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,1,3 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.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.2-3 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.6-9 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.7-19 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, 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 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). 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...
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, ensuring that stakeholders can decide whether they support the intentions, indications for use, and goals of any algorithmic system.50 61  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. 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

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

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.

Explainability and interpretability 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.

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.

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 (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. 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 socio-political 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 unmeasureable.

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 (eg, undocumented or criminalised youth). Therefore, while completeness entails inclusivity, inclusion should always be precipitated by dialogue and collaboration.

ER195

**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
Sikstrom L, et al. BMJ Health Care Inform 2022;29:e100459. doi:10.1136/bmjhci-2021-100459

Open access

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

---

**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 | Interpretable    | 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 65. 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. 181–189 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). 180 |

ML, machine learning.
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.

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, Scheidgger 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, Manylak 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 Barcos S, Hardt M, Narayanan A. Fairness in machine learning. Nips tutorial 2017;1:2017.
10 Darlington RB. Another look at “cultural fairness”. 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, Moos 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;3:153–63.
15 Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. Minds Mach 2020;30:99–120.
16 Srivastava M, Heidarf 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 Berlin 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 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 Kettell 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.
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.

Boag W, Kovaleva O, McCoy TH, et al. Hard for humans, hard for machines: prediction admission after psychiatric hospitalization using narrative notes. Transl Psychiatry 2021;11:32.

Suchting R, Green CE, Glazier SM, et al. A data science approach to predicting patient aggressive events in a psychiatric hospital. Psychiatric Res 2018;268:217–22.

Vijayan JL, Cochrane DM, Johnson MR. Do risk assessment tools help manage and reduce risk of violence and reoffending? A systematic review. Law Hum Behav 2018;42:181–214.

Connor M, Armbruster M, Hurley K, et al. Diagnostic sensitivity of the dynamic appraisal of situational aggression to predict violence and aggression among behavioral health patients in the emergency department. J Emerg Nurs 2020;46:302–9.

Rawls J. Justice as fairness. Philos Rev 1958;67:164–94.

Carr CL. Rawls J. Justice as fairness. Philos Rev 1958;67:164–94.

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.

Zarsky T. The trouble with algorithmic decisions: an analytic road overview-of-hr-codes-byprovince-final-en pdf. N Engl J Med 2018;379:1874–82.

O’Neil C. How the machine ‘thinks’: Understanding transparency ideal and its application to algorithmic accountability. Sci Technol Human Values 2020;26:1771–96.

Barnett L, Walker J, Nicol J. Machine Learning Fairness Metrics. In: lii HD, editor. Proceedings of the 37th International Conference on Machine Learning, PMLR, 2020; 8377–87.

Lepri B, Oliver N, Letouzé E, et al. Diagnostic sensitivity of algorithmic systems in Wikipedian organizational culture. Big Data Soc 2020;7:205395171986951.

Chandler C, Fairhead J, Kelly A, et al. Measuring Non-Expert Opacity in Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. arXiv [cs.AI], 2017. Available: http://arxiv.org/abs/1708.08296

Gunning D, Aha D. DARPA’s explainable artificial intelligence (XAI) program. AI Mag 2019;40:44–58.

Slack D, Hilgard S, Jia E, et al. How can we fool lime and ShAP? Adversarial Attacks on Post hoc Explanation Methods, 2019. Available: https://openreview.net/forum?id=NTOHao__v0B [Accessed 19 Jul 2021].

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.

Watson DS, Floridi L. The explanation game: a formal framework for interpretable machine learning. Synthese 2021;198:9211–42.

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

Babić B, Gerke S, Evgeniou T, et al. Beware explanations from AI in health care. Science 2021;372:284–6.

Geiger RS. Beyond opening up the black box: investigating the role of nominal clinical data for predicting Hospital readmission. KI - Künstliche Intelligenz 2015;29:153–9.

Lepri B, Oliver N, Letouzé E, et al. Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 2018;31:611–27.

Char DS, Abramoff MD, Feudtner C. Identifying ethical considerations for machine learning healthcare applications. Am J Bioeth 2020;20:7–17.

Brundage M, Avin S, Wang J. Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv, 2020. Available: http://arxiv.org/abs/2004.07213

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.

Geiger RS. Beyond opening up the black box: investigating the role of algorithmic systems in Wikipedian organizational culture. Big Data Soc 2017;4:205395171773073.

Bellamy RKE, Dey K, Hind M. AI fairness 360: an EXTENSIBLE toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. Available: https://github.com/lbm/aif360

Babovic V, Zafirou S, Ishola P, et al. New information on the prescription of contraception in emergency contraception devices. Contraception 2009;68:1711–9.

Hengstler M, Enkel E, Duelli S. Applied artificial intelligence and its role in explainable machine learning. Science 2018;359:1079–83.

Chandler C, Fairhead J, Kelly A, et al. Ebola: limitations of predicting cholera control campaigns. Soc Sci Med 1996;43:1007–24.

Jones VK, Roberts P, Allen D. A data science approach to predicting readmission after psychiatric hospitalization. JAMA Psychiatry 2018;76:642–51.

