Predictive care: a protocol for a computational ethnographic approach to building fair models of inpatient violence in emergency psychiatry

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

ABSTRACT

Introduction Managing violence or aggression is an ongoing challenge in emergency psychiatry. Many patients identified as being at risk do not go on to become violent or aggressive. Efforts to automate the assessment of risk involve training machine learning (ML) models on data from electronic health records (EHRs) to predict these behaviours. However, no studies to date have examined which patient groups may be over-represented in false positive predictions, despite evidence of social and clinical biases that may lead to higher perceptions of risk in patients defined by intersecting features (e.g., race, gender). Because risk assessment can impact psychiatric care (e.g., via coercive measures, such as restraints), it is unclear which patients might be underserved or harmed by the application of ML.

Methods and analysis We pilot a computational ethnography to study how the integration of ML into risk assessment might impact acute psychiatric care, with a focus on how EHR data is compiled and used to predict a risk of violence or aggression. Our objectives include: (1) evaluating an ML model trained on psychiatric EHRs to predict violent or aggressive incidents for intersectional bias; and (2) completing participant observation and qualitative interviews in an emergency psychiatric setting to explore how social, clinical and structural biases are encoded in the training data. Our overall aim is to study the impact of ML applications in acute psychiatry on marginalised and underserved patient groups.

Ethics and dissemination The project was approved by the research ethics board at The Centre for Addiction and Mental Health (053/2021). Study findings will be presented in peer-reviewed journals, conferences and shared with service users and providers.

BACKGROUND

In this project, we investigate an emergent medical and technical innovation we call ‘predictive care’. Predictive care is an approach to medicine driven by advances in artificial intelligence (e.g., machine learning (ML)), which enhance our ability to identify patterns and predict outcomes or behaviours. This data-driven approach combines big data (on whole populations) and small data (on individuals) to facilitate more proactive, precise and personalised care. In psychiatry, tools to predict suicide, psychiatric readmission and violence or aggression are in high demand, although few have been successfully deployed. There is an increasing awareness that ML can amplify inequities, such as racial bias, often because ML models are trained on biased datasets. Implementing biased models in health settings can have adverse impacts on patient care, typically for already marginalised or underserved groups. Thus, the purpose of this project is to conduct a computational ethnography to examine how datasets are created, interpreted and used within a specific context; the
emergency department at The Centre for Addiction and Mental Health (CAMH ED).

CAMH is a large urban, mental health and addictions hospital in downtown Toronto, Canada. Each year, the CAMH ED provides 24 hour/7 days per week emergency assessment and treatment for adults with mental health and substance use issues. In this dynamic and acute care context, patient aggression and violence towards self or others are of paramount concern. Patient aggression has a negative impact on the well-being of patients and staff. Further, common practices used to manage patient aggression, including chemical and physical restraints and seclusion, are coercive practices that can be re-traumatising. These practices negatively impact the patient experience of care, reduce the willingness of patients to disclose sensitive information and may increase patient aggression overall. As a result, most coercive interventions, such as agitation, to facilitate timely and non-coercive rates. Because perceptions of risk impact patient care, identifying early warning signs of aggression, such as agitation, to facilitate timely and non-coercive interventions could be critical. At present, clinical staff employ a wide range of techniques and strategies to manage and prevent violence and aggression. One of the most common practices in psychiatric settings is the use of routine structured risk assessments or psychometric scales. Structured risk scales require that staff rate patients on various behavioural antecedents, such as irritability or sensitivity to provocation. For example, at CAMH, ED nurses enter narrative details about patients into the electronic health record (EHR) every 2–3 hours. They might enter: ‘patient ate her cereal and watched TV quietly’. Later, they may add: ‘patient appeared irritable; argued about meds’. The next morning, another nurse will review these narratives and generate a daily numerical score (0–7) called a ‘Dynamic Appraisal of Situational Aggression’ (DASA).

