BMC Medicine

RESEARCH

Open Access

A highly scalable deep learning language model for common risks prediction among psychiatric inpatients

Enzhao Zhu[1†], Jiayi Wang[1†], Guoquan Zhou[2], Chunbo Li[3], Fazhan Chen[4], Kang Ju[5], Liangliang Chen[5], Yichao Yin[5], Yi Chen[6,7], Yanping Zhang[8], Xu Zhang[1], Xinlin Zhou[9], Zongyuan Wang[1], Jianping Qiu[2], Hui Wang[2], Weizhong Shi[10], Feng Wang[4], Dong Wang[4], Zhihao Chen[11], Jiaojiao Hou[12], Hui Li[3*†] and Zisheng Ai[13*†]

Abstract

Background There is a lack of studies exploring the performance of Transformers-based language models in common risks assessment among psychiatric inpatients. We aim to develop a scalable risk assessment model using multidimensional textualized data and test the stability, robustness, and benefit of this approach.

Methods In this real-world cohort study, a deep learning language model was developed and validated using first hospitalized cases diagnosed with schizophrenia, bipolar disorder, and depressive disorder between January 2016 and March 2023 in three hospitals. The algorithm was externally validated on an independent testing cohort comprising 1180 patients. A total of 140 features, including first medical records (FMR), laboratory examinations, medical orders, and psychological scales, were assessed for analysis. The outcomes were short- and long-term impulsivity (STI and LTI), risk of suicide (STSS and LTSS), and need of physical restraint (STPR and LTPR) assessed by qualified nurses or clinicians. Analysis was carried out between August 2024 and June 2024. Models with different architectures and input settings were compared with each other. The area under the receiver operating characteristic curve (AUROC) was used to assess the primary performance of models. The clinical utility was determined by the net benefit under Youden's threshold.

Results Of 7451 patients included in this study, 2982 (47.6%) were male, and the median (interquartile range) age was 42 (28–57) years. The overall incidence of outcomes was 635 (8.5%), 728 (10.5%), 659 (8.8%), 803 (10.8%), 588 (7.9%), and 728 (9.8%) for STPR, LTPR, STSS, LTSS, STI, and LTI, respectively. The multitask semi-structured Transformers-based language (SSTL) model showed more promising AUROCs (STPR: 0.915; LTPR: 0.844; STSS: 0.867; LTSS: 0.879; STI: 0.899; LTI: 0.894) in the prediction of these outcomes than single-tasked or multimodal language models and traditional structured data models. Combining FMR with other data from electronic health records led to significant

†Enzhao Zhu and Jiayi Wang contributed equally to this work and shared first authorship.

†Hui Li and Zisheng Ai contributed equally to this work.

*Correspondence:
Hui Li
lihuindyxs@163.com
Zisheng Ai
azs1966@126.com
Full list of author information is available at the end of the article

improvements in the performance and clinical utility of SSTL models based on demographic, diagnosis, laboratory tests, treatment, and psychological scales.

Conclusions The SSTL model shows potential advantages in prognostic evaluation. FMR is a strong predictor for common risks prediction and may benefit other tasks in psychiatry with minimum requirements for data and data processing.

Keywords Transformers, Deep learning, Suicide risk, Impulsivity, Physical restraint

Background

Impulsivity and suicide are common adverse outcomes among patients with mental illness, especially those diagnosed with schizophrenia spectrum disorders, bipolar disorder (BD), and depressive disorder (DD) [1–4]. These risks are typically interrelated [5–8] and represent a significant public health challenge that carries important clinical implications [8, 9]. Physical restraint is not only a common method for addressing the above risks but also an important indicator of the quality of medical care and has drawn widespread concern worldwide [10]. Hence, the precise prediction and proficient management of these risks, alongside the implementation of physical restraint, is particularly crucial.

Although traditional risk assessment tools such as self-reported questionnaires [11] exist to aid clinical decisions on patient management and interventions like restraints [12], there is no consensus on the best tools for assessing impulsivity and suicide risk, and detailed monitoring of symptom fluctuations is lacking [1]. While previous studies have attempted to use machine learning to predict adverse events in patients [13, 14], there are still some challenges to this endeavor. One limitation lies in handling structured data and missing values in large real-world samples [15], which may lead to inconsistencies between modeling research and clinically available data [16]. Second, despite the continued discovery of specific markers in psychiatry, the available features in the clinic are still limited [17, 18]. Thus, whether using language as a feature in psychiatry [19] can help address these issues and reduce the complexity of dealing with structured data remains a topic worth investigating [20, 21]. Previous studies have explored language models, demonstrating the potential of language features in psychiatric diagnosis [22, 23], prognosis [24], and treatment [25].

The Transformers architecture [26] has laid the foundation for the rapid development of deep learning language models, enabling the development of state-of-the-art technologies such as ChatGPT [27], LLaMA [28], and DeepSeek-R1 [29]. These cutting-edge language models exhibit considerable capabilities in identifying patterns and associations in specific contexts and processing natural language effectively [30]. Psychiatry, a medical field deeply connected with language, has successfully applied natural language processing (NLP) methods in clinical settings to analyze risk factors, symptoms, and diagnoses [31–33]. Electronic health records (EHRs) of psychiatric inpatients provide a rich resource of longitudinal data and large cohort sizes, fostering the increasing application of NLP tools. Although several Transformer-based models for EHR data have been developed globally [34, 35], there remains a lack of effective NLP models capable of leveraging EHR data to address multiple mental illnesses simultaneously in China [36–38].

Incorporating language modeling with Transformers models is expected to address current gaps, particularly the lack of a practical model and the underutilization of valuable language information in psychiatric clinical practice. This approach aims to provide an effective and reliable risk assessment tool with a strong capability to process unstructured language data, establishing a feasible application route to enhance mental health services quality. This study aims to integrate multidimensional EHR data to identify potential future risks of psychiatric inpatients through Transformers-based language models.

Methods

Study participants

Data were collected from three mental health facilities (Tongji University Mental Health Center, Shanghai Putuo District Mental Health Center, and Shanghai Changning District Mental Health Center, and Shanghai Mental Health Center). The database includes all-time medical history, admission information, diagnoses, scale tests, medical orders, laboratory tests, and risk assessments from real-world EHRs in both clinics and hospitalizations. These data are multi-centered, and especially, Shanghai Mental Health Center is one of the four National Medical Centers for Mental Diseases, thus having strong representativeness and reflecting the situation of psychiatric inpatients in China. This retrospective study included detailed and comprehensive medical records of psychiatric patients hospitalized from 2016 to 2023, seeking to predict acute and long-term risks of suicide, impulsivity, and the need for physical restraint. Patient data were extracted based on unique hospitalization codes. The first admission was identified using the earliest traceable hospitalization code. Private

information such as name, ID number, and address were not included, nor were the records of repeated admissions. The eligibility criteria included (a) 18–65 years old, (b) main diagnosis on admission was BD, DD, or schizophrenia, (c) first hospitalization, and (d) resident in China. These three included diseases are the diagnoses that account for the largest proportion of hospitalized patients in China, and the diagnostic standards follow ICD-10. Exclusion criteria included (a) loss of demographic information, (b) excessive loss of hospital records (more than 20% of included features), (c) length of hospitalization is less than 180 days, (d) accompanied with mental retardation, personality disorder or brain organic disease, (e) accompanied with severe somatic disease, (f) long-term history of psychotropic drug use, and (g) pregnancy or lactation. This study was approved by the Ethics Committee of Tongji University Mental Health Center (Grant number: PDJW-IIT-2023-017CS). This study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines [39].

Procedures

Numerical structured data were extracted directly from the database, including patient demographic information, diagnosis, temporary and long-term medical orders, laboratory and other auxiliary examinations, and psychological assessment scales. From the medical orders data, we specifically extracted commonly used drug usage, electroconvulsive therapy, and repetitive transcranial magnetic stimulation records. To ensure the future clinical practicality of the model and maximize the utilization of hospital data, we selected the Zung Self-rating Anxiety Scale (SAS) and Zung Self-rating Depression Scale (SDS) that were used in every hospital, and previous studies have shown that the scores of these scales are significantly associated with adverse events in psychiatric patients [40, 41]. Text records were obtained from the patient's first medical record (FMR). The FMR is recorded on the day of admission and is a semi-structured HyperText Markup Language (HTML)-formatted text composed of different sections required by each facility. We used the following keywords to extract text from different sections to supplement the patient's profile: chief complaint, present and past history, physical examination, and mental examination. See Additional file 1: Table S1 for the definition and explanation of each included variable. A total of 140 features were included for analysis.

Psychological scales, laboratory examinations, and other demographic information were conducted or recorded between 1 week before admission and the day of admission and were extracted according to the examination time or execution time. Considering that psychiatric drugs need a certain amount of time to take effect, data from the medical order records were extracted from 1 month before admission to the day of admission. As some patients may seek medical treatment in more than one hospital, we also extracted the medical order records of all patients in the study hospitals during that time period to reduce the impact of prescriptions in multiple places.

Outcomes

The outcome variables of this study were the risk of impulsivity, risks of suicide, and the need of physical restraints. According to institutional requirements and clinical pathways, impulsivity and suicide risks are recorded in the risk assessment scale system by qualified nurses or clinicians in real-time on the day of admission and at least every week during hospitalization. Physicians can also proactively evaluate patients when deemed necessary. Impulsivity was assessed by the Impulsive Behavior Risk Assessment Scale (IBRAS) made by Chinese experts combining the Modified Overt Aggression Scale (MOAS) and the Violence Risk Screening-10 (V-RISK-10), which contains 7 items, and the score of each item is 0 for no, while the score for yes varies (1, 2, 3, or 5), depending on the degree of severity of the question. A score of ≥ 5 indicates high clinical concern for future impulsive behavior, which was used as the cutoff value for this study [42]. This assessment was completed by trained clinicians or nurses based on direct observation, patient records, and clinical judgment during hospitalization [43]. Suicide risk was assessed by the Nurse's Global Assessment of Suicide Risk (NGASR), which consists of 15 questions, and patients with a total score of ≥ 9 were considered to be at high risk of suicide [44, 45]. The use of physical restraints is documented in the medical order. This study extracted all outcome variables from the day of admission to 180 days of hospitalization. The timing of the outcome variables was set to be within 7 days (short-term) and 180 days (long-term) of hospitalization.