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.

Chandler C, Fairhead J, Kelly A, et al. Measuring Non-Expert Opacity in Explainable Artificial Intelligence: Theory, Algorithms, and Applications. arXiv [cs.AI], 2021. Available: http://arxiv.org/abs/2003.04244

Saravanakumar KK. The Impossibility Theorem of Machine Fairness - A Causal Perspective. arXiv [cs.LG], 2020. Available: http://arxiv.org/abs/2007.06024

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

Hong W, 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

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=NTOHao__v0B [Accessed 19 Jul 2021].

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.

Watson DS, Floridi L. The explanation game: a formal framework for interpretable machine learning. Synthese 2021;198:9211–42.

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

Gunning D, Aha D. DARPA’s explainable artificial intelligence (XAI) program. AI Mag 2019;40:44–58.

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.

Babic B, Gerke S, Evgeniou T, et al. Beware explanations from AI in health care. Science 2021;372:284–6.

Gauke S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021;4:31.

Jacobs M, Pradler MF, McCoy TH, et al. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Transl Psychiatry 2021;11:1–9.

McCradden MD. When is accuracy off-target? Transl Psychiatry 2021;11:369.

Kaler A. Health interventions and the persistence of rumour: the circulation of sterility stories in African public health campaigns. Soc Sci Med 2009;68:171–9.

Nations MK, Monro P. “I’m not dog, no!”: cries of resistance against cholera control campaigns. Soc Sci Med 1996;43:1007–24.

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.

Chandler C, Fairhead J, Kelly A, et al. Ebola: limitations of correcting misinformation. Lancet 2015;385:1275–7.

Saha D, Schumann C, Mcfleish D. Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics. In: lii HD, editor. Proceedings of the 37th International Conference on Machine Learning, PMLR, 2020: 8377–87.

Kellieri 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

et alAnderson A, Dodge J, Sadarangani A. Explaining Reinforcement Learning to More MORTs: An Empirical Study. arXiv [cs.HC], 2019. Available: http://arxiv.org/abs/1903.09708

Dodge J, Iaio OV, 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.

Ho A. Are we ready for artificial intelligence health monitoring in elder care? BMC Genet 2020;20:358.

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.

Peck EM, Ayuso SE, El-Etr O. Data is personal: attitudes and perceptions of data collection and use among hospital physicians. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2019:1–12.
107 Mitchell S, Potash E, Barocas S, et al. Algorithmic fairness: choices, assumptions, and definitions. Annu Rev Stat Appl 2021;8:141–63.

108 Trewin S, Basson S, Muller M, et al. Considerations for AI fairness for people with disabilities. AI Matters 2019;5:40–63.

109 Pavlopoulos G, Ben-Nun A, Vlachos D. Social Biases in NLP Models as Barriers for Persons with Disabilities. arXiv [cs.CL], 2020. Available: http://arxiv.org/abs/2005.00813

110 Goldenstein J, Poschmann P. Analyzing meaning in big data: performing a MAP analysis using grammatical parsing and topic modeling. Social Methodol 2019:39:297–335.

111 Arnold T, Fuller HJA. In: 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.

112 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.

113 Moore RJ, Smith R, Liu DQ. 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.

114 Leahy E. Overseeing research practice: the case of data editing. BMJ Health Care Inform 2009;15:188–197.

115 Engle SM. The social life of measurement Sally Engle Merry. published online first, 2016. Available: http://www.americanbar.com/ou/ndation.org/uploads/cms/documents/merry-lsi_forum_rev.pdf

116 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.

117 Blohmberg 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:e2033220.

118 Sendak MP, Ratiwif 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.

119 Emanuell EJ, Wackman RM. Artificial intelligence in health care: will the value match the hype? JAMA 2019;321:2281–2.

120 Tomlinson A, Furukawa TA, Ethimiou O, et al. Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. J Eval Clin Pract 2020;26:1325–6.

121 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:e2033220.

122 Abrahamson 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.

123 Moore RJ, Smith R, Liu DQ. 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.

124 Leahy E. Overseeing research practice: the case of data editing. BMJ Health Care Inform 2009;15:188–197.

125 Engle SM. The social life of measurement Sally Engle Merry. published online first, 2016. Available: http://www.americanbar.com/ou/ndation.org/uploads/cms/documents/merry-lsi_forum_rev.pdf

126 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.

127 Blohmberg 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:e2033220.

128 Sendak MP, Ratiwif 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.

129 Emanuell EJ, Wackman RM. Artificial intelligence in health care: will the value match the hype? JAMA 2019;321:2281–2.

130 Tomlinson A, Furukawa TA, Ethimiou O, et al. Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. J Eval Clin Pract 2020;26:1325–6.

131 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:e2033220.

132 Abrahamson 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.

133 Moore RJ, Smith R, Liu DQ. 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.

134 Leahy E. Overseeing research practice: the case of data editing. BMJ Health Care Inform 2009;15:188–197.

135 Engle SM. The social life of measurement Sally Engle Merry. published online first, 2016. Available: http://www.americanbar.com/ou/ndation.org/uploads/cms/documents/merry-lsi_forum_rev.pdf

136 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.

