JMIR Publications
The Leading eHealth Publisher

Advancing Psychiatric Safety with MAP: a Mental health Adverse events Prediction Tool

Elham Dolatabadi, Valentina Tamayo Velasquez, Abdul Hamid Dabboussi, David Wen, Jennifer Crawford, Andrea Waddell, Christo El Morr

Submitted to: JMIR Mental Health
on: September 29, 2025

Preprint
JMIR Publications

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript **5**
Supplementary Files **39**
Figures **40**
Figure 1 ..... **41**
Figure 2 ..... **42**
Figure 3 ..... **43**
Figure 4 ..... **44**

Preprint JMIR Publications

Advancing Psychiatric Safety with MAP: a Mental health Adverse events Prediction Tool

Elham Dolatabadi[1, 2] BSc, MSc, PhD; Valentina Tamayo Velasquez[3, 4] HBsc; Abdul Hamid Dabboussi[1] BSc, MSc; David Wen[3] PhD; Jennifer Crawford[3, 5] PhD, CHE; Andrea Waddell[3, 6] MEd, MD; Christo El Morr[1, 7] BA, HBsc, MSc, PhD

[1]School of Health Policy & Management Faculty of Health York University Toronto CA
[2] Vector Institute Toronto CA
[3] Waypoint Centre for Mental Health Care Penetanguishene CA
[4]Institute of Medical Science Temerty Faculty Medicine University of Toronto Toronto CA
[5] Faculty of Health Sciences The University of Ontario Institute of Technology Oshawa CA
[6] Temerty Faculty of Medicine The University of Toronto Toronto CA
[7] North York General Hospital Toronto CA

Corresponding Author:
Elham Dolatabadi BSc, MSc, PhD
School of Health Policy & Management
Faculty of Health
York University
340 Stong College
165 Campus Walk
Toronto
CA

Preprint
JMIR Publications

Abstract

Background: Patient safety incidents are a leading cause of harm in psychiatric settings, yet early warning systems tailored to mental health remain underdeveloped. Traditional risk tools like DASA-IV offer limited predictive accuracy and are reactive rather than proactive.

Objective: We introduce the Mental Health Adverse Event Prediction (MAP) tool, a deep learning-based EWS trained on longitudinal psychiatric EMR data to anticipate adverse events in 24-hour windows.

Methods: A retrospective and prospective cohort study using routinely collected electronic medical record (EMR) data to train and validate ML models for short-term risk prediction. This study took place at Waypoint Centre for Mental Healthcare, a large inpatient psychiatric hospital in Ontario, Canada, serving both high-secure forensic and non-forensic patient populations. A total of 4,651 patients and 403,098 encounters from January 2020 to August 2024. For prospective evaluation, the 2024 test set included 900 patients and 48,313 encounters. MAP was trained using recurrent neural networks with attention mechanisms on multivariate time-series data. The model employed an autoregressive design to forecast risk based on 7 days of prior patient data and was benchmarked against the DASA-IV clinical tool and other machine learning baselines. The primary outcome was the occurrence of an adverse mental health event recorded in the EMR within the next 24 hours. Model performance was assessed using area under the receiver operating characteristic curve (AUC) and recall, alongside subgroup analyses and interpretability assessments using integrated gradients.

Results: The LSTM with attention mechanism achieved the highest predictive performance (AUC = 0.83), outperforming existing tools such as the DASA-IV by 20 AUC points (0.81 vs. 0.61)??, and demonstrating the potential of ML-based models to support proactive risk management in mental health settings.

Conclusions: MAP is one of the first prospectively evaluated deep learning-based early warning systems for psychiatric inpatient care. By outperforming existing clinical tools and providing interpretable, rolling predictions, MAP offers a pathway toward safer, more proactive mental health interventions. Future work should assess its equity implications and integration into routine psychiatric workflows.

(JMIR Preprints 29/09/2025:84318)

DOI: https://doi.org/10.2196/preprints.84318

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Preprint JMIR Publications

Original Manuscript

Preprint
JMIR Publications

Advancing Psychiatric Safety with MAP: a Mental health Adverse events Prediction Tool

Elham Dolatabadi, Valentina Tamayo Velasquez[2,3], Abdul Hamid Dabboussi, David Wen[2], Jennifer Crawford[2,5], Andrea Waddell[2,3]*, Christo El Morr[1]*

* Equal senior authorship in contribution

1. York University, Toronto, Ontario, Canada
2. Waypoint Centre for Mental Healthcare, Penetanguishene, Ontario, Canada
3. University of Toronto, Toronto, Ontario, Canada
4. Vector Institute, Toronto, Ontario, Canada
5. The University of Ontario Institute of Technology, Oshawa, Ontario, Canada

Email of co-authors:
Elham Dolatabadi: edolatab@yorku.ca
Valentina Tamayo Velasquez: valentina.tamayo@mail.utoronto.ca
Abdul Hamid Dabboussi: abdulhamid.dabboussi@gmail.com
David Wen: dwen@waypointcentre.ca
Jennifer Crawford: jcrawford@waypointcentre.ca
Andrea Waddell: awaddell@waypointcentre.ca
Christo El Morr: elmorr@yorku.ca

Corresponding Author:
Elham Dolatabadi
315 Stong College - SC Keele Campus, Toronto, Ontario, Canada M3J1P3
edolatab@yorku.ca

Manuscript word count:
3,000 words

Preprint JMIR Publications

Abstract

Background: Patient safety incidents are a leading cause of harm in psychiatric settings, yet early warning systems tailored to mental health remain underdeveloped. Traditional risk tools like DASA-IV offer limited predictive accuracy and are reactive rather than proactive.

Objective: We introduce the Mental Health Adverse Event Prediction (MAP) tool, a deep learning-based EWS trained on longitudinal psychiatric EMR data to anticipate adverse events in 24-hour windows.

Methods: A retrospective and prospective cohort study using routinely collected electronic medical record (EMR) data to train and validate ML models for short-term risk prediction. This study took place at Waypoint Centre for Mental Healthcare, a large inpatient psychiatric hospital in Ontario, Canada, serving both high-secure forensic and non-forensic patient populations. A total of 4,651 patients and 403,098 encounters from January 2020 to August 2024. For prospective evaluation, the 2024 test set included 900 patients and 48,313 encounters. MAP was trained using recurrent neural networks with attention mechanisms on multivariate time-series data. The model employed an autoregressive design to forecast risk based on 7 days of prior patient data and was benchmarked against the DASA-IV clinical tool and other machine learning baselines. The primary outcome was the occurrence of an adverse mental health event recorded in the EMR within the next 24 hours. Model performance was assessed using area under the receiver operating characteristic curve (AUC) and recall, alongside subgroup analyses and interpretability assessments using integrated gradients.

Results: The LSTM with attention mechanism achieved the highest predictive performance (AUC = 0.83), outperforming existing tools such as the DASA-IV by 20 AUC points (0.81 vs. 0.61), and demonstrating the potential of ML-based models to support proactive risk management in mental health settings.

Conclusions: MAP is one of the first prospectively evaluated deep learning-based early warning systems for psychiatric inpatient care. By outperforming existing clinical tools and providing

interpretable, rolling predictions, MAP offers a pathway toward safer, more proactive mental health interventions. Future work should assess its equity implications and integration into routine psychiatric workflows.

Keywords:
Early warning system, psychiatric adverse events, adverse event prediction, machine learning, clinical decision support, patient safety.

Preprint JMIR Publications

Introduction

Patient and staff safety are top priorities in healthcare, yet patient safety incidents remain the third leading cause of death in Canada.[1] Many of these incidents stem from adverse events such as falls, medication errors, and medical complications.[2] A recent study found that one in four hospital admissions involved adverse events, with a quarter of these deemed preventable.[3] While all healthcare settings face safety risks, psychiatric environments present a distinct set of challenges, including suicide, restraint, and seclusion - events which contribute to continued suffering and injury.[4,5] Despite a higher prevalence of adverse events in mental health, research on patient safety in these settings remains limited compared to other medical fields.[6,7]

These incidents not only worsen patient outcomes but also increase risks for staff.[8] Globally, about 24% of healthcare workers experience physical violence annually, with psychiatric staff at particularly high risk.[9–11] Reducing adverse events through assessment and prediction is crucial for improving staff and patient safety.

Current methods for assessing patient deterioration rely heavily on voluntary reporting,

critical incident reviews, and clinician judgment.[12] While actuarial tools such as the Dynamic Appraisal of Situational Aggression (DASA) and the Brøset Violence Checklist are also used, they have shown limited predictive accuracy and tend to miss early warning signs.[13,14] As a result, many opportunities for timely intervention are lost, especially in high-risk but low-observable cases.

Early warning systems (EWS) are widely used in medicine, leveraging routinely collected clinical data to detect early signs of patient deterioration. Tools such as the National Early Warning Score 2 (NEWS2) have been effectively implemented in acute care settings to support timely interventions.[15–17] At the same time, machine learning (ML) is transforming risk assessment by enabling the analysis of large-scale, high-dimensional healthcare data.[18–20] Predictive ML models are developed using historical patient records combined with expert input to train, test, and refine algorithms for higher performance and clinical relevance.[21,22] Compelling examples include CHARTwatch, developed to predict inpatient deterioration in General Internal Medicine, and Sepsis Watch, designed to identify patients at risk of sepsis before clinical recognition.[23–25] Yet, psychiatric care has not seen comparable innovation, in part due to the complexity of mental health data, lack of validated digital tools, and underrepresentation of psychiatric settings in EWS research.

To address this gap, we introduce a novel ML-based EWS, Mental Health Adverse Event Prediction (MAP) tool. The MAP tool is a deep learning-based early warning system, leveraging longitudinal electronic medical record (EMR) data from a specialized psychiatric hospital. MAP is trained to predict the likelihood of adverse events within 24-hour windows using autoregressive recurrent neural networks, enriched with attention mechanisms and interpretability via integrated gradients. Unlike traditional tools, MAP is capable of continuous, real-time risk forecasting, even in the absence of prior incidents. We benchmarked MAP against DASA and other ML models, demonstrating superior performance, particularly in complex and high-risk subgroups.

