bias in medical artificial intelligence (AI) systems refers to systematic errors that unfairly favor or disadvantage certain groups or outcomes. While often attributed to data quality issues, such as incomplete or unrepresentative datasets, bias also stems from the unexamined assumptions of researchers and developers. These influences, known as positionality, include the unique backgrounds, experiences, and worldviews that shape decisions during AI development. When unaddressed, positionality can lead to inequitable outcomes and suboptimal performance in healthcare. The Team Card (TC) is a tool designed to address these challenges by providing a structured framework for reflecting on and documenting positionality variables, perspectives, and biases throughout AI development. By integrating diverse perspectives, the TC helps research teams identify and address blind spots in development practices that stem from researchers' subjectivity. It also enables teams to assess whether addressing positionality improves fairness and reduces bias in the performance of AI in clinical settings. Research across fields such as business, science, and organizational behavior highlights that diversity fosters better decision-making, enhanced innovation, and improved overall outcomes. A publicly available TC prototype, reflecting our Team Card, is accessible at https://www.teamcard.io/team/demo, offering a practical, adaptable solution for advancing equitable and effective AI systems in healthcare.

1. Introduction

Over 40% of hospitals in the United States have integrated Clinical Decision Support (CDS) software into a broad spectrum of critical functions including diagnostics, disease management, prescription services, and alarm systems, significantly enhancing healthcare delivery and patient care [1]. Despite the challenges presented by factors such as compatibility with existing clinical infrastructure [2], data bias [3] and ethical concerns [4,5,6], CDS software functions employing machine learning (ML) algorithms and other forms of artificial intelligence (AI) currently represent a rapidly expanding use case for AI in healthcare. These AI-driven CDS software functions (medical AI systems) are often described as non-knowledge based, because they generate recommendations through statistical analysis or pattern recognition within electronic health records and other data sources, instead of relying on established medical knowledge [1]. This raises concern for the explainability of clinical recommendations provided by medical AI systems, which can sometimes function as "black-boxes", meaning

that the resulting decisions are not easily understood [4,7]. The increasing deployment of medical AI systems has thus prompted action to establish a basis of trust in these tools by improving their accuracy, robustness, and explainability [8]. To address these concerns in the United States, the Biden administration issued an executive order in October of 2023 on the trustworthy development and safe use of AI [9]. In healthcare, the order calls for the development of an AI assurance policy that encompasses a safety program and relevant standards. This follows earlier action by both the Food and Drug Administration (FDA) and the Department of Health and Human Services (DHHS), in response to calls to regulate the growing prevalence of algorithmic decision making in patient care.

In 2019 [10] and 2022, [11] the FDA issued draft and final guidance, respectively classifying, certain CDS software, clarifying which functions of CDS qualify as Software as a Medical Device (SaMD) subject to its regulatory oversight. According to the guidance, high-risk CDS software functions that are intended to inform clinical decisions and that meet the FDA's definition of a medical device must undergo rigorous premarket review, including the evaluation of supporting clinical data to ensure safety and effectiveness. However, many medical AI systems do not meet the criteria for SaMD and therefore fall outside the FDA's regulatory jurisdiction [12,13]. Furthermore, among the medical AI systems that do qualify as devices, the majority are deemed to present either low risk or substantial equivalence to previously authorized devices, allowing them to gain clearance through expedited pathways, such as the 510(k) process, which often exempts them from requiring clinical review [14,15].

In 2022, the DHHS proposed to update existing provisions of the Affordable Care Act that prohibit discrimination in covered health programs to also prohibit discrimination by clinical algorithms [16]. However, some argue that the proposed framework places excessive demands on healthcare professionals, as it requires them to effectively evaluate all algorithms implemented in their practices for potential discrimination across a broad range of parameters including; race, sex, color, national origin, age, and ability. It is also argued that the framework fails to account for the technical expertise required to audit ML algorithms for potential bias [12,13]. Finally, there is also concern that the proposed standard may simply be unachievable; not only are the algorithms powering medical AI systems typically proprietary and thus inaccessible [17], but also, these algorithms and their results may not be explainable [4,7,18]. In response to these challenges, the American Medical Association has stressed the importance of establishing crosscutting responsibilities for developers and end users of medical AI systems [19], highlighting the need for greater adherence to consensus-driven AI assurance standards [20] further upstream - among those who develop medical AI systems.

Current upstream efforts to avert harmful bias in medical AI systems have focused on better contextualization of the performance characteristics of ML algorithms through data and code sharing [21]. However, a researcher's positionality coincides with biases at every stage of development – not only in the acquisition of data and the development of ML algorithms, but also in the assessment of the resulting medical AI system's performance [22,23]. It thus becomes imperative to situate the positionality of individuals developing these systems, by accounting for potential biases that may stem from researcher subjectivity, throughout the AI development lifecycle. To date, the most significant efforts to situate ML researchers have been limited to scientometrics, which report on characteristics such as the regional location or origin of authors who have works published in scientific journals. While this approach provides some insight [24], there remains no standardized method for communicating the situatedness, as well as the potential resulting biases, of the individuals behind the development of a given medical AI system.

This work introduces the Team Card (TC); an ethical protocol that narrates researcher identity and positionality as essential components of AI development. The protocol assists research and development teams to critically examine their positionality in relation to the

AI they create, thereby mitigating the encoding of harmful bias in medical AI systems. The TC fosters epistemic diversity and inclusion by encouraging the incorporation of orthogonal perspectives in development practices while also promoting accountability by directly linking medical AI systems to their creators. Additionally, the TC provides a structured framework for collecting the data needed to validate diversity's impact on the mitigation of bias in the research and development of medical AI systems.

2. Positionality and epistemic diversity in the research and development of AI

Discussions of positionality in research derive from the premise that scientific knowledge is neither value-neutral nor objective. Rather, it is socially situated and is laden with both values and intent, which serve to reinforce the dominance of established views [25] and hierarchies of power [26]. *Dominant Science*, as observed by feminist and Indigenous scholars, assesses knowledge against its own self-image. In accepting – as scientific – those knowledges that align with established forms of power and eschewing those that do not, it engenders oppressive characteristics [27,28] that presume to adjudicate who can engage in the production of scientific knowledge [29].

Positionality broadly refers to an individual's worldview, which is shaped by gender, ethnic identity, experiences, social milieu, cultural background, and other formative influences. The term has become increasingly central to qualitative research processes, serving to identify a method for quality control [30] by locating the researcher in the scientific process. Positionality statements in research are a mode of reflection on how the researcher's perspective might have shaped the design, implementation, and analysis of the project, illuminating how the situated subjectivity of knowledge can influence outcomes at every stage of the research process. Positionality statements thus communicate the researcher's orientation toward the project through the lens of that individual's worldview and within the context of a broader social milieu.

In this manner, positionality statements engage an ethic of critical reflexivity as a necessary precursor to addressing harmful social biases in scientific knowledge [25]. Current efforts to situate knowledge in research are informed by established standards for qualitative reporting, such as Standards for Reporting Qualitative Research (SRQR) and Consolidated Criteria for Reporting Qualitative Research (COREQ), which aim to transparently convey reflexivity and details about research teams vis a vis qualitative studies. For example, Indigenous scholars have underscored the importance of locating themselves through explicit self and cultural identification with protocols of introduction [31]. Similarly, others have described how by sharing their backgrounds and life experiences, they seek to build trust between researchers and the researched [32], in a process of relational accountability [33].

Developing these ideas, feminist and care-oriented approaches have proposed ethical frameworks for the research and development of AI that begin from the proposition that AI is socially situated and that the data with which ML algorithms are trained are not neutral [34]. These frameworks support research methods [35] that focus on diversity and empowerment to mitigate the potential for representational harm in the AI development lifecycle. They do so by contending that a diversity of perspectives should shape the design and development of the technologies that define society, ensuring that these technologies are attuned to intersecting oppressions such as racism [36], sexism, and classism [34].

Nevertheless, the AI research and development landscape continues to lack diversity [37]. A scientometric analysis of original research in medical AI system development indicates that a concentration of relevant academic papers emanates from distinct knowledge hubs in the US and in China [38]. This centrality of authorship establishes collaboration and co-authorship

dynamics that serve to benefit scientific productivity, but also result in increasingly homogenous research [39] that reflects the positionality of certain research groups to the exclusion of others. The resulting research is vulnerable to potential blind spots; its efficiency and proliferation derives from homogenous research teams that tend toward convergent team processes, quickly aligning on objectives and conclusions [40].

