JAMA Psychiatry | Original Investigation | AI IN PSYCHIATRY

Predicting Diagnostic Progression to Schizophrenia or Bipolar Disorder via Machine Learning

Lasse Hansen, MSc, PhD; Martin Bernstorff, MD, PhD; Kenneth Enevoldsen, MSc, PhD; Sara Kolding, MSc; Jakob Grøhn Damgaard, MSc; Erik Perfalk, MD; Kristoffer Laigaard Nielbo, PhD; Andreas Aalkjær Danielsen, MD, PhD; Søren Dinesen Østergaard, MD, PhD

IMPORTANCE The diagnosis of schizophrenia and bipolar disorder is often delayed several years despite illness typically emerging in late adolescence or early adulthood, which impedes initiation of targeted treatment.

OBJECTIVE To investigate whether machine learning models trained on routine clinical data from electronic health records (EHRs) can predict diagnostic progression to schizophrenia or bipolar disorder among patients undergoing treatment in psychiatric services for other mental illness.

DESIGN, SETTING, AND PARTICIPANTS This cohort study was based on data from EHRs from the Psychiatric Services of the Central Denmark Region. All patients aged 15 to 60 years with at least 2 contacts (at least 3 months apart) with the Psychiatric Services of the Central Denmark Region between January 1, 2013, and November 21, 2016, were included. Analysis occurred from December 2022 to November 2024.

EXPOSURES Predictors based on EHR data, including medications, diagnoses, and clinical notes.

MAIN OUTCOMES AND MEASURES Diagnostic transition to schizophrenia or bipolar disorder within 5 years, predicted 1 day before outpatient contacts by means of elastic net regularized logistic regression and extreme gradient boosting (XGBoost) models. The area under the receiver operating characteristic curve (AUROC) was used to determine the best performing model.

RESULTS The study included 24 449 patients (median [Q1-Q3] age at time of prediction, 32.2 [24.2-42.5] years; 13 843 female [56.6%]) and 398 922 outpatient contacts. Transition to the first occurrence of either schizophrenia or bipolar disorder was predicted by the XGBoost model, with an AUROC of 0.70 (95% CI, 0.70-0.70) on the training set and 0.64 (95% CI, 0.63-0.65) on the test set, which consisted of 2 held-out hospital sites. At a predicted positive rate of 4%, the XGBoost model had a sensitivity of 9.3%, a specificity of 96.3%, and a positive predictive value (PPV) of 13.0%. Predicting schizophrenia separately yielded better performance (AUROC, 0.80; 95% CI, 0.79-0.81; sensitivity, 19.4%; specificity, 96.3%; PPV, 10.8%) than was the case for bipolar disorder (AUROC, 0.62, 95% CI, 0.61-0.63; sensitivity, 9.9%; specificity, 96.2%; PPV, 8.4%). Clinical notes proved particularly informative for prediction.

CONCLUSIONS AND RELEVANCE These findings suggest that it is possible to predict diagnostic transition to schizophrenia and bipolar disorder from routine clinical data extracted from EHRs, with schizophrenia being notably easier to predict than bipolar disorder.

JAMA Psychiatry. doi:10.1001/jamapsychiatry.2024.4702
Published online February 19, 2025.

+ Multimedia

+ Supplemental content

Author Affiliations: Department of Affective Disorders, Aarhus University Hospital—Psychiatry, Aarhus, Denmark (Hansen, Bernstorff, Enevoldsen, Kolding, Damgaard, Perfalk, Danielsen, Østergaard); Department of Clinical Medicine, Aarhus University, Aarhus, Denmark (Hansen, Bernstorff, Kolding, Damgaard, Perfalk, Danielsen, Østergaard); Psychosis Research Unit, Aarhus University Hospital—Psychiatry, Aarhus, Denmark (Danielsen); Center for Humanities Computing, Department of Culture and Society, Aarhus, Denmark (Hansen, Bernstorff, Enevoldsen, Kolding, Damgaard, Nielbo).

