BMJ Health & Care Informatics

Conceptualising fairness: three pillars for medical algorithms and health equity

Laura Sikstrom,[1,2] Marta M Maslej,[1] Katrina Hui,[1,3] Zoe Findlay,[4] Daniel Z Buchman,[1,5] Sean L Hill[1,3]

To cite: Sikstrom L, Maslej MM, Hui K, *et al*. Conceptualising fairness: three pillars for medical algorithms and health equity. *BMJ Health Care Inform* 2022;**29**:e100459. doi:10.1136/bmjhci-2021-100459

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100459).

Received 31 July 2021
Accepted 14 December 2021

Check for updates

[1]Centre for Addiction and Mental Health, Toronto, Ontario, Canada
[2]Department of Anthropology, University of Toronto, Toronto, Ontario, Canada
[3]Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada
[4]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada
[5]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

Correspondence to
Dr Laura Sikstrom;
laura.sikstrom@camh.ca

ABSTRACT

Objectives Fairness is a core concept meant to grapple with different forms of discrimination and bias that emerge with advances in Artificial Intelligence (eg, machine learning, ML). Yet, claims to fairness in ML discourses are often vague and contradictory. The response to these issues within the scientific community has been technocratic. Studies either measure (mathematically) competing definitions of fairness, and/or recommend a range of governance tools (eg, fairness checklists or guiding principles). To advance efforts to operationalise fairness in medicine, we synthesised a broad range of literature.

Methods We conducted an environmental scan of English language literature on fairness from 1960-July 31, 2021. Electronic databases Medline, PubMed and Google Scholar were searched, supplemented by additional hand searches. Data from 213 selected publications were analysed using rapid framework analysis. Search and analysis were completed in two rounds: to explore previously identified issues (a priori), as well as those emerging from the analysis (de novo).

Results Our synthesis identified 'Three Pillars for Fairness': transparency, impartiality and inclusion. We draw on these insights to propose a multidimensional conceptual framework to guide empirical research on the operationalisation of fairness in healthcare.

Discussion We apply the conceptual framework generated by our synthesis to risk assessment in psychiatry as a case study. We argue that any claim to fairness must reflect critical assessment and ongoing social and political deliberation around these three pillars with a range of stakeholders, including patients.

Conclusion We conclude by outlining areas for further research that would bolster ongoing commitments to fairness and health equity in healthcare.

INTRODUCTION

Automated-decision-making systems in medicine (often machine-learning or ML-based) represent an emergent medical and technological innovation we call 'Predictive Care'. Predictive care combines Big Data (on whole populations) and Small Data (on single people) to facilitate proactive, precise, and personalised health interventions. It is widely viewed as the ML tool with the most promise to solve some of the most complex and intractable problems in healthcare.[1] However, according to recent scholarship on algorithmic injustice, there is growing evidence to suggest that ML tools amplify existing inequities, such as racial bias, often because they are trained on biased datasets.[2–5] Therefore, implementation in clinical contexts is concerning because predictive care systems have the potential to discriminate against people based on sociodemographic characteristics such as age, sex or race.[6] These concerns have led to explosive growth in 'fairness-aware ML,' a new field that aims to design fair algorithmic systems[1] by detecting and eliminating bias.[7 8]

In ML discourses, the notion of *fairness* appeared briefly in the late 1960s as a shorthand for a range of procedural and statistical methods designed to track and measure different forms of discrimination.[9–11] Rediscovered recently[12] most current approaches to fairness are technocratic.[13] Studies either approach fairness as a set of (mathematical) techniques,[14] and/or recommend a set of governance procedures that can be used to mitigate against any unintended harms (eg, fairness checklists or guiding principles).[15] However, it remains unclear how exactly current approaches to fairness map onto established ethical frameworks.[7 16–19] For example, the narrow definition of fairness in ML discourses does not fully engage with fairness as an *idiom*, or a mode of expression used to resolve public debates and emotional tensions that emerge alongside questions about what it means to build a good and just society.[20–23] Nor do these techniques or procedures fully address debates about who should/will benefit the most from these advances and why.[19 24–33] Finally, it remains unclear how or which notions of fairness might be used to advance health equity.[34] However, without conceptual clarity, attempts to operationalise fairness will be spurious.

To advance efforts to operationalise fairness in medicine, we synthesised a broad range of literature on fairness in medical algorithms. The results of our synthesis identified three pillars of fairness: transparency, impartiality and inclusion. We draw on these insights to propose a multidimensional conceptual framework to guide empirical research on the operationalisation of fairness in healthcare. We conclude by applying these three pillars to a case use scenario, drawing on examples from psychiatry. Although predictive care systems are not yet widely employed in psychiatry,[35] models to predict suicide,[36] psychiatric readmission,[37] and inpatient violence are in high demand.[38 39] However, the performance of these models are often limited; for instance, most individuals identified with ML as being at high risk do not become violent,[40] introducing a strong potential for bias in false positive predictions for certain groups. Although the future implementation of predictive care models is motivated by the provision of safer and more efficient care, biased predictions can perpetuate health inequities. Thus, predictive care in psychiatry offers a timely example for illustrating the value of our three pillars in advancing the operationalisation of fairness in healthcare. Our overall aim is to invite discussion and spur innovative solutions.

METHODOLOGY: WHAT'S FAIR?

The planning phase of this research included a medical anthropologist (LS) and a computational neuroscientist (SLH). We noted that there are few scholarly works devoted exclusively to understanding what it means to be fair or unfair (for exceptions[41–43]). We hypothesised that this may be because fairness is what sociolinguists call a 'strategically deployable shifter'.[44] The meaning of any shifter depends on how the concept is used, by whom and in what context. Shifters are identifiable because they are often used by both critics and their intended targets. For example, developers of a predictive care model can claim it is "fair" because it pairs most patients with appropriate interventions. Detractors can claim it is 'unfair' because most patients paired with inappropriate interventions belong to protected groups, or a category of people protected by law, policy or similar authority.[45 46] Therefore, our research question for this review was: how do different disciplines define and operationalise fairness in relation to ML in healthcare?

Many health systems are poised to implement the use of Big Data and ML in medicine. Yet, few studies exist that describe the outcome or impact of predictive care tools on the diagnosis, treatment and lived experience of illness. Therefore, we chose an environmental scan over a systematic review so we could survey, document and interpret commonly cited dimensions of fairness related to the use of ML in healthcare in a timely manner.[47] It is particularly useful in contexts where data acquisition is necessary to identify emerging trends in a rapidly evolving research field.[48] Our aim was to foster the responsible interpretation and use of knowledge derived from advances in ML and to ensure that policy uptake is relevant and beneficial for all (see online supplemental appendix 1 for more details).

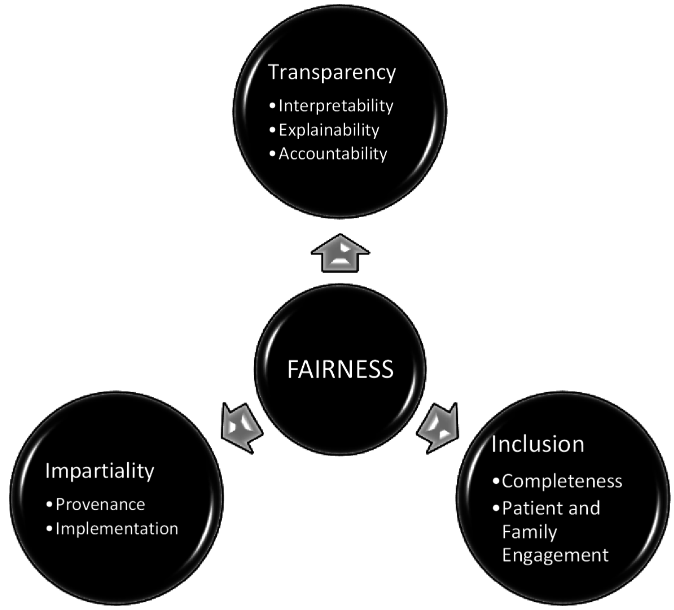

Figure 1 Three pillars of fairness.

RESULTS

Our synthesis of the literature identified three dimensions related to fairness: transparency, impartiality, and inclusion. Each of these dimensions had intertwined attributes (see figure 1). The majority of the literature examined one or two of these pillars in relation to ML in healthcare, while few reported on all three. Rather than report raw numbers, we have indicated the degree to which each dimension of fairness is considered by a discipline (table 1). While not assessing the quality of the studies we extracted, this approach highlights current gaps in the fairness and ML literature. For example, computational scientists were preoccupied with 'bias' and 'bias detection' (eg, provenance), social scientists with transparency and accountability, whereas clinicians were most concerned with implementation (table 1).

Three pillars for fairness and health equity

Although the literature we reviewed details a range of dimensions related to fairness, there is no single conceptual framework that integrates all of them. This article aims to address this gap through developing a conceptual framework for fairness we call 'Three Pillars for Fairness and Health Equity' (see table 2). Below we describe each of these pillars in turn and pay specific attention to the relationship between medical algorithms, predictive care and health equity.