Through this study, we aim to move beyond reactive safety practices toward proactive, data-informed risk mitigation in mental health care, advancing both patient and staff safety in a setting

long underserved by digital innovation.

Methods

Study Design and Data Acquisition

In this study, we used routinely collected clinical data extracted from the EMR at Waypoint Centre for Mental Healthcare (Waypoint), Ontario, Canada. We retrospectively retrieved data from all patients at Waypoint between January 2020 and August 2024, including static and dynamic variables (s-Table 1).

Ethics Statement

This study was approved by the York University Office of Research Ethics (Certificate #: e2023-163), and the Research Ethics Board (REB) of Waypoint Centre for Mental Health Care (Reference #RCRA#23.08.01) with waived informed consent.

Data representation and processing

The baseline data preprocessing included one-hot encoding and normalization of all measures. We implemented a standardized aggregation strategy to address the variability in time-series data arising from differing measurement frequencies, where some clinical parameters were recorded daily and others multiple times per day. Patient encounters were segmented into 24-hour intervals, aligning with clinical workflows that typically operate on daily cycles for alerts. Within each interval, all measures were aggregated to provide a comprehensive snapshot of patient health over the specified time frame. Numerical variables were averaged across the interval, while categorical variables were first encoded numerically based on severity or clinical importance and then summed within the 24-hour period.

We collected admission diagnosis data, based on the DSM-5, selecting the 45 most frequent diagnoses across all patients to prevent overfitting. Medication data included our patient group’s five

most relevant categories, each represented as a binary indicator, denoting whether or not it was administered within the past 24 hours (s-Table 1).

The primary outcome in our study is the occurrence of any mental health adverse event. For our prediction task, each patient encounter was labelled based on whether a logged adverse event occurred in the EMR system within the next 24-hour bin. This binary label (event vs. no event) was used as the target variable for MAP training. Moreover, prior adverse events in the previous 24-hour intervals were also dynamically added to future intervals, referred to as the history of any incident.

Building the MAP Model

To improve psychiatric-medical baselines, we designed a deep learning-based early warning system. Specifically, we developed recurrent neural networks employing the Long Short-Term Memory (LSTM) model that triggers an alert every 24 hours, based on a variable sequence length (3-7 days) of patient data, treated as a hyperparameter. During training, the ground-truth history of any adverse event was provided for every 24-hour interval in the model. In the inference phase, the model operated in an autoregressive mode: it used its own predicted output for the previous 24-hour window as an input signal for the next prediction step. To enhance the model's ability to focus on the most relevant temporal signals within the input sequence, we further explored an LSTM model with attention mechanisms (LSTM+Attn). In this variant, an attention layer was added to the LSTM hidden states. For each 24-hour prediction interval, the attention mechanism dynamically assigned weights to each timestep in the historical input sequence, allowing the model to selectively focus on the most informative data that contributed to the risk signal.

We also evaluated our model against two ML approaches: Light Gradient-Boosting Machine (LightGBM) and Fully Connected Neural Network (FCNN). All time-series features were aggregated over 3-7 days using the same methodology applied to 24-hour intervals. Predictions were then made for the next 24-hour interval, allowing for consistent model evaluation and direct comparison across

different sequence lengths to identify the best-performing approach

The dataset was divided and we created two distinct test sets, (1) a held-out-patient test set, with no patient overlap between the development (3,000 patients) and the test sets (751 patients) and (2) an out-of-time test set, split across time, using 2020-2022 data for training and 2023 data for testing (Table 1). Once the model selection was finalized, we used all data from 2020-2023 to train MAP and 2024 to prospectively test and evaluate MAP's real-world applicability. To explore the factors driving MAP's predictions, we employed integrated gradients to compute feature importance in our LSTM model by computing the path-integrated gradients from the input to the actual input.[26]

Comparison with clinical measures

We compared MAP's predictive performance against DASA-IV, a standardized tool used in our hospital for evaluating risks of aggression.[27] The DASA-IV includes seven items assessing behavioral indicators (i.e., irritability, negative attitudes, verbal threats, etc.), each scored as 0 (not observed) or 1 (observed), with a total scores categorized as low (0-1), moderate (2-3), or high (>3).[28,29] To align DASA-IV with MAP's binary classification, we restructured the risk categories. Moderate and high risk were grouped as 'at risk' (positive prediction), while low risk as 'no risk' (negative prediction). While MAP is designed to predict a broader range of mental health-specific adverse events, this allowed us to compare DASA-IV's performance against MAP's predictions and the ground truth outcomes recorded in patient encounters.

Results

Cohort characteristics

The dataset encompassed 4,651 patients and 403,098 patient encounters over 55 months. The demographic details distribution of the patient cohort is presented in Table 1. For the prospective evaluation of the best performing ML model and comparison against clinical baselines, we used data

from 2024, with detailed breakdowns provided in Table 1.

Adverse event distribution per individual varied across the sample. Figure 1 shows the number of patients grouped by the frequency of adverse events they experienced during their hospital stay. The majority of patients experienced few or no adverse events: 3251 patients had no incidents, and 587 experienced up to 2 incidents. A total of 344 patients had between 3 and 16 events, with a median of 14.5 in this range. A smaller group of 162 patients experienced between 17 and 83 incidents, most of whom (42) had between 17 and 20, while only 22 had more than 83 events. As incident frequency increased, cohort size decreased. The mean number of adverse events across the sample was 2.85, while the mode and median were both 0, highlighting the skewed nature of the data. This imbalance is important to consider, as it affects how the model learns from the hospital's patient population, with most of the training data representing patients with few or no incidents.

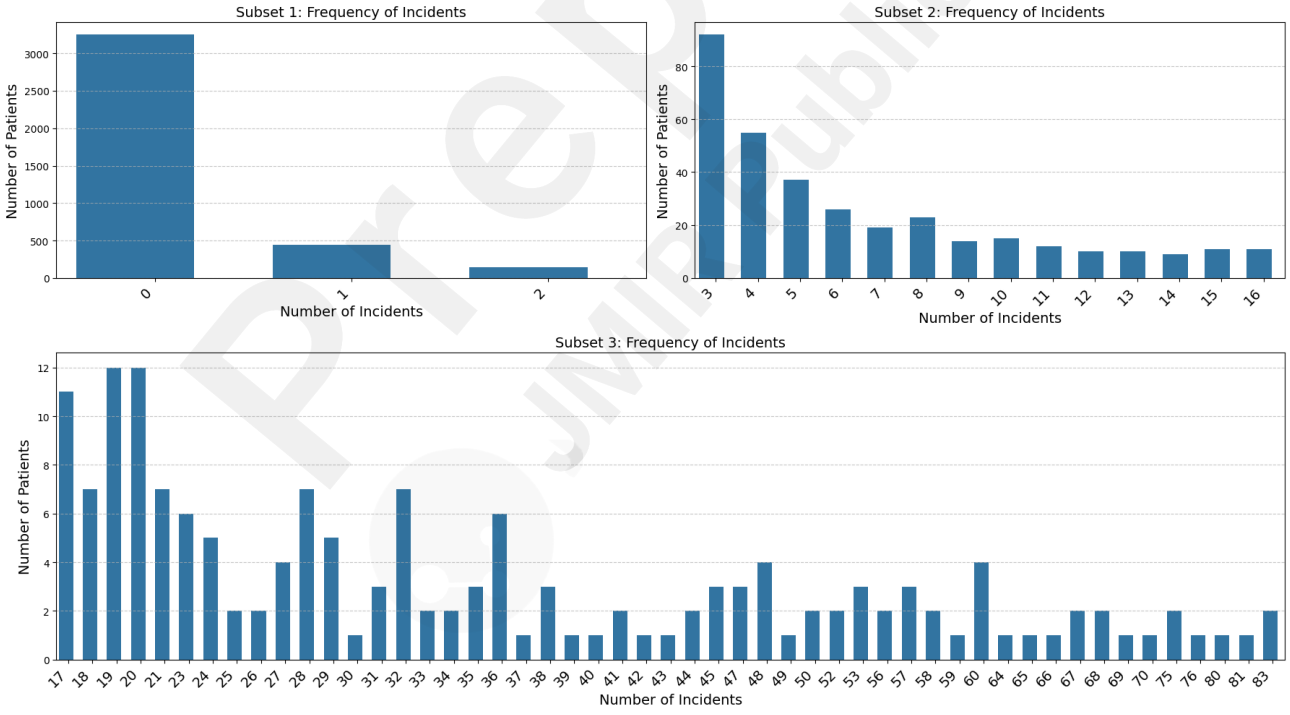


Figure 1. The distribution of the total adverse events across the sample of patients between 2020-2024. On the x-axis is the specific number of adverse events, and on the y-axis is the number of individuals who experienced that many adverse events between the 2020-2024 time period. The different subsets represent ranges of incidents. Subset 1: Cohorts of individuals with 0-2 incidents; Subset 2: 3-16 incidents; Subset 3: 17-83 incidents.

MAP's Prediction

Table 2 presents the performance comparison of the four ML models: GB, FFNN, LSTM, and LSTM+Attn. Each model was trained multiple times using different random seeds, and performance metrics were averaged across runs to ensure robustness. Given the imbalanced nature of the dataset, model performance was evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC) and recall. The LSTM+Attn consistently achieved the highest performance, with an AUC of 0.87 for held-out patients and 0.72 for out-of-time patients. We selected the LSTM+Attn as the final architecture for MAP (Table 2).

