

Original Paper

Advancing Psychiatric Safety With the Predictive Risk Identification for Mental Health Events Tool: Retrospective Cohort Study

Elham Dolatabadi^{1,2}, BSc, MSc, PhD; Valentina Tamayo Velasquez^{3,4}, HBsc; Abdul Hamid Dabboussi¹, BSc, MSc; David Wen⁴, PhD; Jennifer Crawford^{4,5}, PhD; Andrea E Waddell^{4,6}, MEd, MD; Christo El Morr^{1,7}, BA, HBsc, MSc, PhD

¹School of Health Policy and Management, Faculty of Health, York University, Toronto, ON, Canada

²Vector Institute, Toronto, ON, Canada

³Institute of Medical Science, Temerty Faculty Medicine, University of Toronto, Toronto, ON, Canada

⁴Waypoint Centre for Mental Health Care, Penetanguishene, ON, Canada

⁵Faculty of Health Sciences, The University of Ontario Institute of Technology, Oshawa, ON, Canada

⁶Temerty Faculty of Medicine, The University of Toronto, Toronto, ON, Canada

⁷North York General Hospital, Toronto, ON, Canada

Corresponding Author:

Elham Dolatabadi, BSc, MSc, PhD

School of Health Policy and Management

Faculty of Health

York University

Stong College, 340

165 Campus Walk

Toronto, ON, M3J 1P3

Canada

Phone: 1 6477069756

Email: edolatab@yorku.ca

Abstract

Background: Patient safety incidents are a leading cause of harm in psychiatric settings, yet early warning systems (EWS) tailored to mental health remain underdeveloped. Traditional risk tools such as the Dynamic Appraisal of Situational Aggression–Inpatient Version (DASA-IV) offer limited predictive accuracy and are reactive rather than proactive.

Objective: We introduce the Predictive Risk Identification for Mental Health Events (PRIME) tool, a deep learning–based EWS trained on longitudinal psychiatric electronic medical record (EMR) data to anticipate adverse events in 24-hour windows.

Methods: A retrospective cohort study using routinely collected EMR data to train and validate machine learning (ML) models for short-term risk prediction was conducted. This study took place at Waypoint Centre for Mental Health Care, a large inpatient psychiatric hospital in Ontario, Canada, serving both high-security forensic and nonforensic patient populations. A total of 4651 patients and 403,098 encounters from January 2020 to August 2024 were included. For model evaluation, the 2024 test set included 900 patients and 48,313 encounters. PRIME was trained using recurrent neural networks with attention mechanisms on multivariate time-series data. The model used 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 ML baselines. The primary outcome was the occurrence of an adverse mental health event recorded in the EMR within the following 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 long short-term memory with attention mechanism achieved the highest predictive performance (AUC=0.83), outperforming existing tools such as DASA-IV by 0.20 AUC (0.81 vs 0.61) and demonstrating the potential of ML-based models to support proactive risk management in mental health settings.

Conclusions: The PRIME tool is one of the first developed and evaluated deep learning–based EWS for psychiatric inpatient care. By outperforming existing clinical tools and providing interpretable, rolling predictions, PRIME 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

Introduction

Patient and staff safety are top priorities in health care, 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 1 in 4 hospital admissions involved adverse events, with a quarter of these deemed preventable [3]. While all health care settings face safety risks, psychiatric environments present a distinct set of challenges, including suicide, restraint, and seclusion—events that contribute to continued deterioration and injury [4,5]. Despite a higher prevalence of adverse events in mental health, research on patient safety and mental deterioration-related adverse events 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]. Worldwide, approximately 24% of health care 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 and the Brøset Violence Checklist are also used, these 2 primarily target short-term aggression and violence prediction, and 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 with early signs of deterioration that are not easily detected. Additionally, there are other widely validated measures for more specific feature prediction, such as the Historical, Clinical, and Risk Management, also used for violence risk assessment; the Columbia Suicide Severity Rating Scale for the assessment of suicidal ideation and behavior; and many other risk assessment tools [15,16]. We focused on both the Dynamic Appraisal of Situational Aggression and Brøset Violence Checklist measures as they are 2 of the most widely validated and routinely implemented structured risk assessment tools in inpatient psychiatry [17].

Early warning systems (EWSs) 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 have been effectively implemented in acute care settings to support timely interventions [18-20]. At the same time, machine learning (ML) is transforming risk assessment by enabling the analysis of large-scale, high-dimensional health care data [21-23]. 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 [24,25]. 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 [26-28]. However, 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, the Predictive Risk Identification for Mental Health Events (PRIME) tool. The PRIME tool is a deep learning-based EWS leveraging longitudinal electronic medical record (EMR) data from a specialized psychiatric hospital. The goal of the PRIME tool is to predict mental health-specific adverse events, including but not limited to self-harm, suicide attempts, violence toward others, and aggressive behaviors (Multimedia Appendix 1). PRIME is trained to predict the likelihood of these adverse events within 24-hour windows using autoregressive recurrent neural networks enriched with attention mechanisms and interpretability via integrated gradients. Unlike traditional tools, PRIME is capable of continuous, real-time risk forecasting even in the absence of prior incidents. We benchmarked PRIME against Dynamic Appraisal of Situational Aggression–Inpatient Version (DASA-IV) and other ML models, and it demonstrated superior performance, particularly in complex and high-risk subgroups.

Through this study, we aimed 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 Health Care (hereafter referred to as “Waypoint”), Ontario, Canada. We retrospectively retrieved data from all patients at Waypoint between January 2020 and August 2024, including static and dynamic variables (Multimedia Appendix 1).