Transparency

Transparency was cited as a key dimension of fairness with three intertwined attributes: interpretability, explainability and accountability.[49–52] Each encompasses

Table 1 Key dimension of fairness in the literature review by discipline (n=213)

Research field	Fairness dimension	Specific attribute	Volume of articles by specific attributes
Computational sciences (n=68)	Transparency	Interpretability/explainability	+ + +
		Accountability	+
		Provenance	+ + + +
	Impartiality	Implementation	+
		Completeness	+ + +
		Patient and family engagement	+
	Inclusion		
Medicine (n=43)	Transparency	Interpretability/explainability	+
		Accountability	+ + +
		Provenance	+ +
	Impartiality	Implementation	+ + + +
		Completeness	+ + +
		Patient and family engagement	+ +
	Inclusion		
Social sciences (n=73)	Transparency	Interpretability/explainability	+ + +
		Accountability	+ + + +
		Provenance	+ + +
	Impartiality	Implementation	+ + +
		Completeness	+ + +
		Patient and family engagement	+ +
	Inclusion		
Interdisciplinary research teams (n=29)	Transparency	Interpretability/explainability	+ +
		Accountability	+ +
		Provenance	+ + +
	Impartiality	Implementation	+ + + +
		Completeness	+ +
		Patient and family engagement	+
	Inclusion		

++++The majority of the literature reviewed in this field.
+++Several peer reviewed articles (five or more).
++A small number of peer reviewed articles (less than five).
+Little or no known literature (two or less).

methods designed to see, understand and hold complex algorithmic systems accountable. These attributes emerge from the fact that the inner workings of most algorithmic systems are invisible to all but the 'highest priests in their domain: mathematicians and computer scientists,' often making their verdicts, even when harmful, beyond dispute or appeal (O'Neil 2016:3).[33 53 54] Thus, transparency requires that the actions of scientists are easy to assess,[55–57] ensuring that stakeholders can decide whether they support the intentions, indications for use, and goals of any algorithmic system.[58 59] However, the opacity of algorithmic systems requires that we revisit our expectations for transparency in predictive care. For example, novel approaches in ML, such as enhancing feature representations with latent embeddings or applying neural networks, can improve our ability to predict important health outcomes,[60] but they also make models less transparent. Thus, there is a need to establish the degree to which we must be able to interpret and explain model results to clinicians, patients, and families. Crucially, the ability to see inside a system should not be conflated with the ability to govern it.[50 61]

Interpretability and explainability

In the literature we reviewed, interpretability and explainability are often used interchangeably.[49] However, interpretability most often refers to procedures and statistical techniques primarily used by scientists, to test, validate, and replicate findings.[62] In ML, this involves evaluation metrics (eg, accuracy, sensitivity, specificity), which can be used to compare performance across protected groups.[12 34] However, a predictive care model achieving similar performance across samples or settings is interpretable but not necessarily fair. If a predictive care model is biased against a sociodemographic group, this bias may carry over or be amplified in a different setting or sample.[63–66] Moreover, as described by the 'impossibility theorem,' not all fairness criteria can be satisfied at the same time.[6 16 67] For example, a predictive care model can achieve high accuracy (and therefore be interpretable and statistically fair) but can still be discriminatory.[68 69]

This limitation of interpretability may be addressed by explainability, which in part involves understanding how model features contribute to prediction. Various technical tools and procedures exist to address concerns

Table 2 Three pillars for fairness

Fairness pillar	Source of unfairness	Challenge:	Attribute	Key questions
Transparency: A range of methods designed to see, understand and hold complex algorithmic systems accountable in a timely fashion.	'Like Gods, these mathematical models were opaque, their workings invisible to all but the highest priests in their domain: mathematicians and computer scientists' (O'Neil: 3)	How can we foster democratic and sustained debate on the role of AI/ML in healthcare with a range of stakeholders, including patients experiencing complex and serious mental illness and/or addiction?	Interpretable	Are biases from predictive care models carried over across samples and settings?
			Explainable	Which model features are contributing to bias and what kinds of assumptions do they amplify? How does an understanding of these features by stakeholders impact clinical care?
			Accountable	How does predictive care impact stakeholders (patients, families, nurses, social workers)? What governance structures are in place to ensure fair development and deployment? Who is responsible for identifying and reporting potential harms?
Impartiality: Health care should be free from unfair bias and systemic discrimination.	'AI can help reduce bias, but it can also bake in and scale bias' (Silberg and Manyika:2)	How are complex social realities transformed into algorithmic systems, and what kinds of normative assumptions drive these processes?	Provenance	Do predictive care model features reflect socio-economic and political inequities? Might these features contribute to biased performance?
			Implementation	What harms might result from the implementation of predictive care models? Do they disproportionately affect certain groups?
Inclusion: The process of improving the ability, opportunity, and dignity of people, disadvantaged on the basis of their identity, to access health services, receive compassionate care and achieve equitable treatment outcomes.	'Randomised trials estimate average treatment effects for a trial population, but participants in clinical trials often aren't representative of the patient popuation that ultimately receives the treatment' (Chen: 167).	How can we ensure that the benefits of advances in clinical AI accrue to the most structurally disadvantaged?	Completeness	Is information required to detect bias missing? Is there sufficient data to evaluate predictive care models for intersectional bias? Are marginalised groups involved in the collection and use of their data?
			Patient and Family Engagement	Have stakeholders been involved in the development and implementation of predictive care? Do patients perceive models as being fair or positively impacting their care?

ML, machine learning.

about the so-called 'black box problem' of algorithmic systems, such as techniques to identify how models weigh features.[70 71] In the context of fairness however, explainability is only useful if highly-weighted features point to sociodemographic biases in model performance. For example, it may be possible to identify potential sources of bias in a predictive care model by examining whether the nature or availability of important features differ between sociodemographic groups. Moreover, in the literature we reviewed, explainability was also often used to draw attention to the social and communicative processes that surround predictive care tools. For example, these studies emphasised that fairness was not just about conveying accurate and unbiased information, but also about communicating the purpose, relevance and limitations of an algorithmic systems.[72–74]

Even if explainability is possible, it may not yield desirable outcomes.[75 76] According to emerging evidence, many clinicians are susceptible to following incorrect diagnostic advice.[77] This effect is more pronounced when ML-based advice is paired with explanations of features contributing to prediction,[78] suggesting that explainability can adversely impact clinical decision making.[79] Further, the ability to interpret and explain how a model works is not sufficient to mitigate harms. Recent studies of vaccine hesitancy and resistance to ebola campaigns emphasise that trust in public health interventions is often undermined by power differentials between patients and clinicians.[5 80–83] Although little is known about how patients engage with predictive care in clinical contexts,[84] sustained dialogue and shared decision-making between stakeholders that takes their concerns, desires and lived experiences seriously is critical.[79 85–88] Thus, there is an urgent need to develop engaging, effective and user-friendly explanations of predictive care models for clinicians, patients, their caregivers and the general public.[89 90] Explainability, therefore, must encompass both technical and social processes of translating the purpose, relevance and limitations of algorithmic systems to these various stakeholders, and providing targeted guidance on their use (eg, to complement clinical intuition, inform and negotiate care).

Accountability

Interpretability and explainability are described as prerequisites for the third transparency attribute: accountability. Accountability refers to governance structures, procedures, and tools used to evaluate and hold algorithmic systems accountable in a timely manner. Since predictive care often impacts acutely ill, marginalised or vulnerable groups, accountability cannot rest on the agency of a single person to assert their right to fair and equitable care.[91 92] In other words, we cannot expect those impacted by predictive care (patients, families, nurses, social workers) to be the ones to hold it accountable. From a fairness perspective, downloading the responsibility to those primarily impacted—and potentially harmed—by the technology is also ethically worrisome as it places a disproportionate burden on these groups to mobilise change. Rather, the governance structures that measure and track algorithmic systems must operate at multiple scales and be monitored continuously.[93–96] These structures should ensure that the development and implementation of predictive care is responsible and responsive to the needs and perspectives of various stakeholders.[88]

Impartiality

'We shape our tools, and thereafter, our tools shape us'.[97]

One of the most cited dimensions of fairness is that individuals should be free from unfair bias and systemic discrimination.[53] In medicine, both human and non-human actors gather, integrate and curate datasets to support care. As part of this process, (data scientists) aspire to collect unbiased data, but critics point out that data are not inherently fair, objective or impartial.[19] Rather, data reflect widespread biases and historical patterns of exclusion and inequality persisting in society at large,[98 99] which often extend to data on which predictive care models are trained. On the other hand, it is well documented that medical practices without algorithmic systems are far from impartial. Rather arbitrary and idiosyncratic practices in medicine frequently intersect with harmful sexist, racist and classist assumptions about patients.[100 101] From this perspective, algorithmic systems may be more fair because 'biased algorithms are easier to fix than biased people.'[102–104]

At first glance, it might seem like computational scientists and their critics have reached the same conclusion: that poor quality and biased data are likely to perpetuate harm. In the computational sciences, there is a growing assumption that encoding more data about a dataset's origins (metadata) and circumstances (context) surrounding its creation will resolve these issues.[1 105–111] However, as Seaver (2017:1105) and others argue, 'context is the kind of thing that cannot be modelled' since 'contexts are not containers, but… relational properties occasioned through activity.'[112] Rather than side with either perspective, we see this divergence as a vital opportunity for collaboration between computational and social scientists.[113 114] Thus, our conceptualisation of fairness includes two crucial attributes of impartiality that warrant further attention: a dataset's origins very broadly defined—or it's 'provenance' and its end-use—or 'implementation'.

Provenance

The view that encoding metadata will resolve issues of fairness maintains that with enough technical rigour, biases can be separated from the data, defined, contained and managed.[115] Unfortunately, containing or removing bias from training data may not be possible, because biased features are often linked with other features in ways that are not apparent.[105 116] Furthermore, this bias is maintained by social, technical and political systems which persist despite efforts to redress model bias with technical means.[19 30] Accordingly, evidence suggests that interdisciplinary or 'hybrid' teams support fairness-aware ML.[117] Domain experts, such as clinicians, social scientists or patient advocacy groups, have enhanced understandings of context situated bias,[114 116 118] support the curation of salient axes of difference,[119] and improve topic modelling and natural language processing models by aiding social bias detection.[120–122] For example, 'computational ethnography' is an approach to fairness-aware ML that emphasises the importance of a holistic understanding of any given dataset.[123 124] In sum, provenance requires more than a bias assessment that measures predictive accuracy across protected groups. In particular, far less attention has been paid to *how* complex social realities are transformed into algorithmic systems and the normative assumptions that drive these processes.[125–127] For example, rather than define 'fairness' as a fixed attribute, the literature we reviewed emphasised that it is a value-laden social and political determination made by individuals or groups of people within specific contexts. A broader sociotechnical approach to provenance will further support the identification of marginalised subgroups, facilitate meaningful analysis and support fairness-aware predictive care.