Preprint JMIR Publications

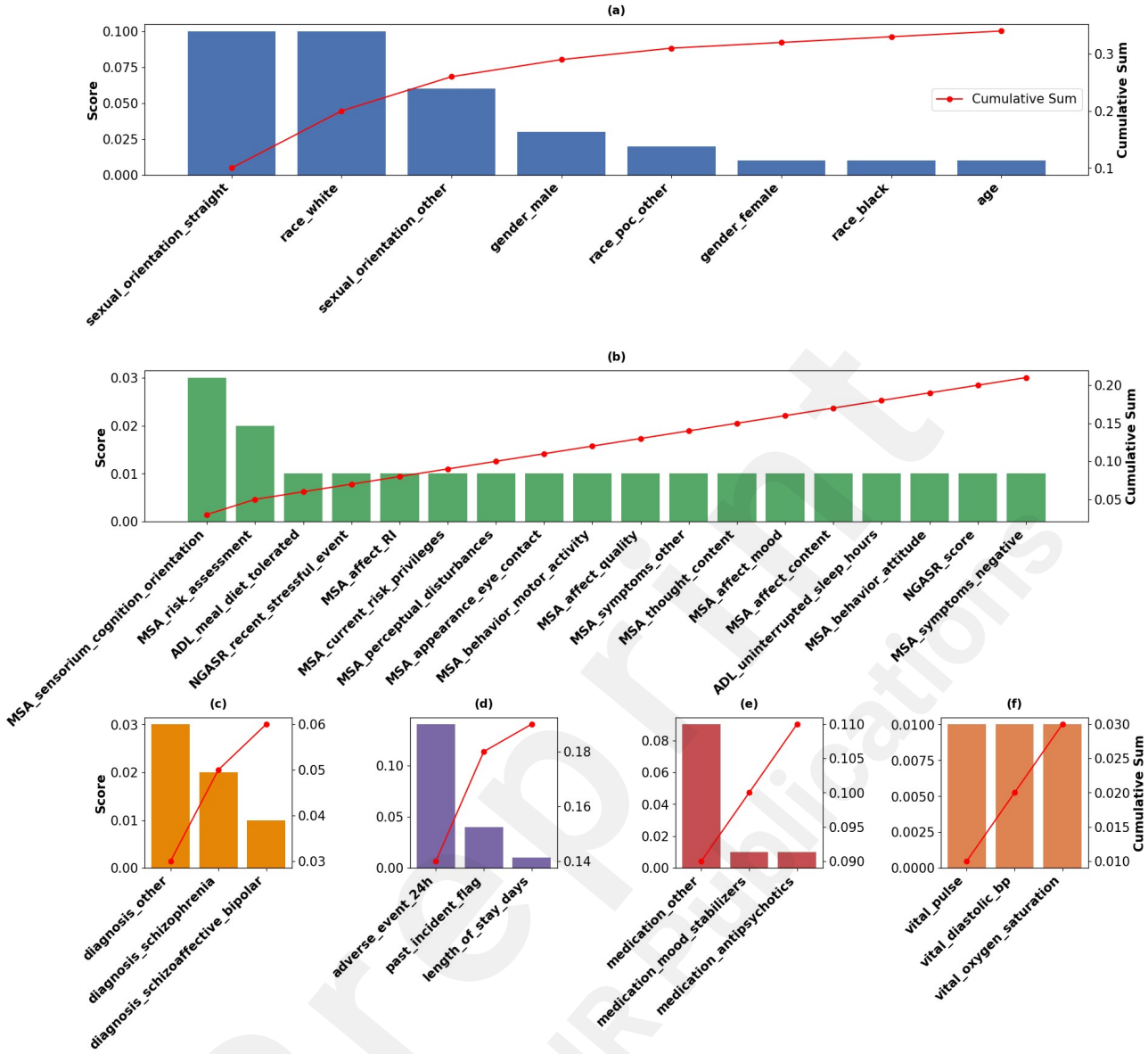


Figure 2. Feature importance across distinct categories in the MAP model: (a) Overall feature importance highlighting the top predictors; (b) Demographic features including gender, race, and sexual orientation; (c) Clinical assessment features such as mental status indicators and functional assessments; (d) Clinical diagnoses including major psychiatric conditions; (e) Clinical variables related to adverse events, incident history, and hospital stay duration; (f) Medication-related features, including mood stabilizers and antipsychotics; and (g) Vital signs, including pulse, blood pressure, and oxygen saturation.

In the prospective evaluation using the dataset from 2024, with 48,313 encounters, 2,106 recorded adverse events, MAP achieved an AUC of 83% (Table 1). The performance varied within

and across subgroups, with AUC ranging from 0.69 (Black patients) to 0.87 (other gender identities), indicating potential biases favoring larger, more represented groups within the sample data. Across racial subgroups, AUC differed by 14%; across gender subgroups, AUC varied by 5%; across sexual orientation subgroups, AUC varied by 2%; across program types, AUC differed by 4%; and across age groups, AUC variation was minimal (<1%).

Among the 40 features included in MAP, the top 14 predictors collectively account for 80% of the model's importance, highlighting a diverse mix of demographic, medical, and psychosocial factors that drive prediction. The most influential feature overall is "adverse events in the past 24 hours," as shown in Figure 2(a). When further analyzing feature contributions within specific categories, demographic factors (Figure 2(b)), such as Straight and other sexual orientations (Sexual Orientation), White (Race), and Male (Gender), indicate the highest importance scores. From the clinical assessments (Figure 2(c)), 15 Mental Status Assessment indicators, along with 'ADL_meal_diet_tolerated' and 'sleep_uninterrupted', are identified as important contributors. In the category of clinical diagnoses (Figure 2(d)), Schizophrenia and Schizoaffective/Bipolar Disorder emerge as the most significant predictors. Among clinical variables (Figure 2(e)), the occurrence of an adverse event in the last 24 hours, the history of a past incident, and the cumulative length of stay in the hospital are the most important. In the medications category (Figure 2(f)), mood stabilizers and antipsychotics contribute meaningfully, though other medications demonstrate even greater overall impact. Finally, among vital signs (Figure 2(g)), features such as pulse, blood pressure, and oxygen saturation meet the 0.90 cumulative importance threshold, although their influence remains relatively modest.

Comparison with the standardized risk assessment tool DASA

MAP demonstrated a 0.2 AUC improvement over the DASA assessment tool when evaluated on the prospective 2024 dataset, with MAP achieving an AUC of 0.81 compared to DASA's AUC of 0.61

(Supplementary Table 2). To further assess MAP's performance across different patient groups, we analyzed its effectiveness based on the historical incidence of adverse events for each individual in the training dataset (previous adverse event history). Figure 3 illustrates the performance differences between MAP and DASA across various patient groups, where each group is defined by the number of adverse events recorded in both the training (past) and evaluation (future) datasets. We categorized individuals into 63 subgroups based on different combinations of past and future incident counts.

The MAP tool significantly outperformed DASA in 40 of 63 subgroups (Wilcoxon test, p-value = 0.007). For individuals with no prior incidents in the training set but up to 58 total incidents in the evaluation period (Figure 3 (a)), MAP achieved an AUC of 0.62, while DASA achieved an AUC of 0.50. For individuals with up to 10 incidents in the past and up to 23 in the future (Figure 3 (b)), DASA outperformed MAP in cases where individuals had 1, 2, or 5 future incidents. However, MAP outperformed DASA in the remaining four subcategories within this range. For individuals with moderate incident frequency (11-99 past incidents), MAP outperformed DASA in 11 of 19 groups (Figure 3 (c)). Among individuals with frequent incidents (>100 incidents past incidents), MAP outperformed DASA in 4 of 5 subgroups (Figure 3 (d)). Notably, MAP's performance is better in edge cases where individuals had a high number of past incidents but only one in the future.

Preprint JMIR Publications

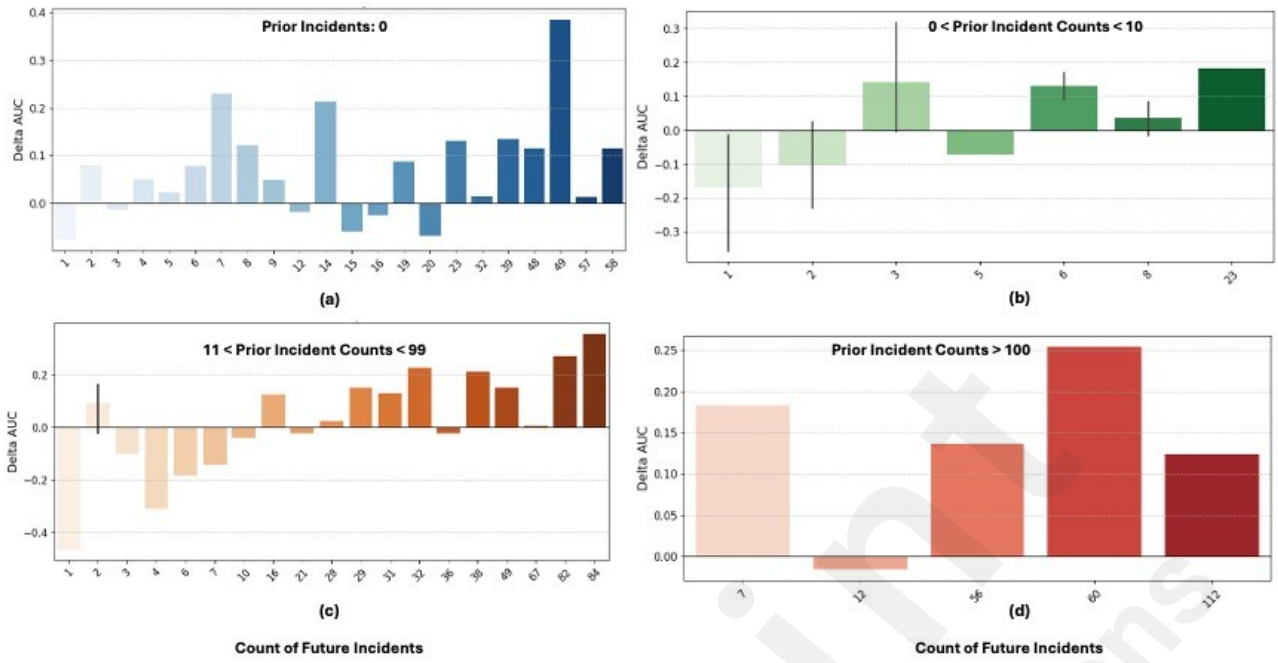


Figure 3. Difference in AUC ROC performance scores (Delta AUC = ML AUC ROC - DASA AUC ROC). A positive delta AUC indicates the ML model outperformed DASA for that specific cohort group. A negative delta AUC indicates DASA outperformed ML for that specific cohort group.