Ethical Considerations

This study was approved by the York University Office of Research Ethics (certificate e2023-163) and the Research Ethics Board of Waypoint Centre for Mental Health Care (reference #RCRA#23.08.01) with waived informed consent. The Research Ethics Board waived the need for informed consent since the data was retrospectively collected in routine practice.

Data Representation and Processing

First, we conducted a literature review to identify factors widely associated with mental health deterioration and adverse events. We collaborated with clinicians, physicians, clinical informatics specialists, and the research team to review these factors and select the variables within our EMR ([Multimedia Appendix 1](#)). 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 were recorded multiple times per day. These factors encompassed a range of clinical and behavioral variables in the following categories: inpatient admission assessments that included demographic and diagnosis data, clinical risk assessments, physiological data, recent behavioral data, and mental status assessment data ([Multimedia Appendix 1](#)). 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, whereas 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 *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, selecting the 45 most frequent diagnoses across all patients to prevent overfitting. Medication data included our patient group's 5 most relevant categories, each represented as a binary indicator denoting whether it was administered within the previous 24 hours ([Multimedia Appendix 1](#)).

The primary outcome in our study was the occurrence of any mental health adverse event. For our prediction task, each patient encounter was labeled based on whether a logged adverse event in the EMR system occurred within the following 24-hour bin. This binary label (event vs no event) was used as the target variable for PRIME 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 PRIME Model

To improve psychiatric-medical baselines, we designed a deep learning-based EWS. Specifically, we developed recurrent neural networks using the long short-term memory (LSTM) model that triggered 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 each 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+attention). In this variant, an attention layer was added to the LSTM hidden states ([Multimedia Appendix 2](#)). For each 24-hour prediction interval, the attention mechanism

dynamically assigned weights to each time step 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 2 ML approaches: light gradient boosting machine (LightGBM) and feedforward neural network (FNN). All time-series features were aggregated over 3 to 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. To support robust evaluation and model selection, the dataset was first partitioned into two distinct test sets: (1) a held-out patient test set with no patient overlap between the development (3000 patients) and the test sets (751 patients) and (2) an out-of-time test set split across time using 2020 to 2022 data for training and 2023 data for testing.

Once the model selection was finalized, we performed a final training phase using all data from 2020 to 2023 to build PRIME. This final model was then evaluated on 2024 data to assess real-world applicability. Model calibration was assessed on the 2024 evaluation cohort using reliability (calibration) curves and the Brier score [29,30]. Reliability curves were generated using 10 uniformly spaced probability bins plotting the mean predicted risk against observed outcome frequency within each bin. The Brier score was computed as the mean squared difference between predicted probabilities and observed binary outcomes, providing a quantitative measure of overall probabilistic accuracy.

To explore the factors driving PRIME's predictions, we used integrated gradients to compute feature importance in our LSTM model by computing the path-integrated gradients from the input to the actual output [31]. To quantify uncertainty, gradients were bootstrapped over 100 resampled datasets.

Comparison With Clinical Measures

We compared PRIME's predictive performance against that of the DASA-IV, a standardized tool used at our hospital to evaluate risks of aggression [32]. The PRIME tool's predictions included all mental health-specific adverse events recorded in the hospital's incident log ([Multimedia Appendix 1](#)). DASA-IV includes 7 items assessing behavioral indicators (ie, irritability, negative attitudes, and verbal threats), each scored as 0 (not observed) or 1 (observed), with a total score categorized as low (0-1), moderate (2-3), or high (>3) [33,34]. To align DASA-IV with PRIME's binary classification, we restructured the risk categories. Moderate and high risk were grouped as "at risk" (positive prediction), whereas low risk was grouped as "no risk" (negative prediction). PRIME is designed to predict a broader range of mental health-specific adverse events, whereas DASA-IV is limited to aggression-related incidents and deterioration. Our goal was to compare the PRIME tool with the current validated tool used in clinical practice. This allowed us to compare DASA-IV's performance against PRIME's predictions and the ground truth outcomes recorded in patient encounters.

Results

Cohort Characteristics

The dataset encompassed 4651 patients and 403,098 patient

encounters over 55 months. The demographic characteristic distribution of the patient cohort is presented in Table 1. For the evaluation of the best-performing ML model and comparison against clinical baselines, we used data from 2024, with detailed breakdowns provided in Table 1.

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

	Model development (2020-2023)				Model evaluation (2024)—evaluation set (48,313)
	Held-out patients		Out-of-time patients		
Data split (number of patient encounters)	Development set (281,022)	Test set (73,763)	Development set (259,257)	Test set (95,528)	
Patients, n (%)	3000 (80)	751 (20)	2851 (70.3)	1202 (29.7)	900 (100)
Period	January 1, 2020, to December 31, 2023	January 1, 2020, to December 31, 2023	January 1, 2020, to December 31, 2022	January 1, 2023, to December 31, 2023	January 1, 2024, to August 19, 2024
LOS ^a , mean (SD)	629.22 (632.67)	530.41 (654.51)	708.04 (676.56)	336.90 (390.40)	134.99 (117.83)
Sex, n (%)					
Female	839 (27.98)	231 (30.81)	845 (29.63)	310 (25.75)	235 (26.05)
Male	2045 (68.17)	502 (66.82)	1912 (67.08)	842 (70.05)	633 (70.37)
Other	116 (3.85)	18 (2.37)	94 (3.29)	50 (4.20)	32 (3.58)
Sexual orientation, n (%)					
Heterosexual	1878 (62.59)	472 (62.90)	1833 (64.31)	699 (58.15)	541 (60.14)
Other	1122 (37.41)	279 (37.10)	1018 (35.69)	503 (41.85)	359 (39.86)
Race, n (%)					
Black	273 (9.10)	36 (4.73)	224 (7.85)	109 (9.03)	73 (8.10)
First Nations	61 (2.05)	18 (2.45)	63 (2.21)	23 (1.92)	19 (2.13)
White	1987 (66.23)	572 (76.20)	2000 (70.15)	763 (63.49)	570 (63.28)
Other races	679 (22.62)	125 (16.62)	564 (19.78)	307 (25.57)	238 (26.49)
Incident prevalence					
Total number of incidents	11,744	2569	10,688	3625	2106
Patients, n (%)	762	209	766	342	266

^aLOS: length of stay in days.