Models

Models were built in Python 3.8 with packages of Pytorch 2.0.0 [46] (Transformers-based language model) and Scikit-learn 1.1.3 [47] (traditional machine learning: XGBoost and logistic regression). In this study, our Transformers-based language model comprises a pre-trained text encoder, a multilayer perceptron (MLP), and a linear classifier. Transformers-based text encoders [26, 48] were used for feature extraction from texts. Several well-established Transformers-based encoders were considered for feature extraction, including XLNet [49], BERT [48], ALBERT [50], RoBERTa [51], Longformer [52], and BigBird [53]. BERT, ALBERT, and RoBERTa are widely adopted for their

high accuracy and bidirectional context understanding. Longformer and BigBird utilize sparse attention for extended sequences. For our task, XLNet (XLNet-base: 12-layer; 768 nodes; 12 self-attention heads; pre-trained on 32.9 billion words) was selected for its strengths in capturing long-range dependencies and processing complex text information efficiently. XLNet's permutation-based training [54], segment recurrence mechanism, and relative positional encoding [55] enhance its ability to process long clinical texts, making it particularly suitable for clinical text analysis. XLNet has shown superior performance in text classification tasks compared to other encoders [49]. A 2-layered MLP with 512 nodes was placed after the text encoder because it can stabilize the extracted features [56]. Then, a linear classifier predicts positive and negative probabilities with a softmax function for each outcome. The maximum input length was set to 1200 Chinese characters during implementation, based on the observed upper quartile of input text length and considerations of computational efficiency. Texts longer than this length were truncated. The multimodal models use an additional 3-layered MLP to receive the structured data, while the latent features are concatenated in a subsequent linear layer. We developed an automatic semi-structured textualization function that transforms structured data into texts to input additional information into the Transformers-based language models. Since the structured information is converted into text and most models have no direct restriction on the length of the input text, the semi-structured Transformers-based language (SSTL) model can accept information of arbitrary length and does not require special handling of missing values during inference. Figure 1 illustrates the design of the semi-structured textualization function and the patient assessment process. To properly construct the semi-structured text, we first developed a feature dictionary to appropriately respond to the required feature index and other necessary information such as lab units and frequency of medical orders. We then searched the original database for the unique patient index for each included patient. If the required data are available in the original database, these data were textualized according to the feature dictionary. The transformed text of these structured data is saved separately according to the classification described by Additional file 1: Table S1, which was considered clinically appropriate by our clinicians. When feeding the models, each part of the text is connected using a special symbol, $<SEP>$, which represents a separator token used to delineate different sequences or sentences within the Transformer models' input. Multitask Transformers-based language models were trained for three outcomes of interest simultaneously, using the same feature extractors and different classifiers. Adopting multitask learning potentially enhances model generalization, improves efficiency, and fosters information and feature sharing across diverse tasks [57]. Schematics of the different Transformers-based language models are shown in Additional file 1: Fig. S1. XGBoost [58] and logistic regression were selected for structured data modeling, due to their wide applications and excellent performance in previous studies. Model codes are available on GitHub (https://github.com/EnzhaoZhu/Common-risks-prediction-among-psychiatric-inpatients).

Patients with more than 20% missing values were excluded. We did not perform any missing value filling for the SSTL model. For traditional machine learning models, missing values were filled with the median within each cohort. Cluster random sampling was used to identify one hospital (Shanghai Putuo District Mental Health Center) as an external testing cohort to test the model's generalization performance in response to potentially differently distributed data. We divided the remaining data into a training cohort and an internal testing cohort in the ratio of 8:2. The models were trained based on tenfold cross-validation repeated ten times in the training cohort for the best replicability [59]. Tuned hyperparameters for Transformers-based language models included the number of MLP layers (tested over 1, 2, 3), number of nodes per layer (128, 256, 512), learning rate ($1e-5$, $1e-4$, $1e-3$), the percentage of dropout (0.1, 0.2, 0.3), and the mini-batch size (8, 16, 32), while number of trees (100, 200, 300), tree depth (3, 6, 9), and learning rate ($1e-3$, $1e-2$, $1e-1$) were tuned for XGBoost. Grid search was employed to find the optimal parameter values based on cross-validation performance in the training cohort. To account for label imbalance, the following approaches were taken: (1) focal loss [60] was used to adaptively adjust the extent to which the sample contributes to the loss, which allows greater weight to be assigned to samples that are inaccurately predicted; (2) the training samples were oversampled to ensure that the negative and positive samples were equally distributed. For the multitask models, we oversampled the joint distribution of labels to ensure that the joint distributions were unchanged after oversampling. In multitask learning, we employed a dynamic strategy to adjust task weights, which emphasizes tasks with higher losses by assigning higher weights while considering the average task importance [61]. To prevent overfitting, the cross-validated loss generated by each iteration is recorded, and training is automatically terminated if the loss does not decrease in 10 iterations (10 * mini-batch size (16) = 160 samples). The maximum number of epochs was set to 10, and the optimizer used was AdamW. Transformers-based

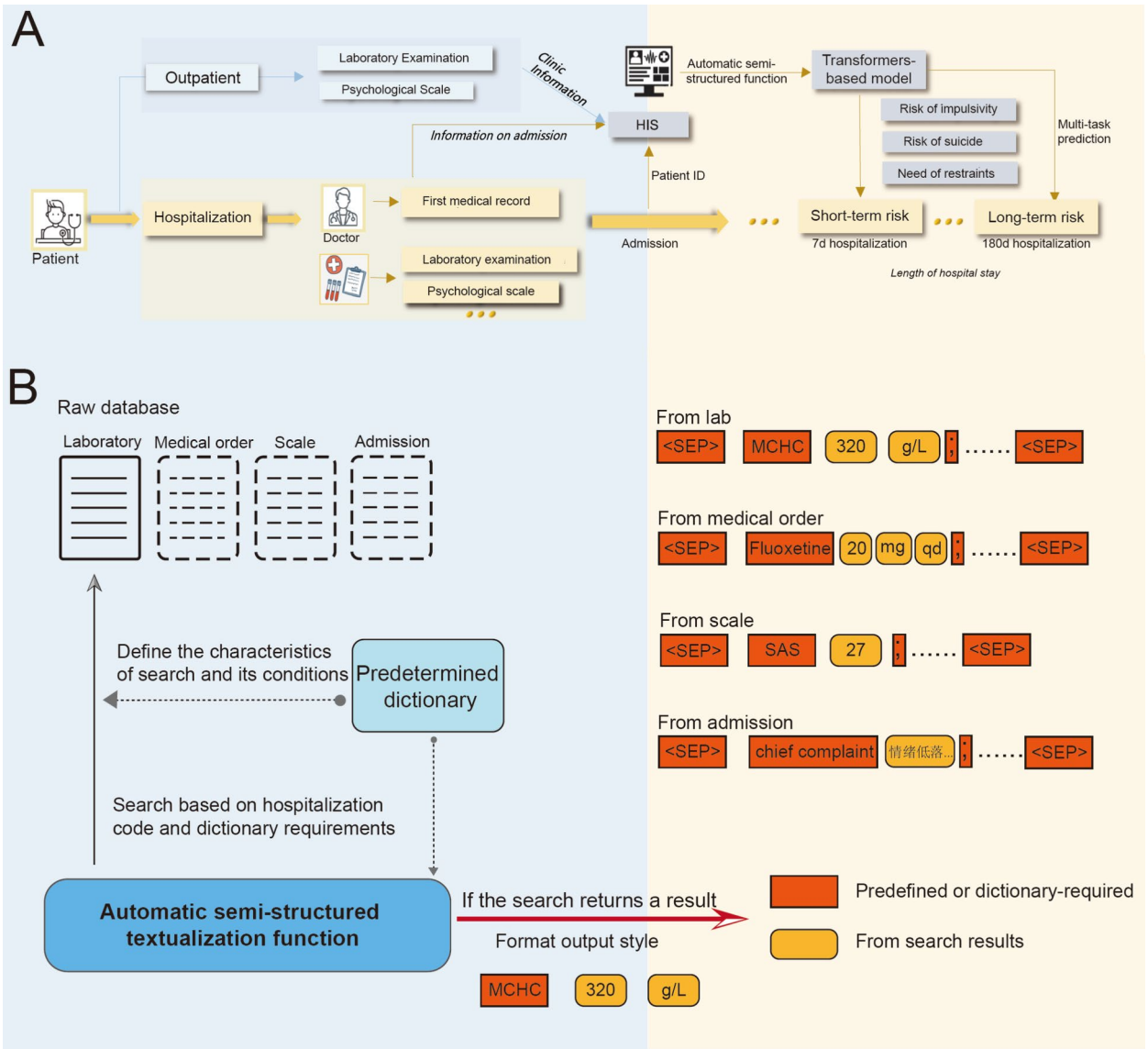


Fig. 1 Schematics of an individual patient assessment and the automatic semi-structured textualization function. **A** Schematics of an individual patient assessment. **B** Schematics of the automatic semi-structured textualization function. HIS, hospital information system

language models were trained in parallel on three graphics processors (RTX 3090 Ti), with a total of 64 gigabytes of graphics memory.