Implementation

Implementation refers to integrating a predictive care model into a clinical setting. The limited evidence available suggests that it is incredibly difficult to replicate the power of a predictive algorithm in real-world settings.[128–130] Significantly, potential uses of algorithmic systems in medicine are limitless. From a clinical perspective, these systems can personalise and optimise care.[131 132] From a health systems perspective, they can be useful tools to support the fair allocation of limited resources.[51 133] However, the integration of any algorithmic system into most clinical settings will require new workflows, which may challenge established hierarchies between doctors and nurses[129 134] and redefine what makes a 'good' clinician.[133 135–139] To fully understand the benefits or harms that could arise within algorithmic systems, it is equally important to consider at the outset of any project how it will be used, by whom, and to what end. Fair implementation foregrounds the clinical context where predictive care models are deployed.[140]

Inclusion

The final dimension of fairness we identified is inclusion. Among data scientists, inclusion often refers to both the representativeness of the dataset and its relative *completeness* (eg, how many features are filled in adequately). In other words, 'high-quality' data is accurate, precise, and collected from sufficiently large and representative samples.[141–143] This approach is concerned with ensuring that any benefits and harms derived from advances in predictive care accrue equally/equitably across sociodemographic groups. Others argue that this approach is an 'illusion'[144] and highlight the importance of building inclusive data infrastructures that prevent the misuse and commodification of marginalised peoples' data by supporting patient and family engagement.[145–148] Combined, these attributes have the potential to hold systems accountable, prevent unintended harms, and support the design and use of robust and fair algorithmic systems that advance health equity.

Completeness

Fairness-aware ML requires access to sociodemographic data. Unfortunately, data required to measure inequities is often absent and collected inconsistently.[118 149–153] Additional legal and social constraints limit access to sensitive sociodemographic data.[154] In Canada, for example, the collection of race/ethnicity data in healthcare settings has been restricted due to a range of historical and sociopolitical forces. For example, Thompson[155] illustrates how the Holocaust in the Second World War shook the foundations of the biological construction of race, which raised serious questions about the ethics of collecting this data.[155 156] Significantly, limited sample sizes among marginalised groups pose a significant problem for predictive care as outputs will be biased towards the majority group.[157–159] In addition, most current approaches to operationalising fairness focus only on legally protected categories, such as race or legal gender.[160] Yet, sexual orientation, gender identity and disability are prototypical instances of unobserved characteristics, because they are frequently unrecorded but also fundamentally unmeasureable.[161 162]

Finally, these challenges are further amplified by the fact that intersectionality—overlapping systems of disadvantage related to intersecting social categories like race or gender—is critical for understanding health outcomes in relation to marginalised identities.[163–167] Unfortunately, intersectional analyses are often limited by data availability; features contributing to intersectional bias may not be measured or the sizes of intersectional groups may be insufficient to generate meaningful performance metrics.[168] At the same time, opacity (the ability to remain unseen by an algorithm) may have political and social value for groups under surveillance (eg, undocumented or criminalised youth).[169] Therefore, while completeness entails inclusivity, inclusion should always be precipitated by dialogue and collaboration.

Patient and family engagement

As we chart the course for predictive care, we must centre the needs and lived experiences of those most likely to be impacted by ML.[31] At present, there is much speculation about how predictive care might enhance or disrupt clinical care work, or the range of therapeutic procedures, processes and outcomes oriented towards 'health and healing' in medicine[170 171] and 'recovery' in psychiatry.[129 134 172 173] However, the research to date has minimally addressed how patients engage with predictive care. According to some studies, patients are interested in contributing to the design of these technologies and having control over the use of their data.[174] Knowledge about patient engagement more broadly may be used to inform future work in this space. In particular, fair inclusion entails much more than diversifying our sampling frames. We must diversify our perspectives and ask those most impacted how predictive care (and their consequences) are experienced.

DISCUSSION

In online supplemental appendix 2, we apply our conceptual framework to consider an urgent issue of fairness in one area of predictive care: risk assessment in inpatient psychiatric settings.[38 39] Preventing and managing violence or aggression in mental healthcare is an ongoing challenge, with negative impacts on both patients and staff. Consequently, there are ongoing efforts to predict which inpatients may be at risk.[38] Over the past several decades, various features have emerged as predictors of this risk.[175] ML-based models trained on patient characteristics, structured assessments and clinical notes have achieved reasonable performance in predicting violence or aggression.[38 40 176] While these models achieve good overall accuracy in distinguishing between individuals who may or may not become violent or aggressive, they show poor performance in identifying the small subset of individuals who will actually exhibit this behaviour. According to one study for example, only 23% of people assigned as high risk became violent,[40] suggesting that many high-risk individuals are 'false positives'. Nevertheless, no studies

Table 3 The three fairness pillars, their attributes and relation to ML-based prediction of inpatient violence in psychiatric settings

Pillar	Attribute	Relation to predictive care
Transparency	Interpretability	ML models achieve high accuracy in predicting violent behaviour in psychiatric settings.[38] If these models achieve similar performance in new settings, they would be considered interpretable. However, if models are biased (ie, generating more false positives for inpatients defined by certain features), interpretability would be maintained even if biases carried forward to new samples.[63]
	Explainability	ML models are often trained on structured risk assessment scores.[38] Scores may be biased against certain groups (eg, recent immigrants due to language barriers or cultural miscommunications), leading to biased models. Pairing predictions with feature explanations can lead clinicians to over-rely on ML models,[78] which can exacerbate adverse impacts when models are biased.
	Accountability	ML models have been trained on actigraphy features to predict aggression in patients with dementia.[178] However, patients should not be expected to advocate for themselves if models seem biased or are not generalisable, given their particularly vulnerable status.
Impartiality	Provenance	Prior conviction and a diagnosis of schizophrenia are predictors of violence.[38 179] Training models on these features could lead to certain groups being disproportionately classified as high-risk (eg, black men, due to residing in more policed areas,[180] or being more likely misdiagnosed with schizophrenia[181]. Since these features are linked to other predictors, removing them does not remove model bias, nor does it address the social and political realities contributing to bias in the training data.[111 182]
	Implementation	ML modelling of violence risk is in part motivated by a desire to allocate staff resources to high-risk patients, but staff-patient interactions are known antecedents to violent behaviours.[183] Most patients classified as high-risk do not become violent;[40] however, pre-emptive interventions involving interactions with staff could precipitate violent behaviours.
Inclusion	Completeness	A focus on legally protected categories may disregard biases related to unobserved characteristics (eg, sexual orientation or disability). Individuals with invisible or undiagnosed disabilities (eg, autism spectrum disorder) may display behaviours interpreted as precursors to violence or aggression.[184–186] Additional marginalised groups might emerge when intersectional identities are taken into account.
	Patient and family engagement	Collaboration in decision making during admission and maximising choice are important values for patients in settings where autonomy is limited.[187–189] Patients may prioritise other aspects of care not captured by ML (eg, the caring relationships built with staff and peers, as compared with therapeutic interventions).[190]

ML, machine learning.

to date have explored whether groups defined by certain features are more likely to have this outcome, despite a strong potential for bias in this domain. In anticipation of further development and implementation of ML-based risk assessment, we demonstrate the value of employing our multidimensional framework as a heuristic tool to facilitate thoughtful and sustained dialogue on different dimensions of fairness in predictive care. In table 3, we summarise considerations related to ML-based prediction of inpatient risk for each fairness attribute. For a detailed discussion of these points, see online supplemental appendix 2.

CONCLUSION

Our literature synthesis demonstrates that scholars and computational scientists alike must broaden their notions of fairness to examine normative assumptions about what it means to build a just society and who decides what is fair. Further, the operationalisation of fairness requires going beyond developing rigorous data processing procedures or deploying sophisticated techniques to detect, mitigate and eliminate bias in ML. Predictions can be fair (eg, accurate) and still amplify inequities.[14 68] A multidimensional framework for fairness entails sustained dialogue with a range of stakeholders in the careful weighing of competing claims to fairness. It also involves proactively designing ML tools with and for marginalised and underserved communities.[5 34 177] Thus, fairness is not an outcome of rigorous and thoughtful research, but the social and political process required to advance health equity.

Critically, medical algorithms are neither 'fair' nor 'unfair;' fairness is not a binary classifier. We have used our conceptual framework of fairness as a heuristic tool to surface normative values embedded into our algorithmic systems to ensure that the opportunities presented by predictive care promote health equity. Current efforts to operationalise fairness have not strengthened our ability to safeguard against the possibility that predictive care tools might 'scale up' health inequities, nor have they provided the means to redress these imbalances once found. Designing fairness-aware predictive care systems requires sociotechnical approaches; interdisciplinary, collaborative and patient-centred research that foregrounds power dynamics and clinical contexts will promote health equity. Further, rather than 'de-bias' or validate algorithms after they have been constructed, we need to pay more attention to how data are collected, what kinds of data make up larger datasets, and how data are interpreted and instrumentalised within algorithmic systems.

