BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Predicting child and adolescent mental health emergency department revisits: a machine-learning approach compared to a clinician-derived baseline

Navjot Kaur Bians[1], Joonsoo Sean Lyeo[1,2], Jeff Gilchrist[1], Christina Honeywell[1,3], Paula Cloutier[1], Allison Kennedy[1,3,4] and Kathleen Pajer[1,3,5*]

Abstract

Background Predicting child and youth mental health (CYMH) emergency department (ED) revisits (RVs) is critical for improving patient outcomes and optimizing use of resources. Previous CYMH ED RV studies have used statistical methods with research cohorts and produced varying results. Our aims were to develop a predictive algorithm incorporating machine learning (ML) with electronic health records (EHR) and validate it against a clinician-driven algorithm in a proof of concept project.

Methods Data were retrospectively collected from a tertiary care pediatric hospital's EHR from November 2017–November 2023, yielding 12,700 ED encounters from 8,696 patients, 8–18 years of age. The feature set comprised patient demographics, visit-level variables, laboratory results, procedure codes, and medication records. A mapping of 230 International Classification of Diseases (ICD)-10 codes into 28 Diagnostic and Statistical Manual (DSM)-5 categories was performed and a logistic regression (LR) ML model developed. Both tasks used clinical expert input. Seven clinical experts then independently assigned weights to 191 variables using a custom-designed application to create a structured clinician-weighted baseline for comparison to the ML algorithm. Both models were evaluated using AUROC and F1 score as primary metrics with precision and recall as secondary. LR coefficients and odds ratios were the primary interpretability outputs, while SHapley Additive exPlanations (SHAP) were used for supplementary visualization across four age strata.

Results The LR machine learning model achieved an AUROC of 0.78, outperforming the structured clinician-weighted baseline (AUROC range: 0.54–0.64) Detailed analysis revealed that predictors such as past ED RV count, psychotherapeutic medication history, substance use history, and prior outpatient MH visits were consistently influential.

Conclusions This proof of concept project demonstrates that ML can provide complementary, clinically interpretable predictions of CYMH ED RV. Alignment between model-derived predictors and clinician-weighted features supports

*Correspondence:
Kathleen Pajer
kpajer@cheo.on.ca

Full list of author information is available at the end of the article

interpretability and lays a foundation for further development. Future steps include enhancing sensitivity, expanding feature sets, and conducting prospective silent-mode validation to refine performance.

Clinical trial registration Not applicable.

Keywords Emergency department revisits, Machine learning, Mental health, Precision health, Precision mental health, Children, Youth

Background

Presentation rates to emergency departments (EDs) for child and youth mental health (CYMH) care grew 2-5-fold from the mid-2000s until the COVID-19 pandemic [1–5]. Rates then dropped from 2020 to 2022 during the pandemic [6, 7], but not as much as non-CYMH visits [6, 7]. In the most up to date publication with data from 2023, Valtuille and colleagues [8] reported that in France, the CYMH pandemic-related decline in ED presentations was followed by a rapid increase in rates from 2022 to 2023, especially for eating disorders, sleeping problems, and intentional self-harm.

Unfortunately, it's difficult to provide most CYMH-disordered patients with the care they need in an ED setting. There are too few CYMH-trained staff, inadequate space, insufficient time to address mental health (MH) problems, and lack of discharge options [9, 10]. Moreover, while a relatively small percentage of pediatric ED patients are seeking MH care, as many as 47% of CYMH patients are revisits (RVs) [4, 11–13], with 13–27% of them returning within six months [2, 11–15]. It is crucial to address this problem, as CYMH return patients often have adverse health outcomes, e.g. higher rates of suicide, self-harm, and overall mortality [3] and add to the swelling demand for care, while concurrently straining scarce resources and space in already burdened EDs [16].

Predicting which CYMH patients will return for care is the first step in designing prevention programs. But predicting RVs is complicated. Leon and colleagues [15] reported from in systematic review that CYMH RVs were characterized by myriad demographic, clinical, care access, and utilization factors. Older age, female sex, lower socioeconomic status, and ethnic minority status were associated with higher rates of return, but traditional measures of clinical severity or triage-based classifications were not. In contrast to conventional wisdom that connecting patients to outpatient CYMH services will reduce RVs, being involved in such services was associated with higher rates of RVs. In fact, being in treatment was one of the strongest predictors of return within six months.

In a more recent review in 10–24-year-olds, Wilson et al. [17] analyzed over 60 studies. They confirmed MH outpatient service usage as a predictor of higher RVs but found no evidence for sex or specific MH disorders. Although behavioral disorders had higher rates of RVs, they concluded that no key factors were uniformly predictive.

Supporting and extending these findings, Cushing and colleagues examined RV data as recent as 2020 from 38 U.S. hospital EDs. In the cohort of 308,264 patients with CYMH ED visits, age, discharge status, and diagnosis were not associated with RVs [2]. But factors such as the presence of multiple psychiatric comorbidities, use of chemical restraints in the ED, having public insurance, decreased access to neighborhood resources, and inpatient psychiatric unit admission were all associated with RVs within 6 months [2]. Unexpectedly, patients with substance use disorders were less likely to return.

Collectively, this body of research indicates that predicting who will return for ED CYMH care is difficult from studies using standard methodologic approaches. A useful alternative may be the use of artificial intelligence (AI) with electronic health records (EHR) to develop predictive algorithms. Machine learning (ML) has shown promise in better predicting adult non-psychiatric disorder ED RVs than standard methods [18, 19], but its application in CYMH ED RVs has only been tried in one study [20].

Using ML to predict ED RVs has traditionally been approached using structured ML models, relying on structured feature, or predictor variable selection, guided by statistical filtering and domain expertise, or deep learning models, which aim to learn patterns from high-dimensional EHR data, reducing reliance on manual feature selection. While deep learning has shown promise in healthcare applications, its use in CYMH RV prediction remains limited due to interpretability challenges, high data demands, and computational constraints. A study by Saggu et al. [20] applied deep learning techniques, a graph neural network, to predict CYMH ED RVs and demonstrated improved predictive performance over traditional models. However, the study also highlighted ongoing challenges in clinical applicability in the interpretation of model outputs.

Despite advancements in ML, a key challenge in clinical adoption is the lack of a meaningful benchmark for evaluating model utility. Without a clinician-derived model to serve as a comparison, or baseline, it is unclear whether ML predictions provide real clinical value or merely identify associations [21]. To our knowledge, no previous studies have sought to syncretize the use of

clinician-selected variables with machine learning to predict ED RVs, although this approach may be fruitful in increasing clinical utility of ML algorithms.

Therefore, we conducted a proof of concept study using both a clinical expertise and a data-driven ML method to predict CYMH ED RVs. Our aims were to use EHR data to:

1. Establish a clinician-driven predictive algorithm based on clinician-assigned weights for the EHR features, which could then serve as a benchmark for ML performance evaluation.
2. Develop an ML algorithm to predict 6-month CYMH ED RVs.
3. Compare performance of these two prediction methods based on area under the receiving operating characteristic curve (AUROC) analyses and feature importances.

Methods

Setting and data

We used EHR records from the Children's Hospital of Eastern Ontario (CHEO), a free-standing pediatric tertiary care hospital in Ontario, Canada. An academic health sciences centre affiliated with the University of Ottawa, CHEO serves patients from eastern Ontario, southwestern Quebec, and Nunavut.

The dataset was obtained from the CHEO MH Data Mart. A CYMH patient was defined as anyone from 0 to 18 years of age with at least one of the following International Classification of Diseases (ICD)-10 [22] group of codes recorded in their EHR: ('F05–F05.00', 'F05.8–F51.9', 'F53–F53.9', 'F54–F99', 'R41–R41.88', 'R44–R49.8', 'R78–R78.9'). These were selected by 9 child psychiatrists to reflect codes characteristic of psychiatric or CYMH disorders, each blind to the choices of each other. ED encounters from November 1, 2017 to November 30, 2023 were included, comprising 31,758 MH patients and their 98,119 ED encounters. Age and sex were the only demographic data available.

This study was submitted to the Children's Hospital of Eastern Ontario Research Ethics Board (REB) (CHEO-REB# 22/11X), which reviewed and approved it, with the requirement to seek consent waived by the REB, in accordance with TCPS2 Article 5.5 A, as this was a retrospective study of de-identified EHR records.

Feature logic

For the purposes of predicting RVs, encounters were labeled as "Revisit" and "No Revisit". An encounter was labelled as "Revisit" only if the same patient had a subsequent ED encounter within six months. If there was no further encounter within six months, it was labelled as "No Revisit." Additionally, for the patients' final encounters within the date range (November 1, 2017 to November 30, 2023), they were labelled "No Revisit" if there were data available for the next 6 months from the date of their final encounter. For example, if a patient's last encounter was in April or May 2023, and we have data available up to November 2023, the patient is considered to be "No Revisit".