Adverse event distribution per individual varied across the sample of patients between 2020 to 2024. When grouping the number of patients by the frequency of adverse events they experienced during their hospital stay, there is a decrease in the number of patients who experience a high count of adverse events. Most patients experienced few or no adverse events: 69.9% (3251/4651) had no incidents, and 12.6% (587/4651) experienced up to 2 incidents. A total of 7.4% (344/4651) of the patients had between 3 and 16 events, with a median of 14.5 (IQR 14.25). A smaller group of 162 patients experienced between 17 and 83 incidents, most of whom (n=42, 25.9%) had between 17 and 20, whereas only 22 (13.6%) had more than 83 events. As incident frequency increased, cohort size decreased. The mean number of adverse events across the sample was 2.85, whereas 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.

PRIME’s Predictions

Table 2 presents the performance comparison of the 4 ML models: light gradient boosting (LightGBM), feedforward neural network (FNN), LSTM, and LSTM+attention. 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+attention model consistently achieved the highest performance, with an AUC of 0.87 for held-out patients and 0.72 for out-of-time patients Multimedia Appendix 3. We

selected the LSTM+attention model as the final architecture for PRIME (Table 2).

Table 2. Performance comparison of 4 machine learning models evaluated using area under the receiver operating characteristic curve (AUC) and recall. Metrics were averaged across multiple runs with different random seeds to ensure robustness.

Category and subcategory	AUC	Recall
Model selection, mean (SD)		
Held-out patients		
LightGBM ^a	0.51 (0.004)	0.02 (0.009)
FNN ^b	0.52 (0.005)	0.05 (0.011)
LSTM ^c	0.87 (0.002)	0.75 (0.02)
LSTM+attention	0.87 (0.002)	0.74 (0.04)
Out-of-time patients		
LightGBM	0.52 (0.002)	0.04 (0.003)
FNN	0.54 (0.01)	0.08 (0.03)
LSTM	0.84 (0.01)	0.72 (0.01)
LSTM+attention	0.85 (0.01)	0.75 (0.02)
PRIME's^d performance		
Sex		
Male	0.83	0.36
Female	0.84	0.29
Intersex	0.87	0.23
Race		
Black	0.69 ^e	0.16
First Nations	0.8	0.16
White	0.84	0.38
Other racial identities	0.81	0.27
Sexual orientation		
Heterosexual	0.82	0.34
Other	0.84	0.34
Program type		
Regional (nonforensic)	0.83	0.34
Provincial (forensic)	0.8	0.27
Age group (years)		
18-65	0.81	0.32
≥65	0.81	0.38
All	0.81	0.3

^aLightGBM: light gradient boosting machine.

^bFNN: feedforward neural network.

^cLSTM: long short-term memory.

^dPRIME: Predictive Risk Identification for Mental Health Events.

^eItalicization indicates significance.

For the evaluation using the dataset from 2024, with 48,313 encounters and 2106 recorded adverse events, PRIME achieved an AUC of 83% (Table 2). The performance varied within and across subgroups, with AUC ranging from 0.69 (Black patients)

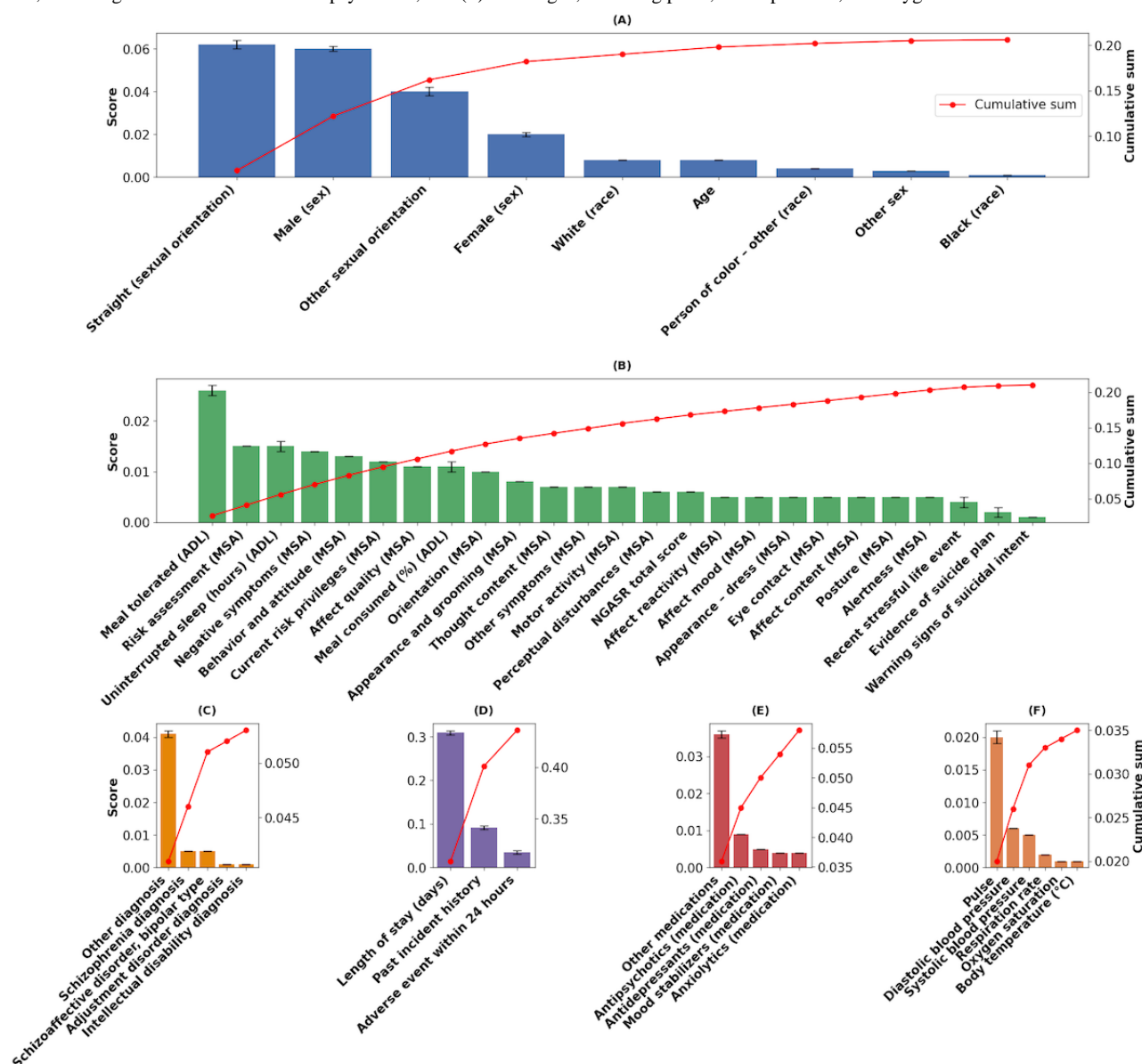
to 0.87 (intersex patients), indicating potential biases favoring larger, more represented groups within the sample data. Across racial subgroups, AUC differed by 14%; across sex subgroups, AUC varied by 5%; across sexual orientation subgroups, AUC