Statistical analysis

Initial data processing was conducted in PostgreSQL 4.2. Model performance metrics were calculated in Python 3.10 with the following packages: Scikit-learn 1.1.3, Pandas 1.5.2 [62], Numpy 1.23.4 [63], and Torchmetrics 1.4.0 [64]. Further data description and analysis were conducted in R 4.1.3. Continuous variables are reported as medians and interquartile ranges (IQR), while categorical variables are presented as numbers and percentages (%). Model performance was evaluated using the area under the receiver operating characteristic curve (AUROC). Secondary performance metrics included the area under the precision-recall curve (AUPRC), Youden's *J* statistic, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score. These metrics collectively provide a comprehensive assessment of predictive performance: AUPRC highlights the model's effectiveness in identifying positive cases, sensitivity and specificity evaluate accuracy across positive and negative outcomes, F1 score balances precision and sensitivity, PPV and NPV indicate prediction reliability, and Youden's *J* statistic optimizes the

Table 1 Overview of the primary performance metric of all models in the external testing cohort

Model	Short-term need of physical restraint	Long-term need of physical restraint	Short-term risk of suicide	Long-term risk of suicide	Short-term risk of impulsivity	Long-term risk of impulsivity
Model 1	0.727 (0.715–0.748)	0.801 (0.793–0.814)	0.827 (0.821–0.835)	0.859 (0.853–0.871)	0.771 (0.76–0.789)	0.801 (0.792–0.815)
Model 1[a]	0.777 (0.724–0.825)	0.702 (0.69–0.719)	0.887 (0.863–0.905)	0.857 (0.827–0.889)	0.824 (0.793–0.863)	0.749 (0.698–0.82)
Model 2	0.830 (0.822–0.844)	0.821 (0.813–0.828)	0.881 (0.876–0.888)	0.816 (0.811–0.826)	0.822 (0.814–0.833)	0.822 (0.812–0.834)
Model 2[a]	0.827 (0.801–0.867)	0.800 (0.792–0.811)	0.852 (0.818–0.875)	0.851 (0.814–0.876)	0.834 (0.793–0.879)	0.822 (0.785–0.864)
Model 3	0.867 (0.861–0.875)	0.860 (0.853–0.869)	0.856 (0.848–0.864)	0.815 (0.808–0.829)	0.776 (0.765–0.794)	0.829 (0.802–0.857)
Model 3[a]	**0.915 (0.91–0.922)**	**0.844 (0.836–0.856)**	**0.867 (0.861–0.875)**	**0.879 (0.873–0.885)**	**0.898 (0.892–0.911)**	**0.894 (0.867–0.918)**
XGBoost	0.760 (0.752–0.774)	0.769 (0.76–0.779)	0.824 (0.819–0.835)	0.836 (0.83–0.847)	0.781 (0.769–0.793)	0.745 (0.734–0.758)
Logistic regression	0.720 (0.711–0.734)	0.701 (0.692–0.713)	0.755 (0.748–0.768)	0.732 (0.722–0.742)	0.685 (0.675–0.703)	0.682 (0.67–0.695)

This table summarizes the area under the receiver operating characteristic curve of all models in the external testing cohort. The 95% confidence interval was calculated using bootstrap with 1000 iterations. Bolded values indicate the best-performing model for each outcome. Model 1, language model with first medical record only; Model 2, language model with multimodal design; Model 3, language model with semi-structured textualization design; XGBoost, XGBoost model with structural data only; Logistic regression, logistic regression model with structural data only; a, language model with multitask design

decision threshold for balanced sensitivity and specificity to maximize accurate predictions. Reported values are averages and 95% confidence intervals (CI) obtained from bootstrapping with 1000 iterations. Clinical utility was assessed using net benefit at thresholds optimized by Youden's J statistic [65]. This approach quantifies the trade-off between true positives and false positives, offering a practical measure of our model's effectiveness in supporting clinical decision-making [66]. In the blank model, which did not include any features, the cut-off was baseline prevalence. Nadeau and Bengio's corrected resampled t-test [67] was used for model metrics comparison, which adjusts for the non-independence of resampled statistics. Various text encoders were evaluated based on their AUROCs. To test the robustness of the model in the face of different proportions of missing values, we randomly removed a certain proportion of text. Pre-trained SimBERT was used to delete text without destroying semantic integrity [68]. To test the added benefit of FMR, we compared the performance and clinical utility of the model with and without FMR. Feature importance was assessed by calculating the reduction in AUROC after removing the corresponding feature set from the external testing cohort. In these tests, only the input data were changed, while the language models were not retrained. The relatedness of three outcomes was calculated using Yule's phi correlation. $P<0.05$ was considered statistically significant. Holm-Bonferroni was used to adjust for multiple testing.

Results

Study overview

The data overview is summarized in Additional file 1: Table S2. A total of 7451 patients were included in this study, identified by unique hospitalization codes. 3757 (50.4%), 940 (12.6%), and 2754 (37.0%) patients were primarily diagnosed with schizophrenia, bipolar disorder, and depression. The median (IQR) age was 42 (28–57) years. 2982 (47.6%) patients were male. The overall incidence of outcomes was 635 (8.5%), 728 (10.5%), 659 (8.8%), 803 (10.8%), 588 (7.9%), and 728 (9.8%) for short-term need of physical restraint (STPR), long-term need of physical restraint (LTPR), short-term risk of suicide (STSS), long-term risk of suicide (LTSS), short-term risk of impulsivity (STI), and long-term risk of impulsivity (LTI), respectively. Compared with the combined training and internal testing cohort, the external testing cohort had older patients (51 [IQR: 39–65] vs. 42 [IQR: 28–57] years), shorter FMR lengths (780 [IQR: 601–965] vs. 858 [IQR: 708–1009] characters), and higher prevalence of schizophrenia (67.1% vs. 47.3%) and physical restraint needs (short-term: 17.6% vs. 6.8%; long-term: 19.9% vs. 8.8%) (all $P<0.001$). The correlation of the three outcomes is summarized in Additional file 1: Tables S3–S4 for short- and long-term outcomes, which show a significant positive correlation between each outcome.

Model performance

As summarized in Table 1 and Additional file 1: Tables S5–S10, all six of our Transformers-based language models and XGBoost achieved AUROCs higher than 0.700 in predicting STPR, LTPR, STSS, LTSS, STI, and LTI, significantly better than the blank model ($P<0.001$). However, the AUPRCs of the structured data model indicated a relatively poor performance on imbalanced data. The best Transformers-based language model had an AUROC of more than 0.8 in the prediction of each outcome, indicating a good classification performance [69].

Additional metrics in the external testing cohort further supplemented these findings, with AUPRC values from 0.577 to 0.684, sensitivity from 0.617 to 0.849, specificity from 0.743 to 0.955, F1 scores from 0.446 to 0.632, PPV from 0.304 to 0.648, and NPV from 0.944 to 0.972. Collectively, these metrics indicate the model's acceptable predictive performance and enhanced clinical applicability over structured data models. We subsequently compared the SSTL model with four other models (Additional file 1: Table S11). The performance of the SSTL model was significantly better than that of other models in 33 (AUROCs) and 28 (AUPRCs) out of 42 tests, showing that using the semi-structured textualization function to transform structured data into text can improve the performance of most models, which is better than using only FMR or simply merging text features with structured data features. Based on the comparison between multitask and single-task language models (Additional file 1: Table S12), the multitask models had superior performance in 17 (AUROCs) and 13 (AUPRCs) out of 36 tests. However, in 4 (AUROCs) and 5 (AUPRCs) tests, it was inferior to single-task models, primarily focusing on the LTI prediction. We compared multitask SSTL models utilizing different text encoders and found that XLNet achieved the best performance in four out of six tasks (Additional file 1: Fig. S2). RoBERTa performed better in the LTSS task, with the highest AUROC (0.88, 95% CI, 0.872–0.891; two-sided $P=0.915$ vs. XLNet), while Longformer showed the best performance in the STI task (0.91, 95% CI, 0.902–0.918; two-sided $P=0.061$ vs. XLNet), although their advantage over XLNet was not statistically significant. RoBERTa and Longformer outperformed BERT and ALBERT across multiple tasks, while BigBird generally showed weaker performance. Thus, XLNet was determined as the best text encoder. The ROC curves of multitask SSTL models are presented in Additional file 1: Fig. S3 (internal testing cohort) and Fig. 2 (external testing cohort), with AUROCs of (STPR: 0.915, standard deviation (SD), 0.009), (LTPR: 0.844, SD, 0.012), (STSS: 0.867, SD, 0.011), (LTSS: 0.879, SD, 0.009), (STI: 0.899, SD, 0.013), and (LTI: 0.894, SD, 0.012) in the external testing cohort.

The robustness tests are demonstrated in Additional file 1: Fig. S4, in which the *P* values and differences were calculated hierarchically (i.e., 20% versus all; 40% versus 20%). When text content was randomly deleted by 20%, the average reduction in AUROCs was 0.098 (95% CI, 0.072–0.121), 0.077 (95% CI, 0.049–0.112), 0.071 (95% CI, 0.042–0.098), 0.052 (95% CI, 0.027–0.077), 0.104 (95% CI, 0.066–0.14), and 0.033 (95% CI, 0.008–0.062) for STPR, LTPR, STSS, LTSS, STI, and LTI. No significant performance degradation of the LTI model was observed when 20% of the text content was removed. The model was insensitive to the deletion of textual content from 40 to 60% and 60 to 80% ($P>0.05$).

Added value of first medical record and feature importance

For short-term outcomes, the clinical utility ranged from 0.51 to 0.553, while for long-term outcomes, it ranged from 0.405 to 0.507 (Additional file 1: Table S13).