Final encounters without data for the following six months from the final encounter date were removed. For example, if a patient's last encounter occurred in July or August 2023, and no data existed for the following 6 months, these encounters were excluded because the cut-off for which we had data was November 30, 2023, which meant we couldn't confirm whether the patient did or did not RV after that encounter. Our choice of a 6-month follow-up period reflects the convention seen in the broader literature on CYMH ED RVs [16, 23, 24] and the observation [12] and [25] that 65–85% of CYMH ED patients have a RV within six months.

Data filtering

To ensure the dataset's quality and relevance for predicting CYMH ED RVs, several filtering steps were applied, shown in Fig. 1. Encounters without MH-related ICD-10 codes as the primary or most responsible diagnosis were excluded. Geographic restrictions were also applied, limiting the dataset to patients residing in Ontario or the neighboring province of Quebec, included because they present to the CHEO ED.

Age-based filtering was implemented to exclude patients younger than 8 years and older than 18 years. Younger patients were removed based on a synthesis of clinician feedback, data indicating low rates of CYMH RVs, and ICD-10 codes or primary complaints different than those seen in 8 years and older. Patients 19 years and older were excluded, as they no longer qualified for treatment at the pediatric hospital and were only admitted in unusual circumstances, e.g., during the pandemic when adult hospitals were running over census, a local or regional emergency with overflow to our hospital, or those with significant neurodevelopmental delay for whom it is difficult to find post-pediatric care.

Additionally, encounters associated with R41 ICD-10 codes, which represent the transient category of 'other symptoms and signs involving cognitive functions and awareness', were excluded to improve the dataset's focus on persistent and clinically relevant MH disorders. This decision was informed by prior expert reviews, indicating that R41 cases were non-representative of the CYMH population. Patients with more than four ED RVs were also excluded to minimize the influence of extreme RV frequencies, which could disproportionately affect model performance. Finally, geographic subregions with fewer than seven patients with RV history were excluded to

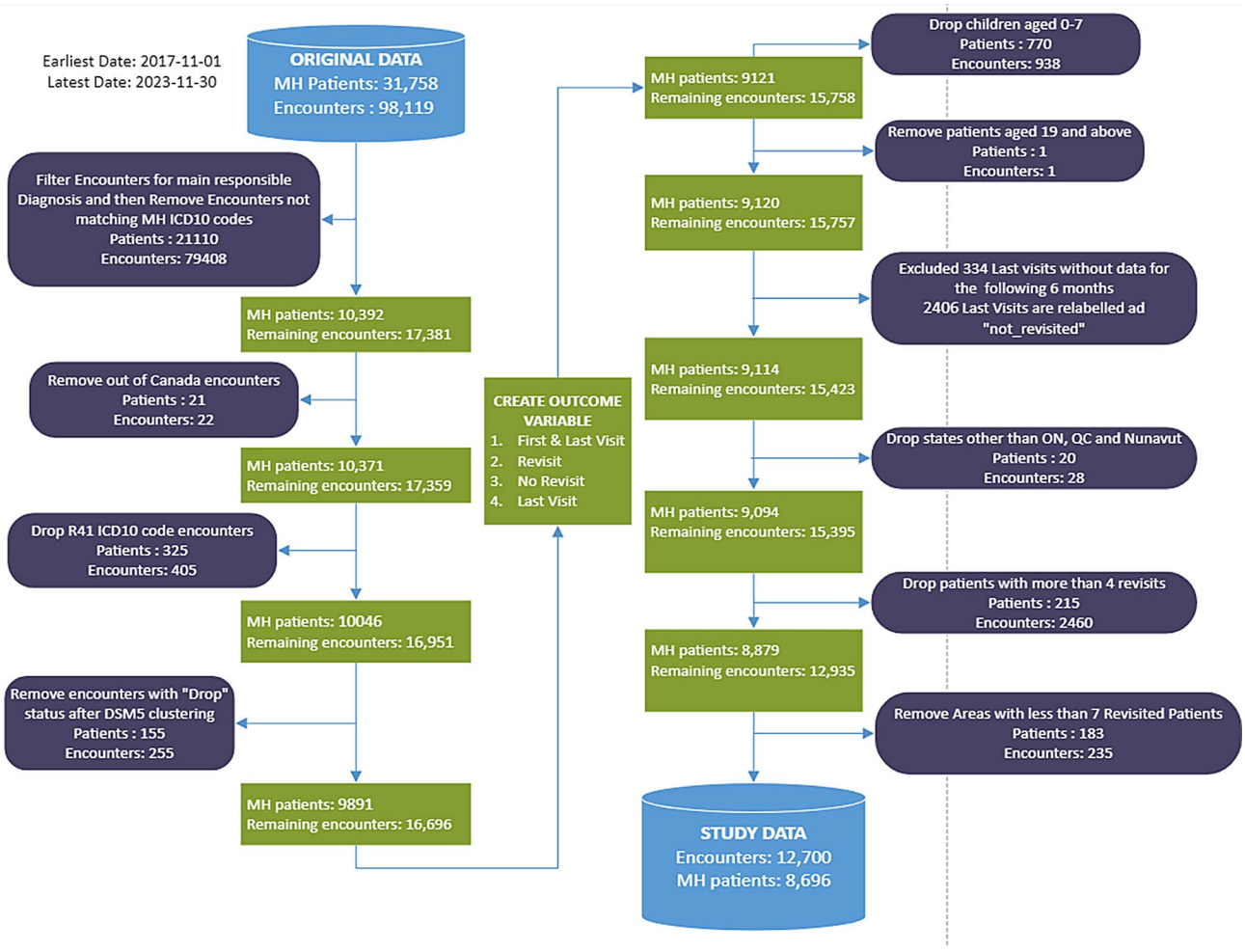

Fig. 1 Data filtering and refinement steps applied to create the final study cohort

ensure sufficient representation across geographic areas and reduce statistical noise. Application of all these filtering steps resulted in a final cohort of 12,700 encounters from 8,696 patients (22% with RVs, 78% with no RVs).

The final dataset comprised the following categories of features drawn from multiple data domains:

- Patient-level variables: Age, sex, living status, address, and MH history.
- Visit-level variables: Primary complaints, ICD-10 codes, the Canadian Triage and Acuity Scale, and arrival method, e.g., ambulance, police.
- Lab procedures: Names/codes of lab tests ordered during the encounter.
- Lab procedure results: Outcomes of ordered lab tests.
- Procedures: Names/codes of non-lab procedures performed or suggested during the encounter.
- Medications: Therapeutic names/codes of drugs prescribed during an encounter.

Feature engineering

To capture clinically meaningful features and reduce the dimensionality of ICD-10 diagnostic codes, 230 codes were mapped by expert psychiatrists into 28 Diagnostic and Statistical Manual (DSM)-5 [26] categories. Two additional categories ("Single Behaviors" and "Single Feeling Symptoms") were created to account for ICD-10 codes representing isolated behavioral or emotional presentations not delineated in DSM-5. This approach balanced interpretability by aligning codes with clinically recognized categories while retaining sufficient granularity for predictive modeling.

In addition, a "Past Revisit Count" feature was generated to quantify historical ED RVs. Binary indicators were also created to denote whether labs were ordered, procedures were performed, or medications were administered during each visit, including discharge medications. Furthermore, medication counts were further stratified by therapeutic class, e.g., psychotherapeutic drugs, antiparkinsonian drug, vitamins to provide granular insight into pharmacological interventions received.

Feature selection and preprocessing

Although mapping ICD-10 codes to DSM-5 categories reduced the number of features, nearly 200 of these features remained (Supplementary Material 1). To identify redundant features, pairwise chi-squared tests were used to evaluate relationships between categorical features, followed by Cramer's V for all significant findings to measure the effect sizes. Most of the features exhibited weak correlations with the outcome, with the notable exception of "Past Revisit Count" which demonstrated a considerably stronger association. Weak correlations are common in healthcare datasets, as outcomes of interest, like RVs, are rarely driven by a single factor. Instead, they are often influenced by the complex interactions between clinical, procedural, and contextual variables [27].

Many features in the dataset were highly correlated with each other, which can cause problems such as multicollinearity. Since correlation only shows patterns and does not prove cause-and-effect, we used feedback from clinical experts to decide which variables would be most useful in practice. When two features were strongly correlated, the less relevant one was removed based on expert judgment. This process resulted in a final set of 30 features for the logistic regression model (Table 1).