While research generally suggests that structured scales such as the DASA are effective de-escalation and prevention tools, high false positive rates are concerning (eg, over 50% of patients rated as high risk do not go on to become violent or aggressive). Given the current limitations of structured risk assessments, there is a growing interest in the possibilities presented by advances in ML to support clinical decision-making in mental health. By leveraging patients’ sociodemographic, clinical and behavioural characteristics recorded in EHRs, ML might provide timelier and more precise measurements of risk than using structured ratings alone. Although EHRs are primarily used for documentation and communication among clinicians, they also contain factors contributing to risk, such as substance use or a history of violent altercations with police. Although ML models have achieved reasonable overall performance in predicting a risk of violence or aggression, no studies to date have examined whether certain patient groups are over-represented in false positive rates. Because perceptions of risk impact patient care (eg, via coercive measures, such as increased surveillance or restraints), it remains unclear which patients could be harmed by the application of ML in risk assessments.

The dearth of research on inequities in assessing a risk for inpatient violence or aggression is particularly troubling, given evidence of social and systemic factors that may bias perceptions of risk in certain patient groups. For example, almost 30% of patients are brought into the CAMH ED by police, and police presence is commonly communicated among clinicians as a factor relevant to risk assessment. However, Black men are more commonly apprehended by police, which may contribute to higher perceptions of risk for this demographic group. Indeed, evidence suggests Black patients receiving emergency psychiatric care are more likely to be held involuntarily, as compared with white patients. If systemic factors, like police presence, increase perceptions of risk for certain patient groups, this bias is likely reflected in EHRs. Moreover, one popular source of training data for modelling violence or aggression is clinical notes. However, ML models trained on note features to predict other psychiatric outcomes, such as rehospitalisation for depression, have been shown to be biased, generating less accurate predictions in lower-income patients. Training ML models on EHRs might therefore result in biased predictions, which can exacerbate inequities for historically marginalised or underserved groups. Without a clear understanding of the factors underlying biased predictions, ML applications are likely to perpetuate existing inequities and harms. Current approaches to mitigating bias in ML focus on debiasing data or models to make fairer predictions. However, removing biased features from training data may not be possible, since it is often unclear how they might be linked to other features. Further, these approaches rarely address the underlying sociotechnical realities that lead to bias in training data. In-depth investigation of such biases is critical but challenging to carry out, since it requires interdisciplinary collaboration and meaningful input from service providers and service users. To address this issue, we pilot a novel computational ethnographic approach to studying bias in emergency psychiatric care, focusing on how patient data is compiled, interpreted and used to predict a risk of violence or aggression.

This project draws on theoretical insights from critical medical anthropology on care and psychiatry, the anthropology of algorithms and ethical debates about the place of ML in highstakes applications such as healthcare. Following Seaver et al, we define our object of study as an ‘algorithmic system’, which in our proposed study is a heterogeneous and diffuse sociotechnical system that can be engaged with empirically. We will investigate how risk assessments are produced through everyday social interactions, data entry practices and institutional standards. Participant observation, semi-structured interviews and ML modelling of EHRs will enable ‘thick description’ of risk assessment practices in an acute psychiatric care setting.
extrapolated from specific examples to generate conclusions about the emerging role of predictive care in clinical settings. To our knowledge, we are the first researchers to document these processes within an emergency psychiatric setting. Our project is also particularly timely, given ongoing efforts to develop and deploy ML-based predictive risk assessment care tools in medicine. 

Objectives
Our project provides the critical context for future ML-based risk assessment tools in psychiatry. Its overall aims are to (a) examine the potential for a predictive care tool to perpetuate inequities in acute psychiatric care; and (b) gain critical insights into the social and systemic sources of inequities. To achieve these aims we will:

1. Conduct participant observation in the ED, focusing on the clinical context that shapes routine risk assessments.
2. Complete interviews with clinicians, patients and knowledge users to contextualise risk assessments.
3. Evaluate a predictive care model trained on EHRs to support risk assessment for evidence of intersectional, sociodemographic bias.