By contrast, divergent team processes that juxtapose differing values and ideas are a hallmark of ethnically diverse teams [41]; studies show that researchers think differently in ethnically diverse groups because they anticipate divergent team process to more rigorously challenge their ideas and to alter small group dynamics [42,43]. The observed result is that socially diverse teams – characterized by greater diversity in ethnicity, gender, age, and institutional affiliation – tend to produce research that is more frequently cited, with socio-ethnic diversity in authoring teams having the greatest impact on increased citations [42,44]. Cognitive science and organizational behavior also demonstrates that diverse teams – that include different kinds of thinkers from varying disciplines – outperform homogenous groups on complex tasks and are better equipped to identify blind-spots, because they bring varied perspectives and approaches to problem solving [44,45,46,47].

Beyond research, diversity delivers tangible benefit across multiple domains. In clinical practice, a more diverse workforce has been shown to improve diagnostic accuracy and outcomes across patient populations [48]. In business, research shows that organizations with more diverse management teams demonstrate higher levels of innovation, better decision-making and achieve superior operating performance due to the variety of perspectives and experiences they bring to problem-solving [49,50,51,52]. Similarly, in governance, nations with inclusive political and economic institutions are seen to be more prosperous [53]. These findings highlight the value of diversity in fostering productive, adaptive, and equitable systems.

Drawing these perspectives together through practical experimentation, we observe that while the social diversity of the authoring team brought a range of situated perspectives to this research, it did not inherently lead to epistemic diversity – the diversity of ways of thinking, knowing, and problem-solving. This distinction is critical for advancing bias mitigation in AI. In our experience, epistemic diversity must be cultivated through an active and collaborative process of critical self-reflection, interdisciplinary engagement, and the contrasting of positionalities within research teams. The TC protocol enables teams to systematically characterize positionality variables by creating structured spaces for reflection on how positionality shapes methodological choices in AI development. This framework allows teams to identify factors that may influence bias, such as social identities, disciplinary backgrounds, and lived experiences. By enabling the documentation and analysis of these variables, the TC helps uncover patterns that contribute to bias, offering actionable insights for its mitigation. This process is particularly valuable for addressing biases in medical AI, which are intersectional in nature and arise from a broad range of factors spanning data quality and model design, as well as social and technological determinants [54,55,56].

While it is reasonable to assume that non-diverse teams may overlook biases due to homogeneity in perspectives, our argument is supported by evidence that diverse teams, particularly those fostering epistemic diversity, are better positioned to identify blind spots – which is essential for addressing the complex nature of bias in AI.

3. The importance of representation in the research and development of medical AI systems

The capacity for AI to perpetuate prejudice in healthcare is increasingly recognized. Bias in clinical algorithms and medical AI systems has engendered discrimination across a range of

demographic variables including race and ethnicity, age, disability, socio-economic status, English language proficiency and gender [57,58,59]. Attention to health disparities in the US has largely focused on racial and ethnic disparities in patient outcomes [60]. These disparities have rooted across all stages of the clinical value chain, and are seen to affect the lives of millions of patients across the US [61]. The issue is most acute in medical diagnostics where, despite a contemporary understanding that race is not a dependable proxy for genetic difference, old beliefs to the contrary remain embedded in medicine. This is evident in the longstanding practice of "correcting" diagnostic algorithms for race. Ubiquitous diagnostic algorithms and clinical practice guidelines that adjust outputs for race include examples in cardiology, nephrology, obstetrics, oncology, endocrinology, pulmonology, and urology [62]. Explicit racial biases in healthcare – such as the aforementioned race corrections – mingle with implicit racial biases arising from learned attitudes, to produce the increasing number of AI failures observed along color lines in clinical settings. The expanding corpus of racial discrimination in medical AI systems spans diagnostic procedures, therapeutic interventions and facility management systems [61,62,63,64,65].

It should be noted that these disparate patient outcomes are typically attributed to deficiencies in the data used to train medical AI systems. The issue of data quality includes situations in which the data do not reflect the true epidemiology of a demographic due, for instance, to historic racial bias in diagnosis [4,66,67]. It also includes situations where unequal access to care has been encoded as statistical forms of bias, as well as situations with sampling concerns - where data sets do not contain enough demographic diversity. These limitations decrease the reliability of resulting medical AI systems in racially diverse settings and are, to varying degrees, the result of existing social bias and inequality. The true scope of the issue however extends beyond data quality, as AI development practices are themselves susceptible to the embedded beliefs and unconscious judgments of research and development teams.

Isolating the locus of harmful bias in medical AI systems is made more complex where there are several digital determinants of health that are socially situated and that will affect system performance. Practical examples of these determinants in the clinical setting might include language concordance, digital literacy, and access to digital infrastructure such as electricity and stable internet connectivity [56]. It can thus be understood that harmful bias in medical AI systems is integrally related to forms of social bias: discriminatory patient outcomes result when existing forms of social inequality are normalized by researchers and encoded into data. These biased outcomes further disadvantage communities that have already been structurally marginalized, so that forms of social bias and statistical bias in AI interact with each other in a recursive manner.

Emerging research on the autodidactic nature of certain medical AI systems presents further considerations: Wawira Gichoya et al. demonstrated that when ML is applied to de-identified medical images such as radiographs, CT scans, and mammograms, it can predict a subject's self-identified racial identity with an accuracy of 80-99% across these imaging modalities [68]. The capability is readily acquired by standard ML algorithms that are trained with diverse data sets and accuracy persists, even after controlling for historic racial proxies like body-mass index, disease distribution and breast density.

These medical AI systems can accurately predict a patient's race from medical images that are corrupted, noisy, or cropped. Moreover, they base racial inferences on data that lies beyond the standard medical variables utilized by radiologists and other health care professionals. Thus, physicians may be unable to monitor and control this behavior when it is undesirable. Additionally, since many models are built using de-identified data sets, the inherent bias of these models may not be readily identifiable. Set against a deep legacy of racism in the U.S. healthcare system [69], the latent ability of medical AI systems to make unprompted

racial inferences is of serious concern. The same phenomenon has also been observed with gender inferences, where an AI system reading only de-identified retinal images was found to accurately predict multiple cardiovascular risk factors, including gender, even though ophthalmologists could not [6,70].

These findings highlight the immense scope for medical AI systems to deepen disparate patient outcomes along social divides, particularly where models can make undetected inferences from a vast mosaic of data. It is thus crucial to facilitate the development of equity-focused medical AI systems by both empowering research and development teams to identify and address the harmful effects of social bias in AI development practices and by holding teams accountable for the differential impact of their AI products [71], thereby ensuring that vulnerable patient populations are not further marginalized by rapid improvements in health innovation [72].

4. Geopolitical considerations for equity in medical AI system development

A primary issue facing the upstream segment of the AI ecosystem is the matter of data representativeness, as the preponderance of training data currently originates from the United States and China. A 2022 analysis of PubMed publications reveals that c.40% of the datasets referenced in medical AI literature are sourced from the United States and c.14% from China [73]. This leads to the development of solutions that generalize less effectively to underrepresented groups. A further concern is the issue of dominant groups in the research and development of AI as interwoven social and professional networks are prevalent in the medical research community, establishing and entrenching niche dominance [74]. There is a distinct lack of diversity within these niches as researchers with higher importance and centrality within their respective niches are seen to be less likely to be female or to come from low-and-middle-income countries (LMICs) [55,75,76].

The resulting centralization of knowledge production in high-income countries has created a power imbalance, where these regions have disproportionate influence on global standards and policies related to AI products. Moreover, the emerging disconnect between AI developers and the contexts in which their products are deployed means that there may be no process of accountability by which to address situations in which AI products cause harm – particularly when products from high-income countries are exported to other regions. International collaboration among regulatory bodies to intensify scrutiny on the datasets used to train medical AI systems, particularly when these systems are deployed in diverse geographic and social contexts [71], is one part of the solution. A comprehensive solution, however, also requires a focus on contextual testing to ensure that AI products are appropriate for the contexts in which they are deployed. This goal can be achieved by leveraging a growing global network of AI research groups through international partnerships. International collaboration among research groups not only fosters diversity and inclusion in medical AI system development by encouraging data exchange and open science practices, but also presents opportunities for capacity building in LMICs.

Although the AI landscape currently lacks coordinated regulation [77], the ecosystem has sought to adopt various initiatives that support a standard of trustworthiness for AI. Successive waves of self-regulation have seen the adoption of Datasheets for Datasets [78], which improves communication between dataset creators and users. The subsequent adoption of Model Cards [21] was a further step to standardize the disclosure of key performance characteristics for trained ML algorithms. These cards detail operating parameters, such as intended use cases and performance evaluation criteria, which helps to reduce the deployment

of trained ML algorithms in scenarios for which they are not adequately designed. Algorithm assurance [79] through model audits has also been proposed as an IT risk management protocol, and the growing number of audits signals a commitment to the development of equity focused AI [22]. However, more is required to realize a standard of trustworthiness for AI in healthcare.