Table 1 Constructs and variables included in the analysis

Construct	Variable Names
Demographics	• 8-to-10-year age group • 11-to-13-year age group • 14-to-15-year age group • 16-to-18 year age group • Sex_Male • StateOrProvince_Quebec
Primary Care Doctor	• HasPcp
Mental Health Problem History	• IsSomatoformChiefComplaint • HasPositiveAsqValue • HasPsychoTherapeuticMedHx • HadFeedingProblemFailureToThrive • Hx HadSubstanceUseHx • HadPastOpMhVisits
Diagnostic and Statistical Manual 5	• anxiety • depressive • substance_addiction
Medical Procedures	• procedure_ecg • procedure_consult • procedure_outpatient_referral • procedure_mental_health_services
Disposition	• IsAdmittedFromEd • Inpatient • EdDisposition_Discharge
Medication Received During Visit	• med_psychotherapeutic_drugs_count • med_elect/caloric/h2o_count • med_antiparkinson_drugs_count • med_antihistamines_count • med_diagnostic_count • med_vitamins_count
Revisit and Outcome	• past_revisit_count • outcome

As can be seen in Table 1, the final set of features comprised continuous variables, e.g., *past_revisit_count* and *Medication Count* during encounters and categorical variables. To prepare these features for modeling, standard scaling was applied to numerical features to normalize their distributions and one-hot encoding was used to convert categorical variables into binary columns, ensuring they were numerically represented without introducing bias. A column transformer was used to integrate both transformations into a unified preprocessing pipeline, ensuring compatibility with the ML model.

Model development

Clinician driven rule-based model

The study employed a dual-model strategy. In machine learning, a baseline model serves as a reference point for evaluating more complex models and determining if the added complexity is beneficial. To establish a structured clinician-weighted baseline, seven clinicians with expertise in pediatric ED mental health care participated, and were compensated for their time.

Each expert was provided with detailed instructions and a pilot-tested custom-designed application to facilitate the weight assignment process. The dataset comprised 191 features with concise descriptions, and experts were instructed to assign weights to these features ranging from −3 (a very low likelihood of predicting ED RVs) to +5 (a very high likelihood of predicting ED RVs), i.e., the two ends of the range were to be regarded as equal indicators of the extremes. The wider positive range only was created to provide finer granularity for marking features associated with increased risk. To avoid potential misinterpretation, the weights were normalized to ensuring symmetry between risk-reducing and risk-increasing features. This design did not imply disproportionate importance of positive over negative weights but was intended to capture the greater clinical relevance of distinguishing degrees of increased risk.

Detailed guidelines and examples were provided during the training phase, ensuring that each expert's interpretation of clinical importance was standardized as much as possible. Independent assessment by each expert minimized group consensus bias. Upon completion of the weight assignment, the weights were normalized to a balanced ±3 scale, then the application computed a weighted sum of the retained features for each patient encounter. These weighted sums were then passed through a nonlinear activation function (tanh) to generate continuous risk scores.

For evaluation, we systematically explored thresholds across the score distribution to classify encounters as 'revisit' or 'no revisit,' calculating accuracy, precision,

recall, specificity, and F1 score at each threshold. The threshold with the highest F1 score was selected as the model's operating point, and AUROC was computed by considering performance across all thresholds. Each expert's weights were used to generate a separate clinician-derived model, resulting in seven distinct models. These scores were then validated against the true outcome labels to evaluate each model's performance.

Logistic regression model and validation

In parallel with the clinician-driven approach, an LR model was developed using a curated set of 30 features. These were derived from the initial pool of 191 features through the feature selection and preprocessing steps described above, ensuring that the ML process began from the same broad feature set as the clinician-driven model. It was trained with balanced class weights to improve the detection of minority-class instances. LR offers several advantages, including interpretability of coefficient estimates and a straightforward mechanism for regularization. In this study, L2 regularization was applied to mitigate overfitting, shrinking rather than eliminating coefficients to maintain transparency about each feature's contribution [28, 29]. A structured ML model was deemed the preferred approach for this study, because it better balances predictive performance with transparency and clinical interpretability [30].

To rigorously evaluate and tune the model, the data were split into an 80% training subset and a 20% test set. On the 80% training data, we conducted 5-fold nested cross-validation: (1) an inner loop with GridSearchCV optimized hyperparameters (penalty type and regularization strength) and (2) an outer loop that assessed generalization performance using unseen folds. The best-performing hyperparameters were then used to retrain the model on the entire training set.

Evaluation was conducted on the reserved 20% test set. For interpretability, we report global logistic regression coefficients and odds ratios with 95% confidence intervals as the primary measures.

We also conducted an exploratory analysis of age-related variation. The test set was stratified into four age groups (8–10, 11–13, 14–15, and 16–18 years), and SHapley Additive exPlanations (SHAP) [31] values used as a supplementary visualization to illustrate how the global model's feature contributions distribute across age strata.

The performances of both the clinician-driven and logistic regression models were evaluated on the test set using AUROC, F1, precision, and recall. AUROC and F1 were prioritized as robust measures for imbalanced datasets, while precision and recall provided complementary insights into false positive and false negative behavior. Definitions and clinical interpretations of the metrics are summarized in Table 2. All analyses were conducted in Python using scikit-learn [32] and other standard libraries. For threshold-dependent metrics (precision, recall, F1), we applied the default probability cutoff of 0.5 as implemented in scikit-learn.

Table 2 Performance evaluation metrics and definitions

Performance metric	Definition
Area Under the Receiving Operator Characteristic Curve (AUROC)	Quantifies the algorithm's overall ability to distinguish between revisiting and non-revisiting patients across all possible prediction thresholds.
Precision	Proportion of patients predicted to revisit that do revisit. High precision means a low rate of false positives (Hand et al., 2021).
Recall (Sensitivity)	Proportion of those who did revisit and were correctly identified by the model. High recall indicates a low rate of false negatives, i.e., fewer actual revisits are missed.
F1-Score	The harmonic mean of precision and recall. It provides a single score that balances both precision and recall, when there is an uneven distribution between classes (revisiting and non-revisiting patients) ensuring a balanced performance in identifying revisits while maintaining the reliability of those predictions).

Results

Dataset overview and train-test splitting

The final dataset consisted of 12,700 ED encounters from 8,696 pediatric patients aged 8–18 years, with 22% of encounters classified as RVs and 78% as non-RVs. To ensure balanced representation between train/test data and to minimize sampling bias, we assessed the distribution of all finally selected features across the full dataset, training, and testing subsets. The age-stratified train-test split further reinforced model robustness, allowing for a structured evaluation of RV risk across pediatric developmental stages while maintaining generalizability to broader clinical populations.

Table 3 presents a comparative profile of patient characteristics, like medical history and demographics, for different age groups (8–10, 11–13, 14–15, and 16–18), highlighting how these characteristics differ across age ranges. This table also confirms that the distribution of these characteristics is consistent between the training and testing datasets, ensuring that machine learning models built on these data can be reliably evaluated.

Model performance

Performance comparison: clinician-derived vs. logistic regression

In medical decision-making, the AUROC serves as a widely accepted measure of a model's ability to distinguish between two outcomes, here: RV or no RV for MH

Table 3 Comparison of *filtered dataset* (1), *training dataset* (2), and *test dataset* (3) patient features by age group

