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 5,1,3 Sean L Hill1,3

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. 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. 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. These concerns have led to explosive growth in ‘fairness-aware ML,’ a new field that aims to design fair algorithmic systems by detecting and eliminating bias.

In ML discourses, the notions 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. Rediscovered recently, most current approaches to fairness are technocratic. Studies either approach fairness as a set of (mathematical) techniques, and/or recommend a set of governance procedures that can be used to mitigate against any unintended harms (eg, fairness checklists or guiding principles). However, it remains unclear how exactly current approaches to fairness map onto established ethical frameworks. 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. Nor do these techniques or procedures fully address debates about who should/will benefit the most from these advances and why. Finally, it remains unclear how or which notions of fairness might be used to advance health equity. 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, models to predict suicide, psychiatric readmission, and inpatient violence are in high demand. 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, introducing a strong potential for bias in false positive predictions for certain groups. Although the future implementation of predictive care model 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). We hypothesised that this may be because fairness is what sociolinguists call a ‘strategically deployable shifter’. 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. 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. It is particularly useful in contexts where data acquisition is necessary to identify emerging trends in a rapidly evolving research field. Our aim was to foster the responsible interpretation and use of knowledge derived from advances in ML.

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. Each encompasses
Open access

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). Thus, transparency requires that the actions of scientists are easy to assess, ensuring that stakeholders can decide whether they support the intentions, indications for use, and goals of any algorithmic system. 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, 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.

**Interpretability and explainability**

In the literature we reviewed, interpretability and explainability are often used interchangeably. However, interpretability most often refers to procedures and statistical techniques primarily used by scientists, to test, validate, and replicate findings. In ML, this involves evaluation metrics (eg, accuracy, sensitivity, specificity), which can be used to compare performance across protected groups. 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. Moreover, as described by the ‘impossibility theorem,’ not all fairness criteria can be satisfied at the same time. For example, a predictive care model can achieve high accuracy (and therefore be interpretable and statistically fair) but can still be discriminatory.

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

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

++++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).
Table 2  Three pillars for fairness

<table>
<thead>
<tr>
<th>Fairness pillar</th>
<th>Source of unfairness</th>
<th>Challenge:</th>
<th>Attribute</th>
<th>Key questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transparency: A range of methods designed to see, understand and hold complex algorithmic systems accountable in a timely fashion.</td>
<td>‘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)</td>
<td>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?</td>
<td>Interpretable</td>
<td>Are biases from predictive care models carried over across samples and settings?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Explainable</td>
<td>Which model features are contributing to bias and what kinds of assumptions do they amplify?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Accountable</td>
<td>How does an understanding of these features by stakeholders impact clinical care?</td>
</tr>
<tr>
<td>Impartiality: Health care should be free from unfair bias and systemic discrimination.</td>
<td>‘AI can help reduce bias, but it can also bake in and scale bias’ (Silberg and Manyika:2)</td>
<td>How are complex social realities transformed into algorithmic systems, and what kinds of normative assumptions drive these processes?</td>
<td>Provenance</td>
<td>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?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Implementation</td>
<td>What harms might result from the implementation of predictive care models? Do they disproportionately affect certain groups?</td>
</tr>
<tr>
<td>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.</td>
<td>‘Randomised trials estimate average treatment effects for a trial population, but participants in clinical trials often aren’t representative of the patient population that ultimately receives the treatment’ (Chen: 167).</td>
<td>How can we ensure that the benefits of advances in clinical AI accrue to the most structurally disadvantaged?</td>
<td>Completeness</td>
<td>Is information required to detect bias missing? Is there sufficient data to evaluate predictive care models for intersectional bias?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Are marginalised groups involved in the collection and use of their data?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Patient and Family Engagement</td>
<td>Have stakeholders been involved in the development and implementation of predictive care? Do patients perceive models as being fair or positively impacting their care?</td>
</tr>
</tbody>
</table>

ML, machine learning.

about the so-called ‘black box problem’ of algorithmic systems, such as techniques to identify how models weigh features. 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.

Even if explainability is possible, it may not yield desirable outcomes. According to emerging evidence, many clinicians are susceptible to following incorrect diagnostic advice. This effect is more pronounced when ML-based advice is paired with explanations of features contributing to prediction, suggesting that explainability can adversely impact clinical decision making. 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.  

Although little is known about how patients engage with predictive care in clinical contexts, sustained dialogue and shared decision-making between stakeholders that takes their concerns, desires and lived experiences seriously is critical. 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. 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. 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. These structures should ensure that the development and implementation of predictive care is responsible and responsive to the needs and perspectives of various stakeholders.

**Impartiality**

‘We shape our tools, and thereafter, our tools shape us.’

One of the most cited dimensions of fairness is that individuals should be free from unfair bias and systemic discrimination. 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. Rather, data reflect widespread biases and historical patterns of exclusion and inequality persisting in society at large, 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. From this perspective, algorithmic systems may be more fair because ‘biased algorithms are easier to fix than biased people.’