Twitter Laura Sikstrom @LauraSikstrom and Daniel Z Buchman @DanielZBuchman

Contributors LS and SLH conceptualised, designed and analysed the data for this review. LS took the lead on the 'Three Pillars for Fairness' framework. MMM took the lead on the Case Scenario in Psychiatry. KH and ZF provided critical insights on the framework and the case scenario from a clinical perspective. DZB provided critical insights from bioethics on the conceptual framework and case scenario. All authors critically reviewed, edited and approved the final manuscript.

Funding This work was supported by the Dalla Lana School of Public Health Interdisciplinary Data Science Seed Grant (DZB and SLH), AMS Fellowship in Compassion and Artificial Intelligence (DZB), Canadian Institutes of Health Research Health Systems Impact Fellowship (LS and MMM) and the Social Sciences and Humanities Research Council Insight Development Grant (LS, MMM and KH, #430-2021-01166).

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as online supplemental information.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Daniel Z Buchman http://orcid.org/0000-0001-8944-6647

REFERENCES

1 Neff G. Why Big Data Won't Cure Us. *Big Data* 2013;1:117–23.
2 Singh JP, Grann M, Fazel S. A comparative study of violence risk assessment tools: a systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clin Psychol Rev* 2011;31:499–513.
3 Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
4 Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med* 2020;26:16–17.
5 Benjamin R. Assessing risk, automating racism. *Science* 2019;366:421–2.
6 Friedler SA, Scheidegger C, Venkatasubramanian S. On the (im) possibility of fairness. *arXiv preprint arXiv:*1609 07236, 2016. Available: https://arxiv.org/abs/1609.07236
7 Silberg J, Manyika J. Notes from the AI frontier: tackling bias in AI (and in humans) (June 2019), 2019. McKinsey global Institute. Available: https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/MGI-Tackling-bias-in-AI-June-2019.pdf
8 Green B, Hu L. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In: *Proceedings of the machine learning: the debates workshop*, 2018. https://econcs.seas.harvard.edu/files/econcs/files/green_icml18.pdf
9 Barocas S, Hardt M, Narayanan A. Fairness in machine learning. *Nips tutorial* 2017;1:2017.
10 Darlington RB. Another look at "cultural fairness"1. *J Educ Meas* 1971;8:71–82.
11 Cleary TA. Test bias: prediction of grades of Negro and white students in integrated colleges. *J Educ Meas* 1968;5:115–24.
12 Dwork C, Hardt M, Pitassi T. Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, New York, NY, USA, 2012:214–26.
13 TM L, Mosse D. Rendering society technical. In: *Adventures in Aidland: the anthropology of professionals in international development*. , 2011: 6, 57.
14 Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 2017;5:153–63.
15 Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. *Minds Mach* 2020;30:99–120.
16 Srivastava M, Heidari H, Krause A. Human perception of fairness: a descriptive approach to fairness for machine learning. in association for computing machinery. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, 2019:2459–68.
17 Zou J, Schiebinger L. Design AI so that it's fair. *Nature* 2018;559:324–6.
18 Ferryman K, Pitcan M. Fairness in precision medicine. *Data & Society*, 2018. Available: https://kennisopenbaarbestuur.nl/media/257243/datasociety_fairness_in_precision_medicine_feb2018.pdf
19 boyd danah, Crawford K. Critical questions for big data. *Information, Communication & Society* 2012;15:662–79.
20 Rawls J. Fairness to Goodness. *Philos Rev* 1975;84:536–54.
21 Powers M. Assistant Professor of Philosophy Madison Powers. In: *Social justice: the moral foundations of public health and health policy*. Oxford University Press, 2006.
22 Nuhrat Y. Fair to swear? gendered formulations of fairness in football in turkey. *Journal of Middle East Women's Studies* 2017;13:25–46.
23 Berliner D, Lambek M, Shweder R, *et al*. Anthropology and the study of contradictions. *HAU: Journal of Ethnographic Theory* 2016;6:1–27.
24 Meegan D. *America the fair: using brain science to create a more just nation*. Cornell University Press, 2019.
25 Wolsink M. Wind power implementation: The nature of public attitudes: Equity and fairness instead of 'backyard motives'. *Renewable and Sustainable Energy Reviews* 2007;11:1188–207.
26 Butler R. Children making sense of economic insecurity: Facework, fairness and belonging. *Journal of Sociology* 2017;53:94–109.
27 Deeming C. Social democracy and social policy in neoliberal times. *J Sociol* 2014;50:577–600.
28 Benjamin R. Race after technology: Abolitionist tools for the new jim code. *Soc Forces* 2020;98:1–3.
29 Perez CC. *Invisible women: exposing data bias in a world designed for men random house*, 2019.
30 Crawford K. *Atlas of AI*. Yale University Press, 2021.
31 Kalluri P. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 2020;583:169.
32 Katell M, Young M, Dailey D. Toward situated interventions for algorithmic equity: lessons from the field. In: *Association for computing machinery*. New York, NY, USA, 2020: 45–55.
33 Eubanks V. *Automating inequality: how high-tech tools profile, police, and Punish the poor*. St. Martin's Publishing Group, 2018.
34 Rajkomar A, Hardt M, Howell MD, *et al*. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–72.
35 Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. *NPJ Digit Med* 2019;2:77.

36 Belsher BE, Smolenski DJ, Pruitt LD, *et al*. Prediction models for suicide attempts and deaths: a systematic review and simulation. *JAMA Psychiatry* 2019;76:642–51.
37 Boag W, Kovaleva O, McCoy TH, *et al*. Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Transl Psychiatry* 2021;11:32.
38 Suchting R, Green CE, Glazier SM, *et al*. A data science approach to predicting patient aggressive events in a psychiatric hospital. *Psychiatry Res* 2018;268:217–22.
39 Viljoen JL, Cochrane DM, Jonnson MR. Do risk assessment tools help manage and reduce risk of violence and reoffending? A systematic review. *Law Hum Behav* 2018;42:181–214.
40 Connor M, Armbruster M, Hurley K, *et al*. Diagnostic sensitivity of the dynamic appraisal of situational aggression to predict violence and aggression by behavioral health patients in the emergency department. *J Emerg Nurs* 2020;46:302–9.
41 Rawls J. Justice as fairness. *Philos Rev* 1958;67:164–94.
42 Carr CL. *On fairness. Canadian centre for diversity and inclusion (January 2018*. Routledge, 2017.
43 Fairness sWJ. Respect and the egalitarian ethos revisited. *J Ethics* 2010;14:335–50.
44 Silverstein M. The indeterminacy of contextualization: when is enough enough. *The contextualization of language* 1992;22:55–76.
45 Barnett L, Walker J, Nicol J. *An examination of the duty to accommodate in the Canadian human rights context*. Library of Parliament, 2012.
46 Chun J, Gallagher-Louisy C. *Overview of human rights codes by Province and Territory in Canada*. Canadian Centre for Diversity and Inclusion, 2018. https://ccdi ca/media/1414/20171102-publications-overview-of-hr-codes-byprovince-final-en pdf
47 Rowel R, Moore ND, Nowrojee S, *et al*. The utility of the environmental scan for public health practice: lessons from an urban program to increase cancer screening. *J Natl Med Assoc* 2005;97:527–34.
48 Charlton P, Kean T, Liu RH, *et al*. Use of environmental scans in health services delivery research: a scoping review. *BMJ Open* 2021;11:e050284.
49 Krishnan M. Against Interpretability: a critical examination of the Interpretability problem in machine learning. *Philos Technol* 2020;33:487–502.
50 Ananny M, Crawford K. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc* 2018;20:973–89.
51 Zarsky T. The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. *Sci Technol Human Values* 2016;41:118–32.
52 Floridi L, Cowls J, King TC, *et al*. How to design AI for social good: seven essential factors. *Sci Eng Ethics* 2020;26:1771–96.
53 Barocas S, Selbst AD. Big data's disparate impact. Calif Law Rev, 2016. Available: https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/calr104&section=25
54 O'Neil C. *Weapons of math destruction: how big data increases inequality and threatens democracy*. Crown, 2016.
55 Sandvig C, Hamilton K, Karahalios K. Auditing algorithms: research methods for detecting discrimination on Internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 2014;22:4349–57.
56 Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data Soc* 2016;3:205395171562251.
57 Lepri B, Oliver N, Letouzé E, *et al*. Fair, transparent, and accountable algorithmic decision-making processes. *Philos Technol* 2018;31:611–27.
58 Char DS, Abràmoff MD, Feudtner C. Identifying ethical considerations for machine learning healthcare applications. *Am J Bioeth* 2020;20:7–17.
59 Brundage M, Avin S, Wang J. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv*, 2020. Available: https://arxiv.org/abs/2004.07213
60 Krompaß D, Esteban C, Tresp V, *et al*. Exploiting latent Embeddings of nominal clinical data for predicting Hospital readmission. *KI - Künstliche Intelligenz* 2015;29:153–9.
61 Geiger RS. Beyond opening up the black box: investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data Soc* 2017;4:205395171773073.
62 Bellamy RKE, Dey K, Hind M. Ai fairness 360: an EXTENSIBLE toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. Available: https://github.com/ibm/aif360
63 et alXu Z, Liu J, Cheng D. Assessing the Fairness of Classifiers with Collider Bias. arXiv [cs.LG], 2020. Available: http://arxiv.org/abs/2010.03933
64 Darlow L, Jastrzębski S, Storkey A. Latent Adversarial Debiasing: Mitigating Collider Bias in Deep Neural Networks. arXiv [cs.LG], 2020. Available: http://arxiv.org/abs/2011.11486
65 Prosperi M, Guo Y, Sperrin M, *et al*. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell* 2020;2:369–75.
66 et alChou Y-L, Moreira C, Bruza P. Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications. arXiv [cs.AI], 2021. Available: http://arxiv.org/abs/2103.04244
67 Saravanakumar KK. The Impossibility Theorem of Machine Fairness - A Causal Perspective. arXiv [cs.LG], 2020. Available: http://arxiv.org/abs/2007.06024
68 Kleinberg J. Inherent trade-offs in algorithmic fairness. *Abstracts of the 2018 ACM International Conference* 2018 https://dl.acm.org/doi/abs/10.1145/3219617.3219634
69 Wang C, Han B, Patel B. In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction. arXiv [stat.ML], 2020. Available: http://arxiv.org/abs/2005.04176
70 Slack D, Hilgard S, Jia E. How can we fool lime and SHAP?Adversarial Attacks on Post hoc Explanation Methods, 2019. Available: https://openreview.net/forum?id=nTHOa8__v0B [Accessed 19 Jul 2021].
71 Lundberg SM, Erion G, Chen H, *et al*. From local explanations to global understanding with Explainable AI for trees. *Nat Mach Intell* 2020;2:56–67.
72 Watson DS, Floridi L. The explanation game: a formal framework for interpretable machine learning. *Synthese* 2021;198:9211–42.
73 Samek W, Wiegand T, Müller K-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. arXiv [cs.AI], 2017. Available: http://arxiv.org/abs/1708.08296
74 Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. *AI Mag* 2019;40:44–58.
75 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;1:206–15.
76 Babic B, Gerke S, Evgeniou T, *et al*. Beware explanations from AI in health care. *Science* 2021;373:284–6.
77 Gaube S, Suresh H, Raue M, *et al*. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med* 2021;4:31.
78 Jacobs M, Pradier MF, McCoy TH, *et al*. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Transl Psychiatry* 2021;11:1–9.
79 McCradden MD. When is accuracy off-target? *Transl Psychiatry* 2021;11:369.
80 Kaler A. Health interventions and the persistence of rumour: the circulation of sterility stories in African public health campaigns. *Soc Sci Med* 2009;68:1711–9.
81 Nations MK, Monte CM. "I'm not dog, no!": cries of resistance against cholera control campaigns. *Soc Sci Med* 1996;43:1007–24.
82 Fairhead J. Understanding social resistance to the Ebola response in the forest region of the Republic of guinea: an anthropological perspective. *Afr Stud Rev* 2016;59:7–31.
83 Chandler C, Fairhead J, Kelly A, *et al*. Ebola: limitations of correcting misinformation. *Lancet* 2015;385:1275–7.
84 Saha D, Schumann C, Mcelfresh D. Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics. In: *Iii HD, Singh A, eds. Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020: 8377–87.
85 Kelliher A, Barry B. Designing Therapeutic Care Experiences with AI in Mind. In: *2018 AAAI Spring Symposium Series*. aaai.org, 2018. Available: https://www.aaai.org/ocs/index.php/SSS/SSS18/paper/viewPaper/17585
86 et alAnderson A, Dodge J, Sadarangani A. Explaining Reinforcement Learning to Mere Mortals: An Empirical Study. arXiv [cs.HC], 2019. Available: http://arxiv.org/abs/1903.09708
87 Dodge J, Liao QV, Zhang Y. Explaining models: an empirical study of how explanations impact fairness judgment. in association for computing machinery. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, New York, NY, USA, 2019:275–85.
88 Ho A. Are we ready for artificial intelligence health monitoring in elder care? *BMC Geriatr* 2020;20:358.
89 Hengstler M, Enkel E, Duelli S. Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technol Forecast Soc Change* 2016;105:105–20.
90 Peck EM, Ayuso SE, El-Etr O. Data is personal: attitudes and perceptions of data visualization in rural Pennsylvania. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2019:1–12.