Discussion

Despite the growing number of adverse events in mental health settings, deep learning (ML) tools that leverage routinely collected electronic medical record (EMR) data to predict patient deterioration remain limited. Our model, MAP, represents a first-of-its-kind approach tailored specifically to psychiatry and demonstrates strong predictive performance, achieving an AUC-ROC of 0.83. Leveraging autoregressive LSTM with attention mechanisms, MAP operates in a rolling prediction mode, enabling 24-hour forecasts even in the absence of recent incident data. Notably, the history of prior incidents emerged as one of the most informative features, reinforcing the predictive value of temporal continuity in patient risk trajectories. Furthermore, the inclusion of patients from both forensic and non-forensic acute care programs contributes to the model's generalizability across diverse mental health populations.

Currently, no ML-based predictive alerting tools are deployed in mental health settings. Instead, clinicians rely on actuarial tools such as the DASA-IV to assess risks related to violence and aggression.[30] On the same dataset, MAP outperformed the DASA-IV (AUC: 0.83 vs. 0.61). While DASA has reported AUCs between 0.61-0.82 in other studies, it is important to note that MAP and DASA target different outcomes.[31] MAP captures a broader spectrum of deterioration events, including suicide, self-harm, and clinical decompensation, whereas DASA is limited to aggression-related outcomes. The lower AUC for DASA in our dataset likely reflects these differences in scope. Nonetheless, MAP's ability to deliver significantly stronger performance across a wider range of adverse events underscores its versatility and robustness. In clinical practice, focusing solely on aggression is insufficient; risks of suicide and self-harm are equally critical. By encompassing a more comprehensive set of risks, MAP provides clinicians with a holistic and actionable risk assessment framework, supporting earlier and more effective interventions. MAP also showed strong performance even in patients with no prior recorded incidents, addressing a critical limitation of traditional tools that rely heavily on observable behavior or clinician judgment.

The feature "adverse event in the past 24 hours" emerged as a strong predictor of future deterioration, consistent with findings from acute care settings where recent clinical instability is a key driver of risk. Similar patterns have been observed in inpatient deterioration models, where temporal proximity to prior events significantly enhances predictive accuracy.[32,33] Beyond clinical history, our results indicate that a wide array of features, including demographic variables, mental status assessments, clinical diagnoses, medications, and vital signs, contribute meaningfully to risk prediction. This multidimensional pattern aligns with emerging work suggesting that accurate prediction of psychiatric outcomes requires integrating different types of structured medical data and psychosocial factors.[34–36] Overall, these findings underscore the importance of using holistic patient representations to capture the complex drivers of risk in mental health, a direction that has been underexplored in existing ML applications in psychiatry.

A limitation of this study, as previously noted, is the underrepresentation of certain demographic subgroups, which affected the model's predictive performance. We observed up to an 18% variation in AUC across subpopulations, indicating disparities in performance. Notably, the model was less accurate for two racial subgroups: Black and First Nations individuals, with AUC scores 14% and 3% lower, respectively, than the overall model performance. Additionally, both groups had a recall of 0.16, which was lower than that of all other subgroups, suggesting a higher rate of false negatives and an undercalling of risk. These disparities likely stem from the low representation of these groups in the dataset, Black individuals comprising less than 10% of the sample, and First Nations individuals less than 3%.

Ongoing monitoring and evaluation of MAP are needed to assess its real-world performance and potential biases. Future work will focus on mitigating these biases through bias-aware data augmentation and fairness-aware learning algorithms (e.g., adversarial debiasing) to improve representation across subgroups.[37,38,39] Piloting the MAP tool in a live clinical setting is the next step in validating its performance and efficacy and informing the next steps toward broader clinical deployment.

In this study, we developed and evaluated an LSTM model that could predict patients at risk for an adverse event. The model showed good performance across different subgroup populations, and our findings suggest that the model would outperform currently used risk assessment tools. Its autoregressive design, prospective evaluation, and near real-time operation position it for real-world clinical integration. By generating dynamic forecasts without dependence on manual clinician input, MAP can augment existing workflows and support earlier interventions in settings where mental health staff face high demands and elevated safety risks.

Declarations

Acknowledgements

This work was made possible through access to data provided by Waypoint Centre for Mental Health Care, and we gratefully acknowledge the support and collaboration of the Waypoint research team. We also extend our appreciation to York University and its dynamic research community at the intersection of health, data science, and machine learning.

Funding

This work was funded by the Healthcare Insurance Reciprocal of Canada (HIROC). E.D. 's research is supported by the Canadian Institutes of Health Research (CIHR) Pandemic Preparedness and Health Emergencies Research Initiative and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant.

Conflicts of Interest

None declared.

Data Availability

The datasets generated and analysed during the current study are not publicly available due to privacy policies and ethical restrictions but are available from the corresponding author on reasonable request.

Authors' contributions

ED, AW, & CEM contributed to the conceptualization and methodology of the study. ED, CEM, JC, AHD, and DW contributed to the coding and the model development. VTV and ED conducted the data analysis and the writing of the manuscript. All authors reviewed and approved the final version of the manuscript.

Abbreviations

- Artificial intelligence (AI)
- Area under the curve (AUC) Receiver-operating characteristic curve (ROC) (AUC ROC)
- The Bröset Violence Checklist (BVC)
- Dynamic Appraisal of Situational Aggression (DASA)
- Electronic medical record (EMR)
- Early Warning System (EWS)
- Feed-forward neural network (FFNN)
- Gradient boosting (GB)
- General internal medicine (GIM)
- Intensive care unit (ICU)
- Long-short-term memory (LSTM)
- Machine learning (ML)
- The National Early Warning Score 2 (NEWS2)
- Research ethics board (REB)

Preprint © JMIR Publications

Ethics approval and consent to participate

This study was approved by the York University Office of Research Ethics (Certificate #: e2023-163), and the Research Ethics Board (REB) of Waypoint Centre for Mental Health Care (Reference #RCRA#23.08.01) with waived informed consent.

Supplementary Materials

S-Table 1. List of baseline and temporal features used for the MAP tool.

Included Variables	Variable (Location of data element in EMR)	Static/ Dynamic Measure	Type of Variable
Demographics			
Inpatient Admission Assessment	Sexual Orientation: Straight, Other sexual orientations	Static	Binary (TRUE or FALSE)
	Gender: Male, Female, Other gender identity		Binary (TRUE or FALSE)
	Race: White, Black, People of colour, Indigenous		Binary (TRUE or FALSE)

Preprint JMIR Publications

	Age		Discrete (Numerical)
Clinical			
Medication (Administration)	Generic mood stabilizers	Dynamic	Binary (0 or 1)
	Generic antipsychotics		
	Generic antidepressants		
	Generic anxiolytics		
	Other		
Vital Signs	Systolic blood pressure	Dynamic	Continuous (Numerical)
	Diastolic blood pressure		
	Pulse		Continuous (Numerical)
	Temperature in Celsius		Continuous (Numerical)
	Respiration rate		Continuous (Numerical)
	Oxygen saturation level		Continuous (Numerical)
Length of stay	Length of stay (days)	Dynamic (input in model development from dataset was static)	Discrete (Numerical)
Primary Diagnosis	Schizophrenia	Dynamic	Categorical

Preprint JMIR Publications

(Top 45)	Autism spectrum disorder		
	Intellectual developmental disorde (intellectual disability), Severe		
	Schizoaffective disorder, Bipolar type		
	Intellectual developmental disorder (intellectual disability), Moderate		
	Adjustment disorders, with mixed disturbance of emotions and conduct		
	Intellectual developmental disorder (intellectual disability), Mild		
	Major neurocognitive disorder due to multiple etiologies, Moderate, with agitation (code first the other medical etiologies)		
	Major neurocognitive disorder due to multiple etiologies, Severe, with agitation (code first the other medical etiologies)		
	Bipolar I disorder, Current or most recent episode depressed, in full remission		
	Major neurocognitive disorder due to multiple etiologies, with behavioral disturbance		
	Bipolar I disorder, Current or most recent episode unspecified		

Preprint JMIR Publications

	Bipolar I disorder, Current or most recent episode manic, Moderate		
	Unspecified schizophrenia spectrum and other psychotic disorder		
	Schizoaffective disorder, Depressive type		
	Major neurocognitive disorder due to Huntington's disease, with behavioral disturbance (code first G10 Huntington's disease)		
	Bipolar I disorder, Current or most recent episode depressed, in partial remission		
	Alcohol use disorder, Severe		
	Bipolar I disorder, Current or most recent episode manic, with psychotic features		
	Bipolar I disorder, Current or most recent episode depressed, Moderate		
	Major neurocognitive disorder probably due to vascular disease, Moderate, with agitation (no additional medical code)		
	Borderline personality disorder		
	Other specified schizophrenia spectrum and other psychotic disorder		

