Appendix 1: Developing a Conceptual Framework for Fairness

Methodology

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

Results

eTable 1: Categories and search terms used in literature searches

	Primary Search Term	Secondary Search Term(s)

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

Appendix 2: Case Scenario: Risk Assessment in Psychiatry

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

Transparency

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

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

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

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

This potential for amplifying harms in predictive care makes it necessary to reconsider accountability, since those who are primarily impacted should not be expected to identify biases and mobilize change. Many inpatients in psychiatric facilities are acutely ill or marginalized. They can suffer from chronic and complex health needs related to historical and intergenerational traumas, homelessness, and encounters with the law. These patients might be unaware of potential biases in predictive care and would not have the means to redress any harm caused by this bias. As another example, using wearable sensors to monitor patients with dementia has been explored as a way to reduce burden on staff and caregivers (and promote

increased independence for patients). Various documented associations between actigraphy and aggressive behaviour introduce an opportunity to train ML models to predict aggression in patients with dementia.[9] However, few efforts have been made to evaluate models across samples and settings leading to issues of generalizability. If a subset of patients is likely to be incorrectly classified, who will know? Even if models are interpretable and explainable, who should advocate for patients, given their particularly vulnerable status?

Impartiality

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

Identifying and removing biased features from training datasets, or encoding more data about them, does not necessarily result in fair prediction of inpatient violence, in part because these features are inextricably linked to other predictors. For example, given historically racist

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

Deploying biased models can perpetuate harmful outcomes for already disadvantaged groups when model predictions are used to make clinical decisions. The development of ML models to predict violence in psychiatric settings is primarily motivated by an aim to improve patient and staff safety.[3] However, there is also an underlying desire to more effectively allocate staff and hospital resources, given limited operating costs.[3,19,20] The logic behind this secondary motivation is that surveillance or pre-emptive intervention (e.g., restricting freedoms or privileges) can be focused on subsets of patients at highest risk of violence. However, clinician-patient interactions, particularly those that limit patient freedoms or deny requests, are known antecedents to violent behaviours, precipitating almost 40% of incidents in inpatient settings.[21] Most patients identified with ML to be at risk do not become violent.[22] However, if pre-emptive interventions were implemented,[3] these patients could become aggressive or violent due to increased interactions with clinicians. Clinicians and staff might also respond to ML-

based risk classification by treating these patients differently, altering their trajectory of care. Critically, once a predictive care model is deployed, what is defined as "fair" (accurate, efficient, cost-effective) from a health systems perspective may not be perceived as fair (unbiased) from the patient's point of view. Thus, considerations related to implementation must involve balancing among multiple and sometimes competing interests (e.g., of clinicians, hospital administrators, patients).

Inclusion

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

In addition to ensuring that predictive care has equitable impacts across sociodemographic groups, inclusion also involves engaging those impacted groups in all aspects of the ML process – from data compilation to model development and implementation. Qualitative data from acute psychiatric inpatient settings emphasizes that collaboration in decision-making during admission and maximizing choice are important values for patients, particularly in settings where autonomy

is somewhat limited.[28–30] While current efforts to operationalize fairness emphasize prediction performance across protected groups, patients may have different notions of fairness or prioritize other aspects of care altogether beyond what is captured by ML. For example, Hejmanek (2016) found that Black youth in psychiatric custody attributed their recovery to the caring relationships that they built with staff and peers, not to State-mandated therapeutic interventions, like group therapy.[31] Moreover, typical institutional responses to aggression adversely impact the experience of care,[32,33] reduce patients' willingness to disclose sensitive information, and undermine treatment adherence.[34,35] Thus, in service of achieving inclusion in ML-based risk assessments, we must seek to understand and include the perspectives, preferences, and experiences of patients in acute psychiatric settings.

References Cited

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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