Research hypotheses
We hypothesise that assessments of violence or aggression in psychiatric ED settings are biased based on demographic (eg, race, housing) and contextual factors (eg, police involvement), leading to inequities in care for samples defined by these intersecting features. Specifically, we expect to find higher structured risk assessment scores (ie, DASA ratings) in racially marginalised groups, foreign-born individuals, and potentially for patients admitted to the ED by police. Extending on previous work, narratives or clinical notes describing these patients may contain more negative sentiment. Accordingly, ML models trained on this data, in combination with sociodemographic and clinical features, might overestimate risks for these patients, exhibiting higher ‘false positive’ predictions. Such findings could imply that racialisation, language barriers and police involvement can bias risk assessments, potentially through perceptions of these groups as being more aggressive. Furthermore, if individuals are incorrectly identified as being at high risk based on sociodemographic or contextual factors (eg, racial marginalisation, police involvement), the deployment of predictive care could amplify bias and increase inequities.

We also anticipate that our ethnographic findings will identify structural factors impacting risk assessments that are not normally accounted for in approaches that attempt to debias datasets after they are created. For example, we anticipate that many ‘unquantifiable remainders’ (eg, frustration when language barriers make it difficult to communicate, alert fatigue, patient surges and/or unobserved characteristics such as autism spectrum disorder) are likely to impact the material realities of data collection more broadly. Accounting for both the quantifiable and unquantifiable dimensions that bias training datasets is critical to the construction of fair predictive care models.

METHODS AND ANALYSIS

Approach
Computational ethnography is an approach that extends ethnography’s toolkit to include computational methods, such ML. It continues to leverage the strength of ethnography in terms of understanding the messy realities that surround the collection, use and interpretation of datasets, while also documenting the connections from everyday work practices to broader organisational and political processes (eg, the impact COVID-19 might have on risk management). In particular, drawing on the insights from institutional ethnography, we will observe how texts, such as textual data contained in EHRs, are ‘activated’ when they are read, completed or filled in by staff as they go about their work. Traditionally, ‘texts’ referred to memos, emails and reports produced by institutions. In psychiatric settings, however, the management of risk is coordinated using digital texts, primarily through EHRs. The scale and velocity of textual data contained in EHRs make it difficult for ethnographers to analyse texts using established tools (eg, content analysis). Thus, to fully grasp the impact of predictive care for clinical practice and health equity, we extend on Cury et al. ‘hybrid methodology’. Hybrid methodologies combine (and challenge) the analytical frameworks and methods from multiple disciplines; in our case, computer science, medicine, anthropology, bioethics and psychology. Our hybrid approach emphasises the importance of sharing perspectives, reflecting on tensions and being open to new ideas.

Setting
Within emergency settings, there are a range of situational factors that may precipitate violence. For example, CAMH has the busiest ED in Canada, reporting more than 14500 patient visits in 2021–2022, an increase of nearly 110% over the last decade. Further, in Ontario a recent study found a high and increasing prevalence of involuntary admissions (70.7% in 2009, in 2013 74.1%). Involuntary admissions occur when an individual with mental illness is admitted to hospital under Ontario’s Mental Health Act against their will owing to a perceived imminent danger to the individual or others. Involuntary admission is a known antecedent of violence or aggression in emergency settings as it disrupts the patient–provider relationship and has a negative impact on the patient’s perception of care. Incidents are also more likely to occur during the first few days of admission, prior to the stabilisation of acute symptoms. Staff shortages due to the COVID-19 pandemic and a steep increase in police acting as first responders to mental health crises are additional stressors for both staff and patients (see table 1). At CAMH, clinical staff
have tremendous expertise with violence prevention, making it an excellent setting to observe risk assessment best practices and high-quality care.

Study design
This study has two phases: (1) data collection and analysis and (2) knowledge mobilisation.

Phase 1: data collection and analysis (July 2022 to August 2023)

Data collection
We will collect and analyse data from three sources: (1) EHRs (ie, patient data, structured and narrative risk assessments), (2) participant observation in the CAMH ED and 3) semistructured interviews with service providers, service users and key stakeholders. Findings from our quantitative analysis of EHRs will refine ethnographic observations, interview guides and purposive sampling strategies. Findings from our ethnographic analysis will provide critical contextual data that will be used to better understand our quantitative results. Data analysis will be conducted iteratively alongside data collection.