Corresponding Author: Lasse Hansen, MSc, PhD, Department of Clinical Medicine, Aarhus University, Palle Juul-Jensens Boulevard 175, 8200 Aarhus C, Denmark (lasse.hansen@clin.au.dk).

Schizophrenia and bipolar disorder are severe mental disorders that often impair the ability to lead a normal life.[1,2] Indeed, both disorders have severe negative consequences for social functioning, work ability, and lifespan.[1-3] Despite typically emerging in late adolescence or early adulthood, diagnosis is often delayed several years.[4,5] Timely and accurate diagnosis is crucial because diagnostic delay impedes the initiation of targeted treatment. Furthermore, the longer the duration of untreated illness, the worse the prognosis becomes.[4,5] However, timely diagnosis of schizophrenia and bipolar disorder is challenging due to the prodromal phase, in which patients do not yet meet full diagnostic criteria, and due to symptom overlap with other disorders such as anxiety and depression.[1,6,7] In fact, many patients who eventually receive a diagnosis of schizophrenia or bipolar disorder have previously received treatment for other and less severe mental disorders.[8,9]

Machine learning applied to electronic health record (EHR) data likely holds great promise for assisting in the diagnosis of complex psychiatric conditions such as schizophrenia and bipolar disorder.[10] Clinical notes in EHRs are presumably particularly valuable in this context because they contain comprehensive descriptions of symptoms, treatment responses, and patient-clinician interactions. Due to the sheer amount of unstructured text in these notes, often covering several years, it is difficult for clinicians to harness and utilize the comprehensive information embedded within them efficiently. Using methods from natural language processing and deep learning, it may be possible to extract and synthesize data from clinical notes, uncovering patterns that could indicate an impending progression from less severe conditions to schizophrenia or bipolar disorder.[11]

This study investigates whether machine learning models trained on routine clinical data from EHRs can predict the risk of diagnostic progression to schizophrenia or bipolar disorder among patients undergoing treatment in psychiatric services. Early diagnosis enabled by machine learning models could potentially reduce the duration of untreated illness in schizophrenia and bipolar disorder, leading to better prognoses and improved illness trajectories. Building upon prior research such as Irving et al,[12] Wang et al,[13] and Raket et al,[14] which focused on predicting progression to psychosis and bipolar disorder from a single index date, the present study advances the field in 2 key aspects. First, we introduce a clinically relevant prediction modeling framework that dynamically issues predictions at multiple critical time points throughout a patient's course of contact with psychiatric services. Second, we conduct a comprehensive analysis of various methods for integrating text from clinical notes into EHR-based prediction models.

Methods

The use of EHRs from the Central Denmark Region (CDR) for this cohort study without informed consent was approved by the Legal Office of the CDR in accordance with the Danish Health Care Act §46, Section 2. According to the Danish Committee Act, ethical review board approval is not required for studies based solely on data from EHRs. Reporting follows the Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis With Artificial Intelligence (TRIPOD+AI)[15] guideline. An illustration of the methods used in this study is shown in **Figure 1**. Additional information can be found in the eMethods in Supplement 1.

Key Points

Question Can diagnostic progression to schizophrenia or bipolar disorder be accurately predicted from routine clinical data extracted from electronic health records?

Findings In this cohort study of 24 449 patients aged 15 to 60 years with a total of 398 922 outpatient contacts to the Psychiatric Services of the Central Denmark Region between 2013 and 2016, progression to schizophrenia was predicted with higher accuracy than bipolar disorder, which proved to be a more difficult target.

Meaning These findings suggest that detecting progression to schizophrenia through machine learning based on routine clinical data is feasible, which may reduce diagnostic delay and duration of untreated illness.

Data

The study included data from an updated version of the Psychiatric Clinical Outcome Prediction cohort,[17] spanning routine EHR data from all individuals with at least 1 contact with the Psychiatric Services of the CDR (catchment population of approximately 1.3 million) in the period from January 1, 2011, to November 22, 2021 (Figure 1A).