Characteristic	8 to 10 YOA Types of Datasets			11–13 YOA Types of Datasets			14–15 YOA Types of Datasets			16–18 YOA Types of Datasets		
	1	2	3	1	2	3	1	2	3	1	2	3
Male	53.1%	53.2%	52.9%	29.9%	29.9%	29.9%	33.7%	33.4%	34.6%	36.2%	36.2%	36.3%
From Quebec	8.3%	8.2%	8.7%	6.8%	7.1%	5.7%	4.5%	4.6%	4.2%	4.0%	4.1%	3.5%
Past ED Revisit Count	4.9%	5.0%	4.7%	12.4%	12.8%	10.7%	13.7%	13.7%	13.9%	11.7%	11.2%	13.6%
Primary Care Provider	98.6%	98.9%	97.5%	98.7%	98.7%	98.9%	97.8%	97.7%	98.2%	97.2%	97.3%	96.8%
Positive ASQ	3.7%	3.2%	5.4%	49.7%	50.2%	47.8%	48.9%	48.3%	51.3%	45.0%	45.1%	44.8%
Psychotherapeutic Med History	58.4%	58.5%	58.0%	70.2%	70.6%	68.8%	72.7%	72.2%	74.6%	71.2%	71.7%	68.8%
Feeding/Failure to Thrive History	9.5%	9.1%	10.9%	15.9%	15.9%	15.9%	14.5%	14.7%	13.7%	11.8%	12.3%	10.2%
Substance Use History	1.9%	2.0%	1.8%	12.4%	12.3%	12.7%	23.6%	23.3%	24.4%	29.2%	28.7%	31.2%
Past Outpatient Mental Health Visits	58.5%	58.6%	58.0%	68.7%	68.9%	67.7%	58.4%	57.6%	61.2%	50.7%	50.8%	50.6%
Somatoform Complaint	8.5%	8.5%	8.3%	7.5%	7.3%	8.1%	6.5%	6.6%	6.0%	6.3%	6.0%	7.4%
Anxiety Complaint	29.2%	29.5%	28.3%	26.1%	26.2%	26.1%	24.0%	24.0%	24.0%	23.6%	24.0%	21.9%
Depression Complaint	14.0%	13.9%	14.1%	34.1%	34.4%	32.8%	38.6%	38.4%	39.4%	36.2%	36.0%	37.0%
Substance Addiction Complaint	1.0%	1.0%	1.1%	4.0%	3.9%	4.4%	11.1%	11.1%	10.7%	14.9%	14.9%	15.2%
Admitted from ED	4.6%	4.6%	4.7%	13.1%	13.0%	13.4%	16.6%	17.0%	15.3%	19.0%	19.3%	17.8%
Inpatient Medication	11.9%	11.8%	12.3%	14.8%	14.7%	15.5%	18.0%	18.2%	17.2%	23.1%	23.1%	23.3%
Psychotherapeutic Med Count	10.4%	10.2%	11.6%	13.1%	13.4%	11.8%	15.8%	15.6%	16.3%	19.0%	19.3%	17.8%
Electrolyte/Caloric/Hydration Med Count	4.5%	3.8%	7.2%	10.8%	10.9%	10.6%	16.0%	15.9%	16.5%	21.4%	21.3%	22.2%
Antiparkinson Med Count	0.1%	0.2%	0.0%	0.3%	0.3%	0.4%	0.1%	0.1%	0.1%	0.7%	0.7%	0.7%
Antihistamines Med Count	1.0%	1.1%	0.7%	0.8%	0.8%	1.1%	0.8%	0.9%	0.6%	1.1%	1.3%	0.7%
Diagnostic Med Count	0.2%	0.1%	0.7%	0.6%	0.6%	0.5%	0.7%	0.8%	0.3%	0.6%	0.6%	0.7%
Vitamins Med Count	1.3%	1.3%	1.4%	2.1%	2.0%	2.5%	2.3%	2.3%	2.0%	3.2%	3.2%	3.2%
ED Disposition: Discharge	91.9%	91.9%	92.0%	83.5%	83.9%	82.1%	78.5%	78.1%	80.0%	75.0%	75.0%	75.1%
ECG Procedure	5.9%	5.9%	6.2%	12.2%	11.5%	14.8%	13.2%	13.4%	12.4%	14.3%	13.8%	16.2%
Consultation Procedure	7.6%	7.6%	7.6%	14.6%	14.7%	14.3%	16.4%	16.5%	16.0%	19.4%	20.3%	15.5%
Outpatient Referral	3.5%	3.9%	1.8%	3.3%	3.1%	3.9%	3.3%	3.4%	3.0%	3.7%	3.5%	4.4%
Mental Health Services Referral	0.0%	0.0%	0.0%	0.4%	0.4%	0.4%	1.0%	1.0%	0.8%	1.1%	1.3%	0.2%
Revisited Patient	15.5%	15.7%	14.9%	23.8%	24.0%	22.7%	23.9%	23.7%	24.8%	20.7%	20.6%	21.0%

Table 4 Performance of clinician-derived baseline compared to logistic regression model

Metric	Clinician-Driven Baseline							Logistic Regression Model
	Experts							
	1	2	3	4	5	6	7	
AUROC	0.577	0.615	0.541	0.638	0.627	0.601	0.610	0.78
Precision	0.26	0.29	0.24	0.33	0.34	0.37	0.36	0.44
Recall	0.85	0.74	0.87	0.64	0.54	0.39	0444	0.59
F1-Score	0.39	0.42	0.37	0.44	0.42	0.38	0.39	0.51

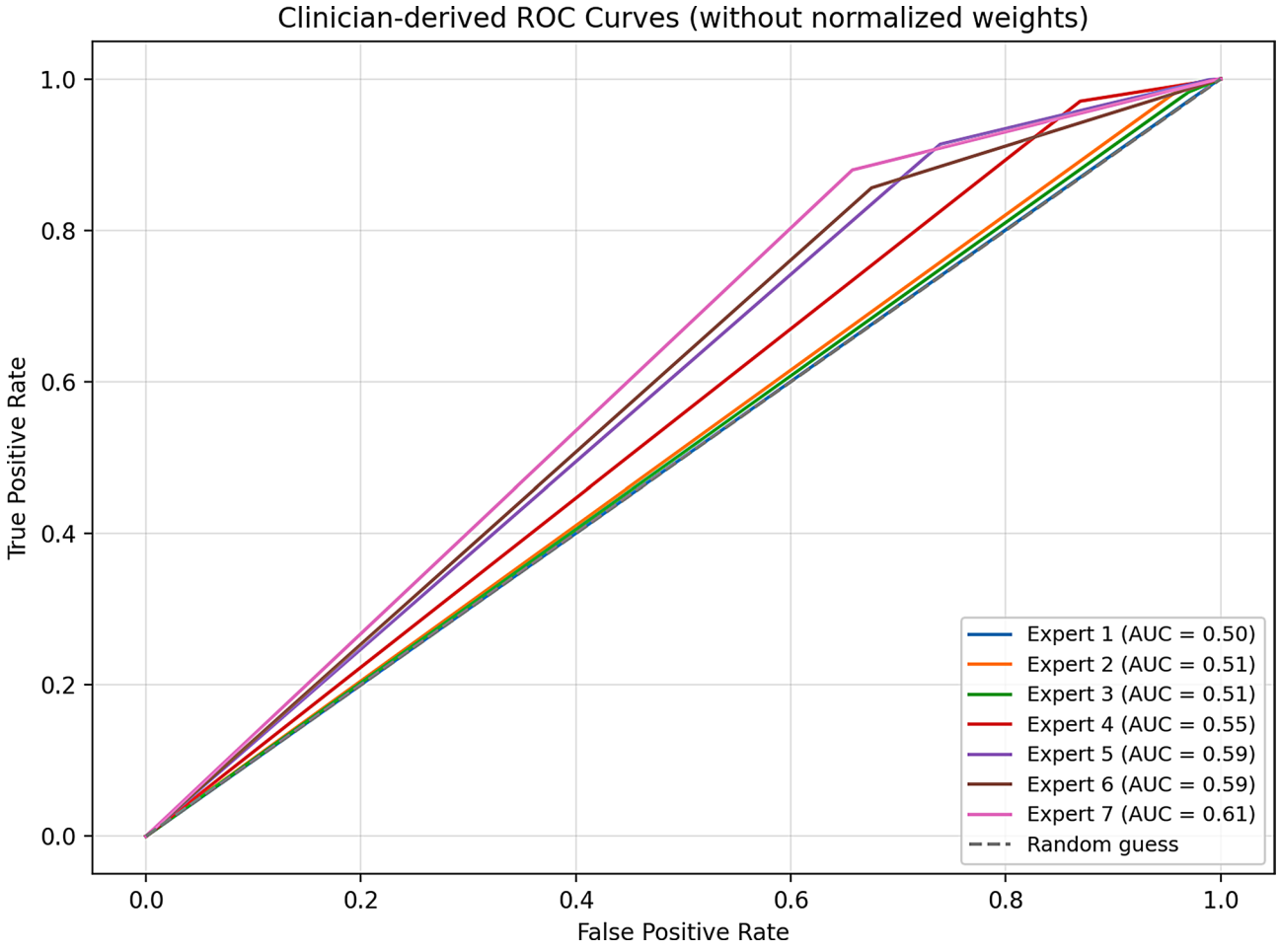

Fig. 2 ROC curves for the seven clinician-derived models (Experts 1–7) generated without weight normalization

concerns. An AUROC of 0.5 indicates performance at the level of random guessing, offering no better predictive value than a coin flip. Scores below 0.5 suggest performance that is in effect, worse than random. By contrast, as the AUROC approaches 1.0, the model demonstrates increasingly perfect discriminatory power.

Table 4 contrasts the performance metrics of our two approaches when applied to the test dataset. The clinician-derived model yielded modest AUROCs, ranging from 0.54 to 0.64 (across seven raters), indicating modest but above-chance discrimination. Despite relatively high recall (0.39–0.87) the models had limited precision (0.24–0.37), producing a substantial number of false positives. F1 scores ranged from 0.37 to 0.44, underscoring the trade-off between capturing most revisit cases and misclassifying non-revisits. Figures 2 and 3 present the ROC curves of the seven clinician-derived models. Figure 2 shows results generated without weight normalization, while Fig. 3 displays results with weight normalization applied.