5. Core attributes of the TC

TCs can provide relevant information about the authorship of medical AI systems. While they are not intended to be prescriptive, they should include core information that enables stakeholders to understand salient positionality attributes of the contributors behind a given medical AI system. We note that guidelines established by the SRQR and COREQ have been criticized for being overly rigid and for adopting a focus on checklist detail, rather than retaining focus on the holistic outcomes of relevant disclosures [80]. This protocol considers the shortcomings of comparable frameworks and instead invites teams to nominate the most relevant elements for disclosure in each situation. Written statements, diagrams, illustrations, audio-visual and multi-modal content are all viable media for the expression of positionality in this disclosure protocol.

Core attributes of the TC broadly comprise a discussion of team positionality and disclosure of team composition and are detailed below (Table 1). They can be thought of as a synthesis of two primary tenants of AI assurance and governance: (i) a self-reflective component to support the effective mitigation of harmful bias and (ii) a disclosure component to encourage greater accountability. We maintain that teams, and individuals within them, should retain control over what they elect to disclose and should not be compelled to reveal private aspects of their positionality. We however emphasize that while teams are free to add or omit information as they deem relevant, a culture of transparency is essential for the protocol's practical success and for fostering more responsible practices in medical AI system development.

Like the teams they describe, TCs are not intended to remain static. They should be updated periodically to ensure they accurately reflect the evolving composition and positionality of teams. We also recommend that prior contributors, who may no longer be active team members, be appropriately acknowledged to support the fair recognition of contributions over time and to further cultivate a culture of inclusion. Finally, while we believe that TC disclosures should primarily include information typically disclosed in research endeavors, we acknowledge that publicly sharing sensitive personal information may pose privacy risks. Teams should carefully consider these factors when compiling TC disclosures.

6. A presentation of our TC

The authoring team experimented with the TC protocol to investigate whether social diversity within our research team could effectively contribute to epistemic diversity and ultimately assist us to interrogate areas of potential bias in the development of this manuscript. Although we did not engage in AI production, our objective was to apply the TC framework experimentally, as a lens for identifying and addressing potential biases in our research. We present the TC compiled by the authors of this manuscript as a tangible example of how the protocol can be interpreted to convey salient information about team composition and the positionality attributes conferred by that composition.

We find that the social diversity of our team, combined with the various disciplines and competences among authors, as well as a shared commitment to reflexive dialogue on how these lenses shape our understanding of bias in AI, together enriched epistemic diversity in our research and manuscript preparation. We posit that this enhanced epistemic diversity

Table 1. Core attributes of the TC.

CORE ATTRIBUTE		DISCLO-SURE PARAMETER	GOVERNANCE STANDARD	GOVERNANCE OBJECTIVES & ACTION STEPS
Discussion of positionality		[1] Positionality	Transparency, accountability	• A discussion of the perspective from which the problem is being approached, including reflections on appointed methodology and data selection, as well as team composition. Areas of concern for potential harmful bias and the measures taken to mitigate these risks are highlighted
Disclosure of team composition	**Team structure**	[2] Roles & expertise	Ethical oversight	• Include team functions that retain a focus on ethical considerations
			End-user advocacy	• Include perspectives that advocate for the end-user (e.g. physicians and patient advocates)
			Regulatory compliance	• Maintain compliance with regulatory standards by including dedicated or consulted compliance functions
		[3] Institutional affiliation	Multi-disciplinary collaboration	• Consider cross-disciplinary collaborations to include varied perspectives and expertise
			Regulatory body engagement	• Maintain a dialogue with relevant regulatory bodies to enhance compliance functions
			Ethical review	• Design AI system development process that anticipates the formal review and approval of an appropriately constituted committee
		[4] Context/ location	Contextual testing and validation	• Test and validate AI tools in the geographic contexts in which they will be deployed
	Team identity	[5] Identity	Inclusion	• Inclusion is dynamic in that it recognizes dimensions of diversity that emerge from identities, as they pertain to team objectives, including the intersection of socioeconomic status, ethnicity, cultural background, gender, ability, and sexual orientation - but goes beyond to ensure belonging, respect, and success. Appropriate governance standards should comprise elements of bias mitigation as well as those that take into account inclusivity in medical AI design • Consider modes of team engagement and TC presentation that promote accessibility of the content and accommodate effective collaboration across a potential range of abilities. This might include consideration of formats such as audio and illustration • Consider documenting the rationale behind the problem definition and the selection of proxy variables, reflect on how the framing aligns with intended outcomes, evaluate how closely the proxies approximate the true variables of interest, and implement measures to mitigate foreseeable risks

https://doi.org/10.1371/journal.pdig.0000495.t001

could play a meaningful role in supporting effective bias mitigation strategies in the development of medical AI systems, and advocate for the adoption of TCs in other teams aiming to address similar challenges. This approach not only supports bias mitigation but also establishes a foundation for empirically assessing the value of diversity in shaping ethical and effective medical AI systems. Finally, we emphasize that a cultural shift in AI development that shows commitment to meaningful engagement across perspectives, is essential to the scuccess of the TC and similar bias mitigation strategies.

We re-iterate that the TC presented here serves as an example relevant to our specific context—preparing a manuscript on the TC protocol itself—and thus does not include the technical considerations that would apply to AI development. We acknowledge that different interpretations, whether through text, visual, or mixed media, could also effectively implement the protocol. We later discuss two illustrative cases to further demonstrate the protocol's practical implementation and benefits.

[1] Discussion of the authoring team's positionality

This area of research was prompted by the observations of physician-scientists on the team, who noted structural inequality in healthcare outcomes across patient populations. These observations raised concern for the potential deepening of disparities along social fault lines, with the layering of AI on legacy healthcare systems. The diverse positionalities within the team include perspectives informed by lived experiences of intersecting oppressions,

providing a nuanced understanding of how individuals may encounter discrimination in healthcare. These perspectives also highlight the influence of social determinants on health outcomes across patient groups. The following discussion synthesizes the positionality statements prepared by individual team members. The statements can be accessed at the following link: https://www.teamcard.io/team/demo.

The authoring team's positionality reflects the diverse experiences and commitments of a research team dedicated to addressing bias and promoting equity in medical AI systems. Members bring unique, firsthand perspectives from varied global contexts, including clinical experience in the U.S., Brazil, the Philippines, Iran, China, the UK, and Tasmania - which drives the team's focus on health equity and access for marginalized populations. Others on the team emphasize reflexivity and transparency in AI research, advocating for inclusive methodologies informed by their backgrounds in machine learning, anthropology, and public health. The team's commitment to ethically responsible AI development is further strengthened by the contributions of members with expertise in bioethics, and feminist philosophy, who focus on mitigating harm and bias in healthcare applications. Additionally, members of the team share a dedication to matters of social justice, with personal and professional experiences of discrimination shaping their advocacy for accountability and governance in the development of medical AI systems. By integrating technical, social, and ethical dimensions, we believe the mix of competencies and perspectives on the team facilitated robust and critical assessment of potential blind spots in the research.

[2] Roles and expertise

Authorship was engaged collaboratively by a team of 16 individuals, 12 of whom identify as clinicians, ML engineers, data scientists or an intersection of these descriptors. This allowed for self-reporting and advocacy from perspectives across the medical AI value chain, ranging from AI product developers to healthcare professionals implementing medical AI systems as end-users. Remaining authors contributed expertise as anthropologists, social science researchers, bioethicists and healthcare technology investors, each providing an unique lens on the social determinants of equitable AI. Additionally, 2 members of the team contributed ethical review and regulatory compliance considerations, informed by their engagements with institutional review boards.

[3] Institutional affiliations

Institutional affiliations of the authoring team include Duke University, Harvard Medical School, Mass General Brigham (MGB), Massachusetts Institute of Technology, University of Oregon, Technical University of Munich School of Medicine and Health, the Beth Israel Deaconess Medical Center (BIDMC), and the University of the Philippines Manila. Engagement across the team delivered multi-disciplinary collaboration by integrating expertise across clinical practice, computer science and engineering and the social sciences. Although the preparation of this manuscript is not subject to regulation, direct engagement with research oversight and ethical review bodies was solicited through the contributions of sitting members of the MGB IRB, who form part of the authoring team. We acknowledge that our understanding of the regulatory landscape for medical AI systems could be further strengthened by consulting relevant individuals or bodies as these frameworks evolve. In the interim, we have relied on the appropriate experience of members on the team and desktop research. No funding affiliation is relevant to this manuscript.