91 Floridi L, Cowls J, Beltrametti M, *et al*. AI4People-An ethical framework for a good AI Society: opportunities, risks, principles, and recommendations. *Minds Mach* 2018;28:689–707.
92 Carr S. 'AI gone mental': engagement and ethics in data-driven technology for mental health. *J Ment Health* 2020;29:125–30.
93 Carroll SR, Rodriguez-Lonebear D, Martinez A. Indigenous data governance: strategies from United States native nations. *Data Sci J* 2019;18:31.
94 Reddy S, Allan S, Coghlan S, *et al*. A governance model for the application of AI in health care. *J Am Med Inform Assoc* 2020;27:491–7.
95 Baker DB, Kaye J, Terry SF. Governance through privacy, fairness, and respect for individuals. *EGEMS* 2016;4:1207.
96 Eaneff S, Obermeyer Z, Butte AJ. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *JAMA* 2020;324:1397–8.
97 Culkin JM. J S. A SCHOOLMAN'S GUIDE TO MARSHALL McLUHAN, 1967. Available: https://www.etalon-walk.com/s/JOHN_CULKIN.pdf [Accessed 27 Oct 2021].
98 Noble SU. *Algorithms of Oppression: how search Engines reinforce racism*. NYU Press, 2018.
99 Pierson E, Cutler DM, Leskovec J, *et al*. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med* 2021;27:136–40.
100 Higashi RT, Tillack A, Steinman MA, *et al*. The 'worthy' patient: rethinking the 'hidden curriculum' in medical education. *Anthropol Med* 2013;20:13–23.
101 Oldani MJ. Uncanny scripts: understanding pharmaceutical emplotment in the Aboriginal context. *Transcult Psychiatry* 2009;46:131–56.
102 Mullainathan S. Biased algorithms are easier to fix than biased people, 2019. Ny times. Available: https://www.cis.upenn.edu/~mkearns/teaching/ScienceDataEthics/nyt.pdf
103 Kleinberg J, Ludwig J, Mullainathan S. A guide to solving social problems with machine learning. *Harv Bus Rev*, 2016. Available: https://www.homeworksmontana.com/wp-content/uploads/edd/2019/07/kleinberg_ludwig_mullainathan_2016_hbr_a_guide_to_solving_social_problems_with_machine_learning.pdf
104 Kleinberg J, Ludwig J, Mullainathan S, *et al*. Algorithms as discrimination detectors. *Proc Natl Acad Sci U S A* 2020;117:30096–100.
105 Dourish P. What we talk about when we talk about context. *Pers Ubiquitous Comput* 2004;8:19–30.
106 de Paula R, Holanda M, Gomes LSA, *et al*. Provenance in bioinformatics workflows. *BMC Bioinformatics* 2013;14 Suppl 11:S6.
107 Mayernik MS. Open data: accountability and transparency. *Big Data Soc* 2017;4:205395171771885.
108 Wercelens P, da Silva W, Castro K. Data Provenance Management of Bioinformatics Workflows in Federated Clouds. In: *2019 IEEE International Conference on bioinformatics and biomedicine*, 2019: 750–4.
109 Goble C, Cohen-Boulakia S, Soiland-Reyes S, *et al*. Fair computational Workflows. *Data Intelligence* 2020;2:108–21.
110 Leavy S, Siapera E, O'Sullivan B. Ethical Data Curation forAI: An Approach based on Feminist Epistemology and CriticalTheories of Race. *aies-conference.com*. Available: https://www.aies-conference.com/2021/wp-content/posters/171_%20Ethical%20Data%20Curation%20for%20AI_%20An%20Approach%20based%20on%20Feminist%20Epistemology%20and%20Critical%20Theories%20of%20Race.pdf
111 ES J, Gebru T. Lessons from archives: strategies for collecting sociocultural data in machine learning. *arXiv [cs.LG]* 2019.
112 Seaver N. Algorithms as culture: some tactics for the ethnography of algorithmic systems. *Big Data & Society* 2017.
113 Zajko M. Conservative AI and social inequality: conceptualizing alternatives to bias through social theory. *AI Soc* 2007 https://arxiv.org/ftp/arxiv/papers/2007/2007.08666.pdf
114 Neff G, Tanweer A, Fiore-Gartland B, *et al*. Critique and contribute: a practice-based framework for improving critical data studies and data science. *Big Data* 2017;5:85–97.
115 Mayernik MS. Metadata accounts: achieving data and evidence in scientific research. *Soc Stud Sci* 2019;49:732–57.
116 Selbst AD, Boyd D, Friedler SA. Fairness and abstraction in Sociotechnical systems. in association for computing machinery. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 2019:59–68.
117 Cury M, Whitworth E, Barfort S. Hybrid methodology: combining ethnography, cognitive science, and machine learning to inform the development of Context-Aware personal computing and assistive technology. *Ethnographic Praxis in Industry Conference Proceedings*, 2019:254–81.
118 Mitchell S, Potash E, Barocas S, *et al*. Algorithmic fairness: choices, assumptions, and definitions. *Annu Rev Stat Appl* 2021;8:141–63.
119 Trewin S, Basson S, Muller M, *et al*. Considerations for AI fairness for people with disabilities. *AI Matters* 2019;5:40–63.
120 et alHutchinson B, Prabhakaran V, Denton E. Social Biases in NLP Models as Barriers for Persons with Disabilities. arXiv [cs.CL], 2020. Available: http://arxiv.org/abs/2005.00813
121 Goldenstein J, Poschmann P. Analyzing meaning in big data: performing a MAP analysis using grammatical parsing and topic modeling. *Sociol Methodol* 2019;49:83–131.
122 Arnold T, Fuller HJA. *In search of the user's language: Natural language processing, computational ethnography, and error-tolerant interface design. In: Advances in Usability, User Experience and Assistive Technology*. Cham: Springer International Publishing, 2019: 36–43.
123 Abramson CM, Joslyn J, Rendle KA, *et al*. The promises of computational ethnography: improving transparency, replicability, and validity for realist approaches to ethnographic analysis. *Ethnography* 2018;19:254–84.
124 Moore RJ, Smith R, Liu Q. Using computational ethnography to enhance the curation of real-world data (RWD) for chronic pain and invisible disability use cases. *SIGACCESS Access. Comput.* 2020:1–7.
125 Leahey E. Overseeing research practice: the case of data editing. *Sci Technol Human Values* 2008;33:605–30.
126 Engle SM. The social life of measurement Sally Engle Merry. published online first:, 2016. Available: http://www.americanbarfoundation.org/uploads/cms/documents/merry-lsi_forum_rev.pdf
127 Friedler SA, Scheidegger C, Venkatasubramanian S. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun ACM* 2021;64:136–43.
128 Blomberg SN, Christensen HC, Lippert F, *et al*. Effect of machine learning on Dispatcher recognition of out-of-hospital cardiac arrest during calls to emergency medical services: a randomized clinical trial. *JAMA Netw Open* 2021;4:e2032320.
129 Sendak MP, Ratliff W, Sarro D, *et al*. Real-World integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med Inform* 2020;8:e15182.
130 Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype? *JAMA* 2019;321:2281–2.
131 Tomlinson A, Furukawa TA, Efthimiou O, *et al*. Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. *Evid Based Ment Health* 2020;23:52–6.
132 Thieme A, Belgrave D, Doherty G. Machine learning in mental health: a systematic review of the HCI literature to support the development of effective and Implementable ml systems. *ACM Trans Comput-Hum Interact* 2020;27:1–53.
133 Bauer M, Monteith S, Geddes J, *et al*. Automation to optimise physician treatment of individual patients: examples in psychiatry. *Lancet Psychiatry* 2019;6:338–49.
134 Mateescu A, Elish MC. Ai in context: the labor of integrating new technologies, 2019. Available: https://www.voced.edu.au/content/ngv:81783
135 Kemp J, Zhang T, Inglis F, *et al*. Delivery of compassionate mental health care in a digital Technology-Driven age: Scoping review. *J Med Internet Res* 2020;22:e16263.
136 Chin-Yee B, Upshur R. Clinical judgement in the era of big data and predictive analytics. *J Eval Clin Pract* 2018;24:638–45.
137 Hunt LM, Bell HS, Baker AM, *et al*. Electronic health records and the disappearing patient. *Med Anthropol Q* 2017;31:403–21.
138 Topol EJ. High-Performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
139 Glasziou P, Moynihan R, Richards T, *et al*. Too much medicine; too little care. *BMJ* 2013;347:f4247.
140 McCradden MD, Stephenson EA, Anderson JA. Clinical research underlies ethical integration of healthcare artificial intelligence. *Nat Med* 2020;26:1325–6.
141 All of Us Research Program Investigators, Denny JC, Rutter JL, *et al*. The "All of Us" Research Program. *N Engl J Med* 2019;381:668–76.
142 Biruk C. Seeing like a research project: producing "high-quality data" in AIDS research in Malawi. *Med Anthropol* 2012;31:347–66.
143 Ghassemi M, Naumann T, Schulam P, *et al*. Practical guidance on artificial intelligence for health-care data. *Lancet Digit Health* 2019;1:e157–9.
144 Fox K. The Illusion of Inclusion: Large Scale Genomic Data Sovereignty and Indigenous Populations. In:Association for Computing Machinery. *Proceedings of the 26th ACM SIGKDD*