varied by 2%; across program types, AUC differed by 4%; and, across age groups, AUC variation was minimal (<1%).

Calibration analysis demonstrated that PRIME produced well-aligned risk estimates. The reliability curve closely followed the identity line across predicted probability bins (Multimedia Appendix 4), indicating good agreement between predicted and observed event rates. The model achieved a Brier score of 0.036 on the evaluation set, reflecting strong overall calibration performance given the low event prevalence.

Feature importance was aggregated across time steps and encounters and summarized at the feature level (Figure 1). Integrated gradient attributions were bootstrapped over 100 resampled evaluation datasets, with the resulting variability visualized as error bars. Ranking stability was assessed using Spearman correlation, demonstrating near-perfect robustness ($\rho=0.99, -0.0001$ to $+0.0001$). Among the 40 features included in PRIME, the top 16 predictors accounted for approximately 80% of the model's total importance, reflecting a diverse combination of demographic, medical, and psychosocial factors that drive risk prediction.

Figure 1. Feature importance across distinct categories in the PRIME model: (A) Demographic features including gender, race, and sexual orientation; (B) Clinical assessment features such as mental status indicators and functional assessments; (C) Clinical diagnoses including major psychiatric conditions, including schizophrenia diagnosis; (D) Clinical variables related to adverse events, incident history, and hospital stay duration; (E) Medication-related features, including mood stabilizers and antipsychotics; and (F) Vital signs, including pulse, blood pressure, and oxygen saturation.



When further analyzing feature contributions within specific categories, demographic factors (Figure 1A) and indicators, with heterosexual sexual orientation and male sex showing the largest individual contributions, followed by *other* sexual orientation categories and female sex. Race-related variables and age demonstrated comparatively smaller effects. From the

clinical assessments (Figure 1B), meal tolerated (ADL), risk assessment (MSA) and uninterrupted sleep (ADL) were identified as important contributors. In the category of clinical diagnoses (Figure 1C), schizophrenia and schizoaffective or bipolar disorder emerged as the most significant predictors. Among clinical variables (Figure 1D), length of hospital stay

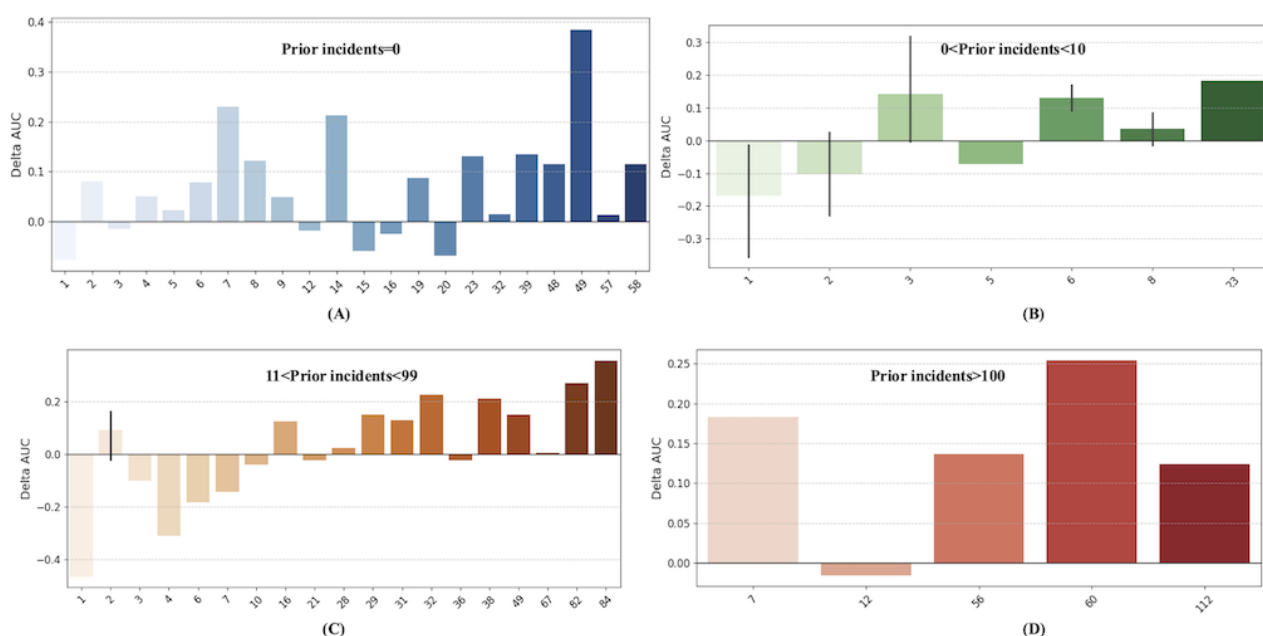
emerged as the most influential contributor, followed by history of past incidents and adverse events in the previous 24 hours. In the medication category (Figure 1E), the other medications category exhibited the largest overall contribution, followed by antipsychotics and antidepressants, with mood stabilizers and anxiolytics contributing more modestly. Finally, among vital signs (Figure 1F), features such as pulse, blood pressure, and oxygen saturation met the 0.90 cumulative importance threshold, although their influence remained relatively modest.