Combining FMR with other recorded patient data from EHR led to significant improvements in the performance (Fig. 3) and clinical utility (Fig. 4) of multitask SSTL models based on demographic, diagnosis, laboratory tests, treatment plans, SAS, and SDS, with the importance of FMR being evident across six outcomes in the external testing cohort. In the ALL model, which consisted of all semi-structured text information, FMR significantly improved model performance (STPR: 0.153; LTPR: 0.253; STSS: 0.128; LTSS: 0.289; STI: 0.085; LTI: 0.256) and clinical utility (STPR: 2.661; LTPR: 3.242; STSS: 0.192; STI: 0.381; LTI: 1.815). For short-term outcomes, the relative contribution of FMR to multitask SSTL models varied from 0.007 to 0.407 (weighted AUROCs) and 0.007 to 0.401 (clinical utility), while that of long-term outcomes were 0.008 to 0.340 (weighted AUROCs) and 0.010 to 0.549 (clinical utility). When FMR was excluded, the AUROCs of ALL models were 0.762 (STPR), 0.591 (LTPR), 0.739 (STSS), 0.590 (LTSS), 0.813 (STI), and 0.639 (LTI).

The feature importance heatmap (Fig. 5) shows that FMR was the strongest predictor across all tasks, while treatment plans, psychological scales, diagnosis, and demographic information also have a statistically significant impact on the model's performance (most of the two-sided $P<0.01$). Additionally, laboratory biomarkers such as blood routine, liver function, and kidney function also influence the prediction accuracy of certain outcomes in the SSTL models.

(See figure on next page.)

Fig. 2 The receiver operating characteristic curve of semi-structured Transformers-based language model in the external testing cohort. **A** The receiver operating characteristic curve of short-term impulsivity. **B** The receiver operating characteristic curve of long-term impulsivity. **C** The receiver operating characteristic curve of short-term risk of suicide. **D** The receiver operating characteristic curve of long-term risk of suicide. **E** The receiver operating characteristic curve of short-term need of physical restraint. **F** The receiver operating characteristic curve of long-term need of physical restraint. SD, standard deviation; AUROC, the area under the receiver operating characteristic curve

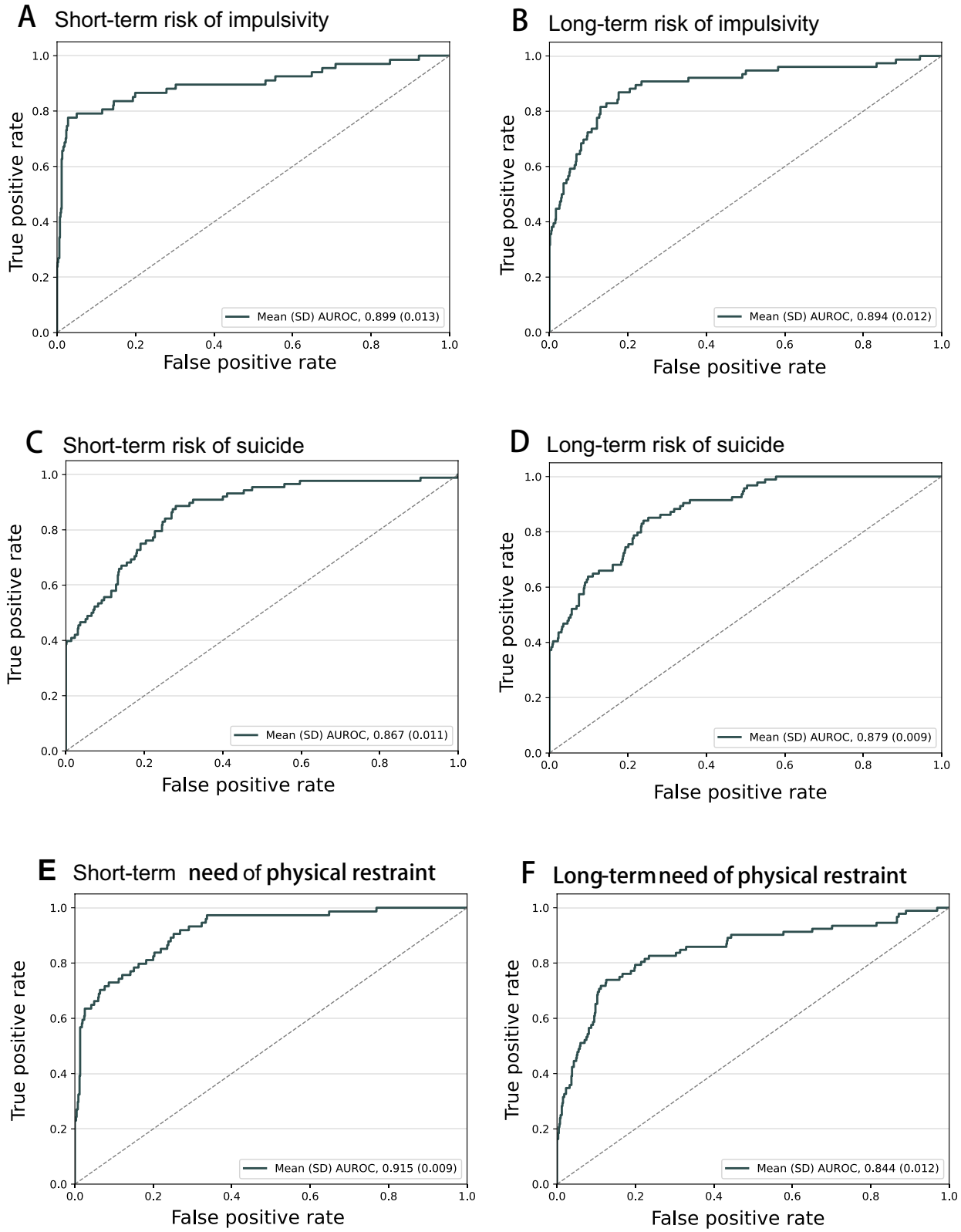
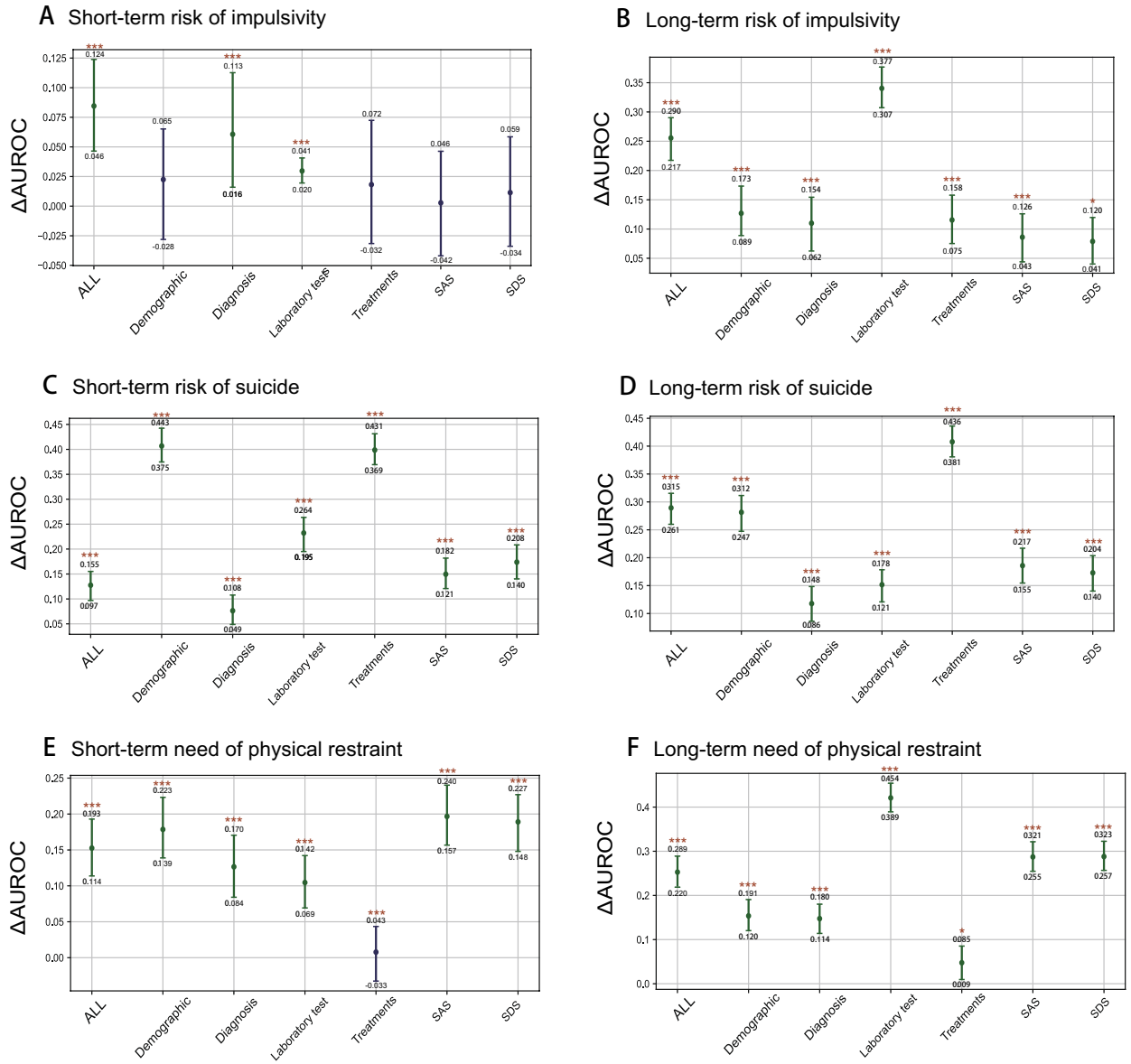


Fig. 2 (See legend on previous page.)

