BMJ Health & Care Informatics

Conceptualising fairness: three pillars for medical algorithms and health equity

Laura Sikstrom, Marta M Maslej, Katrina Hui, Zoe Findlay, Daniel Z Buchman, Sean L Hill

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 systems1 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 recently12 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.


© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

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

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

Received 31 July 2021
Accepted 14 December 2021

BMJ Health Care Inform: first published as 10.1136/bmjhci-2021-100459 on 10 January 2022. Downloaded from http://informatics.bmj.com/ on October 25, 2022 by guest. Protected by copyright.
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 exceptions41–43). 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.
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...
about the so-called ‘black box problem’ of algorithmic systems, such as techniques to identify how models weigh features.\textsuperscript{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.\textsuperscript{72-74}

Even if explainability is possible, it may not yield desirable outcomes.\textsuperscript{75, 76} According to emerging evidence, many clinicians are susceptible to following incorrect diagnostic advice.\textsuperscript{77} This effect is more pronounced when ML-based advice is paired with explanations of features contributing to prediction,\textsuperscript{78} suggesting that explainability can adversely impact clinical decision making.\textsuperscript{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
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.94 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.98

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.93 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.99 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.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 For example, ‘computational ethnography’ is an approach to fairness-aware ML that emphasises the importance of a holistic understanding of any given dataset.121 122 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.123–127 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.\textsuperscript{131-132} From a health systems perspective, they can be useful tools to support the fair allocation of limited resources.\textsuperscript{31, 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.\textsuperscript{129, 134} and redefine what makes a ‘good’ clinician.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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’,\textsuperscript{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.\textsuperscript{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.\textsuperscript{118, 149-153} Additional legal and social constraints limit access to sensitive sociodemographic data.\textsuperscript{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\textsuperscript{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.\textsuperscript{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.\textsuperscript{157-159} In addition, most current approaches to operationalising fairness focus only on legally protected categories, such as race or legal gender.\textsuperscript{160} Yet, sexual orientation, gender identity and disability are prototypical instances of unobserved characteristics, because they are frequently unrecorded but also fundamentally unmeasureable.\textsuperscript{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.\textsuperscript{163-165} 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.\textsuperscript{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).\textsuperscript{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.\textsuperscript{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\textsuperscript{170, 171} and ‘recovery’ in psychiatry.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{35} Over the past several decades, various features have emerged as predictors of this risk.\textsuperscript{175} ML-based models trained on patient characteristics, structured assessments and clinical notes have achieved reasonable performance in predicting violence or aggression.\textsuperscript{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,\textsuperscript{36} 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 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. 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.