Electronic health records
To train our ML model, we will compile any available structured and unstructured EHR data on risk assessment across inpatient acute care units at the hospital since 2016. These include: the DASA, a 7-point scale assessing irritability, willingness to follow directions and other relevant patient variables (completed each morning a patient is on the unit) and the Functional Monitoring Tool (FMT), a narrative description of a patient’s clinical status and behaviour (completed every 2-3 hours). We will also extract relevant patient factors collected at admission into the ED, such as diagnoses, sociodemographic characteristics (ie, gender, race, income, citizenship and housing status), admission information (eg, accompanied by police, involuntary admission under the Ontario Mental Health Act), as well as wait times, which may contribute to risk for aggression or violence.76 Finally, we will extract any available data on our prediction target or outcome. These include violent or aggressive incidents documented in the hospital’s reporting tool (ie, physical, verbal and sexual aggression). Following prior work,4 we also include documented restraints (ie, physical, pharmacological, seclusion) as an outcome, since these interventions are only employed when serious instances of violence or aggression are deemed imminent by staff; however, it is possible that some of these restraint incidents reflect situations in which patients would not have become violent or aggressive. All data extracted from EHRs are listed in online supplemental appendix 1.

Participant observation
Participant observation will involve shadowing clinical volunteers who provide written consent to be observed,70–75 to examine the clinical context that shapes the production and use of the DASA. Participant observations will focus on routine clinical interactions, and morning handovers where clinicians discuss how they define and interpret aggression (eg, ‘posturing’), adverse events and risk mitigation strategies. The aim is to capture and document clinicians’ expert knowledge about the work they do to manage violence and aggression.51 74 This will include contextualising the challenges faced by clinical staff conducting and documenting risk assessments in a dynamic work environment. Rough ‘scratch’ notes taken during observations will be anonymised and integrated into a digital database after each session to minimise any intrusiveness.76 ‘Think alouds’ will be used to ask clarification questions at mutually agreed-on times.77

Semistructured interviews
In the beginning of phase 1, we will conduct semistructured interviews with former CAMH ED patients. Interview questions will be open ended to capture the range of concerns and experiences of participants (see online supplemental appendix 2). Topics will include experiences within a psychiatric emergency department and narratives of individual safety and well-being.78 Near the end of phase 1, we will conduct semistructured interviews with clinicians (see online supplemental appendix 3). Topics will include experiences with aggression, violence and risk assessments. Interviews with other key stakeholders (eg, with police or institutional leaders) will supplement knowledge gaps. We will document positive and negative statements about risk assessments. Each subgroup will likely have unique insights into challenges and opportunities presented by risk assessment. All interviews will be digitally recorded, transcribed and deidentified at the time of transcription.

Data analysis
Step 1: identifying biases
Based on methods from prior work,4 29 we will train ML models on EHRs to predict violence or aggression. Training features will include patient clinical and sociodemographic characteristics, admission information and risk assessments, that is, DASA scores and FMT texts (see online supplemental appendix 1). We will use two approaches to address missingness for sociodemographic variables, where we anticipate most missing data and as much as 50% (eg, for sexual orientation, income, or education). We do not expect these data are missing at

| Table 1 | Increases in police use for mental health services
t| Toronto Police Service (TPS) interactions | 2014 (count) | 2020 (count) | % increase from 2014 to 2020 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TPS mental health calls</td>
<td>22 229</td>
<td>33 059</td>
<td>49%</td>
<td></td>
</tr>
<tr>
<td>TPS mental health apprehensions</td>
<td>7393</td>
<td>11 707</td>
<td>58%</td>
<td></td>
</tr>
</tbody>
</table>
random, since patients with more severe illness or distress may be less likely to complete demographic questionnaires, or some patients may not wish to disclose certain demographic characteristics. Thus, following prior work, we will create a new category for any variables with more than 20% missing data and coding missing values into this category. Variables with 20% or less missing data will be imputed (eg, with MissForest). All categorical variables will be one-hot encoded prior to modelling. To extract features from the textual data, we will train topic models with Latent Dirichlet Allocation on FMTs to determine the probabilities of topics related to violence, aggression, or non-compliance within each text. To capture how patients are perceived by clinicians and staff, we will also quantify text sentiment. We will derive additional representations of the FMTs with transformer-based large language models. Our outcome will be represented as a binary outcome, indicating the presence (1) or absence (0) of a violent or aggressive incident and/or restraint event.