Comparison With the Standardized Risk Assessment Tool DASA-IV

PRIME demonstrated a 0.2 AUC improvement over the DASA-IV assessment tool when assessed on the 2024 evaluation dataset, with PRIME achieving an AUC of 0.81 compared to

DASA-IV's AUC of 0.61 (Multimedia Appendix 5). To further assess PRIME'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 2 illustrates the performance differences between PRIME and DASA-IV 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. To examine the model's performance compared to that of DASA-IV, we defined subgroups based on all the unique combinations of adverse event occurrences observed in the training and test datasets. This yielded 63 unique subgroups representing different patterns and combinations of past and future incident frequencies across the datasets.

Figure 2. 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. AUC: area under the receiver operating characteristic curve; DASA: Dynamic Appraisal of Situational Aggression; ML: machine learning; ROC: receiver operating characteristic curve.



The PRIME tool significantly outperformed DASA-IV in 40 of the 63 subgroups (Wilcoxon test; $P=.007$). For individuals with no prior incidents in the training set but up to 58 total incidents in the evaluation period (Figure 2A), PRIME achieved an AUC of 0.62, whereas DASA-IV achieved an AUC of 0.50. For individuals with up to 10 incidents in the past and up to 23 in the future (Figure 2B), DASA-IV outperformed PRIME in cases in which individuals had 1, 2, or 5 future incidents. However, PRIME outperformed DASA-IV in the remaining 4 subcategories within this range. For individuals with moderate incident frequency (11-99 past incidents), PRIME outperformed DASA-IV in 11 of the 19 groups (Figure 2C). Among individuals with frequent incidents (>100 past incidents), PRIME outperformed DASA-IV in 4 of the 5 subgroups (Figure 2D). Notably, PRIME's performance was better in cases

in which individuals had a high number of past incidents but only 1 in the future.

Discussion

Despite the growing number of adverse events in mental health settings, deep learning tools that leverage routinely collected EMR data to predict patient deterioration remain limited. Our model, PRIME, represents a first-of-its-kind approach tailored specifically to psychiatry and demonstrated strong predictive performance, achieving an AUC of 0.83. Leveraging autoregressive LSTM with attention mechanisms, PRIME 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 nonforensic acute care programs contributes to the model's generalizability across diverse mental health populations. The strong calibration performance observed for PRIME is particularly important for clinical deployment, where accurate probability estimates are essential for risk stratification and decision support. Well-calibrated predictions enable clinicians to interpret PRIME scores as meaningful risk estimates rather than solely as ranking signals.

Currently, no ML-based predictive alerting tools are deployed in mental health settings. Instead, clinicians rely on actuarial tools such as DASA-IV to assess risks related to violence and aggression [35]. On the same dataset, PRIME outperformed DASA-IV (AUC=0.83 vs 0.61). While DASA-IV has reported AUCs between 0.61 and 0.82 in other studies, it is important to note that PRIME and DASA-IV target different outcomes [36]. PRIME captures a broader spectrum of deterioration events, including suicide, self-harm, and clinical decompensation, whereas DASA-IV is limited to aggression-related outcomes. The lower AUC for DASA-IV in our dataset likely reflects these differences in scope. Nonetheless, PRIME'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, PRIME provides clinicians with a holistic and actionable risk assessment framework, supporting earlier and more effective interventions. PRIME 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 one of the predictors 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 [37,38]. 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 [39-41]. 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 2 racial subgroups: Black and First Nations individuals, with AUC scores 14% and 3% lower, respectively, than those for 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, with Black individuals comprising less than 10% (309/3751) of the sample in the training set and less than 10% (73/900) in the evaluation set. Similarly, First Nations individuals comprise less than 3% (79/3751) in the training set and less than 3% (19/900) in the evaluation set. Furthermore, this study did not assess the intersectional effects, such as whether the demographic factors had any effect or potential differences between the forensic and nonforensic programs. These represent important assessments for future work to further evaluate whether predictive models such as PRIME are unbiased and generalizable across different clinical settings and patient populations.

Additionally, while the PRIME tool demonstrated high predictive performance, the complexity that is inherently present in deep learning models may limit clinical interpretability. Ensuring clinician confidence and understanding of the model's prediction is critical for successful implementation. Ongoing monitoring and evaluation of PRIME are needed to assess its real-world performance and potential biases.