In contrast, the LR model achieved an AUROC of 0.78, well above random guessing and reflective of good discriminative capacity. Recall was 0.59, meaning that approximately 60% of true revisit cases were correctly identified, while precision was 0.44, indicating that just under half of flagged cases were true revisits. Although

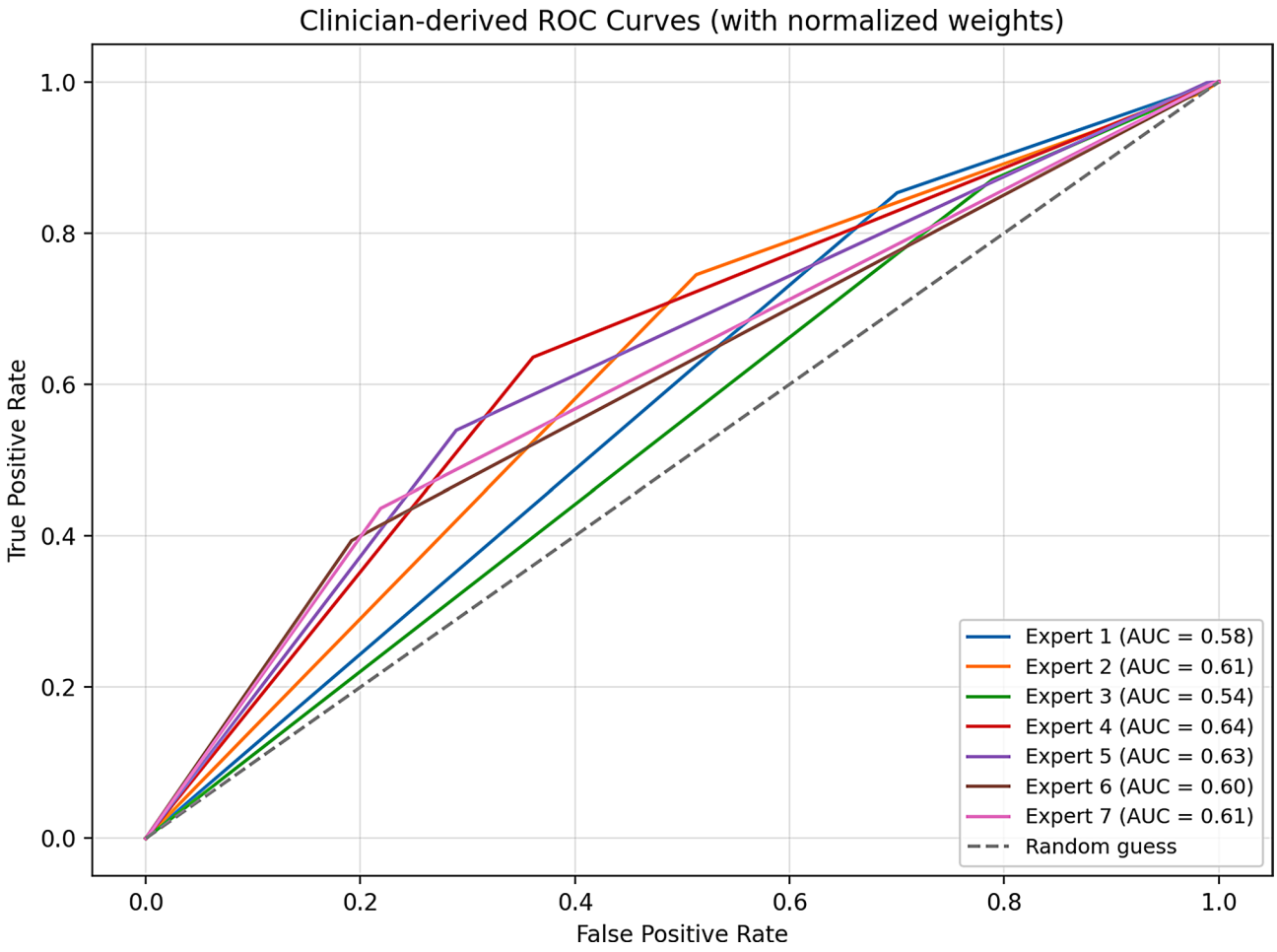

Fig. 3 ROC curves for the same clinician-derived models (Experts 1–7) after applying weight normalization to balance the positive and negative scales

precision was modest, this trade-off enabled the model to capture most revisit cases, a meaningful advance over the current situation in the ED, i.e., no systematic risk stratification.

The AUROC of 0.78 further demonstrates that, overall, the model discriminates revisit risk substantially better than chance. Importantly, these results highlight recall as the good dimension of model performance, which is particularly meaningful in the context of post-discharge follow-up initiatives where false positives translate to additional outreach, while false negatives represent missed opportunities to intervene. To assess the generalizability of these results, we performed five-fold cross-validation, with detailed results reported in Supplementary Material 2. We further compared model performance with and without class weighting. Without class weighting (results in Supplementary Material 3), the model achieved slightly higher AUROC (0.795 vs. 0.789) and specificity (0.950 vs. 0.799), along with improved precision (0.680 vs. 0.454). However, this came at the cost of lower recall (0.470 vs. 0.592), meaning that nearly half of revisit cases were missed. With class weighting, recall improved to ~60%, identifying most children who returned to the ED, but with more false positives. Given that the current model is best suited for post-discharge applications, recall was prioritized over precision, and balanced class weights were selected as the more clinically meaningful operating point.

Exploratory age-based logistic regression model performance analysis

To assess how model performance varies by developmental stage, the test dataset was stratified across four age groups. Across all groups, the LR model maintained strong generalizability, with AUROC values ranging from 0.74 to 0.80. Performance was highest in the 11–13 age group (AUROC = 0.80). AUROC values for 14–15 (0.77) and 16–18 years (0.78) were consistent with the overall model performance, indicating that predictive features remain relatively stable in older adolescents. In contrast, the 8–10 age group exhibited the lowest AUROC (0.74), suggesting slightly greater variability in RV risk factors, potentially influenced by caregiver decision-making and a higher prevalence of somatic complaints.

Figure 4 shows the ROC curves for each age group, confirming that predictive accuracy remains robust across all pediatric stages, with minor variations in discrimination ability.

Feature importance analysis

Clinical expert feature importance and consensus

We compared the weight assignments provided independently by seven ED MH clinicians to examine areas of consensus and disagreement (Supplementary Material 4). The number of past ED revisits, history of psychotherapeutic medication use, history of substance use, and substance addiction as a presenting complaint were most frequently assigned positive weights, reflecting their perceived importance in predicting revisits. In contrast, variables such as male sex, having a primary care provider (PCP), and referrals to outpatient care were commonly assigned negative weights, suggesting these factors were viewed as protective. Additionally, a few variables, such as number of antiparkinsonian medications administered and number of electrolyte or hydration therapies given, were consistently assigned zero weights, indicating consensus that they carried little to no predictive value in this context.

Logistic regression-derived feature importance

Odds ratios with 95% confidence intervals (Supplementary Material 5) confirmed that past ED revisit count (OR = 8.52, 95% CI: 7.97–9.40, $p < 0.001$) was the dominant predictor, with an effect size far greater than any other feature. Other significant risk factors included psychotherapeutic medication history (OR = 1.70, $p < 0.001$), substance use history (OR = 1.34, $p < 0.001$), past outpatient mental health visits (OR = 1.24, $p < 0.001$), and feeding/failure to thrive history (OR = 1.12, $p = 0.015$). Patients from Quebec (OR = 0.76, $p < 0.001$) were less likely to revisit, consistent with differences in follow-up care pathways due to Quebec's healthcare system. Substance addiction as a presenting complaint (OR = 0.67, $p < 0.001$) also showed a negative direction, likely reflecting that these patients are more frequently admitted or diverted to external addiction pathways rather than returning to the ED. Several other features, such as admission from the ED and inpatient medication use, showed borderline associations but did not reach statistical significance. The majority of remaining features including age-group indicators, primary care provider status (OR = 1.07, 95% CI: 0.91–1.32), somatoform complaint (OR = 1.08, 95% CI: 0.96–1.20), anxiety complaint (OR = 0.96, 95% CI: 0.88–1.02), vitamin medication use (OR = 0.95, 95% CI: 0.87–1.02), and electrolyte/hydration medications (OR = 0.97, 95% CI: 0.89–1.04) had odds ratios close to 1.0 with confidence intervals spanning the null value. These findings indicate limited predictive utility and suggest that these variables do not meaningfully differentiate revisit risk in this dataset.