[4] Context/ geographic location

The authoring team's members are active researchers in Australia, Brazil, Germany, Indonesia, Ireland, Spain, the U.S. and the U.K.. They have each contributed perspectives on harmful bias

mitigation that are relevant to the AI research and development ecosystems prevailing in their respective regions, as well as the clinical contexts in which medical AI systems are adopted in these regions. Additionally, authors based outside of their home countries have contributed relevant impressions that attune to the contexts of their countries of origin, which include China, Iran, Nigeria and South Africa. We believe that this geographic diversity enables a rich contextual evaluation of the TC protocol, which we observe to be robust across various institutional settings, spanning academia and industry, in the regions considered.

We however note, as a key limitation, that the authoring team's research activity and clinical practice is predominantly U.S. based. Moreover, most members of the team are institutionally affiliated with a U.S. based research entity. To illustrate this point, the research clusters shown in Fig 1, represent the manner in which information flow, areas of focus and ideas are organized in the authoring team. They converge to L.C., who heads the Laboratory for Computational Physiology at the Massachusetts Institute of Technology and is central to the formation of the authoring team. This centrality of influence implies an inherent perspective skew towards U.S. institutional norms in the team's perception of an effective TC. While the team has made efforts to incorporate diverse regional perspectives, this limitation, in our view, remains only partially mitigated.

[5] Identity

The authoring team represents a diversity of cultural backgrounds and ages with appropriate gender representation. Self-reported ethnicities and cultural identities of our team members include: African-American, Arabian, East Asian, European, Igbo, Jewish American, KōKwel, Persian, Southeast Asian, Tswana, White American and White British. Members range in age from their 20s to their 50s, comprise self-identified men and women, and include members of the LGBTQIA+ community. Additionally, the team brings firsthand experience with the implementation of medical AI systems in clinical settings, as well as expertise across the development lifecycle of these systems, contributing to discussions on the need for greater accountability and diversity in AI research and development.

However, a key limitation is that the collective perspective of the team may be skewed toward that of a socioeconomic group with high levels of educational attainment and professional ties to prominent medical institutions. This bias is particularly relevant to the team's
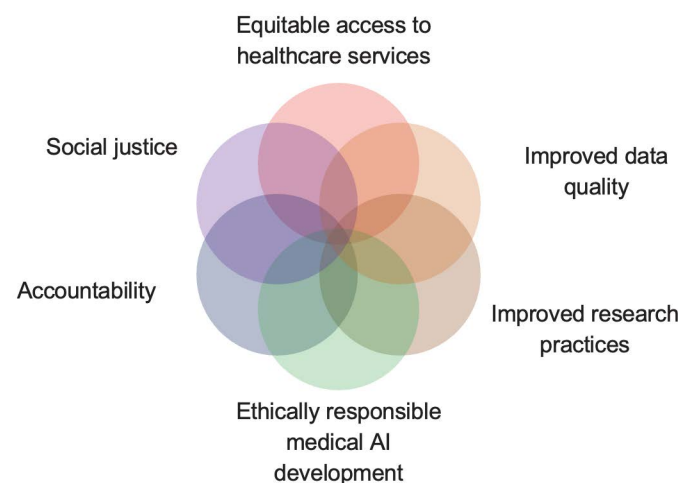

Fig 1. Perspectives from which this research was approached.

https://doi.org/10.1371/journal.pdig.0000495.g001

perception of AI and its medical applications, which we acknowledge may be positively skewed. Consequently, the narrative presented in this manuscript could be enriched by the inclusion of socioeconomic perspectives that have less proximity, and differing levels of access, to the medical establishment. See Fig 2 for our team's relational network.

We note that the inclusion of photographs with corresponding pronouns and LGBTQIA+ affiliation could improve the disclosure by providing a more complete visual representation of our social diversity. This level of disclosure has not been pursued by the team in recognition of the potentially sensitive nature of these attributes, due to the risk of potential discrimination. We have aggregated certain descriptors for the purposes of this illustration but could provide more granular disclosure to oversight bodies as needed. While efforts have been taken to balance dual commitments to individual privacy as well as to transparency and accountability, individual attributes that are not directly interpretable from the illustration prepared, could be vulnerable to reidentification in the context of subsequent TC disclosures.
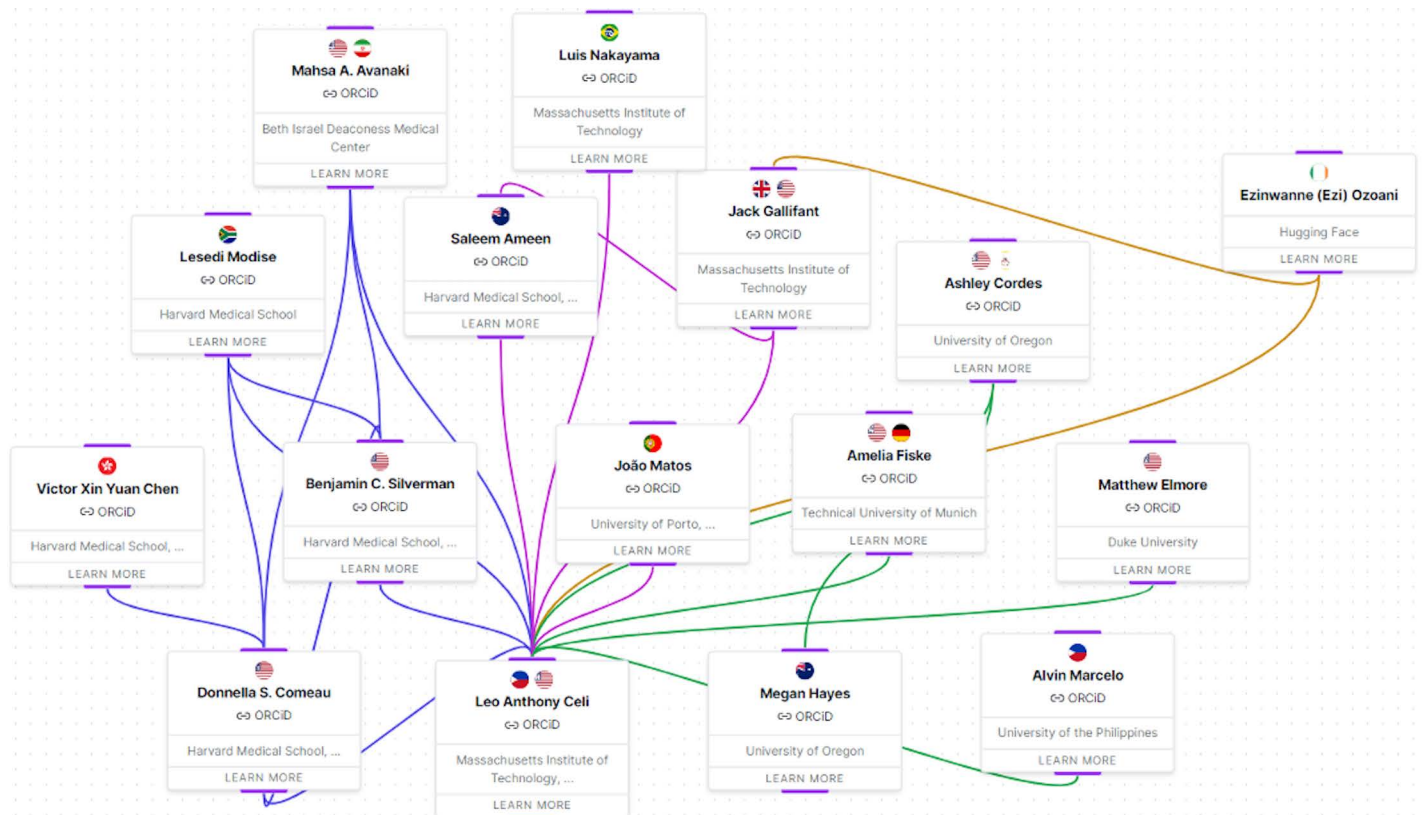
Fig 2. **Relational network prepared by the authoring team.** Our team's relational network is visually represented as a series of spatial arrangements, communicating the overall team structure. Additionally, team members are interconnected with color-coded bands that denote functional clusters within the group. *Figure representing relationships among team members found at:* https://www.teamcard.io/team/demo. *It utilizes a forced-directed graph to display the multifaceted relationships and backgrounds of the team, in a clear, visually discernible way. The visualization was rendered using React Flow (a node-based graph visualization library), and the layout algorithm was implemented with the support of D3.js (a toolkit for data-driven documents). The graph consists of two fundamental components: nodes that represent team members, and edges that represent the relationships between team members. Prior to rendering, nodes were labeled according to their group affiliation, so that clusters of closely connected individuals can be displayed within close proximity to one another. After labeling, a structured circular layout for each cluster was implemented, such that each node within a cluster is evenly spaced to form a radial distribution of equal segments that are proportional to the number of nodes within the given cluster. This initial placement served as a starting point for the dynamic force simulation, which is implemented to iteratively refine node positions to an equilibrium state that represents the natural interconnections among team members.*