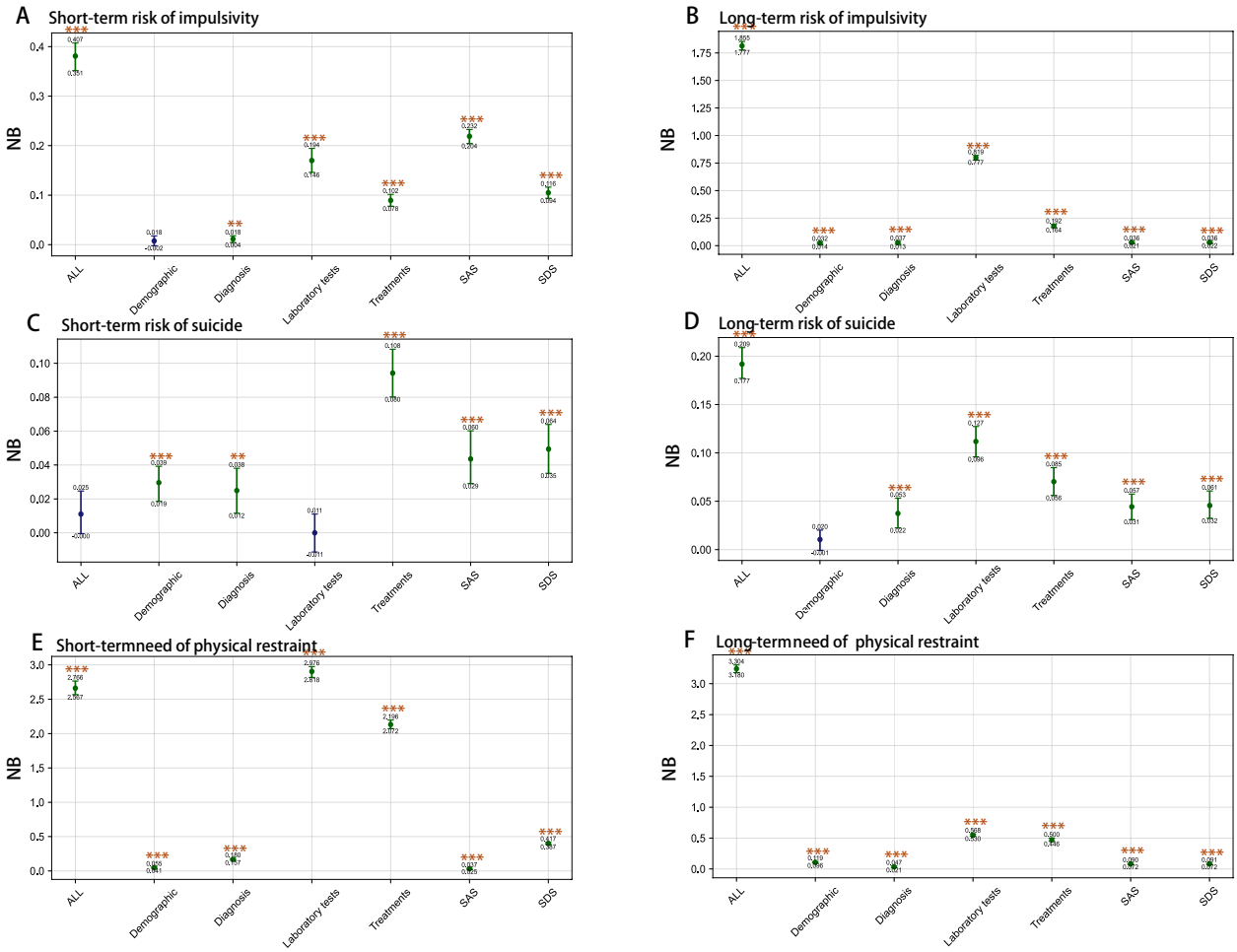


*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$

Fig. 3 The added primary performance of first medical records. **A** The added primary performance in predicting short-term impulsivity. **B** The added primary performance in predicting long-term impulsivity. **C** The added primary performance in predicting short-term risk of suicide. **D** The added primary performance in predicting long-term risk of suicide. **E** The added primary performance in predicting short-term need of physical restraint. **F** The added primary performance in predicting long-term need of physical restraint. The *P* values were single-sided, calculated using Nadeau and Bengio's corrected resampled *t*-test. The ALL model contained all semi-structured text information. SAS, Zung Self-rating Anxiety Scale; SDS, Zung Self-rating Depression Scale; ΔAUROC, the difference of the area under the receivers operating characteristic curve

Discussion

Our study suggests that Transformers-based language models can accurately predict both short- and long-term common psychiatric risks among inpatients, including the use of physical restraint, risk of suicide, and impulsivity. The clinical utility of the model, measured as net benefit under Youden's threshold, ranged from 0.405 to 0.553 in the external testing cohort. This indicates that, per 100 patients assessed, the model could support 41 to 55 net positive decisions, reflecting a favorable balance between true positives and false positives. Metrics such as AUPRC, sensitivity, and specificity underscore the



*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$

Fig. 4 The added clinical utility of first medical records. **A** The added clinical utility in predicting short-term impulsivity. **B** The added clinical utility in predicting long-term impulsivity. **C** The added clinical utility in predicting short-term risk of suicide. **D** The added clinical utility in predicting long-term risk of suicide. **E** The added clinical utility in predicting short-term need of physical restraint. **F** The added clinical utility in predicting long-term need of physical restraint. The clinical utility was determined using the net benefit under Youden's best cutoff. The *P* values were single-sided, calculated using Nadeau and Bengio's corrected resampled *t*-test. The ALL model contained all semi-structured text information. SAS, Zung Self-rating Anxiety Scale; SDS, Zung Self-rating Depression Scale; NB, net benefit under Youden's best cutoff

model's effectiveness in accurately identifying positive cases, even within imbalanced data contexts, affirming its reliability in high-stakes clinical applications. Furthermore, the consistency in PPV, NPV, and F1 scores demonstrates reliable, well-rounded predictive performance across outcomes. Enhanced by FMR and the Transformers architecture, our model surpasses traditional methods by offering a balanced, clinically attuned predictive capacity [14, 70]. Clinically, this robust predictive ability enhances the early identification of high-risk patients and facilitates early, proactive, individualized interventions.

Incorporating FMR into psychiatric prediction models based on other individual data significantly enhances their ability to distinguish between patients who may develop risks in future hospitalization and could lead to clinically relevant improvements. The performance can be further improved by converting the common accessible structured data from the EHR into inputs for the Transformers-based language model. This approach allows our models to evaluate patients with minimal required data and reduces the need for extensive preprocessing steps, such as handling missing values and performing feature engineering on structured data, which are often required by traditional machine learning models [71]. In clinical practice, unavailable data or excessively heavy data preprocessing is a common problem that greatly reduces the utilization

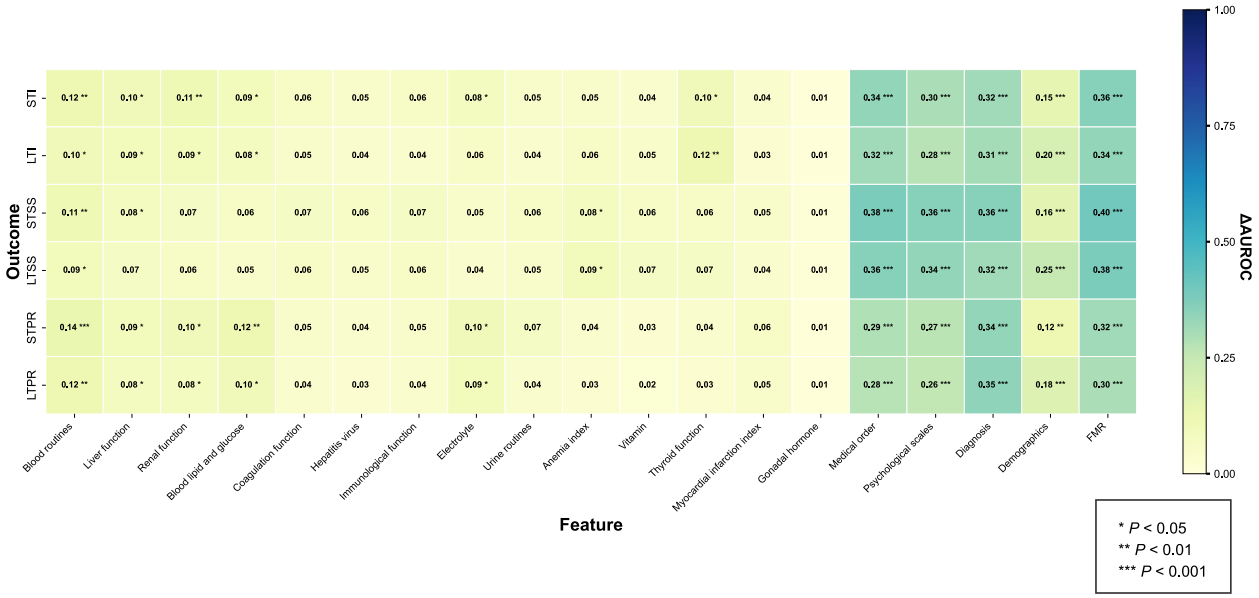


Fig. 5 Feature importance heatmap. STPR, short-term need of physical restraint; LTPR, long-term need of physical restraint; STSS, short-term risk of suicide; LTSS, long-term risk of suicide; STI, short-term risk of impulsivity; LTI, long-term risk of impulsivity. The *P* values were two-sided, calculated using Nadeau and Bengio's corrected resampled *t*-test

of prediction models in practice [72], which further underscores the clinical practicality of our approach.

Moreover, the feature importance analysis reinforces the crucial role of FMR, treatment plans, psychological scales, demographic information, and diagnoses in prediction. This aligns with existing evidence that psychiatric symptoms, medication usage, and diagnostic categories are closely linked to common risks in psychiatric inpatients. Additionally, laboratory biomarkers also serve as risk predictors, possibly due to their association with metabolic dysregulation, inflammation, organ dysfunction, and medication side effects [73–76]. These pathophysiological alterations may contribute to mood disorders, cognitive impairment, and behavioral disorders, which are important factors of psychiatric risks [77–84]. Our finding demonstrates the contribution of different features to the model performance and also highlights the value of integrating multimodal clinical data into risk prediction models.