Preprint JMIR Publications

	Alcohol use disorder, Moderate		
	Posttraumatic stress disorder		
	Antisocial personality disorder		
	Major neurocognitive disorder due to probable frontotemporal degeneration, with behavioral disturbance (code first G31.09 frontotemporal degeneration)		
	Major neurocognitive disorder due to probable Alzheimer's disease, with behavioral disturbance (code first G30.9 Alzheimer's disease)		
	Cannabis-induced psychotic disorder, with moderate or severe use disorder		
	Major neurocognitive disorder due to another medical condition, with behavioral disturbance		
	Persistent depressive disorder		
	Major neurocognitive disorder due to possible Alzheimer's disease, Severe, with agitation (no additional medical code)		
	Major depressive disorder, Recurrent episode, Moderate		
	Major neurocognitive disorder due to prion disease, Unspecified severity, Without accompanying behavioral or psychological disturbance (code first A81.9		

Preprint JMIR Publications

	prion disease)		
	Major depressive disorder, Recurrent episode, with psychotic features		
	Adjustment disorders, with depressed mood		
	Major neurocognitive disorder due to multiple etiologies, Unspecified severity, Without accompanying behavioral or psychological disturbance (code first the other medical etiologies)		
	Bipolar I disorder, Current or most recent episode hypomanic, Unspecified		
	Major neurocognitive disorder possibly due to vascular disease, Moderate, with other behavioral or psychological disturbance (no additional medical code)		
	Major depressive disorder, Recurrent episode, Severe		
	Major neurocognitive disorder due to another medical condition, Moderate, with agitation (code first the other medical condition)		
	Bipolar I disorder, Current or most recent episode manic, Severe		
	Major neurocognitive disorder due to probable frontotemporal degeneration, Unspecified severity, Without accompanying		

Preprint ©JMIR Publications

	behavioral or psychological disturbance (code first G31.09 frontotemporal degeneration)		
	Adjustment disorders, with mixed anxiety and depressed mood		
	Cocaine use disorder, Moderate		
Assessments			
Mental Status Assessment (MSA)	Dress (Appearance)	Dynamic	Discrete (Numerical)
	Eye contact (Appearance)		Discrete (Numerical)
	Posture (Appearance)		Discrete (Numerical)
	Grooming (Appearance)		Discrete (Numerical)
	Orientation (Sensorium and cognition)		Discrete (Numerical)
	Alertness (Sensorium and cognition)		Discrete (Numerical)
	Attitude during interaction (Behaviour)		Discrete (Numerical)
	Motor activity (Behaviour)		Discrete (Numerical)
	Affect congruent to content (Affect)		Discrete (Numerical)

Preprint JMIR Publications

	Affect congruent to mood (Affect)		Discrete (Numerical)
	Affect quality (Affect)		Discrete (Numerical)
	Affect range and intensity (Affect)		Discrete (Numerical)
	Current risks (Current risks)		Discrete (Numerical)
	Outcome privileges (Current risks)		Discrete (Numerical)
	Negative symptoms (Symptoms checklist)		Discrete (Numerical)
	Other mood disturbance indications (Symptoms checklist)		Discrete (Numerical)
	Thought content (Thought content)		Discrete (Numerical)
	Perceptual disturbances (Perceptual disturbance)		Discrete (Numerical)
Activities of Daily Living (ADL)	Meal percent consumed (Meal)	Dynamic	Continuous (Numerical)
	Meal tolerance (Meal)		Discrete (Numerical)
	Hours of uninterrupted sleep (Sleep)		Discrete (Numerical)
Nurses' Global Assessment of	Evidence of a plan to commit suicide	Dynamic (*does not change daily)	Discrete (Numerical)

Preprint JMIR Publications

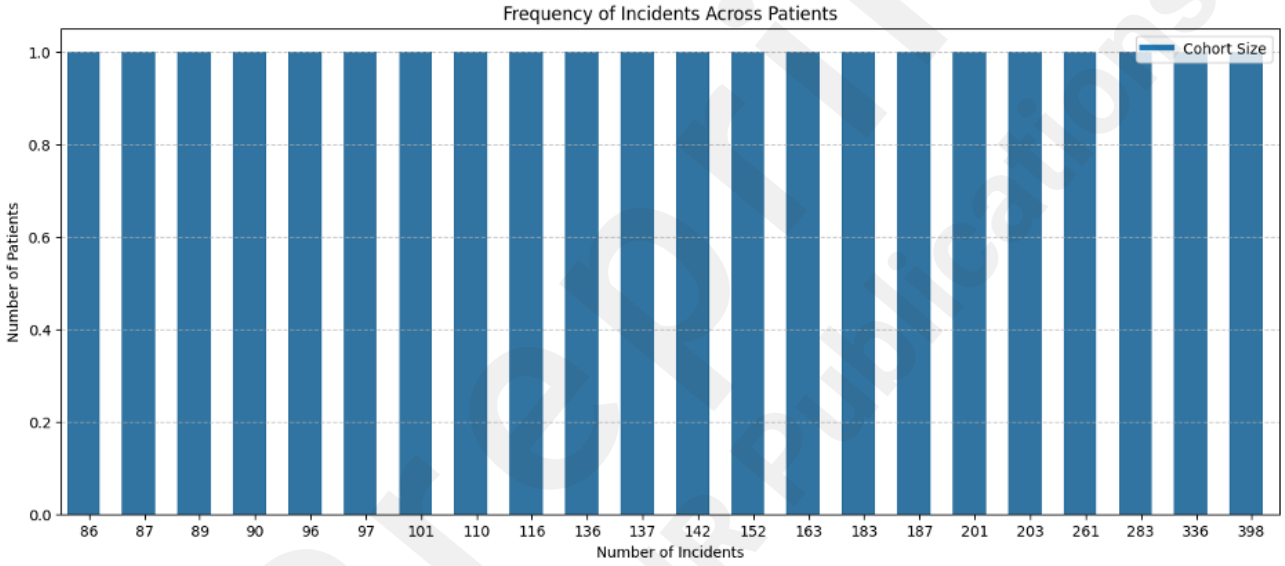
Suicidal Risk (NGASR)	Warning of suicidal intent		Discrete (Numerical)
	Recent stressful life events – e.g., job loss, financial worries, pending court action, ect.		Discrete (Numerical)
	NGASR Score		Continuous (Numerical)
Adverse event log	• Aggressive behaviour • Self harm • Elopement • Behaviour R/T Index off • Inappropriate behaviour • Breach of disposition • Violence towards others • Substance use • Contraband possession • Environmental • Weapons • Threatening behaviour • Choking • Fall • External • Adverse drug reaction • Medical device incident		
24-Hour-Log	Incident log (Whether a patient has had an adverse event/ incident or not in the past 24 hours)	Dynamic	Binary (0 or 1)

Preprint JMIR Publications

Historic Incident	Patient's entire LOS history of an incident (Sum of incidents log)	Dynamic	Discrete (Numerical)

S-Table 2. Clinical baseline model: DASA's Predictive Performances across different data splits (AUCs)

Data Timelines	**DASA Performance (AUC)**
2020-2024	0.6136
2020-2023	0.6165
2024	0.5934



S-Figure 1. The distribution of the total adverse events across the sample of patients between 2020-2024. On the x-axis is the specific number of adverse events and on the y-axis is the number of individuals that experienced that many adverse events between 2020-2024 time period. This graph represents individuals with the range of 86-398 incidents.

Model Training and Hyperparameters

We trained two variants of our model: a baseline LSTM and an LSTM enhanced with attention (LSTM+Attn). The input sequences were of length 4, with a batch size of 128 for all experiments. Both models were trained for a maximum of 100 epochs using the Adam optimizer and an initial learning rate of 0.0001. Learning rate scheduling was applied using a ReduceLROnPlateau strategy with a patience of 3 epochs and a reduction factor of 0.1. Early stopping was also used with a patience of 5 epochs, based on validation loss. The classification threshold was set at 0.5 throughout.

For the LSTM, we used a single-layer configuration with a hidden size of 4 and a dropout probability of 0.4. The LSTM+Attn model shared the same architecture but included an additional attention mechanism with 4 attention heads.

S-Table 3 summarizes the hyperparameter configurations used for each model. Model

training and evaluation were conducted using a consistent data split across training, validation, and test sets.

S-Table 3. Hyperparameters for the LSTM model and the LSTM+Attn model. A comparison of the hyperparameter configurations for the two top performing models evaluated.

Hyperparameters	LSTM	LSTM+Attn
Hidden Size	4	4
Number of Heads	N/A	4
Number of Layers	1	1
Drop Out Probability	0.4	0.4
Initial Learning Rate	0.0001	0.0001
ReduceLROnPlateau Patience	3	3
ReduceLROnPlateau Factor	0.1	0.1
Number of Epochs	100	100
Default Threshold	0.5	0.5
Early Stopping Patience	5	5

Table S-4 presents the performance metrics for the LSTM model over 18 epochs. Notably, the model's AUC-ROC and F1 score steadily improved until early stopping was triggered at epoch 19. The AUC-ROC reached a peak of 0.9292, with an F1 score of 0.4611, suggesting robust discriminatory ability despite the compact model architecture.

S-Table 4. Epoch Training Progress. A summary of the model's performance metrics across 18 epochs. Early stopping triggered at epoch 19.