https://doi.org/10.1371/journal.pdig.0000495.g002

7. Case studies

We further illustrate how the TC framework can be implemented for practical bias mitigation in the development of medical AI systems, through appropriately selected examples. We discuss cases of actual medical AI systems with identified bias, to demonstrate how the integration of the TC protocol could promote reflective development practices that better address and mitigate harmful biases. We also note that TCs are not intended to only be used in cases of demonstrated harm; all AI systems encode the possibility of harmful bias and can benefit from improved reflexivity among the teams developing these systems.

Illustrative case 1: neurology - medical AI systems for the diagnosis of dementia

Several systematic reviews of studies applying ML techniques to automatic speech and language processing for early dementia detection reveal that the use of speech datasets lacking in gender and age diversity is common practice in the field [81,82,83,84]. This reduces the accuracy and generalizability of resulting applications across demographic groups. Although there is recognition that commonly used dementia datasets should appropriately reflect the effects of age and gender on speech patterns [85], significant blind spots in the research persist. Key demographic factors that may impact dementia diagnosis—such as ethnicity, socioeconomic status, first language, and education level—are frequently overlooked in model training [84]. Additional challenges include an over-representation of Alzheimer's disease relative to other dementia types in commonly used datasets [81], limited accent diversity in English-language datasets, insufficient language diversity, and a focus on languages primarily spoken in high-income regions [86]. Integrating TC protocols into research on ML techniques for automatic speech and language processing would support a shift away from standard practices with observed bias.

- Reflexivity: the reflective components of the TC protocol would be of particular relevance to the pre-processing and data compilation stages of the AI development lifecycle for dementia detection applications. Integrating the TC protocol would require research teams to openly reflect on the contextual factors shaping both the focus of their investigation and the datasets they select. This includes transparent discussions on data appropriateness and quality, identification of potential gaps, and actions taken to address these shortcomings.

- Representation: the protocol's focus on appropriate representation in research teams would assist in surfacing often-overlooked factors, such as educational attainment and socioeconomic status, as well as culturally specific parameters like dialect, accent, and variations in syntax and intonation across languages. These variables significantly influence speech patterns and could impact the accuracy of automated dementia detection applications.

- End-user advocacy and ethical oversight: including patient advocate perspectives would enhance the interpretive resources available to researchers, allowing for the integration of important aspects of dementia's lived experience into AI-powered early detection systems. These perspectives would help researchers better understand communication patterns in aging individuals, thereby mitigating the risk of unjust outcomes, such as misinterpretations of healthy speech patterns as signs of neurological decline.

Illustrative case 2: facility management - algorithmic risk prediction in clinical settings

Obermeyer et al [61]. demonstrate that algorithmic risk prediction tools commonly used in healthcare facilities across the U.S., result in substantial racial disparities in the program enrolment of over 200 million individuals annually. The tools are used to identify patients with complex care needs and accordingly, determine which should be enrolled in intensive care management programs. These tools have widely adopted historic health care spending, derived from insurance claim records, as a proxy for needed care. This inaccurately assumes that patients with lower historic healthcare costs are healthier than equally sick patients with higher spending and introduces a bias that favors White patients, who generally have better access to care and spend more on medical services than other racial groups in the U.S. This research underscores that algorithms designed to optimize healthcare expenditure can perpetuate racial biases, even when race is not explicitly observed as a variable. This occurs because the underlying problem definition — whether to prioritize the reduction of healthcare costs or to increase access to care — reflects and reinforces existing structural inequities embedded in historical practices [87].

Similarly, Sarkar et al. [65] find that clinical prediction models used during the COVID-19 pandemic to inform ICU triaging and decisions on the continuation of mechanical ventilation in resource-limited settings exhibit risk scoring biases across ethnic groups. Several commonly used models overpredict mortality in Black and Latino patients in the U.S., raising concerns about the disproportionate withholding or withdrawal of treatments (such as ventilators) in these demographics during the pandemic and in future crises.

These examples, along with other cases of algorithmic bias in clinical risk prediction, not only reflect but also exacerbate existing health inequalities by promoting unequal access to potentially life-saving treatments underscoring the need for a robust assurance protocol that could include the TC protocol and other prudential measures.

- Reflexivity: Since risk prediction algorithms are often applied to electronic health record (EHR) systems — where researchers may have limited control over data sources — the reflexive aspects of the TC protocol can help ensure greater scrutiny of algorithm design and specification. By incorporating the TC, research teams would need to assess the suitability of a given algorithm for EHR data, considering data quality, missing variables, and the potential impact of these limitations on predictive accuracy and bias. The protocol also encourages researchers to document their rationale for problem definition and the selection of proxy variables, reflect on how problem framing aligns with intended outcomes, evaluate how closely the proxies approximate the true variables of interest, and implement measures to mitigate any foreseeable risks. This approach promotes the more considered use of EHRs in algorithmic decision-making.

- Representation: A component of the harmful bias observed in risk prediction algorithms is attributable to researcher assumptions regarding what can be inferred from EHRs. Diverse representation within research teams can thus help bring to light potential limitations in these assumptions, particularly regarding how certain rules or variables may unintentionally exclude or disadvantage specific groups.

- Contextual testing and ethical review: Given the sensitive context in which clinical risk prediction algorithms are deployed, rigorous contextual testing prior to solution deployment is critical to the early identification of potential harmful bias. This testing should form part of an ethical review conducted by an IRB or a suitably qualified committee that can assess the social determinants influencing patterns in EHR data. An ethical review would assist researchers to identify and mitigate the effects of systemic inequality in available data, reducing the risk of deepening disparities in patient outcomes through algorithmic decision-making.

8. Considerations and limitations for operationalizing the TC and for ensuring diversity in the research and development of medical AI systems at scale

The TC's practical approach to addressing unconscious bias in AI development through reflexivity has the potential to support a necessary cultural shift in the research and development of medical AI systems. While it has clear strengths, including its emphasis on ethical alignment, transparency, and accountability, its limitations – such as the lack of standard format and challenges to implementation – highlight areas for further refinement and adaptation.

Key strengths of the protocol

- **Effective bias mitigation through reflexivity.** The primary strength of the TC protocol is that it serves to mitigate harm caused by unconscious bias, by putting the people behind the research and development of medical AI systems in transparent and reflective conversation with the performance of those systems in the field.

- **Practical integration of ethical considerations.** Another key strength of the protocol is that it integrates ethical considerations into AI research and development practices, serving to align the development of medical AI systems with established ethical standards for medical research [88].

- **Enhanced transparency and accountability**. By enhancing transparency regarding the potential biases embedded in medical AI systems, the TC protocol allows for more informed and discerning interaction with these products. This empowers physicians and other end users to act as more effective agents in mitigating discrimination by clinical algorithms. Thus, a further strength of the TC is that it aligns with the proposed regulation [9] for accountability in the development of medical AI systems.

- **Fair recognition of contributors.** By reporting on team composition with an emphasis on inclusion, the TC provides a practical mechanism for acknowledging the collective effort involved in the development of medical AI systems. It promotes the transparent recognition of contributors who might otherwise be overlooked and advocates for the inclusion of potential collaborators who might not traditionally be solicited for contribution.

Key limitations of the protocol and possible mitigants

- **The lack of a standard format for TCs**. As the protocol is intended to be a transparent and informative expression of reflexivity in clinical AI research and the AI development process, the form and content of TCs may vary substantially across a diversity of applications and

team structures in the ecosystem. While this flexibility has its advantages, it also challenges the establishment of fixed guidelines for TC reporting. As a mitigant, we note that positionality statements in qualitative research serve as a starting point in understanding the objectives and the necessary elements of a TC disclosure. From this basis, teams should interpret the cards as a conduit for self-reflection, acknowledgment of potential bias, and evaluation of the impact that any bias may have on the research under consideration. While the most appropriate format of this expression should be left to teams' discretion, the content of the disclosure should respect a common understanding of the principles above. Additionally, asynchronous AI governance processes could prompt marked variation in the quality and the orientation of TC disclosures across regions. Here we believe that auxiliary initiatives to promote transparent, accurate, and useful reporting may include the implementation of validation protocols that are designed to identify and flag critical omissions by teams. One such validation protocol could be a medical AI development process that anticipates the formal review and approval of an IRB or an appropriately constituted ethics committee.