Among the text encoders we tested, XLNet performed best, likely due to its ability to effectively model bidirectional context and long-range dependencies, which may have been particularly beneficial for the variety of tasks in this study. RoBERTa and Longformer showed better performance, possibly due to RoBERTa's improved pretraining and Longformer's efficiency with longer sequences. In contrast, BigBird's large number of parameters may have been less suited to our relatively small dataset, potentially explaining its weaker performance. Although current advanced pre-trained Transformers-based language models represent a significant improvement over previous models in many tasks [26, 30], their substantial computational resource requirements for both training and inference remain a critical challenge for deployment in real-world settings, particularly in resource-constrained environments such as healthcare [85]. To improve computational efficiency and optimize performance across multiple tasks, we designed an integrated model that shares feature extractors across tasks while employing task-specific classifiers. In future endeavors, constructing a general-purpose language model specifically for psychiatry could further streamline clinical applications. Such a model could require no further training or only a few-shot training by strengthening the generalizable feature extraction capabilities of text encoders [85] in psychiatric-specific texts. Techniques such as self-supervised learning [65] and domain-adaptive pretraining [86] may play a pivotal role in achieving this goal, enabling the development of efficient, high-performing models tailored to psychiatric data.

This research is the first to predict multiple risk outcomes for various mental disorders using an NLP model and multicenter data in China. NLP models have been employed in several studies on mental disorders, such as predicting depression from social media text [87, 88] and exploring the neural mechanisms of schizophrenia through semantic processing methods [31]. EHR data also offer great potential due to their high professionalism

and extensive volume. Previous studies have primarily focused on structured data or specific risk-related fields, underutilizing the extensive and highly relevant textual content within the FMR [36, 38]. Our study demonstrates that incorporating FMR significantly enhances model performance and clinical utility, highlighting the importance of deeply mining this textual information in future research.

Strengths and limitations

Our analysis has several strengths. The data for the model development and validation were derived from multiple hospitals and contained a substantial sample size. An independent external testing cohort demonstrates that the model has similar predictive power with potentially different distributions of text data. The predictors used in our models are clinically accessible, which further enhances their usefulness. The present study provides insights for future work on the use of texts and FMR as features based on language modeling to achieve more accurate prognostic assessment and is expected to be potentially extended to other goals such as diagnosis, symptom recognition, and bias control.

Limitations include the fact that the data come from only one city, which leads to the possibility that the model may not be applicable in areas with greater language differences. The use of FMR also presents several challenges. Firstly, since FMRs are retrospectively written by physicians, the quality and narrative style may vary across institutions. Nevertheless, their structural consistency is supported by national documentation standards in China which mandate uniform formats and structured history-taking across all clinical levels [89, 90]. Additionally, robustness tests, the independent external testing cohort, and the large-scale, multi-centered dataset demonstrated the model's reliability in handling variability. Future work may benefit from direct quantification of structural consistency across institutions using corpus-level similarity or section alignment metrics. Second, there is the risk of feature drift, where changes in language features over time due to evolving documentation practices, diagnostic criteria, or terminologies could affect the model's performance [91, 92]. Lastly, the use of textual data, which clinicians manually write, could introduce biases related to gender, race, or pre-existing assumptions, potentially undermining the system's utility. These problems highlight the necessity for prospective validation in future studies to mitigate biases and improve the model's applicability in diverse clinical settings. Moreover, due to inconsistencies in the usage of scales across different hospitals, our study included a limited number of psychiatric scales and did not utilize scales related to psychosis or mania. Furthermore, current language models, including models like XLNet, also face limitations in efficiently processing long-text sequences [93] and optimizing computational efficiency [54]. Future research should prioritize developing solutions that effectively tackle these challenges, potentially by integrating advanced methodologies such as the Mamba architecture [94], knowledge distillation [95, 96], and causal inference [97]. Lastly, interpretability is essential for clinical applications. While providing detailed interpretability insights is currently unfeasible due to ethical considerations, the reduction in AUROC after removing each feature set was evaluated to understand the contribution of each feature to the overall model performance. Future efforts are still needed to further address this limitation through prospective evaluations and case reports to enhance interpretability in a clinically applicable and ethically responsible manner.

Conclusions

This study incorporates language as a predictor and constructs a deep learning prediction model for common psychiatric risks based on real-world clinical data. By converting structured data into texts, the model can leverage the additional information to achieve more accurate predictions and minimize the requirements of input. Our results emphasize the potential of FMR to successfully complement prognostic assessments for psychiatric inpatients and provide insights for subsequent modeling studies.

Abbreviations

AUPRC	Area under the precision-recall curve
AUROC	Area under the receiver operating characteristic curve
HTML	HyperText Markup Language
IBRAS	Impulsive Behavior Risk Assessment Scale
ICD-10	International Classification of Diseases, 10th Revision
MLP	Multilayer perceptron
MOAS	Modified Overt Aggression Scale
NGASR	Nurse's Global Assessment of Suicide Risk
NLP	Natural language processing
NPV	Negative predictive value
PPV	Positive predictive value
ROC	Receiver operating characteristic
SAS	Self-rating Anxiety Scale
SDS	Self-rating Depression Scale
SSTL	Semi-structured Transformers-based language model
STPR	Short-term need of physical restraint
LTPR	Long-term need of physical restraint
STSS	Short-term risk of suicide
LTSS	Long-term risk of suicide
STI	Short-term impulsivity
LTI	Long-term impulsivity
FMR	First medical records

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12916-025-04150-7.

Additional file 1: Tables S1–S13 and Figures S1–S4. Table S1 Assessment of included features. Table S2 Data overview. Tables S3–S4 Correlations among short- and long-term outcomes. Tables S5–S10 Model performance metrics for each outcome. Tables S11–S12 Statistical comparisons of model variants. Table S13 The overall clinical utility. Fig. S1 Schematics of language models. Fig. S2 Comparisons of different text encoders. Fig. S3 The ROC curve of semi-structured in the internal testing cohort. Fig. S4 Robustness test.

Acknowledgements
The authors of this study thank the hospital information department staff and directors including Dr Zhihao Chen (East China University of Science and Technology), Dong Wang (Tongji University Mental Health Center), Feng Wang (Tongji University Mental Health Center), Yi Gu (Shanghai Putuo District Mental Health Center), Yu Mei (Shanghai Mental Health Center), Yifan Liu (Shanghai Mental Health Center), and Yichao Yin (Shanghai Changning Mental Health Center) for their work on data collection, management, and curation; hospital and institutional directors including Dr Hua Wang (Shanghai Putuo District Mental Health Center), Fazhan Chen (Tongji University Mental Health Center), Dianxu Feng (Shanghai Putuo District Health Committee), Guoquan Zhou (Shanghai Putuo District Mental Health Center), Weizhong Shi (Shanghai Hospital Development Center), Liangliang Chen (Shanghai Changning District Mental Health Center), Hui Li (Shanghai Mental Health Center), Chunbo Li (Shanghai Mental Health Center), Hong Qiu (Shanghai Mental Health Center), Lei Wang (Tongji University), Gang Zhu (Shanghai Municipal Finance Bureau), Guozhen Lin (Shanghai Ruijin Hospital), Yanping Zhang (Shanghai Jinshan District Mental Health Center), Xuehui Li (Shanghai Ruijin Hospital) for their work on project administration. They also thank other staff affiliated with the data source hospitals and colleges for their work on patient assessment, data collection, and technical guidance.

Authors' contributions
EZ and ZA had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors made substantial contributions to the study conception and design, data acquisition or analysis, and manuscript preparation. Concept and design: EZ, JW, GZ, JH, HL, ZA Acquisition, analysis, or interpretation of data: EZ, JW, GZ, CL, FC, KJ, LC, YZ1, YY, JQ, HW, FW, DW, ZC, HL, YC Drafting of the manuscript: EZ, JW, XZ, XZ2, ZW, JH Critical review of the manuscript for important intellectual content: EZ, JW, GZ, CL, WS, HL, ZA Statistical analysis: EZ, JW, ZA Obtained funding: GZ, LC, XZ, HL Administrative, technical, or material support: EZ, JW, GZ, CL, FC, KJ, LC, YZ1, YY, HW, JH, HL, ZA Supervision: GZ, CL, FC, HL, ZA All authors read and approved the final manuscript.

Funding
This work was supported by projects from Shanghai Putuo District Municipal Commission of Health (ptkwws202413); Shanghai Municipal Health Commission (202340018); Data Sharing and Emulation of Clinical Trials, CCS-DASET (SHDC2024CRI008); Shanghai Changning District Municipal Commission of Health (CNWJXY026); and School of Innovation & Entrepreneurship, Tongji University (S202310247388, X2024085, and X2024048).

Data availability
The data analyzed in this study are not available to the public in accordance with national legislation (Mental Health Law of the People's Republic of China). Requests for data should be made through the corresponding author upon reasonable cause, subject to data license agreements with School of Medicine of Tongji University and data source hospitals.

Declarations

Ethics approval and consent to participate
This study was approved by the Ethics Committee of Tongji University Mental Health Center (Approval No. PDJW-IIT-2023-017CS). The requirement for informed consent was waived due to the retrospective and de-identified nature of the data. All procedures were conducted in accordance with the Declaration of Helsinki and relevant local regulations.

Consent for publication
Not applicable. This study does not contain any individual person's data in any form (including individual details, images, or videos).

Competing interests
The authors declare no competing interests.