First, we will split our data into train (70%) and test (30%) sets, ensuring that data from the same visit and patient are not split between the two sets. Because risk assessments (ie, DASA scores, narrative notes) are documented daily (see online supplemental appendix 2), classifiers will be trained to predict the outcome on each day the patient is in the acute care unit. The prediction window will be based on DASA ratings compiled in the morning and texts concatenated over the course of each day, starting from the first DASA and until the outcome occurs or until a next DASA is recorded (about 24 hours). Since we expect most outcomes to occur on the first 3 days, we will include up to 3 days or prediction windows for each visit. Figure 1 shows the prediction window and timing of measures included in the ML model. If one or more outcomes occur during a given visit, we will only include data collected until the first occurrence, since interventions used to manage the outcome may alter risk. Since clinical, sociodemographic and admission data is only collected once for each visit, it will be repeated across the 3 days for each visit and patient.

We will evaluate the performance of various classifiers on the training set, including regularised logistic regression, random forest, gradient boosting machines and support vector machines. Classifiers will be trained with 10-fold cross validation to estimate performance, with a focus on area under the receiver operating curve (AUC). Hyperparameters (eg, regularisation strength, decision trees/features) will be tuned on the training set, optimising for AUC. Since violence or aggression tend to be rare (occurring on 2%-5% of days), we expect an imbalanced outcome. We will examine approaches to class rebalancing (eg, over-sampling, under-sampling, importance sampling). The classifier, parameters and sampling approach with the best cross-validated performance on the training set will be applied to generate predictions on the test set. Based on prior work, we expect this classifier to achieve a performance of AUC=0.76–0.80.

Our analysis aims to examine whether ML predictions of violence or aggression are biased against certain patient groups, and not to develop and implement an ML model to automate or support risk assessment. Thus, adapting methods from prior work, we will evaluate model bias by carrying out a fairness assessment of test set predictions generated with the best-performing classifier, focusing on rates of false positive and false negative predictions. We will calculate false positive parity and false negative parity for groups of patients defined by intersections of race, gender, housing, immigration, citizenship and method of admission, using groups expected to be at lowest risk of bias as reference. We anticipate that certain groups (eg, immigrants, racialised men) may have more false positive predictions, as compared with other groups. Findings from this analysis will be used to refine ethnographic observations, interview guides and purposive sampling strategies. Specifically, we will explore potential sources of bias for any groups with parities>1.

Figure 1 Prediction window for machine learning (ML) modelling. DASA, Dynamic Appraisal of Situational Aggression; ED, emergency department.

Note. OBS, observation; *If the outcome (1) occurs on a given day, no further risk assessment data (i.e., narrative notes or DASA scores) will be included in ML modelling for that visit.
Further quantitative analyses will complement ethnographic efforts to explore potential sources of intersectional bias in predictive care. Specifically, we will examine associations between sociodemographic, clinical, and admission features and three outcomes reflecting perceptions of risk: total DASA scores, narrative risk assessment features and restraint rates. On data used for ML modelling, we will generate multilevel models with patient features as predictors of each outcome, clustering multiple observations corresponding to each visit and multiple visits corresponding to each patient (three levels). Based on prior work, DASA scores and restraint rates may be higher for racialised patients, English not as primary language speakers, and patients brought into hospital by police. Based on previous work, narratives describing these patients may also contain more negative sentiment and evidence of non-compliance, agitation and other behavioural indicators of risk. For models involving continuous DASA scores and narrative note features, we will inspect $R^2$ to determine how much variance in the risk assessment is explained by the combined influence of sociodemographic, clinical and admission predictors.

All quantitative analyses will be carried out in R. We will use the caret package for ML modelling, the fairness package for bias assessments and the lmer package for multilevel modelling.