- **A required cultural shift.** TCs may be challenging to operationalize due to the voluntary and self-reported nature of the protocol. As the TC emphasizes epistemic practices—requiring self-reflection by the individuals behind medical AI systems regarding decisions about data inclusion, system performance, and development priorities—adoption may be inconsistent across the AI landscape, where reflective practices are often undervalued. Addressing this challenge necessitates a cultural shift toward reflexive approaches as essential components of bias mitigation. This shift can be facilitated by demonstrating the limitations of existing data and product focused protocols, such as Datasheets for Datasets, Model Cards and Algorithm Assurance, which currently offer only partial solutions to addressing bias in AI.

- **Privacy and integrity concerns.** The protocol may raise data privacy concerns and prompt pushback from team members regarding the public reporting of their personal information, with respect to unintended consequences such as potential tokenization and stigmatization. In response to this limitation, we maintain that while the TC is intended to promote greater transparency, it is nevertheless bound by standard privacy protection regulations such as the Health Insurance Portability and Accountability Act or the General Data Protection Regulation [89], which should allow teams to retain autonomy in their personal disclosures. Additionally, we contend that much of the information required for the TC is already disclosed by teams, through various professional and social media. However, this information is scattered across various platforms (e.g.,: websites, publications, presentations) and media (e.g.,: video and print) to little social benefit, whereas a protocol that centralizes all relevant disclosures by a given researcher, would greatly enhance transparency and accountability in the development of medical AI systems. The TC is thus an opportunity to consolidate existing metadata and to systematize the collection of further information. Future adaptations including integration of standard identifiers such as ORCID IDs in the protocol would serve to streamline the acquisition and evaluation of researcher information (ORCID is a non-profit organization that provides researchers with a unique digital identifier).

- **Practical impediments to implementation at scale.** We also note that the current scale of institutional research and development of medical AI systems presents a practical impediment to the implementation of the TC; certain disclosure parameters may become cumbersome to track where projects involve multiple teams collaborating across various stages of the AI development lifecycle. Here we might suggest multiple TCs for a single medical AI

system, prepared in condensed and modular fashion, for discrete stages in the AI lifecycle. These disclosures could be accompanied by a clear record of the AI system's progression, linking and narrating all relevant TC disclosures. The same approach could be adopted in instances where pre-existing models are leveraged by researchers, including instances in which open-source materials are used. The reflective elements of the TC module prepared for such cases would thus require a transparent discussion that interrogates the necessity and appropriateness of adopting pre-existing materials, selection criteria for these materials, as well as further considerations pertaining to possible implications for the research. These considerations should address the potential for legacy bias in the research and should consider actions to mitigate these effects.

- **The limitations of team diversity.** While we maintain that social diversity alone is insufficient to achieve the epistemic diversity necessary for effective bias mitigation in AI, it remains a valuable starting point for assembling orthogonal perspectives that can broaden the scope of inquiry and help identify potential blind spots. The TC provides a practical framework to facilitate this process by enabling teams to critically reflect on the positionality introduced by their composition and offers a system for tracking the relationship between diversity and epistemic outcomes, creating opportunities to better understand the value of diverse perspectives in AI research and development. We acknowledge that a focus on social diversity in teams has the potential to crowd out other priorities. This may be observed in instances where skill-based vetting of team members is revised to support diversity and inclusion. More broadly, existing social inequalities that perpetuate demographic skews in the market for talent in AI research and development will indeed limit the extent to which the social diversity promoted by the protocol can be operationalized in the landscape. Thus, in teams where there is a perceived lack of social diversity, the reflective elements of the protocol can focus on transparently acknowledging positionality skews within the team, considering their potential implications for the research – such as anticipated biases – and implementing measures to mitigate these risks. Tools such as Implicit Association Testing can further support this reflection by identifying unconscious biases that may otherwise go unnoticed but could influence the research outcomes.

9. Conclusion

The TC holds the promise of facilitating a deeper understanding of how diversity—both social and epistemic—shapes the development of ethical and robust medical AI systems. It provides concrete suggestions for embedding reflexivity into AI development workflows, equipping teams with tools to systematically document positionality variables and critically examine their influence on bias and decision-making.

This approach reframes the challenge of bias mitigation, shifting focus from narrow technical fixes to the subjective human dynamics that underlie AI design. Our experience with the TC underscores its potential to reveal and address the biases often encoded in medical AI systems, by fostering epistemic diversity and critical self-reflection in AI development practices.

As a prototype, the TC is a proof of concept—a tool designed to gather qualitative and quantitative data on team dynamics, with the goal of informing and refining future practices. Research from fields such as business, science, and organizational behavior provides compelling evidence that prioritizing diversity improves decision-making, fairness, and performance. The TC builds on this foundation, laying the groundwork for a cultural shift in AI research and development. Future iterations will expand its utility, adapting it to diverse contexts and scaling its impact to ensure inclusivity and accountability in the medical AI systems we build.

Author contributions

Conceptualization: Lesedi Mamodise Modise, Mahsa Alborzi Avanaki, Saleem Ameen, Leo A. Celi, Victor Xin Yuan Chen, Ashley Cordes, Matthew Elmore, Amelia Fiske, Jack Gallifant, Megan Hayes, Alvin Marcelo, Joao Matos, Luis Nakayama, Ezinwanne Ozoani, Benjamin C. Silverman, Donnella S. Comeau.

Data curation: Saleem Ameen, Jack Gallifant.

Methodology: Saleem Ameen.

Project administration: Lesedi Mamodise Modise, Jack Gallifant, Donnella S. Comeau.

Resources: Leo A. Celi.

Software: Lesedi Mamodise Modise, Mahsa Alborzi Avanaki, Saleem Ameen, Leo A. Celi, Jack Gallifant, Donnella S. Comeau.

Supervision: Leo A. Celi, Benjamin C. Silverman, Donnella S. Comeau.

Visualization: Lesedi Mamodise Modise, Mahsa Alborzi Avanaki, Saleem Ameen, Leo A. Celi, Jack Gallifant, Donnella S. Comeau.

Writing – original draft: Lesedi Mamodise Modise, Mahsa Alborzi Avanaki, Saleem Ameen, Leo A. Celi, Ashley Cordes, Matthew Elmore, Amelia Fiske, Jack Gallifant, Megan Hayes, Joao Matos, Luis Nakayama, Benjamin C. Silverman, Donnella S. Comeau.

Writing – review & editing: Lesedi Mamodise Modise, Mahsa Alborzi Avanaki, Saleem Ameen, Leo A. Celi, Ashley Cordes, Matthew Elmore, Amelia Fiske, Jack Gallifant, Megan Hayes, Luis Nakayama, Benjamin C. Silverman, Donnella S. Comeau.

References

1. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3(1):17. https://doi.org/10.1038/s41746-020-0221-y PMID: 32047862

2. Zhan A. Towards AI-assisted healthcare: System design and deployment for machine learning based clinical decision support [dissertation]. Baltimore (MD): The Johns Hopkins University; 2018.