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. 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.’ Rather than side with either perspective, we see this divergence as a vital opportunity for collaboration between computational and social scientists. 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. 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. Furthermore, this bias is maintained by social, technical and political systems which persist despite efforts to redress model bias with technical means. Accordingly, evidence suggests that interdisciplinary or ‘hybrid’ teams support fairness-aware ML. Domain experts, such as clinicians, social scientists or patient advocacy groups, have enhanced understandings of context situated bias, support the curation of salient axes of difference, and improve topic modelling and natural language processing models by aiding social bias detection.

For example, ‘computational ethnography’ is an approach to fairness-aware ML that emphasises the importance of a holistic understanding of any given dataset. 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. 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. Significantly, potential uses of algorithmic
systems in medicine are limitless. From a clinical perspective, these systems can personalise and optimise care. From a health systems perspective, they can be useful tools to support the fair allocation of limited resources. However, the integration of any algorithmic system into most clinical settings will require new workflows, which may challenge established hierarchies between doctors and nurses and redefine what makes a ‘good’ clinician. 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.

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 (e.g., how many features are filled in adequately). In other words, ‘high-quality’ data is accurate, precise, and collected from sufficiently large and representative samples. 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’ 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. 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. Additional legal and social constraints limit access to sensitive sociodemographic data. 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 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. Significantly, limited sample sizes among marginalised groups pose a significant problem for predictive care as outputs will be biased towards the majority group. In addition, most current approaches to operationalising fairness focus only on legally protected categories, such as race or legal gender. Yet, sexual orientation, gender identity and disability are prototypical instances of unobserved characteristics, because they are frequently unrecorded but also fundamentally unmeasurable. 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. 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. At the same time, opacity (the ability to remain unseen by an algorithm) may have political and social value for groups under surveillance (e.g., undocumented or criminalised youth). 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. 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 and ‘recovery’ in psychiatry. 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. 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. 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. Over the past several decades, various features have emerged as predictors of this risk. ML-based models trained on patient characteristics, structured assessments and clinical notes have achieved reasonable performance in predicting violence or aggression. 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, suggesting that many high-risk individuals are ‘false positives’. Nevertheless, no studies
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. A multi-dimensional 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. 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

---

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

<table>
<thead>
<tr>
<th>Pillar</th>
<th>Attribute</th>
<th>Relation to predictive care</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transparency</td>
<td>Interpretability</td>
<td>ML models achieve high accuracy in predicting violent behaviour in psychiatric settings. 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.</td>
</tr>
<tr>
<td>Explainability</td>
<td></td>
<td>ML models are often trained on structured risk assessment scores. 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, which can exacerbate adverse impacts when models are biased.</td>
</tr>
<tr>
<td>Accountability</td>
<td></td>
<td>ML models have been trained on actigraphy features to predict aggression in patients with dementia. However, patients should not be expected to advocate for themselves if models seem biased or are not generalisable, given their particularly vulnerable status.</td>
</tr>
<tr>
<td>Impartiality</td>
<td>Provenance</td>
<td>Prior conviction and a diagnosis of schizophrenia are predictors of violence. 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, or being more likely misdiagnosed with schizophrenia). 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.</td>
</tr>
<tr>
<td>Implementation</td>
<td></td>
<td>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. Most patients classified as high-risk do not become violent; however, pre-emptive interventions involving interactions with staff could precipitate violent behaviours.</td>
</tr>
<tr>
<td>Inclusion</td>
<td>Completeness</td>
<td>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. Additional marginalised groups might emerge when intersectional identities are taken into account. Collaboration in decision making during admission and maximising choice are important values for patients in settings where autonomy is limited. 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).</td>
</tr>
<tr>
<td>Patient and family engagement</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 review. LS took the lead on the ‘Three Pillars for Fairness’ framework. MMM took review. LS took the lead on the ‘Three Pillars for Fairness’ framework. MMM took the lead on the ‘Three Pillars for Fairness’ framework. MMM took the lead on the ‘Three Pillars for Fairness’ framework.

Funding This work was supported by the Delta 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).

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 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.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is

ORCID iD

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

REFERENCES

31 Kalluri P. Don’t ask if artificial intelligence is good or fair, ask how it shifts power. Nature 2020;583:169.
International Conference on Knowledge Discovery & Data Mining, New York, NY, USA, 2020:3591.


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

<table>
<thead>
<tr>
<th>eTable 1: Categories and search terms used in literature searches</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary Search Term</strong></td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
</tbody>
</table>

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)
<table>
<thead>
<tr>
<th>Step 1: Broad Search</th>
<th>Step 2: Targeted Search</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fairness</strong></td>
<td><strong>Health</strong></td>
</tr>
<tr>
<td><strong>AND</strong></td>
<td><strong>AND</strong></td>
</tr>
<tr>
<td><strong>Machine Learning OR Artificial Intelligence</strong></td>
<td><strong>Health</strong></td>
</tr>
<tr>
<td>“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”</td>
<td>“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”</td>
</tr>
</tbody>
</table>
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


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