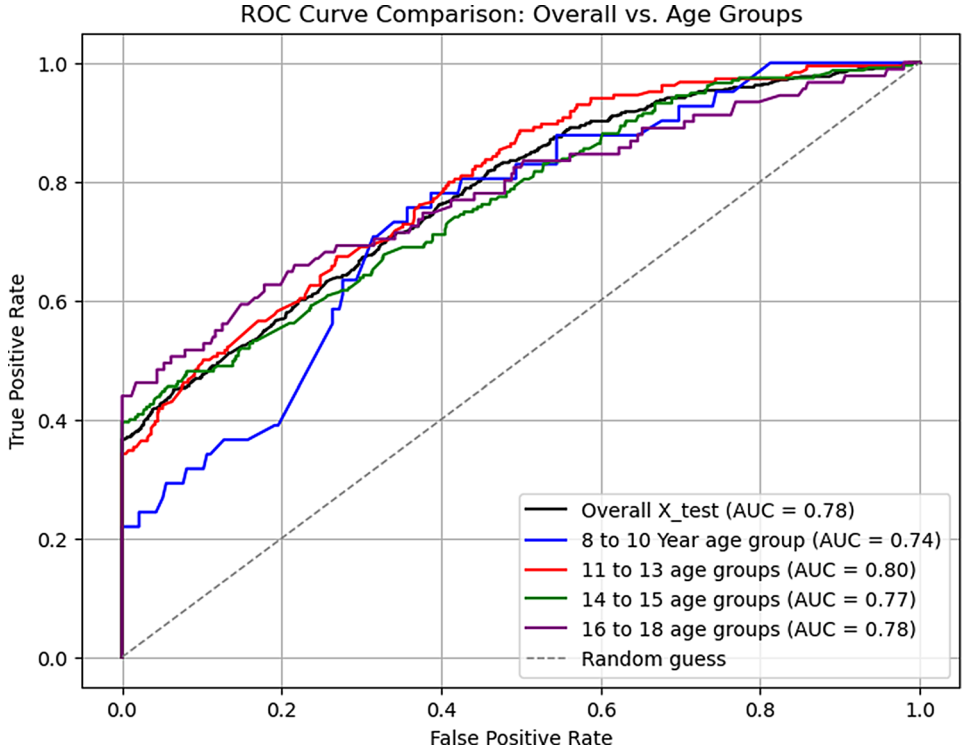

Fig. 4 Receiver Operator Characteristic (ROC) curve comparison: overall vs. age groups

The LR model yielded coefficient estimates that further differentiate features by their strength and direction, as visualized in Fig. 5. High positive coefficients were observed for features such as "past ED RV count," "psychotherapeutic medication history," "substance use history," and "past outpatient MH visits". In addition, several predictors exhibited moderate positive effects, including "feeding/failure to thrive history," "somatoform complaint," and "inpatient medication use," while others, e.g., "having a primary care provider" and "depression as a presenting complaint", showed only mild positive associations. Conversely, negative coefficients were noted for features including "male sex," "MH services referrals," "consultation procedures," "substance addiction as a primary complaint," "admission from the ED," "psychotherapeutic medication count," "anxiety as a presenting complaint,", "Patients from Quebec" and "vitamin medication administration," suggesting these factors are associated with a reduced likelihood of RVs. Furthermore, A variance-based exploration of these variables further confirmed that some predictors with negligible coefficients stemmed from low prevalence or limited variability, not true clinical irrelevance, thus agreeing with the odds ratios.

Because revisit patterns may also vary developmentally, we generated supplementary SHAP visualizations stratified by age groups. These illustrate how the global model's feature importance distributed when applied to test sets from different age strata. While broadly consistent with the overall findings, these analyses are presented as exploratory and are included in Table 5. Table 5 presents global importance and age-specific feature importance rankings, highlighting predictors that remain stable across age groups and those that vary developmentally.

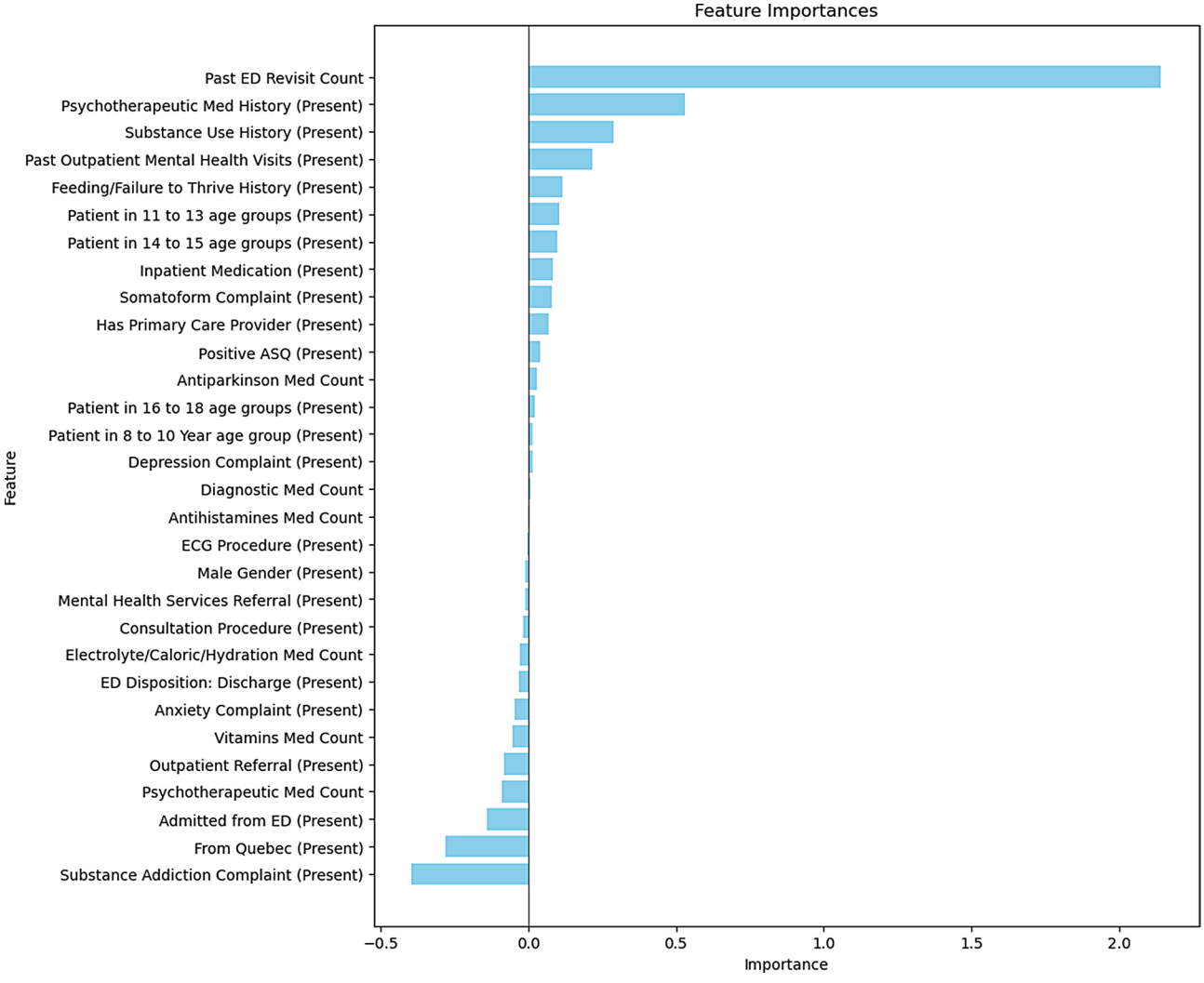

Fig. 5 Feature importance from logistic regression model predicting MH ED RVs

Table 5 Relative importance of age-specific logistic regression features

Feature	Global Importance[1]	SHAP[2] (8–10 YOA[3])	SHAP (11–13 YOA)	SHAP (14–15 YOA)	SHAP (16–18 YOA)
Past ED Revisit Count	++++	+++	+++	+++	+++
Psychotherapeutic Med History	++++	+++	+++	+++	+++
Substance Use History	++++	+++	+++	+++	+++
Past Outpatient Mental Health Visits	++++	+++	+++	+++	+++
Feeding/Failure to Thrive History	++++	+++	++	++	++
11 to 13 Age Group	++++	N/A	0	N/A	N/A
14 to 15 Age Group	++++	N/A	N/A	0	N/A
Inpatient Medication	++++	+++	++	++	++
Somatoform Complaint	++++	+++	++	++	++
Has Primary Care Provider	+++	+	0	0	+
Positive ASQ	++	0	0	0	++
Antiparkinson Med Count	++	0	++	+++	+++
16 to 18 Age Group	++	N/A	N/A	N/A	0
8 to 10 Year Age Group	++	0	N/A	N/A	N/A
Inpatient Medication	++++	+++	++	++	++
Somatoform Complaint	++++	+++	++	++	++
Depression Complaint	++	0	++	0	0
Diagnostic Med Count	0	0	++	+	+
Antihistamines Med Count	0	0	0	0	0
ECG Procedure	-	0	0	0	0
Male Gender	-	0	0	0	0
Mental Health Services Referral	-	0	0	0	0
Consultation Procedure	-	0	0	0	0
Electrolyte/Caloric/Hydration Med Count	--	--	---	--	--
ED Disposition: Discharge	--	0	0	0	-
Anxiety Complaint	--	0	0	--	0
Vitamins Med Count	--	---	---	---	---
Outpatient Referral	---	--	--	--	-
Psychotherapeutic Med Count	---	---	---	---	---
Admitted from ED	---	---	--	--	---
From Quebec	----	---	---	---	---
Substance Addiction Complaint	----	--	--	---	---
Mental Health Services Referral	-	0	0	0	0
Consultation Procedure	-	0	0	0	0
Electrolyte/Caloric/Hydration Med Count	--	--	---	--	--
ED Disposition: Discharge	--	0	0	0	-
Anxiety Complaint	--	0	0	--	0
Vitamins Med Count	--	---	---	---	---
Outpatient Referral	---	--	--	--	-
Psychotherapeutic Med Count	---	---	---	---	---
Admitted from ED	---	---	--	--	---
From Quebec	----	---	---	---	---
Substance Addiction Complaint	----	--	--	---	---