International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, 2020:3591.
145 Snipp CM. *What does data sovereignty imply: what does it look like Indigenous data sovereignty toward an agenda*, 2016: 39–55.
146 Kukutai T, Taylor J O. *Data sovereignty for Indigenous peoples: current practice and future needs*. ANU Press, 2016.
147 Zakaria C, Balan R, Lee Y. StressMon: scalable detection of perceived stress and depression using passive sensing of changes in work Routines and group interactions. *Proc ACM Hum-Comput Interact* 2019;3:1–29.
148 Vincent N, Li H, Tilly N. Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies. In:Association for Computing Machinery. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 2021:215–27.
149 Pinto AD, Glattstein-Young G, Mohamed A, *et al*. Building a Foundation to reduce health inequities: routine collection of sociodemographic data in primary care. *J Am Board Fam Med* 2016;29:348–55.
150 Mc Kenzie K, Antwi BAM, Tuck A. The case for diversity, 2016. Available: https://pdfs.semanticscholar.org/ebcd/d11cd033afc93b389e8a28c331155869d0a7.pdf [Accessed 23 Mar 2021].
151 Kiran T, Sandhu P, Aratangy T, *et al*. Patient perspectives on routinely being asked about their race and ethnicity: qualitative study in primary care. *Can Fam Physician* 2019;65:e363–9.
152 Andrus M, Spitzer E, Brown J. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In:Association for Computing Machinery. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 2021:249–60.
153 Cristina Mora G. *Making Hispanics*. University of Chicago Press, 2021.
154 Holmes JH, Beinlich J, Boland MR, *et al*. Why is the electronic health record so challenging for research and clinical care? *Methods Inf Med* 2021;60:032–48.
155 Thompson D. *The schematic state*. Cambridge University Press, 2016.
156 Saini A. *Superior: the return of race science*. Beacon Press, 2019.
157 Treviranus J. The Value of Being Different. In: *Proceedings of the 16th International Web for All Conference*. New York, NY, USA: : Association for Computing Machinery 2019:1–7.
158 Krawczyk B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 2016;5:221–32.
159 He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng* 2009;21:1263–84.
160 Willen SS. How is health-related "deservingness" reckoned? Perspectives from unauthorized im/migrants in Tel Aviv. *Soc Sci Med* 2012;74:812–21.
161 Tomasev N, McKee KR, Kay J. Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. arXiv [cs.CY], 2021. Available: http://arxiv.org/abs/2102.04257
162 Guo A, Kamar E, Vaughan JW. Toward fairness in AI for people with disabilities: A research roadmap. arXiv [cs.CY], 2019. Available: http://arxiv.org/abs/1907.02227
163 Ford CL, Airhihenbuwa CO, Theory CR. Critical race theory, race equity, and public health: toward antiracism praxis. *Am J Public Health* 2010;100 Suppl 1:S30–5.
164 Seng JS, Lopez WD, Sperlich M, *et al*. Marginalized identities, discrimination burden, and mental health: empirical exploration of an interpersonal-level approach to modeling intersectionality. *Soc Sci Med* 2012;75:2437–45.
165 Rosenfield S. Triple jeopardy? mental health at the intersection of gender, race, and class. *Soc Sci Med* 2012;74:1791–801.
166 Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Proceedings of the 1st conference on fairness, accountability and transparency*. New York, NY, USA: PMLR, 2018: 77–91.
167 Crenshaw K. Mapping the margins: intersectionality, identity politics, and violence against women of color. *Stanford Law Rev* 1991;43:1241.
168 Koppe G, Meyer-Lindenberg A, Durstewitz D. Deep learning for small and big data in psychiatry. *Neuropsychopharmacology* 2021;46:176–90.
169 Bratich J. Digital age| occult (ING) transparency: an Epilogue. *Int J Commun Syst*2016 https://ijoc.org/index.php/ijoc/article/download/4896/1526
170 Kleinman A. From illness as culture to caregiving as moral experience. *N Engl J Med* 2013;368:1376–7.
171 Kleinman A. *The Illness Narratives: Suffering, Healing, And The Human Condition. Basic Books*, 2020.
172 Csordas TJ, Kleinman A. *The therapeutic process medical anthropology contemporary theory and method*, 1996: 3–20.
173 Topol E. Others. The Topol Review. In: *Preparing the healthcare workforce to deliver the digital future*, 2019: 1–48.
174 Cuperfain AB, Hui K, Berkhout SG, *et al*. Patient, family and provider views of measurement-based care in an early-psychosis intervention programme. *BJPsych Open* 2021;7.
175 Dack C, Ross J, Papadopoulos C, *et al*. A review and meta-analysis of the patient factors associated with psychiatric in-patient aggression. *Acta Psychiatr Scand* 2013;127:255–68.
176 Menger V, Spruit M, van Est R, *et al*. Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. *JAMA Netw Open* 2019;2:e196709.
177 Abebe R, Barocas S, Kleinberg J. Roles for computing in social change. *arXiv [cs.CY]* 2019.
178 Khan SS, Ye B, Taati B, *et al*. Detecting agitation and aggression in people with dementia using sensors-A systematic review. *Alzheimers Dement* 2018;14:824–32.
179 Appelbaum PS, Robbins PC, Monahan J. Violence and delusions: data from the MacArthur violence risk assessment study. *Am J Psychiatry* 2000;157:566–72.
180 Meerai S, Abdillahi I, Poole J. An introduction to anti-black Sanism Intersectionalities: a global Journal of social work analysis, research, Polity, and practice 2016;5:18–35.
181 Olbert CM, Nagendra A, Buck B. Meta-Analysis of black vs. white racial disparity in schizophrenia diagnosis in the United States: do structured assessments attenuate racial disparities? *J Abnorm Psychol* 2018;127:104–15.
182 Kohler-Hausmann I. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw UL Rev* 2018;113:1163.
183 Papadopoulos C, Ross J, Stewart D, *et al*. The antecedents of violence and aggression within psychiatric in-patient settings. *Acta Psychiatr Scand* 2012;125:425–39.
184 Matthews M, Bell E. Assessment of risk of violent offending for adults with intellectual disability and/or autism spectrum disorder. In: *The Wiley Handbook of what works in violence risk management*, 2020: 349–66.
185 Gerson R, Malas N, Mroczkowski MM. Crisis in the emergency department: the evaluation and management of acute agitation in children and adolescents. *Child Adolesc Psychiatr Clin N Am* 2018;27:367–86.
186 Fernandes NA, Sawyer A, Zaheer J, *et al*. Adults with intellectual and developmental disabilities presenting to a psychiatric emergency department: a descriptive analysis and predictors of admission. *J Ment Health Res Intellect Disabil* 2020;13:384–95.
187 Valenti E, Giacco D, Katasakou C, *et al*. Which values are important for patients during involuntary treatment? A qualitative study with psychiatric inpatients. *J Med Ethics* 2014;40:832–6.
188 McGuinness D, Murphy K, Bainbridge E, *et al*. Individuals' experiences of involuntary admissions and preserving control: qualitative study. *BJPsych Open* 2018;4:501–9.
189 Hui K, Cooper RB, Zaheer J. Engaging patients and families in the ethics of involuntary psychiatric care. *Am J Bioeth* 2020;20:82–4.
190 Care HKR. closeness, and becoming 'better': Transformation and therapeutic process in American adolescent psychiatric custody. *Ethos* 2016;44:313–32.