3. House Committee on Ways & Means. Fact versus fiction: clinical decision support tools and the (mis) use of race. Congressional Publications [Internet]. 2022 Sep 29. [cited 2024 Jul 28]; Available from: https://perma.cc/25NR-YBY7

4. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med. 2018;15(11):e1002689. https://doi.org/10.1371/journal.pmed.1002689 PMID: 30399149

5. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337–40. https://doi.org/10.1038/s41591-019-0548-6 PMID: 31427808

6. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866–72. https://doi.org/10.7326/M18-1990 PMID: 30508424

7. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA. 2018;320(21):2199–200. https://doi.org/10.1001/jama.2018.17163 PMID: 30398550

8. Liu H, Wang Y, Fan W, Liu X, Li Y, Jain S, et al. Trustworthy AI: a computational perspective. ACM Trans Intell Syst Technol. 2022;14(1):1–59. https://doi.org/10.1145/3546872

9. The White House. FACT SHEET: President Biden issues executive order on safe, secure, and trustworthy artificial intelligence. 2023 Oct 30 [cited 2024 Mar 24]; Available from: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

10. U.S. Food and Drug Administration. Clinical decision support software: Draft guidance for industry and Food and Drug Administration staff. 2019. [cited 2024 Jul 28]; Available from: https://www.regulations.gov/document/FDA-2017-D-6569-0041

11. U.S. Food and Drug Administration. *Clinical decision support software: Guidance for industry and Food and Drug Administration staff.* [cited 2024 Mar 25] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software 2022 [cited 2024 Mar 25]; Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software

12. Shachar C, Gerke S. Prevention of bias and discrimination in clinical practice algorithms. JAMA. 2023;329(4):283–4. https://doi.org/10.1001/jama.2022.23867 PMID: 36602791

13. Goodman KE, Morgan DJ, Hoffmann DE. Clinical algorithms, antidiscrimination laws, and medical device regulation. JAMA. 2023;329(4):285–6. https://doi.org/10.1001/jama.2022.23870 PMID: 36602795

14. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3(1):118. https://doi.org/10.1038/s41746-020-00324-0 PMID: 32984550

15. Lee JT, Moffett AT, Maliha G, Faraji Z, Kanter GP, Weissman GE. Analysis of devices authorized by the FDA for clinical decision support in critical care. JAMA Intern Med. 2023;183(12):1399–401. https://doi.org/10.1001/jamainternmed.2023.5002 PMID: 37812404

16. The Department of Health and Human Services, Centers for Medicare & Medicaid Services. Nondiscrimination in health programs and activities: Proposed rule. Federal Register [Internet]. 2022 Aug 4 [cited 2024 Jul 28]; 87 FR 47824. Available from: https://www.federalregister.gov/documents/2022/08/04/2022-16217/nondiscrimination-in-health-programs-and-activities

17. Giovanola B, Tiribelli S. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. AI Soc. 2023;38(2):549–63. https://doi.org/10.1007/s00146-022-01455-6 PMID: 35615443

18. Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. Can J Cardiol. 2022;38(2):204–13. https://doi.org/10.1016/j.cjca.2021.09.004 PMID: 34534619

19. Crigger E, Reinbold K, Hanson C, Kao A, Blake K, Irons M. Trustworthy augmented intelligence in health care. J Med Syst. 2022;46(2):12. https://doi.org/10.1007/s10916-021-01790-z PMID: 35020064

20. Mello MM, Shah NH, Char DS. President biden's executive order on artificial intelligence-implications for health care organizations. JAMA. 2024;331(1):17–8. https://doi.org/10.1001/jama.2023.25051 PMID: 38032634

21. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Association for Computing Machinery, editor. FAT* '19: Proceedings of the conference on fairness, accountability, and transparency; 2019 Jan 29; New York, USA. New York: Assoc Computing Machinery; 2019.p. 220–229. https://doi.org/10.1145/3287560.3287596

22. Charpignon M-L, Byers J, Cabral S, Celi LA, Fernandes C, Gallifant J, et al. Critical bias in critical care devices. Crit Care Clin. 2023;39(4):795–813. https://doi.org/10.1016/j.ccc.2023.02.005 PMID: 37704341

23. Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health. 2023;2(6):e0000278. https://doi.org/10.1371/journal.pdig.0000278 PMID: 37347721

24. Gallifant J, Zhang J, Whebell S, Quion J, Escobar B, Gichoya J, et al. A new tool for evaluating health equity in academic journals; the diversity factor. PLOS Glob Public Health. 2023;3(8):e0002252. https://doi.org/10.1371/journal.pgph.0002252 PMID: 37578942

25. Harding S. Whose science? Whose knowledge?: Thinking from women's lives. Ithaca (NY): Cornell University Press; 2016.

26. Tuhiwai SL. Decolonizing methodologies: research and indigenous peoples. London (UK): Zed Books; 2021. https://doi.org/10.5040/9781350225282

27. Haraway D. Situated knowledges: the science question in feminism and the privilege of partial perspective. Feminist Studies. 1988;14(3):575. https://doi.org/10.2307/3178066

28. Adas M. Machines as the measure of men: Science, technology, and ideologies of Western dominance. Ithaca (NY): Cornell University Press; 2015.

29. Chilisa B. Indigenous research methodologies. Thousand Oaks (CA): Sage Publications; 2020.

30. Berger R. Now I see it, now I don't: researcher's position and reflexivity in qualitative research. Qual Res. 2013;15(2):219–34. https://doi.org/10.1177/1468794112468475

31. Liboiron M. Pollution Is colonialism. Durham (NC): Duke University Press; 2021. https://doi.org/10.1215/9781478021445

32. Liamputtong P. Researching the vulnerable: a guide to sensitive research methods. London (UK): Sage Publications; 2007. https://doi.org/10.4135/9781849209861

33. Wilson S. Research is ceremony: Indigenous research methods. Halifax (NS): Fernwood Publishing; 2008.

34. Gray J, Witt A. A feminist data ethics of care for machine learning: The what, why, who and how. First Monday [Internet]. 2021 Dec;26(12): 1. https://doi.org/10.5210/fm.v26i12.11833

35. Franks M. Feminisms and cross-ideological feminist social research: standpoint, situatedness and positionality – developing cross-ideological feminist research. J. Int. Women's Stud. 2002 May [cited 2024 Jul 28];3(2):38–50. Available from: https://vc.bridgew.edu/cgi/viewcontent.cgi?article=1601&context=jiws

36. Cave S, Dihal K. The whiteness of AI. Philos Technol. 2020;33(4):685–703. https://doi.org/10.1007/s13347-020-00415-6

37. Freire A, Porcaro L, Gómez E. Measuring diversity of artificial intelligence conferences. arXiv. 2021. https://doi.org/10.48550/arxiv.2001.07038

38. Zhang J, Whebell S, Gallifant J, Budhdeo S, Mattie H, Lertvittayakumjorn P, et al. An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research. Lancet Digit Health. 2022;4(4):e212–3. https://doi.org/10.1016/S2589-7500(22)00032-2 PMID: 35337638

39. Newman MEJ. Co-authorship networks and patterns of scientific collaboration. Proc Natl Acad Sci U S A. 2004;101(1):5200–5. https://doi.org/10.1073/pnas.0307545100 PMID: 14745042

40. Stahl GK, Maznevski ML, Voigt A, Jonsen K. Unraveling the effects of cultural diversity in teams: A meta-analysis of research on multicultural work groups. J Int Bus Stud. 2009;41(4):690–709. https://doi.org/10.1057/jibs.2009.85

41. Davison SC, Ekelund BZ. Effective team processes for global teams. In Lane HW, Maznevski ML, Mendenhall ME, McNett T, editors. The Blackwell handbook of global management. Oxford (UK): Blackwell Publishing; 2017.

42. Freeman RB, Huang W. Collaboration: Strength in diversity. Nature. 2014;513(7518):305. https://doi.org/10.1038/513305a PMID: 25230634

43. Apfelbaum EP, Phillips KW, Richeson JA. Rethinking the baseline in diversity research: should we be explaining the effects of homogeneity?. Perspect Psychol Sci. 2014;9(3):235–44. https://doi.org/10.1177/1745691614527466 PMID: 26173261

44. Swartz TH, Palermo A-GS, Masur SK, Aberg JA. The science and value of diversity: closing the gaps in our understanding of inclusion and diversity. J Infect Dis. 2019;220(220 Suppl 2):S33–41. https://doi.org/10.1093/infdis/jiz174 PMID: 31430380

45. Page SE. The difference: how the power of diversity creates better groups, firms, schools, and societies. Princeton (NJ): Princeton University Press; 2007.