Epoch Number	Training Loss	Validation Loss	Recall	AUC-ROC	F1 Score
1/100	0.4777	0.1586	0.3156	0.5186	0.0745
2/100	0.2034	0.1398	0.0270	0.6394	0.0332
3/100	0.1823	0.1356	0.0258	0.7262	0.0321
4/100	0.1689	0.1336	0.0071	0.7720	0.0124

5/100	0.1565	0.1322	0.0018	0.8063	0.0036
6/100	0.1453	0.1295	0.0038	0.8347	0.0076
7/100	0.1329	0.1274	0.0830	0.8649	0.1496
8/100	0.1210	0.1318	0.2269	0.8894	0.3455
9/100	0.1122	0.1286	0.2968	0.9066	0.4164
10/100	0.1067	0.1287	0.3335	0.9167	0.4474
11/100	0.1030	0.1297	0.3470	0.9235	0.4579
12/100	0.1012	0.1291	0.3516	0.9265	0.4613
13/100	0.1008	0.1291	0.3500	0.9271	0.4586
14/100	0.1005	0.1290	0.3497	0.9279	0.4596
15/100	0.0999	0.1291	0.3514	0.9290	0.4608
16/100	0.1000	0.1293	0.3514	0.9292	0.4598
17/100	0.1001	0.1291	0.3526	0.9288	0.4611
18/100	0.1001	0.1291	0.3498	0.9288	0.4586

Preprint JMIR Publications

References

1. *Patient Harm in Canadian Hospitals? It Does Happen.* Canadian Institute for Health Information; 2024. Accessed February 13, 2024. https://www.cihi.ca/en/patient-harm-in-canadian-hospitals-it-does-happen

2. Leape LL, Brennan TA, Laird N, et al. The Nature of Adverse Events in Hospitalized Patients: Results of the Harvard Medical Practice Study II. *N Engl J Med*. 1991;324(6):377-384. doi:10.1056/NEJM199102073240605

3. Bates DW, Levine DM, Salmasian H, et al. The Safety of Inpatient Health Care. *N Engl J Med*. 2023;388(2):142-153. doi:10.1056/NEJMsa2206117

4. Hilton NZ, Ham E, Rodrigues NC, Kirsh B, Chapovalov O, Seto MC. Contribution of Critical Events and Chronic Stressors to PTSD Symptoms Among Psychiatric Workers. *PS*. 2020;71(3):221-227. doi:10.1176/appi.ps.201900226

5. Chieze M, Hurst S, Kaiser S, Sentissi O. Effects of Seclusion and Restraint in Adult Psychiatry: A Systematic Review. *Front Psychiatry*. 2019;10:491. doi:10.3389/fpsyt.2019.00491

6. Waddell AE, Gratzer D. Patient Safety and Mental Health-A Growing Quality Gap in Canada. *Can J Psychiatry*. 2022;67(4):246-249. doi:10.1177/07067437211036596

7. Velasquez VT, Chang J, Waddell A. The development of early warning scores or alerting systems for the prediction of adverse events in psychiatric patients: a scoping review. *BMC Psychiatry*. 2024;24(1):742. doi:10.1186/s12888-024-06052-z

8. Lee JR, Kim EM, Kim SA, Oh EG. A Systematic Review of Early Warning Systems' Effects on Nurses' Clinical Performance and Adverse Events Among Deteriorating Ward Patients. *J Patient Saf*. 2020;16(3):e104-e113. doi:10.1097/PTS.0000000000000492

9. Liu J, Gan Y, Jiang H, et al. Prevalence of workplace violence against healthcare workers: a systematic review and meta-analysis. *Occup Environ Med*. 2019;76(12):927-937. doi:10.1136/oemed-2019-105849

10. Hesketh KL, Duncan SM, Estabrooks CA, et al. Workplace violence in Alberta and British Columbia hospitals. *Health Policy*. 2003;63(3):311-321. doi:10.1016/s0168-8510(02)00142-2

11. Hiebert BJ, Care WD, Udod SA, Waddell CM. Psychiatric Nurses' Lived Experiences of Workplace Violence in Acute Care Psychiatric Units in Western Canada. *Issues in Mental Health Nursing*. 2022;43(2):146-153. doi:10.1080/01612840.2021.1956656

12. Hibbert PD, Molloy CJ, Schultz TJ, Carson-Stevens A, Braithwaite J. Comparing rates of adverse events detected in incident reporting and the Global Trigger Tool: a systematic review. *International Journal for Quality in Health Care*. 2023;35(3):mzad056. doi:10.1093/intqhc/mzad056

13. Sammut D, Hallett N, Lees-Deutsch L, Dickens GL. A Systematic Review of Violence Risk Assessment Tools Currently Used in Emergency Care Settings. *Journal of Emergency Nursing*. 2023;49(3):371-386.e5. doi:10.1016/j.jen.2022.11.006

14. Ogonah MGT, Seyedsalehi A, Whiting D, Fazel S. Violence risk assessment instruments in forensic psychiatric populations: a systematic review and meta-analysis. *Lancet Psychiatry*. 2023;10(10):780-789. doi:10.1016/S2215-0366(23)00256-0

15. NHS. National Early Warning Score (NEWS). NHS England. 2017. NEWS2 has seen widespread uptake across the NHS in England – at present 100% of ambulance trusts and 76% of acute trusts are using NEWS2, with other early warning scores in place in other areas.

16. Wibisono E, Hadi U, Bramantono, et al. National early warning score (NEWS) 2 predicts hospital mortality from COVID-19 patients. *Annals of Medicine & Surgery*. 2022;76. doi:10.1016/j.amsu.2022.103462

17. Zaidi H, Bader-El-Den M, McNicholas J. Using the National Early Warning Score (NEWS/NEWS 2) in different Intensive Care Units (ICUs) to predict the discharge location of patients. *BMC Public Health*. 2019;19(1):1231. doi:10.1186/s12889-019-7541-3

18. El Morr C, Jammal M, Ali-Hassan H, EI-Hallak W. *Machine Learning for Practical Decision*

Making: A Multidisciplinary Perspective with Applications from Healthcare, Engineering and Business Analytics. Vol 334. Springer International Publishing; 2022. doi:10.1007/978-3-031-16990-8

19. Menger V, Spruit M, van Est R, Nap E, Scheepers F. Machine Learning Approach to Inpatient Violence Risk Assessment Using Routinely Collected Clinical Notes in Electronic Health Records. *JAMA Netw Open*. 2019;2(7):e196709. doi:10.1001/jamanetworkopen.2019.6709

20. Valli I, Marquand AF, Mechelli A, et al. Identifying Individuals at High Risk of Psychosis: Predictive Utility of Support Vector Machine using Structural and Functional MRI Data. *Front Psychiatry*. 2016;7. doi:10.3389/fpsyt.2016.00052

21. Van Der Vegt AH, Campbell V, Mitchell I, et al. Systematic review and longitudinal analysis of implementing Artificial Intelligence to predict clinical deterioration in adult hospitals: what is known and what remains uncertain. *Journal of the American Medical Informatics Association*. 2024;31(2):509-524. doi:10.1093/jamia/ocad220

22. Rubinger L, Gazendam A, Ekhtiari S, Bhandari M. Machine learning and artificial intelligence in research and healthcare. *Injury*. 2023;54:S69-S73. doi:10.1016/j.injury.2022.01.046

23. Pou-Prom C, Murray J, Kuzulugil S, Mamdani M, Verma AA. From compute to care: Lessons learned from deploying an early warning system into clinical practice. *Front Digit Health*. 2022;4:932123. doi:10.3389/fdgth.2022.932123

24. Verma AA, Stukel TA, Colacci M, et al. Clinical evaluation of a machine learning–based early warning system for patient deterioration. *CMAJ*. 2024;196(30):E1027-E1037. doi:10.1503/cmaj.240132

25. Sendak MP, Ratliff 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(7):e15182. doi:10.2196/15182

26. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. Published online 2017. doi:10.48550/ARXIV.1703.01365

27. Moscovici M, Farrokhi F, Vangala L, Simpson AIF, Kurdyak P, Jones RM. Violence risk prediction in mental health inpatient settings using the Dynamic Appraisal of Situational Aggression. *Front Psychiatry*. 2024;15:1460332. doi:10.3389/fpsyt.2024.1460332

28. Ogloff JRP, Daffern M. The dynamic appraisal of situational aggression: an instrument to assess risk for imminent aggression in psychiatric inpatients. *Behavioral Sci & The Law*. 2006;24(6):799-813. doi:10.1002/bsl.741

29. Griffith JJ, Daffern M, Godber T. Examination of the predictive validity of the D ynamic A ppraisal of S ituational A ggression in two mental health units. *Int J Mental Health Nurs*. 2013;22(6):485-492. doi:10.1111/inm.12011

30. Ramesh T, Igoumenou A, Vazquez Montes M, Fazel S. Use of risk assessment instruments to predict violence in forensic psychiatric hospitals: a systematic review and meta-analysis. *Eur psychiatr*. 2018;52:47-53. doi:10.1016/j.eurpsy.2018.02.007

Preprint JMIR Publications

31. Chu CM, Hoo E, Daffern M, Tan J. Assessing the risk of imminent aggression in institutionalized youth offenders using the dynamic appraisal of situational aggression. *J Forens Psychiatry Psychol*. 2012;23(2):168-183. doi:10.1080/14789949.2012.668207

32. Goldhill DR, White SA, Sumner A. Physiological values and procedures in the 24 h before ICU admission from the ward. *Anaesthesia*. 1999;54(6):529-534. doi:10.1046/j.1365-2044.1999.00837.x

33. Steitz BD, McCoy AB, Reese TJ, et al. Development and Validation of a Machine Learning Algorithm Using Clinical Pages to Predict Imminent Clinical Deterioration. *J GEN INTERN MED*. 2024;39(1):27-35. doi:10.1007/s11606-023-08349-3

34. Hahn T, Nierenberg AA, Whitfield-Gabrieli S. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol Psychiatry*. 2017;22(1):37-43. doi:10.1038/mp.2016.201

35. Hansen L, Bernstorff M, Enevoldsen K, et al. Predicting Diagnostic Progression to Schizophrenia or Bipolar Disorder via Machine Learning. *JAMA Psychiatry*. Published online February 19, 2025. doi:10.1001/jamapsychiatry.2024.4702

36. Wolff J, Gary A, Jung D, et al. Predicting patient outcomes in psychiatric hospitals with routine data: a machine learning approach. *BMC Med Inform Decis Mak*. 2020;20:21. doi:10.1186/s12911-020-1042-2

37. Zhang BH, Lemoine B, Mitchell M. Mitigating Unwanted Biases with Adversarial Learning. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’18. Association for Computing Machinery; 2018:335-340. doi:10.1145/3278721.3278779

38. Bellamy RKE, Dey K, Hind M, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J Res & Dev*. 2019;63(4/5):4:1-4:15. doi:10.1147/JRD.2019.2942287

39. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Curran Associates Inc.; 2017:4768-4777.