Appendix 1: Developing a Conceptual Framework for Fairness

Methodology

Informed by our approach to the notion of fairness as a shifter, our data extraction template focused on four broad disciplinary points of view: the computational sciences, medicine, the social sciences and interdisciplinary perspectives (research teams containing at least one member from two of the above categories). We searched Medline, PubMed and Google Scholar using the terms provided in eTable 1, supplemented by additional hand searches for other relevant sources. We extracted the following thematic features from 213 English language papers from 1960 to July 31, 2021: how the notion of fairness was being used; who was advocating for fairness and why; dimension and/or attribute of fairness of focus; key source of unfairness; relevance to ML in healthcare; and future areas of research. Two research team members (LS, SLH) analyzed the data using rapid Framework Analysis[1] augmented by Smith's (2014) approach to textual analysis.[2] Search and analysis were completed in two rounds: to explore previously identified issues (*a priori*), as well as those emerging from the analysis (*de novo*). Additional team members (MM, DB, KH, ZF) refined our framework and developed the case scenario in psychiatry outlined in the final section of this paper. The search categories and specific terms used are presented in eTable 1 below.

Results

eTable 1: Categories and search terms used in literature searches

	Primary Search Term	Secondary Search Term(s)

Step 1: Broad Search	Fairness	“Health equity” AND/OR “Social Justice” AND/OR “Healthcare” AND/OR “public health”
		“Machine learning” AND/OR “Artificial Intelligence” AND/OR “Algorithms” AND/OR “Big Data”
Step 2: Targeted Search	Fairness AND Health AND Machine Learning OR Artificial Intelligence	“Transparency” AND/OR “Interpretability” AND/OR “Explainability” AND/OR “Accountability” AND/OR “Responsibility” AND/OR “Governance” AND/OR “Data Sovereignty” AND/OR “Privacy” AND/OR “Stewardship” AND/OR “Trust” OR “Trustworthy”
		“Impartiality” AND/OR “Bias” AND/OR “Provenance” AND/OR “Metadata” AND/OR “Context” AND/OR “Discrimination” AND/OR “Deployment”
		“Inclusion” OR “Diversity” AND/OR “Completeness” OR “Missingness” AND/OR “Data quality” AND/OR “Small Data” AND/OR “Data Sovereignty” AND/OR “Patient Engagement” OR “Patient and Family Engagement” OR “Equity”

Appendix 2: Case Scenario: Risk Assessment in Psychiatry

In this Appendix, we consider fairness as it relates to risk assessment in inpatient psychiatric settings.[3,4] We provide examples of considerations for each pillar in this expanding area of predictive care (summarized in Table 3).

Transparency

Often, algorithms used in predictive care are not transparent; an algorithm may classify an inpatient as being at high risk of violence during their stay, but the nature of the algorithm can make it difficult to understand how classification is derived.[5] Thus, revisiting our expectations for transparency involves fostering open, democratic and sustained debate on the development and implementation of predictive care models in psychiatric settings with various stakeholders, including patients experiencing complex and serious mental illness and/or substance use disorders.

Among those developing predictive care models, interpretability involves calculating metrics to determine whether model performance is consistent across samples and settings. These metrics are also used to gauge fairness, which can demonstrate how models that are interpretable can nevertheless be unfair. In the context of risk assessment, suppose a model trained on patient characteristics is correct 78% of the time in predicting violent behaviour at a community psychiatric facility.[3] If applied to a new set of patients (e.g., at a different facility or county), the model would be interpretable if it achieved similar performance (~78%). However, if the model generated more false positives for inpatients defined by certain features (e.g., sex, race, or immigration status) at the community psychiatric facility, its interpretability would be maintained if the bias were carried forward to a new set of inpatients at this or another facility.[6]

Understanding how these model features contribute to biased predictions can address this limitation of interpretability. For example, predictions of inpatient violence improve with the use of structured risk assessments, which direct clinicians to rate behavioural antecedents of violence, such as irritability and unwillingness to follow directions, and guide action to prevent aggression.[7] Given that scores from risk assessments are highly predictive for ML modelling,[3] finding that recent immigrants tend to be given higher scores could suggest that language barriers or cultural miscommunications contribute to perceptions of risk, potentially leading to more false positive predictions for this group.

However, explainability is only useful if features point to sociodemographic biases in model performance. Otherwise, pairing predictions with feature explanations can lead clinicians to over-rely on ML models,[8] particularly when explanations of predictive features appear reasonable (e.g., high irritability and non-compliance noted in a structured risk assessment). Unfortunately, if the prediction is biased (e.g., generating more false positive predictions in recent immigrants), this overreliance on ML systems could result in adverse impacts being disproportionately allocated to already structurally disadvantaged groups.

This potential for amplifying harms in predictive care makes it necessary to reconsider accountability, since those who are primarily impacted should not be expected to identify biases and mobilize change. Many inpatients in psychiatric facilities are acutely ill or marginalized. They can suffer from chronic and complex health needs related to historical and intergenerational traumas, homelessness, and encounters with the law. These patients might be unaware of potential biases in predictive care and would not have the means to redress any harm caused by this bias. As another example, using wearable sensors to monitor patients with dementia has been explored as a way to reduce burden on staff and caregivers (and promote increased independence for patients). Various documented associations between actigraphy and aggressive behaviour introduce an opportunity to train ML models to predict aggression in patients with dementia.[9] However, few efforts have been made to evaluate models across samples and settings leading to issues of generalizability. If a subset of patients is likely to be incorrectly classified, who will know? Even if models are interpretable and explainable, who should advocate for patients, given their particularly vulnerable status?

Impartiality

Although efforts are being made to manage biases in data (e.g., provenance), they are often the result of social or political systems which cannot be contained or addressed with ML modelling. In the psychiatric literature for instance, predictors of violent behaviour in inpatient settings have consistently included prior conviction for assault and a diagnosis of schizophrenia, making these factors important features for ML modelling.[3,10] At the same time, these factors could result in certain groups being disproportionately classified as at high risk. Black men may reside in more policed areas than white men, making them more likely to have prior convictions.[11] Furthermore, Black men with affective disorders may also be more likely misdiagnosed with schizophrenia than white men.[12] Since schizophrenia positively predicts inpatient violence, whereas affective disorder is a negative predictor,[205] deploying a model trained on these psychiatric comorbidities could increase the rate of false positive risk classifications in Black men.

Identifying and removing biased features from training datasets, or encoding more data about them, does not necessarily result in fair prediction of inpatient violence, in part because these features are inextricably linked to other predictors. For example, given historically racist attitudes toward Black people, Black men may exhibit behaviours (e.g., paranoia, agitation, or frustration) that could be perceived as symptoms of psychosis in clinical interviews;[13,14] these might also be regarded as antecedents to violent or aggressive behaviour in encounters with police or during structured risk assessments in inpatient settings. However, these behaviours might stem from experiences of racism and a resulting mistrust in legal or healthcare settings.[14] This mistrust can lead to delays in seeking treatment and worsening illness, in turn contributing to misdiagnosis and perceptions of risk. Delays in treatment might also lead to self-medication and substance use, which is another predictor of inpatient violence.[15,16] Even if ML models were redressed to prevent Black men from being misclassified as at high risk (e.g., with adversarial learning), this would not address the underlying social and political realities contributing to bias in the training data.[17,18]

Deploying biased models can perpetuate harmful outcomes for already disadvantaged groups when model predictions are used to make clinical decisions. The development of ML models to predict violence in psychiatric settings is primarily motivated by an aim to improve patient and staff safety.[3] However, there is also an underlying desire to more effectively allocate staff and hospital resources, given limited operating costs.[3,19,20] The logic behind this secondary motivation is that surveillance or pre-emptive intervention (e.g., restricting freedoms or privileges) can be focused on subsets of patients at highest risk of violence. However, clinician-patient interactions, particularly those that limit patient freedoms or deny requests, are known antecedents to violent behaviours, precipitating almost 40% of incidents in inpatient settings.[21] Most patients identified with ML to be at risk do not become violent.[22] However, if pre-emptive interventions were implemented,[3] these patients could become aggressive or violent due to increased interactions with clinicians. Clinicians and staff might also respond to ML-based risk classification by treating these patients differently, altering their trajectory of care. Critically, once a predictive care model is deployed, what is defined as "fair" (accurate, efficient, cost-effective) from a health systems perspective may not be perceived as fair (unbiased) from the patient's point of view. Thus, considerations related to implementation must involve balancing among multiple and sometimes competing interests (e.g., of clinicians, hospital administrators, patients).