46. Freeman RB, Huang W. Collaborating with people like me: ethnic co-authorship within the United States. J Labor Econ. 2015;33(S1):S289–318. https://doi.org/10.1086/678973

47. Rhoten D, Parker A. Education. Risks and rewards of an interdisciplinary research path. Science. 2004;306(5704):2046. https://doi.org/10.1126/science.1103628 PMID: 15604393

48. Gomez LE, Bernet P. Diversity improves performance and outcomes. J Natl Med Assoc. 2019;111(4):383–92. https://doi.org/10.1016/j.jnma.2019.01.006 PMID: 30765101

49. Hunt V, Layton D, Prince S. Diversity matters. McKinsey & Company [Internet]. 2015 [cited 2024 Jul 23]. Available from: https://www.mckinsey.com/insights/organization/~/media/2497d4ae-4b534ee89d929cc6e3aea485.ashx

50. Hunt V, Yee L, Prince S, Dixon-Fyle S. Delivering through diversity. McKinsey & Company [Internet]. 2018 [cited 2024 Jul 23]. Available from: https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/delivering-through-diversity#/

51. Hunt V, Prince S, Dixon-Fyle S, Yee L. Diversity wins: How inclusion matters. McKinsey & Company [Internet]. 2020 [cited 2024 Jul 23]. Available from: https://www.mckinsey.com/featured-insights/diversity-and-inclusion/diversity-wins-how-inclusion-matters#/

52. BlackRock. Lifting financial performance by investing in women: Long-term capitalism at BlackRock. 2023 Nov [cited 2025 Jan 23]. Available from: https://www.blackrock.com/corporate/literature/whitepaper/lifting-financial-performance-by-investing-in-women.pdf

53. Acemoglu D, Robinson JA. Why nations fail: the origins of power, prosperity, and poverty. New York (NY): Random House; 2012 [cited 2025 Jan 23]. Available at http://www.tinyurl.com/27cosoab

54. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM CSUR. 2021;54(6):1–35.

55. Charpignon M-L, Celi LA, Cobanaj M, Eber R, Fiske A, Gallifant J, et al. Diversity and inclusion: a hidden additional benefit of open data. PLOS Digit Health. 2024;3(7):e0000486. https://doi.org/10.1371/journal.pdig.0000486 PMID: 39042705

56. Gallifant J, Celi LA, Pierce RL. Digital determinants of health: opportunities and risks amidst health inequities. Nat Rev Nephrol. 2023;19(12):749–50. https://doi.org/10.1038/s41581-023-00763-4 PMID: 37626271

57. Center for Health Law and Policy Innovation. Comments submitted to Secretary Xavier Becerra regarding RIN 0945-AA17: Nondiscrimination in health and health education programs or activities. 2022 Oct 3 [cited 2024 Jul 28]. Available from: https://chlpi.org/wp-content/uploads/2022/10/Section-1557-Clinical-Algorithms-Comment.pdf

58. Hoffman S, Podgurski A. Artificial intelligence and discrimination in health care. Yale J Health Policy Law Ethics [Internet]. 2020;19(3):1. Case Legal Studies Research Paper No. 2020-29. [cited 2025 Jan 23]. Available from: https://ssrn.com/abstract=3747737

59. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care?. AMA J Ethics. 2019;21(2):E167–179. https://doi.org/10.1001/amajethics.2019.167 PMID: 30794127

60. Braveman P. Health disparities and health equity: concepts and measurement. Annu Rev Public Health. 2006;27167–94. https://doi.org/10.1146/annurev.publhealth.27.021405.102103 PMID: 16533114

61. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53. https://doi.org/10.1126/science.aax2342 PMID: 31649194

62. 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):874–82. https://doi.org/10.1056/NEJMms2004740 PMID: 32853499

63. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns (NY). 2021;2(10):100347. https://doi.org/10.1016/j.patter.2021.100347 PMID: 34693373

64. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol. 2018;154(11):1247–8. https://doi.org/10.1001/jamadermatol.2018.2348 PMID: 30073260

65. Sarkar R, Martin C, Mattie H, Gichoya JW, Stone DJ, Celi LA. Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. Lancet Digit Health. 2021;3(4):e241–9. https://doi.org/10.1016/S2589-7500(21)00022-4 PMID: 33766288

66. Neighbors HW, Jackson JS, Campbell L, Williams D. The influence of racial factors on psychiatric diagnosis: a review and suggestions for research. Community Ment Health J. 1989;25(4):301–11. https://doi.org/10.1007/BF00755677 PMID: 2697490

67. Meints SM, Cortes A, Morais CA, Edwards RR. Racial and ethnic differences in the experience and treatment of noncancer pain. Pain Manag. 2019;9(3):317–34. https://doi.org/10.2217/pmt-2018-0030 PMID: 31140916

68. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen L-C, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health. 2022;4(6):e406–14. https://doi.org/10.1016/S2589-7500(22)00063-2 PMID: 35568690

69. Matthew DB. Just medicine: a cure for racial inequality in American health care. New York (NY): NYU Press; 2015.

70. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2(3):158–64. https://doi.org/10.1038/s41551-018-0195-0 PMID: 31015713

71. Gallifant J, Nakayama LF, Gichoya JW, Pierce R, Celi LA. Equity should be fundamental to the emergence of innovation. PLOS Digit Health. 2023;2(4):e0000224. https://doi.org/10.1371/journal.pdig.0000224 PMID: 37036866

72. Gallifant J, Kistler EA, Nakayama LF, Zera C, Kripalani S, Ntatin A, et al. Disparity dashboards: an evaluation of the literature and framework for health equity improvement. Lancet Digit Health. 2023;5(11):e831–9. https://doi.org/10.1016/S2589-7500(23)00150-4 PMID: 37890905

73. Celi LA, Cellini J, Charpignon M-L, Dee EC, Dernoncourt F, Eber R, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities-A global review. PLOS Digit Health. 2022;1(3):e0000022. https://doi.org/10.1371/journal.pdig.0000022 PMID: 36812532

74. Matos J, Nakayama L, Charpignon M-L, Gallifant J, Kashkooli M, Carli F, et al. The Medical Knowledge Oligarchies. medRxiv 2023. https://doi.org/10.1101/2023.06.02.23290881

75. Collier ZJ, Kanmounye US, Naidu P, Tapia MF, Bustamante A, Bradley D, et al. 59 Burns in low- and middle-income countries: a scientometric analysis of peer-reviewed research. J Burn Care Res. 2022;43(1):S41–2. https://doi.org/10.1093/jbcr/irac012.062

76. Shambe I, Thomas K, Bradley J, Marchant T, Weiss HA, Webb EL. Bibliometric analysis of authorship patterns in publications from a research group at the London school of hygiene & tropical medicine, 2016-2020. BMJ Glob Health. 2023;8(2):e011053. https://doi.org/10.1136/bmjgh-2022-011053 PMID: 36792228

77. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6(1):120. https://doi.org/10.1038/s41746-023-00873-0 PMID: 37414860

78. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, III HD, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86–92. https://doi.org/10.1145/3458723

79. Boer A, De Beer L, Van Praat F. Algorithm assurance: auditing applications of artificial intelligence. In: Berghout E, Fijneman R, Hendriks L, De Boer M, Butijn BJ, editors. Advanced digital auditing. Cham (Switzerland): Springer International Publishing; 2023. p. 149–83.

80. Dossett LA, Kaji AH, Cochran A. SRQR and COREQ reporting guidelines for qualitative studies. JAMA Surg. 2021;156(9):875–6. https://doi.org/10.1001/jamasurg.2021.0525 PMID: 33825809

81. Garcia S, Ritchie CW, Luz S. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review. J Alzheimers Dis. 2020;78(4):1547–74. https://doi.org/10.3233/JAD-200888 PMID: 33185605

82. Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. J Am Med Inform Assoc. 2020;27(11):1784–97. https://doi.org/10.1093/jamia/ocaa174 PMID: 32929494

83. Balagopalan A, Eyre B, Robin J, Rudzicz F, Novikova J. Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech. Front Aging Neurosci. 2021;13:635945. https://doi.org/10.3389/fnagi.2021.635945 PMID: 33986655

84. Shi M, Cheung G, Shahamiri SR. Speech and language processing with deep learning for dementia diagnosis: a systematic review. Psychiatry Res. 2023;329115538. https://doi.org/10.1016/j.psychres.2023.115538 PMID: 37864994

85. Chen L, Dodge HH, Asgari M. Topic-based measures of conversation for detecting mild cognitive impairment. Proc Conf Assoc Comput Linguist Meet. 2020;2020:63–7. PMID: 33642674

86. García AM, de Leon J, Tee BL, Blasi DE, Gorno-Tempini ML. Speech and language markers of neurodegeneration: a call for global equity. Brain. 2023;146(12):4870–9. https://doi.org/10.1093/brain/awad253 PMID: 37497623

87. Benjamin R. Assessing risk, automating racism. Science. 2019;366(6464):421–2. https://doi.org/10.1126/science.aaz3873 PMID: 31649182

88. Friesen P, Douglas-Jones R, Marks M, Pierce R, Fletcher K, Mishra A, et al. Governing AI-driven health research: are irbs up to the task?. Ethics Hum Res. 2021;43(2):35–42. https://doi.org/10.1002/eahr.500085 PMID: 33683015

89. European Parliament. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation). Off J Eur Union [Internet]. 2016 Apr 27 [cited 2024 Jul 28]. Available from: http://data.europa.eu/eli/reg/2016/679/oj