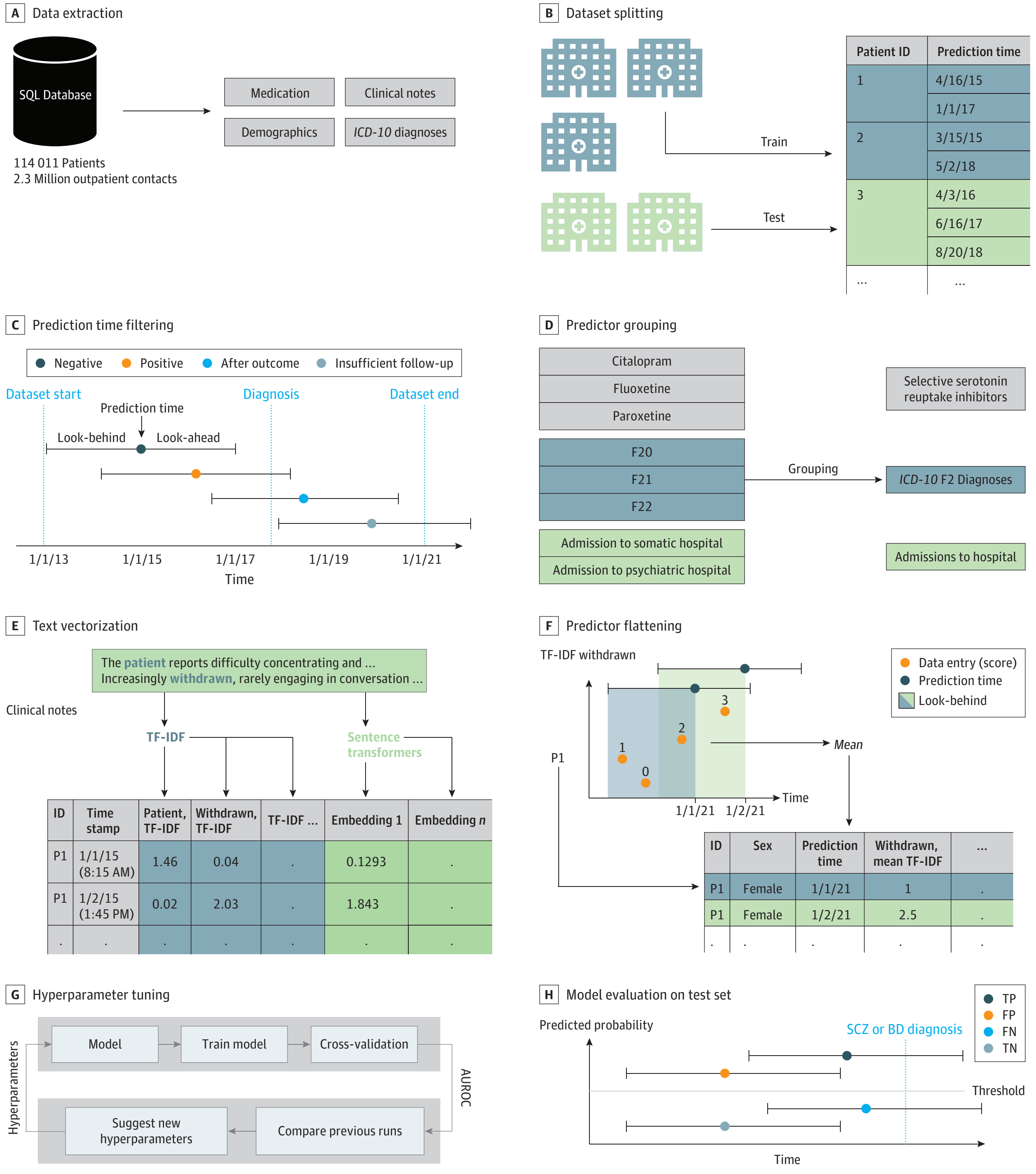
Data Split

The data were split into a training set and an external test set by hospital site. Specifically, patient contacts with the hospitals in the western and eastern part of the CDR (Aarhus, Herning, Holstebro, Randers, Horsens, and Gødstrup) were used for training, while patient contacts with the central part of the region (Silkeborg and Viborg) were used for testing (Figure 1B).