Future work will evaluate the utility, feasibility, and efficacy of the PRIME tool in real-world clinical settings. This future work will also focus on mitigating the previously mentioned biases through bias-aware data augmentation and fairness-aware learning algorithms (eg, adversarial debiasing) to improve representation across subgroups [42-44]. Piloting the PRIME 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 our future pilot and deployment, we plan to use PRIME as a binary risk assessment tool to flag patients at a high risk of adverse events in mental health settings. Finally, although the PRIME tool was developed using data from a single mental health hospital, the model framework and variable-mapping methodological approaches are transferable to other mental health and psychiatric settings. If the PRIME tool is to be implemented in other settings, it will require retraining and validation to account for different patient populations, data sources, and documentation practices.

In this study, we developed and evaluated an LSTM model that could predict patients at risk of 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, model evaluation, and near-real-time operation position it for real-world clinical integration. By generating dynamic forecasts without dependence on manual clinician input, PRIME can augment existing workflows and support earlier interventions in settings where mental health staff face high demands and elevated safety risks.

Acknowledgments

This work was made possible through access to data provided by Waypoint Centre for Mental Health Care, and the authors gratefully acknowledge the support and collaboration of the Waypoint research team. They also extend their 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. ED's research is supported by the Canadian Institutes of Health Research Centre for Research on Pandemic Preparedness and Health Emergencies and a Natural Sciences and Engineering Research Council of Canada Discovery Grant.

Data Availability

The datasets generated or analyzed during this 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, AEW, and CEM contributed to the conceptualization and methodology of the study. ED, CEM, JC, AHD, and DW contributed to coding and model development. VTV and ED conducted the data analysis and writing of the manuscript. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of baseline and temporal features used for the PRIME (Predictive Risk Identification for Mental Health Events) tool.

[\[DOCX File , 4251 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Hyperparameters for the long short-term memory model and the long short-term memory+attention model. A comparison of the hyperparameter configurations for the two top performing models evaluated.

[\[DOCX File , 3742 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

A summary of the model's performance metrics across 18 epochs. Early stopping triggered at epoch 19.

[\[DOCX File , 3743 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Calibration (reliability) curve for the PRIME model evaluated on the 2024 test dataset.

[\[DOCX File , 3917 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Clinical baseline model: Dynamic Appraisal of Situational Aggression's Predictive Performances across different data splits (area under the receiver operating characteristic curve).

[\[DOCX File , 3741 KB-Multimedia Appendix 5\]](#)