Preprint JMIR Publications

Preprint JMIR Publications

Tables

Table 1. Cohort characteristics and dataset splits used for model development and evaluation. After final model selection, the full dataset from 2020–2023 (model development) was used to train MAP, which was prospectively evaluated using 2024 (model evaluation) data to assess real-world performance.

		Model Development (2020–2023)				Model Evaluation (2024)
		Held-out-patient		Out-of-time-patient		Prospective set
		Development set	test set	Development set	test set	
Total size (n)		281,022	73,763	259,257	95,528	48,313
Patients		3000	751	2851	1202	900
Time Period		2020-01-01 to 2023-12-31	2020-01-01 to 2023-12-31	2020-01-01 to 2022-12-31	2023-01-01 to 2023-12-31	2024-01-01 to 2024-08-19
Average LOS		629.22	530.41	708.04	336.90	134.99
Sex	Female (%)	27.98	30.81	29.63	25.75	26.05
	Male (%)	68.17	66.82	67.08	70.05	70.37
	Other (%)	3.85	2.37	3.29	4.20	3.58
Sexual	Straight (%)	62.59	62.90	64.31	58.15	60.14

Orientation	Other (%)	37.41	37.10	35.69	41.85	39.86
Race	Black (%)	9.10	4.73	7.85	9.03	8.10
	First Nation (%)	2.05	2.45	2.21	1.92	2.13
	PoC & Other (%)	22.62	16.62	19.78	25.57	26.49
	White (%)	66.23	76.20	70.15	63.49	63.28
Incident Prevalence	total	11744	2569	10688	3625	2106
	# patients	762	209	766	342	266

Table 2. Performance comparison of four ML models evaluated using AUC and recall. Metrics were averaged across multiple runs with different random seeds to ensure robustness.

Category	Sub-category	AUC	Recall
Model Selection			
Held out Patients	LightGB	0.51 ± 0.004	0.02 ± 0.009
	FFNN	0.52 ± 0.005	0.05 ± 0.011
	LSTM	0.87 ± 0.002	**0.75 ± 0.02**
	LSTM + Attn	**0.87 ± 0.002**	0.74 ± 0.04
Out of Time	LightGB	0.52 ± 0.002	0.04 ± 0.003
	FFNN	0.54 ± 0.01	0.08 ± 0.03
	LSTM	0.84 ± 0.01	0.72 ± 0.01
	LSTM + Attn	**0.85 ± 0.01**	**0.75 ± 0.02**
MAP's performance			
Gender	Male	0.83	0.36
	Female	0.84	0.29
	Other gender identities	0.87	0.23
Race	First Nation	0.8	0.16
	White	0.84	0.38
	People of colour and other racial identities	0.81	0.27
	Black	0.69	0.16
Sexual Orientation	Heterosexual	0.82	0.34
	Other	0.84	0.34
Program Type	Regional (Non-forensic)	0.83	0.34
	Provincial (Forensic)	0.8	0.27
Age Group	18-65	0.81	0.32
	65+	0.81	0.38
All		0.81	0.3

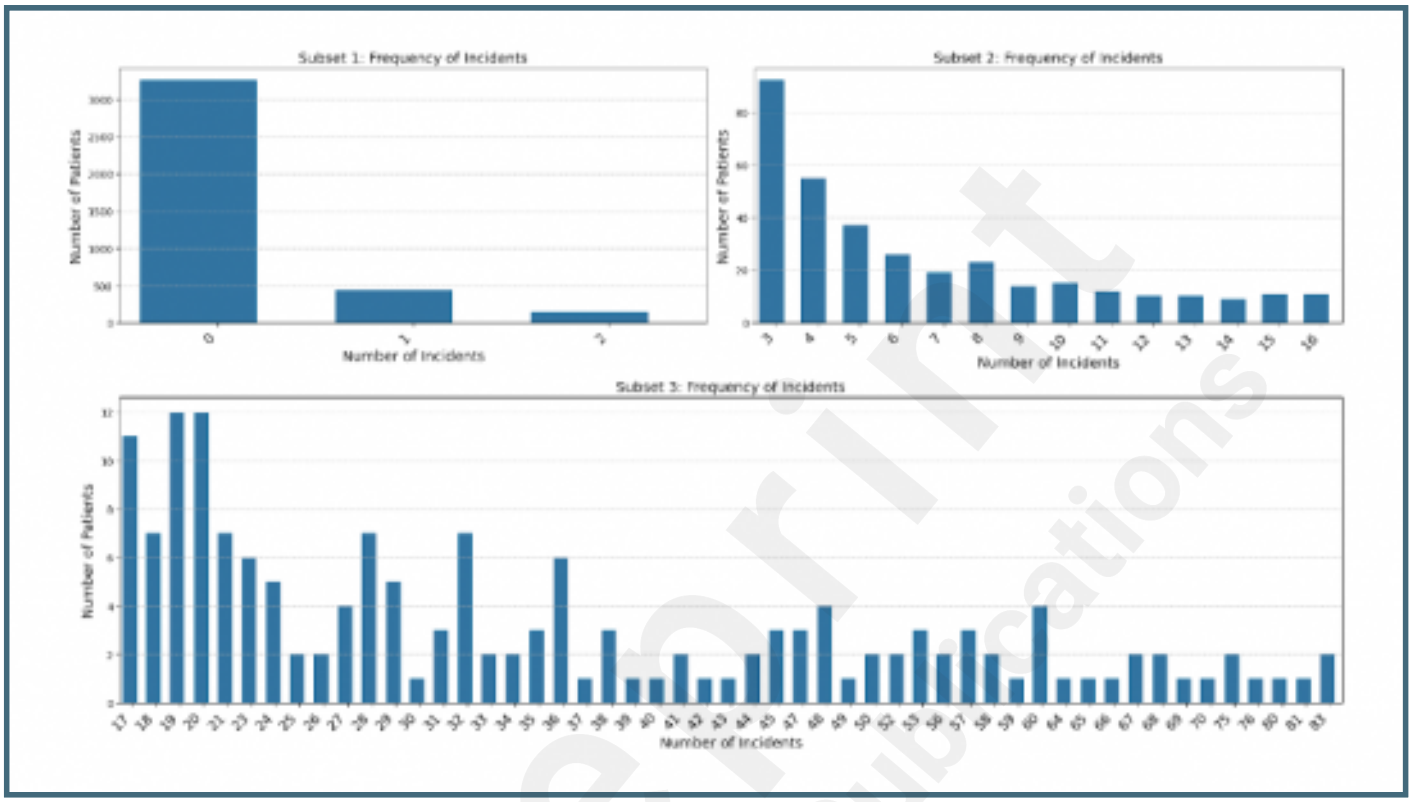
Supplementary Files

Preprint JMIR Publications

Figures

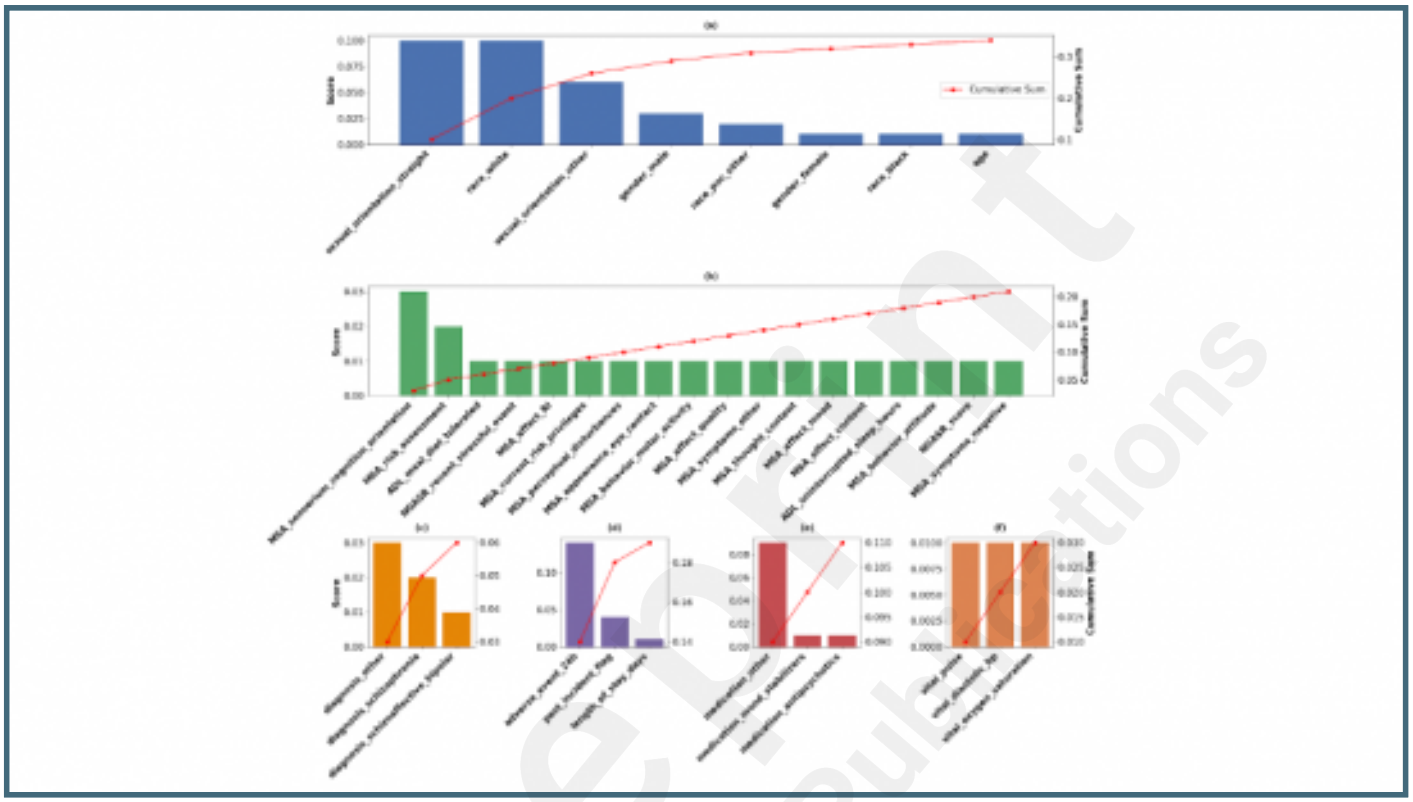
Preprint JMIR Publications

The distribution of the total adverse events across the sample of patients between 2020-2024. On the x-axis is the specific number of adverse events, and on the y-axis is the number of individuals who experienced that many adverse events between the 2020-2024 time period. The different subsets represent ranges of incidents. Subset 1: Cohorts of individuals with 0-2 incidents; Subset 2: 3-16 incidents; Subset 3: 17-83 incidents.