Outcome Definition

Diagnostic progression to schizophrenia was defined as the time of the first *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10)* diagnosis (codes in parentheses) of either schizophrenia (F20) or schizoaffective disorder (F25). Schizoaffective disorder was included because its *ICD-10* definition is very close to that of schizophrenia. Diagnostic progression to bipolar disorder was defined as the time of the first *ICD-10* diagnosis of either a manic episode (F30) or bipolar affective disorder (F31). Models were tested with 3 different outcomes: (1) diagnostic progression to schizophrenia or bipolar disorder (joint model), (2) diagnostic progression to schizophrenia, and (3) diagnostic progression to bipolar disorder. For the joint outcome, the first occurring diagnosis was used. Models were trained to predict the incidence of the outcomes 5 years into the future (5-year look-ahead window).

Figure 1. Data Extraction and Transformation, Model Training, and Model Testing Pipeline



Patient ID	Prediction time
1	4/16/15
	1/1/17
2	3/15/15
	5/2/18
3	4/3/16
	6/16/17
	8/20/18
...	...

ID	Time stamp	Patient, TF-IDF	Withdrawn, TF-IDF	TF-IDF ...	Embedding 1	Embedding *n*
P1	1/1/15 (8:15 AM)	1.46	0.04	.	0.1293	.
P1	1/2/15 (1:45 PM)	0.02	2.03	.	1.843	.
.	.	.	.	.	.	.

ID	Sex	Prediction time	Withdrawn, mean TF-IDF	...
P1	Female	1/1/21	1	.
P1	Female	1/2/21	2.5	.
.	.	.	.	.

This figure was adapted from Bernstorff et al.[16] A, Data were extracted from the electronic health records. B, Data were split into a training and a test set. C, Prediction times occurring after November 21, 2016, or after a diagnosis of schizophrenia (SCZ) or bipolar disorder (BD) were removed. D, Certain predictors were grouped. E, Clinical notes were turned into vectors using term frequency-inverse document frequency (TF-IDF) or sentence transformer models. F, Predictors for each prediction time were extracted by aggregating the data points within the lookbehind for each variable. G, Models were trained and optimized on the training set using 5-fold cross-validation. Hyperparameters were tuned to optimize the area under the receiver operating characteristic curve (AUROC). H, The best models were evaluated on the test set. FN indicates false negative; FP, false positive; *ICD-10*, *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision*; ID, identification; P, patient; SQL, structured query language; TN, true negative; TP, true positive.

Prediction Time Definition

Prediction times were defined as the day before a scheduled outpatient contact; this allows practitioners to prepare possible interventions (eg, a focus on symptoms compatible with progression to schizophrenia or bipolar disorder). At each prediction time, separate models were trained to predict whether the 3 outcomes (schizophrenia or bipolar disorder, schizophrenia only, or bipolar disorder only) occurred within 5 years following the prediction time.

Cohort Definition

The cohort was limited to outpatient contacts occurring after January 1, 2013, due to inconsistencies in the data before 2013, stemming from the gradual implementation of a new EHR system in 2011.[18,19] Only patients aged 15 to 60 years were included due to the low prevalence of schizophrenia and bipolar disorder in younger individuals (eFigure 1 in Supplement 1) and heterogeneous symptoms in late-onset schizophrenia and bipolar disorder.[20,21] Additionally, to avoid flagging patients currently under assessment for one of the disorders, predictions were issued at the earliest 3 months after a patient's first contact with the psychiatric services in the CDR. Outpatient contacts occurring after the diagnosis of schizophrenia or bipolar disorder were removed. If a patient moved to a hospital in a different split (train or test), only information from the first hospital treatment course was used. Prediction times occurring after November 21, 2016, were removed because they did not have the required 5 years of follow-up (Figure 1C). See the eMethods in Supplement 1 for additional details and eFigure 2 in Supplement 1 for a graphical overview.

Predictor Construction

A full list of predictors is shown in the eTable in Supplement 2. Notably, only routine clinical data from the EHRs were considered for predictors. There was no data collection for the purpose of this study.