Author details
[1]School of Medicine, Tongji University, Shanghai, China. [2]Shanghai Putuo Mental Health Center, Putuo District, Shanghai, China. [3]Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center, Shanghai Jiaotong University School of Medicine, Shanghai 200030, China. [4]Clinical Research Center for Mental Disorders, Shanghai Pudong New Area Mental Health Center, School of Medicine, Chinese-German Institute of Mental Health, Tongji University, Shanghai, China. [5]Shanghai Changning Mental Health Center, Changning District, Shanghai, China. [6]Division of Gastrointestinal Surgery, Department of General Surgery, West China Hospital, Sichuan University, Chengdu, China. [7]Department of Infection Control, West China Hospital, Sichuan University, Chengdu, China. [8]Shanghai Jinshan District Mental Health Center, Jinshan District, Shanghai, China. [9]Lakefield College School, Lakefield, ON, Canada. [10]Shanghai Hospital Development Center, Shanghai, China. [11]East China University of Science and Technology, Shanghai, China. [12]University Clinic of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, RWTH Aachen University, Aachen 52074, Germany. [13]Department of Medical Statistics, School of Medicine, Tongji University, Shanghai, China.

Received: 12 August 2024 Accepted: 19 May 2025
Published online: 28 May 2025

References

1. Beaudry G, Canal-Rivero M, Ou J, Matharu J, Fazel S, Yu R. Evaluating the risk of suicide and violence in severe mental illness: a feasibility study of two risk assessment tools (OxMIS and OxMIV) in general psychiatric settings. Front Psychiatry. 2022;13:871213.
2. Miller JN, Black DW. Bipolar disorder and suicide: a review. Curr Psychiatry Rep. 2020;22(2):6.
3. Bai W, Liu ZH, Jiang YY, Zhang QE, Rao WW, Cheung T, Hall BJ, Xiang YT. Worldwide prevalence of suicidal ideation and suicide plan among people with schizophrenia: a meta-analysis and systematic review of epidemiological surveys. Transl Psychiatry. 2021;11(1):552.
4. Liang Y, Wu M, Zou Y, Wan X, Liu Y, Liu X. Prevalence of suicide ideation, self-harm, and suicide among Chinese patients with schizophrenia: a systematic review and meta-analysis. Front Public Health. 2023;11:1097098.
5. Richard-Lepouriel H, Kung AL, Hasler R, Bellivier F, Prada P, Gard S, Ardu S, Kahn JP, Dayer A, Henry C, et al. Impulsivity and its association with childhood trauma experiences across bipolar disorder, attention deficit hyperactivity disorder and borderline personality disorder. J Affect Disord. 2019;244:33–41.
6. Saddichha S, Schuetz C. Impulsivity in remitted depression: a meta-analytical review. Asian J Psychiatr. 2014;9:13–6.
7. Krakowski MI, Czobor P. Depression and impulsivity as pathways to violence: implications for antiaggressive treatment. Schizophr Bull. 2014;40(4):886–94.
8. Hamza CA, Willoughby T, Heffer T. Impulsivity and nonsuicidal self-injury: a review and meta-analysis. Clin Psychol Rev. 2015;38:13–24.
9. May AM, Klonsky ED, Klein DN. Predicting future suicide attempts among depressed suicide ideators: a 10-year longitudinal study. J Psychiatr Res. 2012;46(7):946–52.
10. Bleijlevens MH, Wagner LM, Capezuti E, Hamers JP. Physical restraints: consensus of a research definition using a modified Delphi technique. J Am Geriatr Soc. 2016;64(11):2307–10.
11. Baek IC, Jo S, Kim EJ, Lee GR, Lee DH, Jeon HJ. A review of suicide risk assessment tools and their measured psychometric properties in Korea. Front Psychiatry. 2021;12:679779.

12. Zhong S, Yu R, Cornish R, Wang X, Fazel S. Assessment of violence risk in 440 psychiatric patients in China: examining the feasibility and acceptability of a novel and scalable approach (FoVOx). BMC Psychiatry. 2021;21(1):120.
13. Parsa MPM, Koudys JW, Ruocco AC. Suicide risk detection using artificial intelligence: the promise of creating a benchmark dataset for research on the detection of suicide risk. Front Psychiatry. 2023;14:1186569.
14. Danielsen AA, Fenger MHJ, Østergaard SD, Nielbo KL, Mors O. Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data. Acta Psychiatr Scand. 2019;140(2):147–57.
15. Nesca M, Katz A, Leung CK, Lix LM. A scoping review of preprocessing methods for unstructured text data to assess data quality. Int J Popul Data Sci. 2022;7(1):1757.
16. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data. 2021;8(1):140.
17. Sun J, Dong QX, Wang SW, Zheng YB, Liu XX, Lu TS, Yuan K, Shi J, Hu B, Lu L, et al. Artificial intelligence in psychiatry research, diagnosis, and therapy. Asian J Psychiatr. 2023;87:103705.
18. Zanardi R, Prestifilippo D, Fabbri C, Colombo C, Maron E, Serretti A. Precision psychiatry in clinical practice. Int J Psychiatry Clin Pract. 2021;25(1):19–27.
19. Nour MM, Huys QJM. Natural language processing in psychiatry: a field at an inflection point. Biol Psychiatry Cogn Neurosci Neuroimaging. 2023;8(10):979–81.
20. Bernstorff M, Hansen L, Enevoldsen K, Damgaard J, Hæstrup F, Perfalk E, Danielsen AA, Østergaard SD. Development and validation of a machine learning model for prediction of type 2 diabetes in patients with mental illness. Acta Psychiatrica Scandinavica. 2025;151(3):245–58.
21. Sommer IE. J NdB: How to reap the benefits of language for psychiatry. Psychiatry Res. 2022;318:114932.
22. Yang T, Li F, Ji D, Liang X, Xie T, Tian S, Li B, Liang P. Fine-grained depression analysis based on Chinese micro-blog reviews. Inf Process Manage. 2021;58(6):102681.
23. Amini S, Hao B, Zhang L, Song M, Gupta A, Karjadi C, Kolachalama VB, Au R, Paschalidis IC. Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach. Alzheimers Dement. 2023;19(3):946–55.
24. Wiest IC, Verhees FG, Ferber D, Zhu J, Bauer M, Lewitzka U, Pfennig A, Mikolas P, Kather JN: Detection of suicidality from medical text using privacy-preserving large language models. Br J Psychiatry 2024:1-6. https://doi.org/10.1192/bjp.2024.134.
25. Ewbank MP, Cummins R, Tablan V, Bateup S, Catarino A, Martin AJ, Blackwell AD. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. JAMA Psychiat. 2020;77(1):35–43.
26. Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is All you Need. In: Neural Information Processing Systems: 2017;2017. https://doi.org/10.48550/arXiv.1706.03762.
27. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A et al: Training language models to follow instructions with human feedback. ArXiv 2022, abs/2203.02155. https://doi.org/10.48550/arXiv.2203.02155.
28. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F et al. LLaMA: Open and Efficient Foundation Language Models. ArXiv 2023, abs/2302.13971. https://doi.org/10.48550/arXiv.2302.13971.
29. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J-M, Zhang R, Xu R, Zhu Q, Ma S, Wang P et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. In: 2025; 2025. https://doi.org/10.48550/arXiv.2501.12948.
30. Achiam OJ, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S et al. GPT-4 Technical Report. In: 2023; 2023. https://doi.org/10.48550/arXiv.2303.08774.
31. Nour MM, McNamee DC, Liu Y, Dolan RJ. Trajectories through semantic spaces in schizophrenia and the relationship to ripple bursts. Proc Natl Acad Sci U S A. 2023;120(42):e2305290120.
32. Tan EJ, Sommer IEC, Palaniyappan L. Language and psychosis: tightening the association. Schizophr Bull. 2023;49(Suppl_2):S83-s85.
33. Mowery D, Smith H, Cheney T, Stoddard G, Coppersmith G, Bryan C, Conway M. Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study. J Med Internet Res. 2017;19(2):e48.
34. Meng Y, Speier W, Ong MK, Arnold CW. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. IEEE J Biomed Health Inform. 2021;25(8):3121–9.
35. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. BEHRT: transformer for electronic health records. Sci Rep. 2020;10(1):7155.
36. Cliffe C, Cusick M, Vellupillai S, Shear M, Downs J, Epstein S, Pathak J, Dutta R. A multisite comparison using electronic health records and natural language processing to identify the association between suicidality and hospital readmission amongst patients with eating disorders. Int J Eat Disord. 2023;56(8):1581–92.
37. Irving J, Patel R, Oliver D, Colling C, Pritchard M, Broadbent M, Baldwin H, Stahl D, Stewart R, Fusar-Poli P. Using natural language processing on electronic health records to enhance detection and prediction of psychosis risk. Schizophr Bull. 2021;47(2):405–14.
38. Bittar A, Velupillai S, Roberts A, Dutta R. Using general-purpose sentiment lexicons for suicide risk assessment in electronic health records: corpus-based analysis. JMIR Med Inform. 2021;9(4):e22397.
39. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
40. Sun Y, Li X, Xu L, Ma Z, Yang Y, Yin T, Gao Z, Gong X, Li L, Liu Q, et al. Health-related risky behaviors in Chinese adolescents with autism: a cross-sectional study. Child Adolesc Psychiatry Ment Health. 2021;15(1):39.
41. Zhang L, Yang Y, Li M, Zhou X, Zhang K, Yin X, Liu H. The prevalence of suicide ideation and predictive factors among pregnant women in the third trimester. BMC Pregnancy Childbirth. 2022;22(1):266.
42. Chen Z, Zhang Y, Guo Y, Meng H, Ji J. Screening and nursing of adult psychiatric patients at high risk of impulsive behavior. Chin J Modern Nurs. 2014;32:2.
43. Chen ZZY, Guo Y, Meng H, Ji J. Screening and nursing of adult psychiatric patients at high risk of impulsive behavior. Chin J Modern Nurs. 2014;20(32):4115–6.
44. Li X, Ge H, Zhou D, Wu X, Qi G, Chen Z, Yu C, Zhang Y, Yu H, Wang C. Reduced serum VGF levels are linked with suicide risk in Chinese Han patients with major depressive disorder. BMC Psychiatry. 2020;20(1):225.
45. Cutcliffe JR, Barker P. The Nurses' Global Assessment of Suicide Risk (NGASR): developing a tool for clinical practice. J Psychiatr Ment Health Nurs. 2004;11(4):393–400.
46. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In.; 2019: arXiv:1912.01703. https://doi.org/10.48550/arXiv.1912.01703.
47. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12(null):2825–30.
48. Devlin J, Chang M-W, Lee K, Toutanova K: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: North American Chapter of the Association for Computational Linguistics: 2019; 2019. https://doi.org/10.18653/v1/N19-1423.
49. Yang Z, Dai Z, Yang Y, Carbonell JG, Salakhutdinov R, Le QV XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: Neural Information Processing Systems: 2019; 2019. https://doi.org/10.48550/arXiv.1906.08237.
50. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. ArXiv 2019, abs/1909.11942. https://doi.org/10.48550/arXiv.1909.11942.
51. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv 2019, abs/1907.11692. https://doi.org/10.48550/arXiv.1907.11692.
52. Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. ArXiv 2020, abs/2004.05150. https://doi.org/10.48550/arXiv.2004.05150.
53. Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontañón S, Pham P, Ravula A, Wang Q, Yang L et al. Big Bird: Transformers for Longer Sequences. ArXiv 2020, abs/2007.14062. https://doi.org/10.48550/arXiv.2007.14062.
54. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In.; 2019: arXiv:1906.08237. https://doi.org/10.48550/arXiv.1906.08237.