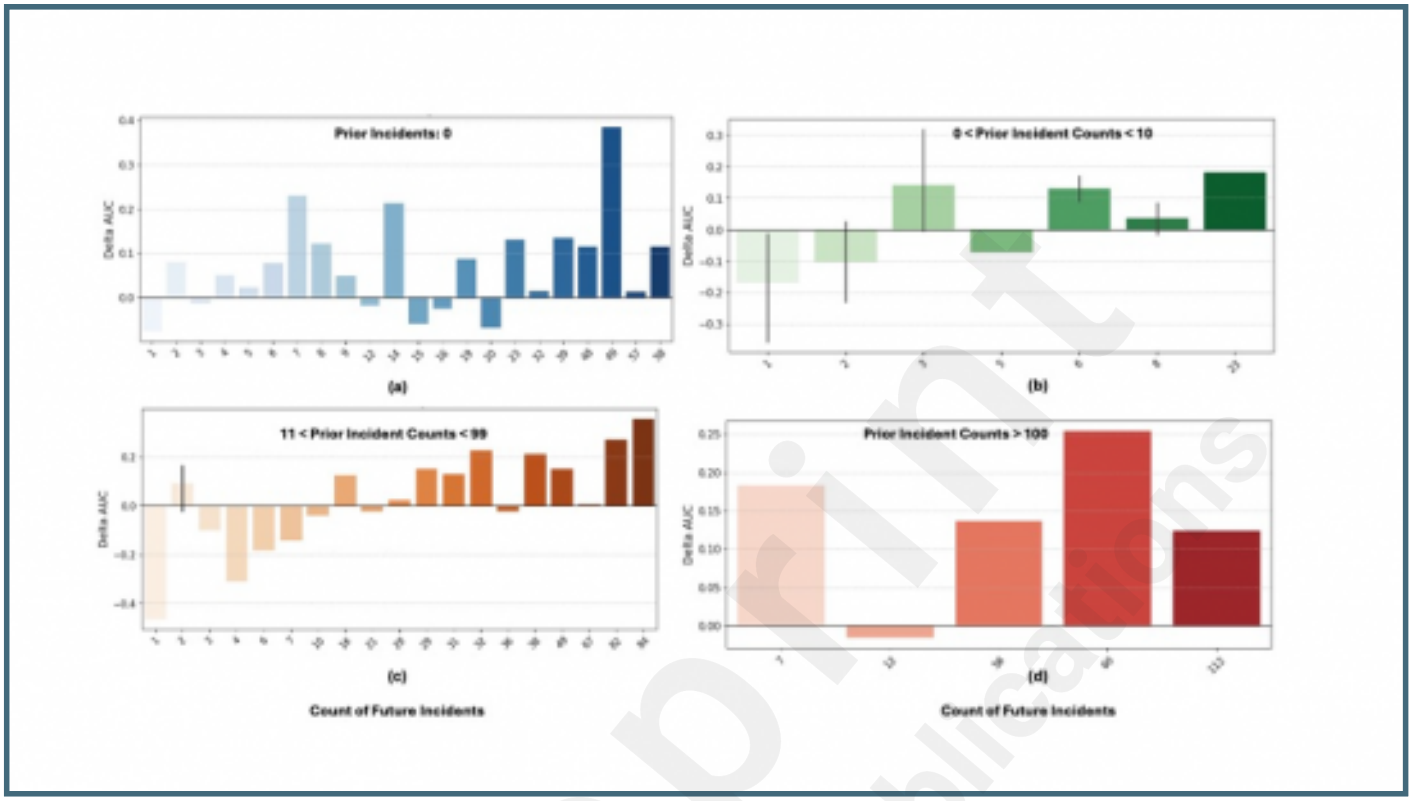


Preprint JMIR Publications

Feature importance across distinct categories in the MAP model: (a) Overall feature importance highlighting the top predictors; (b) Demographic features including gender, race, and sexual orientation; (c) Clinical assessment features such as mental status indicators and functional assessments; (d) Clinical diagnoses including major psychiatric conditions; (e) Clinical variables related to adverse events, incident history, and hospital stay duration; (f) Medication-related features, including mood stabilizers and antipsychotics; and (g) Vital signs, including pulse, blood pressure, and oxygen saturation.

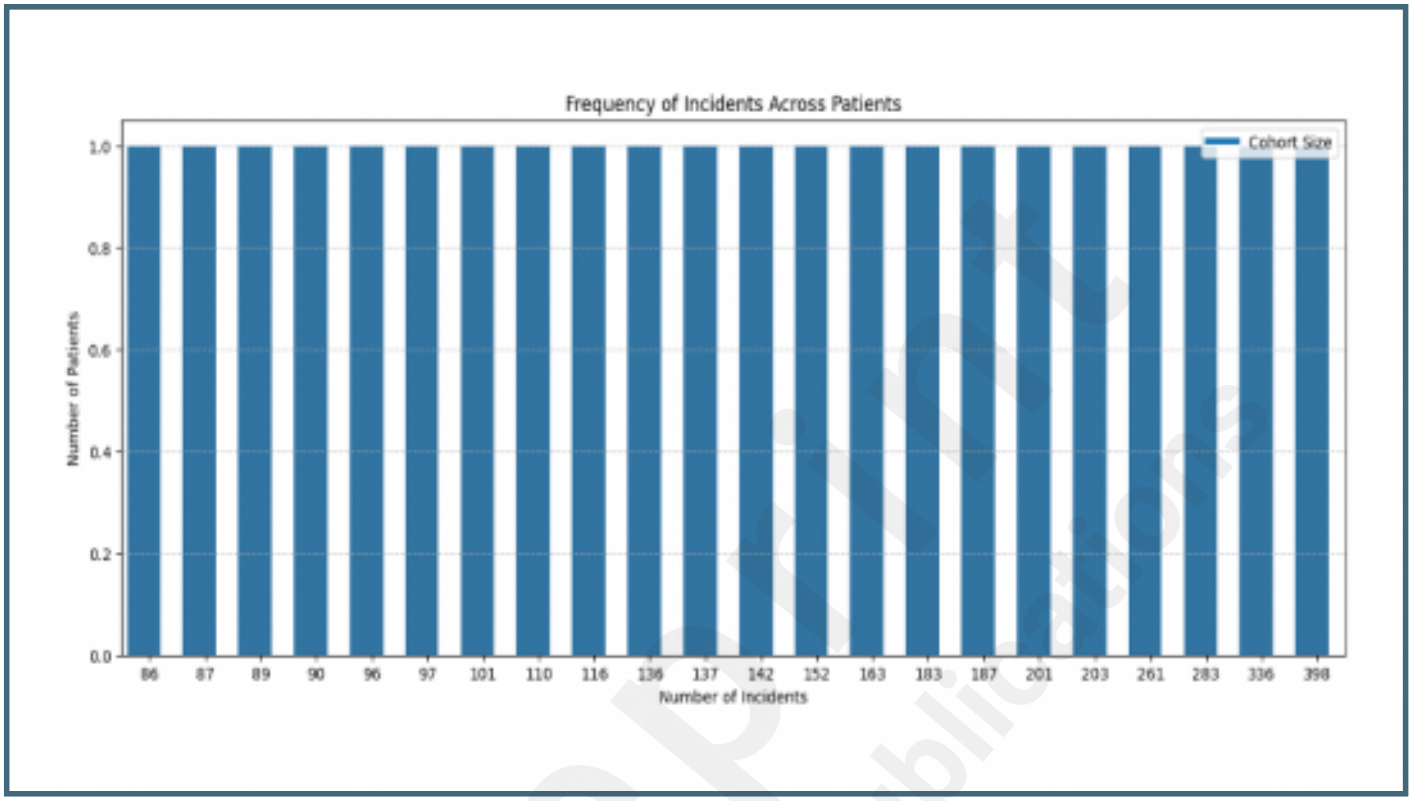


Difference in AUC ROC performance scores (Delta AUC = ML AUC ROC - DASA AUC ROC). A positive delta AUC indicates the ML model outperformed DASA for that specific cohort group. A negative delta AUC indicates DASA outperformed ML for that specific cohort group.



Preprint JMIR Publications

The distribution of the total adverse events across the sample of patients between 2020-2024. On the x-axis is the specific number of adverse events and on the y-axis is the number of individuals that experienced that many adverse events between 2020-2024 time period. This graph represents individuals with the range of 86-398 incidents.



Preprints JMIR Publications