Structured Predictors

Predictors from structured data were constructed by looking back at a specified period (the look-behind window) from each prediction time and extracting a single value for each predictor. When multiple values were present in the look-behind window, we applied an appropriate aggregation function, such as the mean or count. If no values were present in the look-behind window, a fallback value (0 or not a number [NaN]) was used. Predictors were created using look-behind windows of 182 days, 365 days, and 730 days to incorporate different temporal contexts. Predictor construction was conducted using the timeseriesflattener version 2.2.0 package for Python version 3.11.8 (Python Software Foundation),[22] and included demographics, hospital contacts (psychiatric hospital admissions, outpatient contacts with psychiatric services, and somatic hospital contacts), diagnoses, medications (grouped as shown in eTable 1 in Supplement 1) administered during inpatient stays, and scores from rating scales.

Text Predictors

Free-form clinical notes (note types are specified in eTable 2 in Supplement 1) from the EHRs were embedded as numerical feature vectors using 3 different methods: (1) term-frequency of predefined words describing psychopathology, (2) term frequency-inverse document frequency (TF-IDF), and (3) sentence transformers.[23] For further details, see the eMethods in Supplement 1. Text embeddings were processed similarly to structured features; however, only a single look-behind (730 days) was used. See Figure 1E-F for an illustration.

Model Training

Separate models were trained and optimized for each of the 3 outcomes separately (schizophrenia or bipolar disorder [joint model], schizophrenia only, and bipolar disorder only), each following the process outlined in the following sections. Two common state-of-the-art machine learning models, elastic net regularized logistic regression and extreme gradient boosting (XGBoost),[24] were trained and individually hyperparameter tuned using Optuna version 3.4.0 (Akiba et al)[25] (Figure 1G).

Data Augmentation

Data augmentation using synthetically generated data has been proposed to improve performance on multiple classification tasks within health care.[26,27] During training, experiments were therefore conducted to augment the training data with synthetic data generated using 2 methods: tabular denoising diffusion probabilistic models (TabDDPM) and synthetic minority oversampling technique. See the eMethods in Supplement 1 for additional details.

Statistical Analysis

Statistical analysis was conducted from December 2022 to November 2024. The threshold for statistical significance was a 2-tailed $P < .05$. Python version 3.11.8 was used for all analyses. The code for the analyses is available via GitHub.[28]

Model Evaluation

The best-performing model in terms of the highest area under the receiver operating characteristic curve (AUROC) after hyperparameter tuning was retrained on the entire training set and applied to the test set (Figure 1H). All evaluation metrics are based on the test set unless otherwise stated and numbers in parentheses are 95% CIs based on 100 bootstrap resamples. The AUROC was calculated for global performance. Furthermore, we report sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and the median time from first positive prediction to the outcome at specific classification thresholds. Predictor importance was estimated via information gain. As a sensitivity analysis of the best-performing joint model, we tested how it performed in predicting schizophrenia or bipolar disorder separately.

Robustness Analyses

We performed stratified analyses of the stability of model predictions in terms of AUROC over time, demographics, and individual outcomes (schizophrenia or bipolar disorder) on the test set. Additionally, a model using the best performing hyperparameters was trained on an 85% to 15%

Table 1. Descriptive Statistics for Individual Patients and Outpatient Contacts That Were Eligible for Prediction With a 5-Year Look-Ahead Period