137 Blohmberg 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:e2033220.

138 Sendak MP, Ratiwif 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.

139 Emanuell EJ, Wackman RM. Artificial intelligence in health care: will the value match the hype? JAMA 2019;321:2281–2.

140 Tomlinson A, Furukawa TA, Ethimiou O, et al. Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. J Eval Clin Pract 2020;26:1325–6.

141 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:e2033220.

142 Abrahamson 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.

143 Moore RJ, Smith R, Liu DQ. 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.

144 Leahy E. Overseeing research practice: the case of data editing. BMJ Health Care Inform 2009;15:188–197.

145 Engle SM. The social life of measurement Sally Engle Merry. published online first, 2016. Available: http://www.americanbar.com/ou/ndation.org/uploads/cms/documents/merry-lsi_forum_rev.pdf

146 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.

147 Blohmberg 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:e2033220.

148 Sendak MP, Ratiwif 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.

149 Emanuell EJ, Wackman RM. Artificial intelligence in health care: will the value match the hype? JAMA 2019;321:2281–2.

150 Tomlinson A, Furukawa TA, Ethimiou O, et al. Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. J Eval Clin Pract 2020;26:1325–6.

151 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:e2033220.

152 Abrahamson 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.
Snipp CM. What does data sovereignty imply: what does it look like Indigenous data sovereignty toward an agenda, 2016; 39–55.

Kukutai T, Taylor J O. Data sovereignty for Indigenous peoples: current practices and future needs. ANU Press, 2016.

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.

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.

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.

Mc Kenzie K, Antwi BAM, Tuck A. The case for diversity, 2016.

Andrus M, Spitzer E, Brown J. What We Can’t Measure, We Can’t 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.

Cristina Mora G. Making Hispanics. University of Chicago Press, 2021.

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.

Thompson D. The schematic state. Cambridge University Press, 2016.

Saini A. Superior: the return of race science. Beacon Press, 2019.

Trevarinus J. The Value of Being Different. In:Association for Computing Machinery. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 2021:1–7.

Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence 2016;5:221–32.

He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21:1263–84.

Willen SS. How is health-related “deservingness” reckoned? Perspectives from unauthorized im/migrants in Tel Aviv. Soc Sci Med 2012;74:712–21.

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.

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.

Ford CL, Aihinenbuvva CO, Theory CR. Critical race theory, race equity, and public health: toward antiracism praxis. Am J Public Health 2010;100 Suppl 1:S30–5.

Seng JS, Lopez WC, 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.

Rosenfield S. Triple jeopardy? mental health at the intersection of gender, race, and class. Soc Sci Med 2012;74:1791–801.

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.

Crenshaw K. Mapping the margins: intersectionality, identity politics, and violence against women of color. Stanford Law Rev 1991;43:1241.

Koppe G, Meyer-Lindenberg A, Durstewitz D. Deep learning for small and big data in psychiatry. Neuropsychopharmacology 2021;46:176–90.

Bratich J. Digital age occult (ING) transparency: an Epilogue. Int J Commun Syst 2016 https://ijoc.org/index.php/ijoc/article/download/ 4896/1526.

Kleiman A. From illness as culture to caregiving as moral experience. N Engl J Med 2013;368:1376–7.

Kleiman A. The Illness Narratives: Suffering, Healing, And The Human Condition. Basic Books, 2020.

Csordas TJ, Kleiman A. The therapeutic process medical anthropology contemporary theory and method. 1996: 3–20.

Topol E. Others. The Topol Review. In: Preparing the healthcare workforce to deliver the digital future, 2019: 1–48.

Cupertain AB, Hui K, Berkhourt SG, et al. Patient, family and provider views of measurement-based care in an early-psychosis intervention program. BJPsych Open 2021;7.

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.

Shen R, V Snupit M, 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.

Abebe R, Barocas S, Kleinberg J. Roles for computing in social change. arXiv [cs.CY], 2019.

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.

Appelbaum PS, Robbins PC, Monahan J. Violence and delusions: data from the MacArthur violence risk assessment study. Am J Psychiatry 2000;157:566–72.

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.

Olbert CM, Nagartera A, Ahn B. A machine learning 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.

Kohler-Hausmann I. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. Nw UL Rev 2018;113:1163.

Papadopoulos C, Ross J, Stewart D, et al. The antecedents of violence and aggression within psychiatric in-patient settings. Acta Psychiatr Scand 2012;126:425–39.

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.

Giersen 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.

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.

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;40:332–6.

McGuinness D, Murphy K, Bainbridge E, et al. Individuals’ experiences of involuntary admissions and preserving control: qualitative study. BJPsych Open 2018;4:501–9.

Hui K, Cooper RB, Zaheer J. Engaging patients and families in the ethics of involuntary psychiatric care. Am J Bioeth 2020;20:82–4.

Care HKR, closeness, and becoming ‘better’: Transformation and therapeutic process in American adolescent psychiatric custody. Ethos 2016;44:313–32.