[1]Based on logistic regression coefficients, representing the overall average influence of a feature as learned from the training dataset

[2]SHAP (Shapley Additive explanations) values are computed at the instance level for individual predictions within each age subgroup of the test dataset. These instance-level insights are then aggregated to reveal how feature contributions vary across each age group, offering a more nuanced understanding of their impact

[3]YOA: years of age

Note: "+" denotes a positive association with revisit risk and "–" denotes a negative (protective) association; the number of symbols indicates relative magnitude

Comparing clinician-derived and logistic regression feature importance

Supplementary Material 4 provides a comparison of clinician-assigned weights and LR results per feature, highlighting areas of agreement and divergence. There was strong agreement between clinicians and the LR model for "past ED RV count", "psychotherapeutic medication history", and "substance use history", all of which

were consistently identified as key predictors of RVs. However, discrepancies emerged in certain features. For example, "substance addiction as a primary complaint" was considered by clinicians to increase the likelihood of RVs, whereas the LR model suggested it was associated with a decrease in RV probability. Similarly, clinicians viewed having a primary care provider as reducing revisit risk, while the model showed a non-significant slight increase. Other features, such as antiparkinsonian medication count, were assigned zero weights by clinicians but showed small, statistically insignificant associations in the LR model.

Discussion

Using EHR data from a tertiary care pediatric hospital ED during the period of November 2017 to November 2023, we found that our ML algorithm to predict RVs within six months in patients presenting for CYMH care performed well on all indicators. Furthermore, sub-group testing stratified by age showed that the highest prediction was in 11–13-year-olds, suggesting that RV patterns in early adolescence are more clearly defined, likely due to increased psychiatric diagnosis and structured healthcare interactions. Physical symptom manifestations of psychological distress and pharmacological stabilization were identified as being more relevant in younger cohorts. These findings suggest that age-specific tuning or additional contextual features may further enhance model accuracy, particularly for younger children, reinforcing the need for age-stratified models to enhance predictive accuracy and clinical utility by incorporating age-specific risk factors.

We also developed an innovative way to establish a clinician-driven algorithm as a baseline for the ML algorithm by asking clinicians to use an application to estimate positive or negative predictive "weights" for the same features from the EHR used in developing the ML algorithm. The collective clinician-based algorithm's performance was significantly lower than the one generated by ML. Despite achieving high recall, with correctly identifying most RVs, this approach suffered from low precision, generating a substantial number of false positives.

To our knowledge, this is only the second study to use LR ML to predict CYMH ED RVs. The first was by Saggu et al. [20], but they used a different approach. They investigated CYMH ED RVs by comparing predictive accuracy and performance of a graph neural network (GNN) deep learning model with a recurrent neural network (RNN) and a conventional machine learning model. In addition, their algorithms were to predict a 1-month RV rate, and they did not have clinicians participating.

Their study reported an AUROC of 0.66 for logistic regression ML, with 0.70 for their best-performing GNN model. Our higher AUROC (0.78) likely reflects the combination of expert-informed feature engineering and a longer six-month prediction window, which may be more stable and clinically meaningful than shorter horizons.

Our application enabling clinicians to weigh 191 EHR features predicting RVs is unique. Furthermore, to our knowledge, only one other study has compared a clinical rule-based approach and a ML algorithm in the prediction of a CYMH outcome (suicidal behaviors) [33]. Using data from a large longitudinal population survey of adolescents, two types of ML algorithms significantly outperformed a clinical decision rule used as the predictive factor. Although it is premature to draw conclusions from van Vuuren's group or ours, findings such as these support the utility of using ML in the prediction of complex behaviors such as ED RVs or adolescent suicidal behavior.

A closer comparison of key features for the clinician-derived predictions and the ML algorithm showed significant overlap, although there were some discrepancies. The clinicians most frequently selected the number of past ED RVs, psychotropic medication usage, current substance addiction, and history of substance use for positive prediction, with male sex, having a PCP, and discharge referrals to outpatient care being negative predictors. Key features of the ML algorithm were past ED RVs, psychotropic medication usage, and substance use history, while discharge referrals to outpatient care, inpatient admission, and substance addiction disorders were negatively predictive of RVs. The overlap between clinician and computer-derived positive and negative predictors gives some clinical validity to the ML algorithm.

The variability in individual clinician weight assignments further illustrates the subjective nature of clinical judgment and the complexity of predicting revisits. Each clinician's experience shaped their assessments, while the ML model systematically evaluated relationships across a large feature set. The clinician-derived model was intentionally constrained for feasibility, not designed to replicate real-world clinical reasoning, which is far more nuanced and context-dependent. Thus, the observed performance gap reflects methodological design choices rather than the limits of clinician expertise.

There are at least two possible reasons that the group of clinician-driven algorithms did not perform as well as the ML algorithm. First, strengths of ML algorithm development include the capacity to examine the thousands of interactions that occur within a large set of features, and the AI process has nearly limitless weighting and reweighting capabilities. In contrast, the human brain (as well as traditional statistics) cannot process such volumes of interactions and we limited the clinicians' choices to five positive and three negative options by design. Otherwise, making final decisions would be difficult.

Second, as this was the first trial of the ML algorithm, we wanted to see if there were differences between job roles in the predictions of RVs, so we recruited clinicians from each of the pathways in the CYMH ED care process. Therefore, they all had experience in the ED, but these experiences were different and likely affected how they rated the features. ML has no such factor to contend with.

Examining our strong and moderate feature predictors in the context of the previous research on CYMH ED RVs reveals some overlaps with the clinical studies, again suggesting some clinical validity for our algorithm. Leon et al. [15] reported from 11 studies that RVs were positively associated with older age, female sex, lower social class, ethnic minority status, and past or current MH service usage. The results were strongly predicted of reduced RVs, the opposite of the review. This may be due to different definitions of variables in clinical studies and our EHR.

Wilson et al. [17] reported that across 65 studies, no clear set of correlations emerged, but again, contrary to our finding, they reported that previous or current MH outpatient care was a strong positive predictor. They did not find an association with age or sex, but like us, did report that RV history, psychotropic medication usage, and admission were correlated with higher rates of RV. Our finding that substance addiction disorders were a strong negative predictor was similar to the results of Cushing et al. [2].

Limitations

While the proposed model shows promise for identifying children and youth at risk of ED RVs, its utility is subject to several constraints. The model's applicability depends on timely access to high-quality data, yet some predictors (e.g., finalized ICD-10 codes, discharge details) may only be reliably available after the encounter, positioning the tool primarily for post-discharge use rather than use at discharge. However, this could still be a very helpful outcome, as many interventions to prevent RVs, e.g., community health worker for support, occur post-discharge.

The issue of having too few patients from some rural areas led us to exclude those regions from the algorithm development. This may limit the generalizability of our findings to patients from these towns. As we proceed with more levels of validation, we are hopeful that we can find a way to mitigate this problem.

Logistic regression was chosen for its interpretability and transparency, but this choice may limit the ability to capture non-linear or higher-order relationships that more advanced methods could address. Despite applying L2 regularization and cross-validation, overfitting to institutional data remains a concern, underscoring the need for ongoing recalibration and external validation. Finally, findings are based on data from a single tertiary pediatric hospital, which may limit generalizability across other populations and healthcare systems. However, AI algorithms generated from EHR data are often most useful in the hospital where the data were collected and are often used for quality improvment. Features of our final algorithm after more clinical validation may be useful for other hospitals as starting points for them to generate their own algorithms, but generalizability to other sites still remains an open question.

Next steps

Further work will focus on refining and validating the model prior to silent validation. A priority will be exploring alternative thresholds and precision-recall trade-offs to ensure the model's operating point aligns with clinically meaningful use cases. Feature expansion, guided by clinician feedback and external sources such as the Ontario Marginalization Index, alongside evaluation of more flexible algorithms (e.g., random forests, gradient boosting), may further improve predictive power, while retaining interpretability, and we will also evaluate probability calibration.

Conclusions

This study demonstrates that machine learning can meaningfully predict six-month CYMH ED revisits, with logistic regression outperforming a structured clinician-weighted baseline across performance metrics. While the clinician-driven models underperformed overall, there was substantial overlap between clinician-assigned weights and model-derived predictors, reinforcing the clinical plausibility of the ML approach. Areas of divergence, such as the influence of substance addiction or primary care provider access, highlight opportunities for further investigation. At this stage, the model is best positioned as a post-discharge risk stratification tool to support targeted follow-up efforts, with future refinements aimed at improving precision–recall balance, expanding feature sets, and ensuring calibration for eventual clinical integration. Future work will focus on strategies to improve sensitivity without substantially inflating the false-positive burden.