	Patients or outpatient contacts, No. (%)[a]		
Characteristics	Overall	Train	Test
Patient characteristics			
Patients, No.	24 449	20 224	4225
Sex			
Female	13 843 (56.6)	11 332 (56.0)	2511 (59.4)
Male	10 606 (43.4)	8892 (44.0)	1714 (40.6)
Incident BP	1148 (4.7)	911 (4.5)	237 (5.6)
Incident SCZ	841 (3.4)	707 (3.5)	134 (3.2)
Time from first contact to BD diagnosis, median (Q1-Q3), d	771.0 (355.8-1352.0)	818.0 (385.0-1402.0)	593.0 (250.0-1119.0)
Time from first contact to SCZ diagnosis, median (Q1-Q3), d	811.0 (387.0-1492.0)	805.0 (386.0-1491.5)	850.5 (420.2-1505.2)
No. of outpatient contacts, median (Q1-Q3)	9.0 (3.0-21.0)	9.0 (3.0-21.0)	10.0 (4.0-23.0)
No. of admissions, median (Q1-Q3)	3.0 (1.0-6.0)	3.0 (1.0-6.0)	3.0 (1.0-5.0)
Outpatient contact characteristics			
Prediction times (outpatient contacts), No.	398 922	332 818	66 104
Positive prediction times	19 505 (4.9)	15 836 (4.8)	3669 (5.6)
Sex			
Female	257 644 (64.6)	212 579 (63.9)	45 065 (68.2)
Male	141 278 (35.4)	120 239 (36.1)	21 039 (31.8)
Age group, y			
15-18	16 819 (4.2)	15 270 (4.6)	1549 (2.3)
19-20	22 558 (5.7)	17 782 (5.3)	4776 (7.2)
21-30	136 972 (34.3)	112 024 (33.7)	24 948 (37.7)
31-40	100 437 (25.2)	84 150 (25.3)	16 287 (24.6)
41-50	78 922 (19.8)	67 336 (20.2)	11 586 (17.5)
51-60	43 214 (10.8)	36 256 (10.9)	6958 (10.5)
Age, median (Q1-Q3), y	32.2 (24.2-42.5)	32.4 (24.3-42.7)	30.8 (23.4-41.7)
Incident BD within 5 y	11 624 (2.9)	9387 (2.8)	2237 (3.4)
Incident SCZ within 5 y	8319 (2.1)	6841 (2.1)	1478 (2.2)
No. of admissions in prior 2 y, median (Q1-Q3)	0.0 (0.0-0.0)	0.0 (0.0-0.0)	0.0 (0.0-0.0)
ICD-10 diagnosis in prior 2 y			
F0 disorders	7017 (1.8)	5693 (1.7)	1324 (2.0)
F1 disorders	28 536 (7.2)	23 602 (7.1)	4934 (7.5)
F2 disorders	13 802 (3.5)	12 099 (3.6)	1703 (2.6)
F3 disorders	139 808 (35.0)	117 277 (35.2)	22 531 (34.1)
F4 disorders	142 101 (35.6)	122 426 (36.8)	19 675 (29.8)
F5 disorders	23 294 (5.8)	20 404 (6.1)	2890 (4.4)
F6 disorders	75 525 (18.9)	64 022 (19.2)	11 503 (17.4)
F7 disorders	5976 (1.5)	5598 (1.7)	378 (0.6)
F8 disorders	15 268 (3.8)	13 373 (4.0)	1895 (2.9)
F9 disorders	67 306 (16.9)	52 891 (15.9)	14 415 (21.8)

Abbreviations: BD, bipolar disorder; *ICD-10*, International Statistical Classification of Diseases and Related Health Problems, Tenth Revision; SCZ, schizophrenia.

[a] The train set includes the hospital units in Aarhus, Herning, Holstebro, Randers, Horsens, and Gødstrup, Denmark, and the test set covers Silkeborg and Viborg, Denmark.

train-test split, stratified only by patient, not hospital site, to assess performance on unseen patients attending the same psychiatric services.

Results

The cohort consisted of 24 449 unique patients (median [Q1-Q3] age at time of prediction, 32.2 [24.2-42.5] years; 13 843 female [56.6%]) with 398 922 outpatient contacts eligible for prediction. **Table 1** shows an overview of the number of patients and outpatient contacts in each split, along with demographic characteristics. The largest feature set contained 1092 predictors, covering diagnoses, medications, admissions, and embeddings derived from clinical notes (see the eTable in Supplement 2).

Predictor Selection and Data Augmentation

The text-feature set that provided the best predictive performance on the training set was TF-IDF with 1000 features, trained on all note types (eTable 3 in Supplement 1). Consequently, this feature set was used for all subsequent analyses.

The data augmentation method that provided the best predictive performance on the training set was TabDDPM with a 2 × multiplier for the minority class (eTable 4 in Supplement 1). That is, adding synthetic data equivalent to twice the number of positive outcomes (onset of schizophrenia or bipolar disorder within 5 years) yielded the greatest benefits. This configuration was used for all subsequent analyses. See the eMethods in Supplement 1 for further details.