55. Wang K, Huang J, Liu Y, Cao B, Fan J: Combining Feature Selection Methods with BERT: An In-depth Experimental Study of Long Text Classification. In: 2021. Cham: Springer International Publishing; 2021. p. 567-582. https://doi.org/10.1007/978-3-030-67537-0_34.
56. Chen T, Kornblith S, Swersky K, Norouzi M, Hinton GE: Big Self-Supervised Models are Strong Semi-Supervised Learners. ArXiv 2020, abs/2006.10029. https://doi.org/10.48550/arXiv.2006.10029.
57. Caruana RA. Multitask Learning: A Knowledge-Based Source of Inductive Bias: Multitask Learning: A Knowledge-Based Source of Inductive Bias; 1995. https://doi.org/10.1016/B978-1-55860-307-3.50012-5.
58. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. https://doi.org/10.1145/2939672.2939785.
59. Bouckaert RR, Frank E. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In: Advances in Knowledge Discovery and Data Mining: 2004// 2004. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. p. 3-12. https://doi.org/10.1007/978-3-540-24775-3_3.
60. Lin T-Y, Goyal P, Girshick RB, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2017;42:318–27.
61. Verboven S, Chaudhary MH, Berrevoets J, Verbeke W. HydaLearn: Highly Dynamic Task Weighting for Multitask Learning with Auxiliary Tasks. ArXiv 2020, abs/2008.11643. https://doi.org/10.48550/arXiv.2008.11643.
62. McKinney W. Data Structures for Statistical Computing in Python In. Edited by Millman SvdWaJ; 2010. https://doi.org/10.25080/Majora-92bf1922-00a.
63. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanin P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, Del Rio JF, Wiebe M, Peterson P, Gerard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abassi H, Gohlke C, Oliphant TE. Array programming with NumPy. Nature. 2020;585:357–62.
64. Falcon NSDaJBaJSaAHaTKaLDLaDSaCQaMGaW: TorchMetrics - Measuring Reproducibility in PyTorch. J Open Source Software. 2022;1.2.0. https://doi.org/10.21105/joss.04101.
65. Fluss R, Faraggi D, Reiser B. Estimation of the Youden index and its associated cutoff point. Biom J. 2005;47(4):458–72.
66. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–38.
67. Bengio Y, Nadeau C. Inference for the Generalization Error. Machine Learning 1999(99s-25). https://doi.org/10.1023/A:1024068626366.
68. Su J. SimBERT: Integrating Retrieval and Generation into BERT. In.; 2020.
69. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol. 2022;75(1):25–36.
70. Chen Q, Zhang-James Y, Barnett EJ, Lichtenstein P, Jokinen J, D'Onofrio BM, Faraone SV, Larsson H, Fazel S. Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: a machine learning study using Swedish national registry data. PLoS Med. 2020;17(11):e1003416.
71. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol. 2018;14:91–118.
72. Todd A, Alonzo. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating: By Ewout W. Steyerberg. Amjepidemiol; 2009. https://doi.org/10.1007/978-3-030-16399-0.
73. Wasser T, Strockbine B, Uyanwune Y, Kapoor R. Restraint and seclusion practices and policies in U.S. forensic psychiatric hospitals. J Am Acad Psychiatry Law. 2023;51(4):566–74.
74. Belayneh Z, Chavulak J, Lee DA, Petrakis M, Haines TP. Prevalence and variability of restrictive care practice use (physical restraint, seclusion and chemical restraint) in adult mental health inpatient settings: a systematic review and meta-analysis. J Clin Nurs. 2024;33(4):1256–81.
75. Mann JJ, Michel CA, Auerbach RP. Improving suicide prevention through evidence-based strategies: a systematic review. Am J Psychiatry. 2021;178(7):611–24.
76. Weltens I, Bak M, Verhagen S, Vandenberk E, Domen P, van Amelsvoort T, Drukker M. Aggression on the psychiatric ward: prevalence and risk factors. A systematic review of the literature. PLoS One. 2021;16(10):e0258346.
77. Wang Y, Cai X, Ma Y, Yang Y, Pan CW, Zhu X, Ke C. Metabolomics on depression: a comparison of clinical and animal research. J Affect Disord. 2024;349:559–68.
78. Smith ML, Wade JB, Wolstenholme J, Bajaj JS. Gut microbiome-brain-cirrhosis axis. Hepatology. 2024;80(2):465–85.
79. Rose CF, Amodio P, Bajaj JS, Dhiman RK, Montagnese S, Taylor-Robinson SD, Vilstrup H, Jalan R. Hepatic encephalopathy: novel insights into classification, pathophysiology and therapy. J Hepatol. 2020;73(6):1526–47.
80. Drew DA, Weiner DE, Sarnak MJ. Cognitive impairment in CKD: pathophysiology, management, and prevention. Am J Kidney Dis. 2019;74(6):782–90.
81. Pépin M, Klimkowicz-Mrowiec A, Godefroy O, Delgado P, Carriazo S, Ferreira AC, Golenia A, Malyszko J, Grodzicki T, Giannakou K, et al. Cognitive disorders in patients with chronic kidney disease: approaches to prevention and treatment. Eur J Neurol. 2023;30(9):2899–911.
82. Zhuo C, Liu W, Jiang R, Li R, Yu H, Chen G, Shan J, Zhu J, Cai Z, Lin C, et al. Metabolic risk factors of cognitive impairment in young women with major psychiatric disorder. Front Psychiatry. 2022;13:880031.
83. Morozova A, Zorkina Y, Abramova O, Pavlova O, Pavlov K, Soloveva K, Volkova M, Alekseeva P, Andryshchenko A, Kostyuk G, et al. Neurobiological highlights of cognitive impairment in psychiatric disorders. Int J Mol Sci. 2022;23(3):1217.
84. Dragasek J, Minar M, Valkovic P, Pallayova M. Factors associated with psychiatric and physical comorbidities in bipolar disorder: a nationwide multicenter cross-sectional observational study. Front Psychiatry. 2023;14:1208551.
85. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al. Language Models are Few-Shot Learners. ArXiv 2020, abs/2005.14165. https://doi.org/10.48550/arXiv.2005.14165.
86. Misra I. Maaten Lvd: Self-supervised learning of pretext-invariant representations. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020;2019:6706–16.
87. Ma W, Qiu S, Miao J, Li M, Tian Z, Zhang B, Li W, Feng R, Wang C, Cui Y, et al. Detecting depression tendency based on deep learning and multi-sources data. Biomed Signal Process Control. 2023;86:105226.
88. Chiong R, Budhi GS, Dhakal S, Chiong F. A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. Comput Biol Med. 2021;135:104499.
89. National Health Commission of the People's Republic of China: Basic Norms for Medical Record Writing. In.; 2010.
90. Chinese Medical Doctor Association: Standardized residency training program (2022 edition). In.; 2022.
91. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv. 2014;46(4):Article 44.
92. Hansen L, Enevoldsen K, Bernstorff M, Perfalk E, Danielsen AA, Nielbo KL, Østergaard SD. Lexical stability of psychiatric clinical notes from electronic health records over a decade. Acta Neuropsychiatr; 2023:1-11. https://doi.org/10.1017/neu.2023.46.
93. Zaheer M, Guruganesh G, Dubey A, Ainslie J, Alberti C, Ontanon S, Pham P, Ravula A, Wang Q, Yang L et al. Big Bird: Transformers for Longer Sequences. In.; 2020: arXiv:2007.14062. https://doi.org/10.48550/arXiv.2007.14062.
94. Gu A, Dao T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In.; 2023: arXiv:2312.00752. https://doi.org/10.48550/arXiv.2312.00752.
95. Liu X, He P, Chen W, Gao J. Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding. In.; 2019: arXiv:1904.09482. https://doi.org/10.48550/arXiv.1901.11504.
96. Yuan L, Tay FEH, Li G, Wang T, Feng J. Revisiting Knowledge Distillation via Label Smoothing Regularization. In.; 2019: arXiv:1909.11723. https://doi.org/10.48550/arXiv.1909.11723.
97. Feder A, Keith KA, Manzoor E, Pryzant R, Sridhar D, Wood-Doughty Z, Eisenstein J, Grimmer J, Reichart R, Roberts ME et al. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. In.; 2021: arXiv:2109.00725. https://doi.org/10.48550/arXiv.2109.00725.

Publisher's Note