**Step 2: contextualising biases**

In keeping with traditions of qualitative research, analysis will be conducted continually alongside data collection. Both data collection and analysis activities will be documented in a log of study activities. Researchers will read all of the qualitative data recurrently over the course of the study. Selected data will be used for collaborative analysis with all coinvestigators, including lived experience advisors. KH, GS, ZF and JZ will apply their clinical insights to the data. Interviews and core meetings will be digitally recorded, transcribed and coded using NVivo V.12, a qualitative data analysis software. All transcripts will be coded and anonymised by the qualitative research team. We will draw on Braun and Clark’s reflexive framework for coding qualitative data to identify patterns across our dataset in relation to the research question. These themes will be discussed by the research team, for further discussion and analysis across several monthly data integration meetings.

**Phase 2: knowledge mobilisation**

In phase 2 (September 2023 to December 2023), we will use findings from phase 1 will draw on principles of participatory action research. Knowledge mobilisation will involve the cocreation of an intervention that addresses identified gaps or potential systemic or organisational factors contributing to bias (eg, in a series of focus groups with patients, staff, physicians, data scientists and other relevant stakeholders). This application of research is particularly important given that, despite the many advances in ML in medicine over the past few years, there is still a gap when it comes to translating these findings into clinical practice.

**Participant selection**

**Electronic health records**

We will extract any available data for patients visiting CAMH between January 2016 and May 2022, from the ED and nine inpatient units, excluding Forensics units. Only patients who were admitted to acute care units via the ED will be included, unless patients were transferred to the ED from another hospital (since it is unknown whether patients received any interventions used to manage risk of violence or aggression at the hospital).

**Participant observation**

We will conduct participant observation by shadowing a clinical volunteer with at least 3-month experience during their shift. Clinical volunteers may include nurses, programme assistants, social workers and physicians. Notably, given that the ED is a high-volume environment, we will use a two-staged informed consent process to ensure that all staff have the opportunity to opt out of the study (see figure 2 for a description of this process). The clinical staff that have consented to shadowing will inform their patients of the researcher’s presence and clients will have the opportunity to withdraw consent at that time (or anytime during the shift). Clinicians will also use their clinical judgement and expertise to determine whether a researcher’s presence is appropriate and may ask the researcher to step away from the area. Given the acuity of the care environment, observations will be limited to congregate settings (eg, waiting areas).

**Semistructured interviews**

- **Patients**: Sampling will be purposive to reflect the range of biases identified by our ethnographic observations and quantitative findings. Patients will be recruited via flyers distributed to ambulatory and inpatient units. Interested participants will be contacted by the research team. Anyone with lived experience of the ED within the last 5 years will be eligible for the study.

- **Clinicians**: Any clinical staff with at least 3 months of experience working in the ED within the past year will be eligible for the study. Clinic staff will be invited to participate in an individual interview conducted by a research team member. Participants will be sent an invitation via email, along with a copy of the information and consent form. If no response is received within 2 weeks, one follow-up message will be sent.

- **Key informants**: Knowledge gaps will be supplemented by key informant interviews. This could include hospital administrators, security guards and/or police with direct experience of the ED.

**Sample size**

**Participant observation**

We will shadow up to 20 clinicians who volunteer during their regular work shifts at a time and for a duration that...
is convenient to them. Thousands of individuals may also be incidentally present while observations with a clinical staff member take place. Therefore, there is no way to determine the sample size of those observed in this context. This approach is consistent with other observational studies in hospitals.\(^6\) \(^9\) \(^7\) \(^3\) \(^8\) \(^9\) 

**Semistructured interviews**

Although it is difficult to estimate the exact number of interviews in qualitative research, based on the number of staff that work in the ED, we anticipate conducting approximately 50 semistructured interviews with patients, clinical staff and key informants.

**Electronic health records**

We will access data for 11 000 patients across 20 000 admissions to inpatient units since 2016. Since multiple DASAs and FMTs are recorded during each admission, we anticipate over 400 000 DASAs and FMTs to be available for inclusion. Because we are analysing data from the first 3 days, we expect 60 000 observations to be available for training. This sample size exceeds samples used in other studies of ML-based risk prediction.\(^4\) \(^29\) Importantly, the hospital collects detailed sociodemographic information, which will allow for demographic stratification of the risk assessment data. For example, 6%–12% of patients are brought to the CAMH ED by police or law enforcement, and a substantial portion (15%) of patients do not have stable housing, which will allow us to examine whether model prediction and risk assessment ratings differ based on these and other sociodemographic factors.