References

1. The case for investing in patient safety in Canada. Canadian Patient Safety Institute. URL: <https://www.bcit.ca/files/health/pdf/risk-analytica-2017-investing-in-patient-safety-in-canada.pdf> [accessed 2025-05-29]
2. Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, et al. The nature of adverse events in hospitalized patients: results of the Harvard Medical Practice Study II. N Engl J Med. Feb 07, 1991;324(6):377-384. [doi: [10.1056/nejm199102073240605](https://doi.org/10.1056/nejm199102073240605)]
3. Bates DW, Levine DM, Salmasian H, Syrowatka A, Shahian DM, Lipsitz S, et al. The safety of inpatient health care. N Engl J Med. Jan 12, 2023;388(2):142-153. [doi: [10.1056/nejmsa2206117](https://doi.org/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. *Psychiatr Serv*. Mar 01, 2020;71(3):221-227. [doi: [10.1176/appi.ps.201900226](https://doi.org/10.1176/appi.ps.201900226)] [Medline: [31795856](#)]
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. [FREE Full text] [doi: [10.3389/fpsy.2019.00491](https://doi.org/10.3389/fpsy.2019.00491)] [Medline: [31404294](#)]
6. Waddell AE, Gratzner D. Patient safety and mental health-a growing quality gap in Canada. *Can J Psychiatry*. Apr 11, 2022;67(4):246-249. [FREE Full text] [doi: [10.1177/07067437211036596](https://doi.org/10.1177/07067437211036596)] [Medline: [34378413](#)]
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*. Oct 28, 2024;24(1):742. [FREE Full text] [doi: [10.1186/s12888-024-06052-z](https://doi.org/10.1186/s12888-024-06052-z)] [Medline: [39468486](#)]
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*. Sep 26, 2020;16(3):e104-e113. [doi: [10.1097/PTS.0000000000000492](https://doi.org/10.1097/PTS.0000000000000492)] [Medline: [29698354](#)]
9. Liu J, Gan Y, Jiang H, Li L, Dwyer R, Lu K, et al. Prevalence of workplace violence against healthcare workers: a systematic review and meta-analysis. *Occup Environ Med*. Dec 13, 2019;76(12):927-937. [doi: [10.1136/oemed-2019-105849](https://doi.org/10.1136/oemed-2019-105849)] [Medline: [31611310](#)]
10. Hesketh KL, Duncan SM, Estabrooks CA, Reimer MA, Giovannetti P, Hyndman K, et al. Workplace violence in Alberta and British Columbia hospitals. *Health Policy*. Mar 2003;63(3):311-321. [doi: [10.1016/s0168-8510\(02\)00142-2](https://doi.org/10.1016/s0168-8510(02)00142-2)] [Medline: [12595130](#)]
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 Ment Health Nurs*. Feb 11, 2022;43(2):146-153. [doi: [10.1080/01612840.2021.1956656](https://doi.org/10.1080/01612840.2021.1956656)] [Medline: [34379570](#)]
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. *Int J Qual Health Care*. Jul 25, 2023;35(3):mzad056. [FREE Full text] [doi: [10.1093/intqhc/mzad056](https://doi.org/10.1093/intqhc/mzad056)] [Medline: [37440353](#)]
13. Sammut D, Hallett N, Lees-Deutsch L, Dickens GL. A systematic review of violence risk assessment tools currently used in emergency care settings. *J Emerg Nurs*. May 2023;49(3):371-86.e5. [FREE Full text] [doi: [10.1016/j.jen.2022.11.006](https://doi.org/10.1016/j.jen.2022.11.006)] [Medline: [36585335](#)]
14. Ogonah MG, Seyedsalehi A, Whiting D, Fazel S. Violence risk assessment instruments in forensic psychiatric populations: a systematic review and meta-analysis. *Lancet Psychiatry*. Oct 2023;10(10):780-789. [FREE Full text] [doi: [10.1016/S2215-0366\(23\)00256-0](https://doi.org/10.1016/S2215-0366(23)00256-0)] [Medline: [37739584](#)]
15. Dolan M, Blattner R. The utility of the Historical Clinical Risk -20 scale as a predictor of outcomes in decisions to transfer patients from high to lower levels of security-a UK perspective. *BMC Psychiatry*. Sep 29, 2010;10(1):76. [doi: [10.1186/1471-244x-10-76](https://doi.org/10.1186/1471-244x-10-76)]
16. Xiao S, Ge Q, Wang T, Zhang M, Hu A, Zhang X. Psychometric characteristics of the Chinese version of the Columbia-Suicide Severity Rating Scale among people with mental health diagnosis. *BMC Psychiatry*. Aug 21, 2025;25(1):803. [FREE Full text] [doi: [10.1186/s12888-025-07187-3](https://doi.org/10.1186/s12888-025-07187-3)] [Medline: [40842006](#)]
17. Chu CM, Daffern M, Ogloff JR. Predicting aggression in acute inpatient psychiatric setting using BVC, DASA, and HCR-20 Clinical scale. *J Forensic Psychiatry Psychol*. Apr 2013;24(2):269-285. [doi: [10.1080/14789949.2013.773456](https://doi.org/10.1080/14789949.2013.773456)]
18. Suspected Sepsis: Recognition, Diagnosis and Early Management. London, UK. National Institute for Health and Care Excellence; 2024.
19. Wibisono E, Hadi U, Arfijanto MV, Rusli M, Rahman BE, Asmarawati TP, et al. National early warning score (NEWS) 2 predicts hospital mortality from COVID-19 patients. *Ann Med Surg (Lond)*. Apr 2022;76:103462. [FREE Full text] [doi: [10.1016/j.amsu.2022.103462](https://doi.org/10.1016/j.amsu.2022.103462)] [Medline: [35284070](#)]
20. 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*. Sep 05, 2019;19(1):1231. [FREE Full text] [doi: [10.1186/s12889-019-7541-3](https://doi.org/10.1186/s12889-019-7541-3)] [Medline: [31488143](#)]
21. El Morr C, Jammal M, Ali-Hassan H, El-Hallak W. Machine Learning for Practical Decision Making: A Multidisciplinary Perspective with Applications from Healthcare, Engineering and Business Analytics. Cham, Switzerland. Springer; 2022.
22. 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*. Jul 03, 2019;2(7):e196709. [FREE Full text] [doi: [10.1001/jamanetworkopen.2019.6709](https://doi.org/10.1001/jamanetworkopen.2019.6709)] [Medline: [31268542](#)]
23. Valli I, Marquand AF, Mechelli A, Raffin M, Allen P, Seal ML, et al. Identifying individuals at high risk of psychosis: predictive utility of support vector machine using structural and functional MRI data. *Front Psychiatry*. Apr 08, 2016;7:52. [FREE Full text] [doi: [10.3389/fpsy.2016.00052](https://doi.org/10.3389/fpsy.2016.00052)] [Medline: [27092086](#)]
24. van der Vegt AH, Campbell V, Mitchell I, Malycha J, Simpson J, Flenady T, 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. *J Am Med Inform Assoc*. Jan 18, 2024;31(2):509-524. [FREE Full text] [doi: [10.1093/jamia/ocad220](https://doi.org/10.1093/jamia/ocad220)] [Medline: [37964688](#)]