Inclusion

Finally, when operationalizing fairness by focusing on legally protected categories, we might overlook biases in relation to unobserved characteristics, such as sexual orientation or disability. For example, Queer youth in crisis might avoid sharing their sexual orientation or gender identity due to anticipated discrimination.[23,24] In the context of risk assessment, individuals with invisible or undiagnosed disabilities (such as Autism Spectrum Disorder) may also display features that could be interpreted as precursors to violence or aggression.[25–27] Additional marginalized groups might emerge when intersectional identities are taken into account (e.g., defined by sex and race). To our knowledge, there have been no efforts to determine whether ML-based predictions of inpatient violence are biased against protected categories or unobserved characteristics, let alone groups defined by intersectional features (e.g., pertaining to Black men).

In addition to ensuring that predictive care has equitable impacts across sociodemographic groups, inclusion also involves engaging those impacted groups in all aspects of the ML process – from data compilation to model development and implementation. Qualitative data from acute psychiatric inpatient settings emphasizes that collaboration in decision-making during admission and maximizing choice are important values for patients, particularly in settings where autonomy is somewhat limited.[28–30] While current efforts to operationalize fairness emphasize prediction performance across protected groups, patients may have different notions of fairness or prioritize other aspects of care altogether beyond what is captured by ML. For example, Hejmanek (2016) found that Black youth in psychiatric custody attributed their recovery to the caring relationships that they built with staff and peers, not to State-mandated therapeutic interventions, like group therapy.[31] Moreover, typical institutional responses to aggression adversely impact the experience of care,[32,33] reduce patients' willingness to disclose sensitive information, and undermine treatment adherence.[34,35] Thus, in service of achieving inclusion in ML-based risk assessments, we must seek to understand and include the perspectives, preferences, and experiences of patients in acute psychiatric settings.

References Cited

1 Ritchie J, Spencer L, O'Connor W, *et al.* Analysis: practices, principles and processes. *Qualitative Research Practices London: Sage* 2003.

2 Smith DE, Turner SM. *Incorporating Texts into Institutional Ethnographies*. University of Toronto Press 2014. https://play.google.com/store/books/details?id=Xj7AAwAAQBAJ

3 Suchting R, Green CE, Glazier SM, *et al.* A data science approach to predicting patient aggressive events in a psychiatric hospital. *Psychiatry Res* 2018;**268**:217–22. doi:10.1016/j.psychres.2018.07.004

4 Viljoen JL, Cochrane DM, Jonnson MR. Do risk assessment tools help manage and reduce risk of violence and reoffending? A systematic review. *Law Hum Behav* 2018;**42**:181–214. doi:10.1037/lhb0000280

5 Goodman SN, Goel S, Cullen MR. Machine Learning, Health Disparities, and Causal Reasoning. Ann. Intern. Med. 2018;**169**:883–4. doi:10.7326/M18-3297

6 Xu Z, Liu J, Cheng D, *et al.* Assessing the Fairness of Classifiers with Collider Bias. arXiv [cs.LG]. 2020.http://arxiv.org/abs/2010.03933

7 Teo AR, Holley SR, Leary M, *et al.* The relationship between level of training and accuracy of violence risk assessment. *Psychiatr Serv* 2012;**63**:1089–94. doi:10.1176/appi.ps.201200019

8 Jacobs M, Pradier MF, McCoy TH, *et al.* How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Transl Psychiatry*

2021;**11**:1–9. doi:10.1038/s41398-021-01224-x

9 Khan SS, Ye B, Taati B, *et al.* Detecting agitation and aggression in people with dementia using sensors—A systematic review. *Alzheimers Dement* 2018;**14**:824–32.https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2018.02.004

10 Appelbaum PS, Robbins PC, Monahan J. Violence and delusions: data from the MacArthur Violence Risk Assessment Study. *Am J Psychiatry* 2000;**157**:566–72. doi:10.1176/appi.ajp.157.4.566

11 Meerai S, Abdillahi I, Poole J. An Introduction to Anti-Black Sanism. *Intersectionalities: A Global Journal of Social Work Analysis, Research, Polity, and Practice* 2016;**5**:18–35.http://journals.library.mun.ca/ojs/index.php/IJ/article/view/1682 (accessed 23 Mar 2021).

12 Olbert CM, Nagendra A, Buck B. Meta-analysis of Black vs. White racial disparity in schizophrenia diagnosis in the United States: Do structured assessments attenuate racial disparities? *J Abnorm Psychol* 2018;**127**:104–15. doi:10.1037/abn0000309

13 Gara MA, Vega WA, Arndt S, *et al.* Influence of patient race and ethnicity on clinical assessment in patients with affective disorders. *Arch Gen Psychiatry* 2012;**69**:593–600. doi:10.1001/archgenpsychiatry.2011.2040

14 Metzl JM. *The Protest Psychosis: How Schizophrenia Became a Black Disease*. Beacon Press 2010. https://play.google.com/store/books/details?id=L9utU61499UC

15 Amore M, Menchetti M, Tonti C, *et al.* Predictors of violent behavior among acute psychiatric patients: clinical study. *Psychiatry Clin Neurosci* 2008;**62**:247–55. doi:10.1111/j.1440-1819.2008.01790.x

16 Dack C, Ross J, Papadopoulos C, *et al.* A review and meta-analysis of the patient factors associated with psychiatric in-patient aggression. *Acta Psychiatr Scand* 2013;**127**:255–68. doi:10.1111/acps.12053

17 Kohler-Hausmann I. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw UL Rev* 2018;**113**:1163.https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/illlr113&section=38

18 Jo ES, Gebru T. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. arXiv [cs.LG]. 2019. doi:10.1145/3351095.3372829

19 Silver E, Miller LL. A Cautionary Note on the Use of Actuarial Risk Assessment Tools for Social Control. *Crime & Delinquency* 2002;**48**:138–61. doi:10.1177/0011128702048001006

20 Szmukler G, Rose N. Risk assessment in mental health care: values and costs. *Behav Sci Law* 2013;**31**:125–40. doi:10.1002/bsl.2046

21 Papadopoulos C, Ross J, Stewart D, *et al.* The antecedents of violence and aggression within psychiatric in-patient settings. *Acta Psychiatr Scand* 2012;**125**:425–39. doi:10.1111/j.1600-0447.2012.01827.x

22 Connor M, Armbruster M, Hurley K, *et al.* Diagnostic Sensitivity of the Dynamic Appraisal of Situational Aggression to Predict Violence and Aggression by Behavioral Health Patients in the Emergency Department. *J Emerg Nurs* 2020;**46**:302–9. doi:10.1016/j.jen.2019.12.006

23 Abramovich IA. No Safe Place to Go - LGBTQ Youth Homelessness in Canada: Reviewing the Literature. *CJFY* 2012;**4**:29–51. doi:10.29173/cjfy16579

24 Kcomt L, Gorey KM, Barrett BJ, *et al.* Healthcare avoidance due to anticipated discrimination among transgender people: A call to create trans-affirmative environments. *SSM Popul Health* 2020;**11**:100608. doi:10.1016/j.ssmph.2020.100608

25 Matthews M, Bell E. Assessment of risk of violent offending for adults with intellectual disability and/or autism spectrum disorder. The Wiley Handbook of What Works in Violence Risk Management. 2020;:349–66. doi:10.1002/9781119315933.ch17

26 Gerson R, Malas N, Mroczkowski MM. Crisis in the Emergency Department: The Evaluation and Management of Acute Agitation in Children and Adolescents. *Child Adolesc Psychiatr Clin N Am* 2018;**27**:367–86. doi:10.1016/j.chc.2018.02.002

27 Fernandes NA, Sawyer A, Zaheer J, *et al.* Adults with intellectual and developmental disabilities presenting to a psychiatric emergency department: A descriptive analysis and predictors of admission. *J Ment Health Res Intellect Disabil* 2020;**13**:384–95. doi:10.1080/19315864.2020.1822963

28 Valenti E, Giacco D, Katasakou C, *et al.* Which values are important for patients during involuntary treatment? A qualitative study with psychiatric inpatients. *J Med Ethics* 2014;**40**:832–6. doi:10.1136/medethics-2011-100370

29 McGuinness D, Murphy K, Bainbridge E, *et al.* Individuals' experiences of involuntary admissions and preserving control: qualitative study. *BJPsych Open* 2018;**4**:501–9. doi:10.1192/bjo.2018.59

30 Hui K, Cooper RB, Zaheer J. Engaging Patients and Families in the Ethics of Involuntary Psychiatric Care. Am. J. Bioeth. 2020;**20**:82–4. doi:10.1080/15265161.2020.1754511

31 Hejtmanek KR. Care, closeness, and becoming 'better': Transformation and therapeutic process in American adolescent psychiatric custody. *Ethos* 2016;**44**:313–32. doi:10.1111/etho.12135

32 Knowles SF, Hearne J, Smith I. Physical restraint and the therapeutic relationship. *J Forens Psychiatry Psychol* 2015;**26**:461–75. doi:10.1080/14789949.2015.1034752

33 Montreuil M, Thibeault C, McHarg L, *et al.* Moral Experiences of Crisis Management in a Child Mental Health Setting: A Participatory Hermeneutic Ethnographic Study. *Cult Med Psychiatry* 2020;**44**:80–109. doi:10.1007/s11013-019-09639-4

34 Beauford JE, McNiel DE, Binder RL. Utility of the initial therapeutic alliance in evaluating psychiatric patients' risk of violence. *Am J Psychiatry* 1997;**154**:1272–6. doi:10.1176/ajp.154.9.1272

35 Brown P, Calnan M. *Trusting on the Edge: Managing Uncertainty and Vulnerability in the Midst of Serious Mental Health Problems*. Policy Press 2012. https://play.google.com/store/books/details?id=iHZoDwAAQBAJ