Abbreviations

AI	Artificial intelligence
AUROC	Area under the receiving operating characteristic curve
CHEO	Children's Hospital of Eastern Ontario
CYMH	Child and youth mental health
DSM	Diagnostic and Statistical Manual
ED	Emergency department
EHR	Electronic health record
GNN	Graph neural network
ICD	International Classification of Diseases
LR	Logistic regression
MH	Mental health
ML	Machine learning

RNN	Recurrent neural network
RV	Revisit
SHAP	Shapley Additive explanations

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03269-0.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Acknowledgements
We thank Jon Seymour for his assistance with the CHEO RI MH Datamart in the early data collection period. We also thank the clinicians who assisted in feedback on feature selection and those who participated in the creation of the clinician-driven baseline predictive model.

Author contributions
KP conceptualized the study, with details of the design and protocol produced by collaboration with JG, CH, NB, and PC. NB developed the models and analyzed the data. KP and AK organized clinician participation. JSL helped support data collection and organization of manuscript preparation. NB, KP, and JSL led the writing of the manuscript, to which all authors contributed, approving the final draft.

Funding
This study was funded by the CHEO RI Precision Child and Youth Mental Health (PCYMH) Collaboratory, supported by a generous donation to the CHEO Foundation. Neither the Foundation nor the donor had influence on any part of this study.

Data availability
As the data are from the Children's Hospital of Eastern Ontario's EHR, the data cannot be shared because of privacy and ethical restrictions due to potentially sensitive information. Model code is available from the first author upon request.

Declarations

Ethics approval and consent to participate
This study was submitted to the Children's Hospital of Eastern Ontario Research Ethics Board (REB) (CHEOREB# 22/11X), which reviewed and approved it, with the requirement to seek consent waived by the REB, in accordance with TCPS2 Article 5.5 A, as this was a retrospective study of de-identified EHR records. This study was conducted in compliance with the Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Participants.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Children's Hospital of Eastern Ontario (CHEO) Research Institute, 401 Smyth Rd, Ottawa, ON K1H 5B2, Canada
[2]School of Earth, Environment and Society, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada
[3]CHEO, 401 Smyth Rd, Ottawa, ON K1H 8L1, Canada
[4]School of Psychology, University of Ottawa, 136 Jean-Jacques Lussier, Ottawa, ON K1N 6N5, Canada
[5]Department of Psychiatry, University of Ottawa, 1145 Carling Avenue, Ottawa, ON K1Z 7K4, Canada

Received: 16 June 2025 / Accepted: 29 October 2025
Published online: 01 December 2025

References

1. Bommersbach TJ, McKean AJ, Olfson M, Rhee TG. National trends in mental Health–Related emergency department visits among Youth, 2011–2020. JAMA. 2023;329:1469.
2. Cushing AM, Liberman DB, Pham PK, Michelson KA, Festekjian A, Chang TP, et al. Mental health revisits at US pediatric emergency departments. JAMA Pediatr. 2023;177:168.
3. Gardner W, Pajer K, Cloutier P, Currie L, Colman I, Zemek R, et al. Health outcomes associated with emergency department visits by adolescents for self-harm: a propensity-matched cohort study. CMAJ. 2019;191:E1207–16.
4. Mapelli E, Black T, Doan Q. Trends in pediatric emergency department utilization for mental Health-Related visits. J Pediatr. 2015;167:905–10.
5. Sheridan DC, Spiro DM, Fu R, Johnson KP, Sheridan JS, Oue AA, et al. Mental health utilization in a pediatric emergency department. Pediatr Emerg Care. 2015;31:555–9.
6. Hoffmann JA, Carter CP, Olsen CS, Ashby D, Bouvay KL, Duffy SJ, et al. Pediatric mental health emergency department visits from 2017 to 2022: A multicenter study. Acad Emerg Med. 2024;31:739–54.
7. Care for Children and Youth With. Mental Disorders — Data tables. Ottawa, ON: Canadian Institute for Health Information; 2022.
8. Valtuille Z, Trebossen V, Ouldali N, Bourmaud A, Gandré C, Aupiais C, et al. Pediatric hospitalizations and emergency department visits related to mental health conditions and Self-Harm. JAMA Netw Open. 2024;7:e2441874.
9. Havens JF, Marr MC. Pediatric psychiatric emergency Services—Stasis in crisis. JAMA. 2023;329:1453.
10. Hoge MA, Vanderploeg J, Paris M, Lang JM, Olezeski C. Emergency department use by children and youth with mental health conditions: A health equity agenda. Community Ment Health J. 2022;58:1225–39.
11. Cloutier P, Thibedeau N, Barrowman N, Gray C, Kennedy A, Leon SL, et al. Predictors of repeated visits to a pediatric emergency department crisis intervention program. CJEM. 2017;19:122–30.
12. Newton AS, Ali S, Johnson DW, Haines C, Rosychuk RJ, Keaschuk RA, et al. Who comes back? Characteristics and predictors of return to emergency department services for pediatric mental health care. Acad Emerg Med. 2010;17:177–86.
13. Rosic T, Duncan L, Wang L, Eltorki M, Boyle M, Sassi R, et al. Trends and predictors of repeat mental health visits to a pediatric emergency department in Hamilton, Ontario. J Can Acad Child Adolesc Psychiatry. 2019;28:82–90.
14. Hoffmann JA, Krass P, Rodean J, Bardach NS, Cafferty R, Coker TR, et al. Follow-up after pediatric mental health emergency visits. Pediatrics. 2023;151:e2022057383.
15. Leon SL, Cloutier P, Polihronis C, Zemek R, Newton AS, Gray C, et al. Child and adolescent mental health repeat visits to the emergency department: A systematic review. Hosp Pediatr. 2017;7:177–86.
16. Frosch E, dosReis S, Maloney K. Connections to outpatient mental health care of youths with repeat emergency department visits for psychiatric crises. PS. 2011;62:646–9.
17. Wilson R, Jennings A, Redaniel MT, Samarakoon K, Dawson S, Lyttle MD, et al. Factors associated with repeat emergency department visits for mental health care in adolescents: A scoping review. Am J Emerg Med. 2024;81:23–34.
18. Sung C-W, Ho J, Fan C-Y, Chen C-Y, Chen C-H, Lin S-Y, et al. Prediction of high-risk emergency department revisits from a machine-learning algorithm: a proof-of-concept study. BMJ Health Care Inf. 2024;31:e100859.
19. Zhang Z. Early warning model of adolescent mental health based on big data and machine learning. Soft Comput. 2024;28:811–28.
20. Saggu S, Daneshvar H, Samavi R, Pires P, Sassi RB, Doyle TE, et al. Prediction of emergency department revisits among child and youth mental health outpatients using deep learning techniques. BMC Med Inf Decis Mak. 2024;24:42.
21. Habehh H, Gohel S. Machine learning in healthcare. CG. 2021;22:291–300.
22. International statistical classification of diseases. and related health problems. 3: Alphabetical index. 2nd ed. Geneva; 2004.
23. Goldstein AB, Frosch E, Davarya S, Leaf PJ. Factors associated with a Six-Month return to emergency services among child and adolescent psychiatric patients. PS. 2007;58:1489–92.
24. Leon SL, Polihronis C, Cloutier P, Zemek R, Newton AS, Gray C, et al. Family factors and repeat pediatric emergency department visits for mental

health: A retrospective cohort study. J Can Acad Child Adolesc Psychiatry. 2019;28:9–20.
25. Peterson BS, Zhang H, Lucia RS, King RA, Lewis M. Risk factors for presenting problems in child psychiatric emergencies. J Am Acad Child Adolesc Psychiatry. 1996;35:1162–73.
26. American Psychiatric Association. Diagnostic and statistical manual of mental Disorders. Fifth edition. American Psychiatric Association; 2013.
27. Lee CH, Yoon H-J. Medical big data: promise and challenges. Kidney Res Clin Pract. 2017;36:3–11.
28. Demir-Kavuk O, Kamada M, Akutsu T, Knapp E-W. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. BMC Bioinformatics. 2011;12:412.
29. Ng AY. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In: Twenty-first international conference on Machine learning - ICML '04. Banff, Alberta, Canada: ACM Press; 2004. p. 78.
30. Assis A, Dantas J, Andrade E. The performance-interpretability trade-off: a comparative study of machine learning models. J Reliable Intell Environ. 2025;11:1.
31. Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M. The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. Comput Biol Med. 2023;166:107555.
32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
33. Van Vuuren CL, Van Mens K, De Beurs D, Lokkerbol J, Van Der Wal MF, Cuijpers P, et al. Comparing machine learning to a rule-based approach for predicting suicidal behavior among adolescents: results from a longitudinal population-based survey. J Affect Disord. 2021;295:1415–20.

Publisher's note