25. Rubinger L, Gazendam A, Ekhtiari S, Bhandari M. Machine learning and artificial intelligence in research and healthcare. *Injury*. May 2023;54 Suppl 3:S69-S73. [doi: [10.1016/j.injury.2022.01.046](https://doi.org/10.1016/j.injury.2022.01.046)] [Medline: [35135685](#)]
26. 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*. Sep 5, 2022;4:932123. [FREE Full text] [doi: [10.3389/fdgh.2022.932123](https://doi.org/10.3389/fdgh.2022.932123)] [Medline: [36133802](#)]
27. Verma AA, Stukel TA, Colacci M, Bell S, Ailon J, Friedrich JO, et al. Clinical evaluation of a machine learning-based early warning system for patient deterioration. *CMAJ*. Sep 15, 2024;196(30):E1027-E1037. [FREE Full text] [doi: [10.1503/cmaj.240132](https://doi.org/10.1503/cmaj.240132)] [Medline: [39284602](#)]
28. Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med Inform*. Jul 15, 2020;8(7):e15182. [FREE Full text] [doi: [10.2196/15182](https://doi.org/10.2196/15182)] [Medline: [32673244](#)]
29. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Wea Rev*. Jan 1950;78(1):1-3. [doi: [10.1175/1520-0493\(1950\)078<0001:vofeit>2.0.co;2](https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2)]
30. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. [doi: [10.1097/ede.0b013e3181c30fb2](https://doi.org/10.1097/ede.0b013e3181c30fb2)]
31. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *arXiv*. Preprint posted online on March 4, 2017. [FREE Full text] [doi: [10.5260/chara.21.2.8](https://doi.org/10.5260/chara.21.2.8)]
32. Moscovici M, Farrokhi F, Vangala L, Simpson AI, Kurdyak P, Jones RM. Violence risk prediction in mental health inpatient settings using the Dynamic Appraisal of Situational Aggression. *Front Psychiatry*. Dec 10, 2024;15:1460332. [FREE Full text] [doi: [10.3389/fpsy.2024.1460332](https://doi.org/10.3389/fpsy.2024.1460332)] [Medline: [39720430](#)]
33. Ogloff JR, Daffern M. The dynamic appraisal of situational aggression: an instrument to assess risk for imminent aggression in psychiatric inpatients. *Behav Sci Law*. Dec 15, 2006;24(6):799-813. [doi: [10.1002/bsl.741](https://doi.org/10.1002/bsl.741)] [Medline: [17171770](#)]
34. Griffith JJ, Daffern M, Godber T. Examination of the predictive validity of the Dynamic Appraisal of Situational Aggression in two mental health units. *Int J Ment Health Nurs*. Dec 30, 2013;22(6):485-492. [doi: [10.1111/inm.12011](https://doi.org/10.1111/inm.12011)] [Medline: [23363378](#)]
35. 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 Psychiatry*. Aug 2018;52:47-53. [FREE Full text] [doi: [10.1016/j.eurpsy.2018.02.007](https://doi.org/10.1016/j.eurpsy.2018.02.007)] [Medline: [29626758](#)]
36. 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*. Apr 01, 2012;23(2):168-183. [FREE Full text] [doi: [10.1080/14789949.2012.668207](https://doi.org/10.1080/14789949.2012.668207)] [Medline: [25999797](#)]
37. Goldhill DR, White SA, Sumner A. Physiological values and procedures in the 24 h before ICU admission from the ward. *Anaesthesia*. Jun 06, 1999;54(6):529-534. [FREE Full text] [doi: [10.1046/j.1365-2044.1999.00837.x](https://doi.org/10.1046/j.1365-2044.1999.00837.x)] [Medline: [10403864](#)]
38. Steitz BD, McCoy AB, Reese TJ, Liu S, Weavind L, Shipley K, et al. Development and validation of a machine learning algorithm using clinical pages to predict imminent clinical deterioration. *J Gen Intern Med*. Jan 01, 2024;39(1):27-35. [FREE Full text] [doi: [10.1007/s11606-023-08349-3](https://doi.org/10.1007/s11606-023-08349-3)] [Medline: [37528252](#)]
39. Hahn T, Nierenberg AA, Whitfield-Gabrieli S. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol Psychiatry*. Jan 15, 2017;22(1):37-43. [doi: [10.1038/mp.2016.201](https://doi.org/10.1038/mp.2016.201)] [Medline: [27843153](#)]
40. Hansen L, Bernstorff M, Enevoldsen K, Kolding S, Damgaard JG, Perfalk E, et al. Predicting diagnostic progression to schizophrenia or bipolar disorder via machine learning. *JAMA Psychiatry*. May 01, 2025;82(5):459-469. [doi: [10.1001/jamapsychiatry.2024.4702](https://doi.org/10.1001/jamapsychiatry.2024.4702)] [Medline: [39969874](#)]
41. Wolff J, Gary A, Jung D, Normann C, Kaier K, Binder H, et al. Predicting patient outcomes in psychiatric hospitals with routine data: a machine learning approach. *BMC Med Inform Decis Mak*. Feb 06, 2020;20(1):21. [FREE Full text] [doi: [10.1186/s12911-020-1042-2](https://doi.org/10.1186/s12911-020-1042-2)] [Medline: [32028934](#)]
42. 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*. 2018. Presented at: AIES '18; February 2-3, 2018; Orleans, LA. URL: <https://dl.acm.org/doi/10.1145/3278721.3278779> [doi: [10.1145/3278721.3278779](https://doi.org/10.1145/3278721.3278779)]
43. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM J Res Dev*. Jul 1, 2019;63(4/5):4:1-15. [doi: [10.1147/JRD.2019.2942287](https://doi.org/10.1147/JRD.2019.2942287)]
44. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017. Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA. URL: <https://dl.acm.org/doi/10.5555/3295222.3295230>

Abbreviations

AUC: area under the receiver operating characteristic curve

DASA-IV: Dynamic Appraisal of Situational Aggression–Inpatient Version

EMR: electronic medical record

EWS: early warning system

LSTM: long short-term memory

ML: machine learning

PRIME: Predictive Risk Identification for Mental Health Events

Edited by J Torous; submitted 29.Sep.2025; peer-reviewed by H Ryland, SCL Au; comments to author 17.Oct.2025; revised version received 10.Dec.2025; accepted 11.Dec.2025; published 06.Feb.2026

Please cite as:

Dolatabadi E, Tamayo Velasquez V, Dabboussi AH, Wen D, Crawford J, Waddell AE, El Morr C

Advancing Psychiatric Safety With the Predictive Risk Identification for Mental Health Events Tool: Retrospective Cohort Study
JMIR Ment Health 2026;13:e84318

URL: <https://mental.jmir.org/2026/1/e84318>

doi: [10.2196/84318](https://doi.org/10.2196/84318)

PMID:

©Elham Dolatabadi, Valentina Tamayo Velasquez, Abdul Hamid Dabboussi, David Wen, Jennifer Crawford, Andrea E Waddell, Christo El Morr. Originally published in JMIR Mental Health (<https://mental.jmir.org>), 06.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Mental Health, is properly cited. The complete bibliographic information, a link to the original publication on <https://mental.jmir.org/>, as well as this copyright and license information must be included.