---

**Figure 2** Two stage informed consent process. ED, emergency department.
Patient and public involvement statement
The hospital patient engagement team was consulted prior to the formalisation of the research questions and study design. Based on their advice and expertise, we have recruited and hired two patient advisors with lived experience in the hospital’s psychiatric ED. The lived experience advisors will provide guidance throughout the course of the study with a focus on reviewing and interpreting data findings and knowledge mobilisation. Their involvement will be rooted in Canada’s Strategy for Patient-Oriented Research framework (2014), which focuses on inclusion, support, mutual respect and cobuilding to ensure authentic and meaningful patient engagement.91

ETHICS AND DISSEMINATION
This study protocol has been reviewed and approved by the CAMH Research Ethics Board (#053/2021).

Data management and confidentiality
We have minimised the privacy risks associated with accessing the EHRs by processing and analysing this data according to established protocols and safeguards, co-developed with CAMH’s Privacy Department and Research Ethics Board.92 Briefly, these protocols involve hospital data warehouse staff extracting numerical features from the unstructured or text data before it is shared with the study team, since textual data can contain personal and identifiable information. In this way, this sensitive data (in its raw form) never leaves dedicated hospital clinical servers and is never accessed directly by the study team.

For qualitative methods, identifiable data will be gathered to schedule interviews or focus groups, but these will not be linked to data for analysis. All research participants will receive a unique participant ID. The master code list will be kept securely stored on the hospital server and only the research team will have access to it. Interviews and focus groups could potentially include identifiable data. However, no identifiable data will be transcribed, and once analysis is complete, the audio recording will be deleted. Although researchers might overhear personal health information during participant observation, this information will not be recorded, documented or verbally shared.

Patient interviews
Participants may experience distress during the interviews. If in distress, a list of resources will be made available for participants and interviewers will also be able to direct patients to a clinician from the research team, if needed. Researchers will also remind participants that they can stop the interview or skip a question at any time.

Staff perception of risk
Staff may have concerns about the presence of a researcher on the unit conducting participant observation. Researchers will continuously remind staff that all information collected is confidential and will not be shared outside the research team, including with unit management. Additionally, researchers will emphasise to staff that they may opt out of the study at any time.

Dissemination
Results from this study will be prepared for a final internal report, conference presentations and manuscripts for publication in open access and peer-reviewed journals. Findings will also be shared with all participants online or in in-person community forums held at study sites and with providers during regularly scheduled staff meetings. Lived experience advisors will play an active role in the dissemination of study results, including coauthorship on papers, presenting study results within both the scientific and general community. The research team will also create a resource summarising the development of a hybrid computational ethnography to be shared with data scientists, clinicians and social scientists engaged in developing predictive care tools in medicine.

Author affiliations
1The Krembil Centre for Neuroinformatics, The Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
2Anthropology, University of Toronto, Toronto, Ontario, Canada
3The Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
4Information Management Group, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
5Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada
6Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada
7Gerald Sheff and Shantilla Kachan Emergency Department, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
8Office of Education, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
9Dalhousie University, School of Public Health, University of Toronto, Toronto, Ontario, Canada

Twitter Laura Sikstrom @laurasikstrom and Juveria Zaheer @juveriazaheer

Contributors All authors conceptualised and designed the research protocol. LS, MM and ZF drafted the manuscript. SLH, JZ, KH, GS, ZF and DZB revised the manuscript by providing critical domain expertise. All authors critically reviewed, edited and approved the final manuscript.

Funding This work was supported by a Dalla Lana School of Public Health Interdisciplinary Data Science Seed Grant (D2B and SLH), Canadian Institutes of Health Research Health Systems Impact Fellowship (LS and MM), the Social Sciences and Humanities Research Council Insight Development Grant (LS #430-2021-01166) and a Google Award for Inclusion Research (LS and SLH, no award/grant number).

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods and analysis section for further details.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ). BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines,
REFERENCES


5 Panch T, Mattie H, Celai LA. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2:77:77::.


7 Ghassemi M, Nsoesie EO. In medicine, how do we machine learning anything real? Patterns 2022;3:100392.