Model Training

The performance of the joint model (ie, the model trained to predict the first occurring onset of either schizophrenia or bipolar disorder) approached the performance of the separate models when evaluated on each outcome separately on the training set. Specifically, for schizophrenia, the joint model achieved an AUROC of 0.75 (95% CI, 0.74-0.75) compared with 0.78 (95% CI, 0.77-0.78) for the schizophrenia-only model. For bipolar disorder, the joint model achieved an AUROC of 0.66 (95% CI, 0.65-0.66) compared with 0.67 (95% CI, 0.67-0.68) for the bipolar disorder–only model. The performance of the joint model when predicting the first diagnosis of either schizophrenia or bipolar disorder in the training phase is shown in eTable 5 in Supplement 1. Specifically, eTable 5 in Supplement 1 shows that the feature set including structured data, text, and synthetic data performed slightly better than the other feature sets on the training set, with an AUROC of 0.70 (95% CI, 0.70-0.70). The feature set including structured data and text and the feature set only using text features both achieved an AUROC 0.69 (95% CI, 0.69-0.69). Using only structured data yielded an AUROC of 0.66 (95% CI, 0.65-0.66). At a threshold of the 4% highest risk predictions marked as positive, the median lead time for the best model to flag patients who would develop schizophrenia or bipolar disorder was 0.7 years (on the training set). XGBoost was superior to logistic regression in all cases (eTable 6 in Supplement 1). For details on the hyperparameter search spaces, optimal values and imputation methods for the XGBoost and logistic regression models on the feature set, see eTable 7 in Supplement 1.

Joint Model (Predicting Schizophrenia or Bipolar Disorder) Testing

When applied to the test set (**Figure 2**), the XGBoost joint model using only text features that performed best achieved an AUROC of 0.64 (95% CI, 0.63-0.65) (Figure 2A). Figure 2B shows the confusion matrix using this model and a threshold based on a 4% predicted positive rate. The PPV was 13.0%, indicating that for every 7.7 positive predictions, 1 prediction time was followed by a diagnosis of schizophrenia or bipolar disorder within 5 years. The sensitivity at the level of prediction times was 9.3%, the specificity was 96.3%, and 13.5% of all patients who received a diagnosis of schizophrenia or bipolar disorder (47 of 347 patients) were predicted positive at least once (eTable 8 in Supplement 1). The median (Q1-Q3) time from the first positive prediction to the outcome was 1.1 (0.2-2.2) years (see Figure 2D). As shown in Figure 2C, for the joint model, sensitivity was generally higher for predicting schizophrenia compared with bipolar disorder. The joint model achieved an AUROC of 0.74 (95% CI, 0.73-0.75) for the schizophrenia-only outcome and 0.57 (95% CI, 0.56-0.58) for the bipolar disorder–only outcome on the test set (eTable 9 in Supplement 1).

A table of the 10 most important features for the joint model according to information gain is shown in **Table 2**. Notably, text embeddings of words, including *voices* and *admission*, were found to be highly influential for the model.

Models Trained to Predict Either Schizophrenia or Bipolar Disorder

The results for the models trained to predict either schizophrenia or bipolar disorder separately are shown in eFigure 3, eFigure 4, eTable 10, and eTable 11 in Supplement 1. The models predicting schizophrenia obtained the best performance, with an AUROC on the test set of 0.80 (95% CI, 0.79-0.81) for the best model. At a 4% predicted positive rate, sensitivity was 19.4%, specificity was 96.3%, and the PPV was 10.8% on the test set. Bipolar disorder proved more difficult, with the best model achieving an AUROC of 0.62 (95% CI, 0.61-0.63) on the test set. At a 4% predicted positive rate, sensitivity was 9.9%, specificity was 96.2%, and the PPV was 8.4% on the test set. As shown in eTable 12 in Supplement 1, for the separate outcomes, inclusion of text and synthetic features seemed to improve performance on both the training and test set.

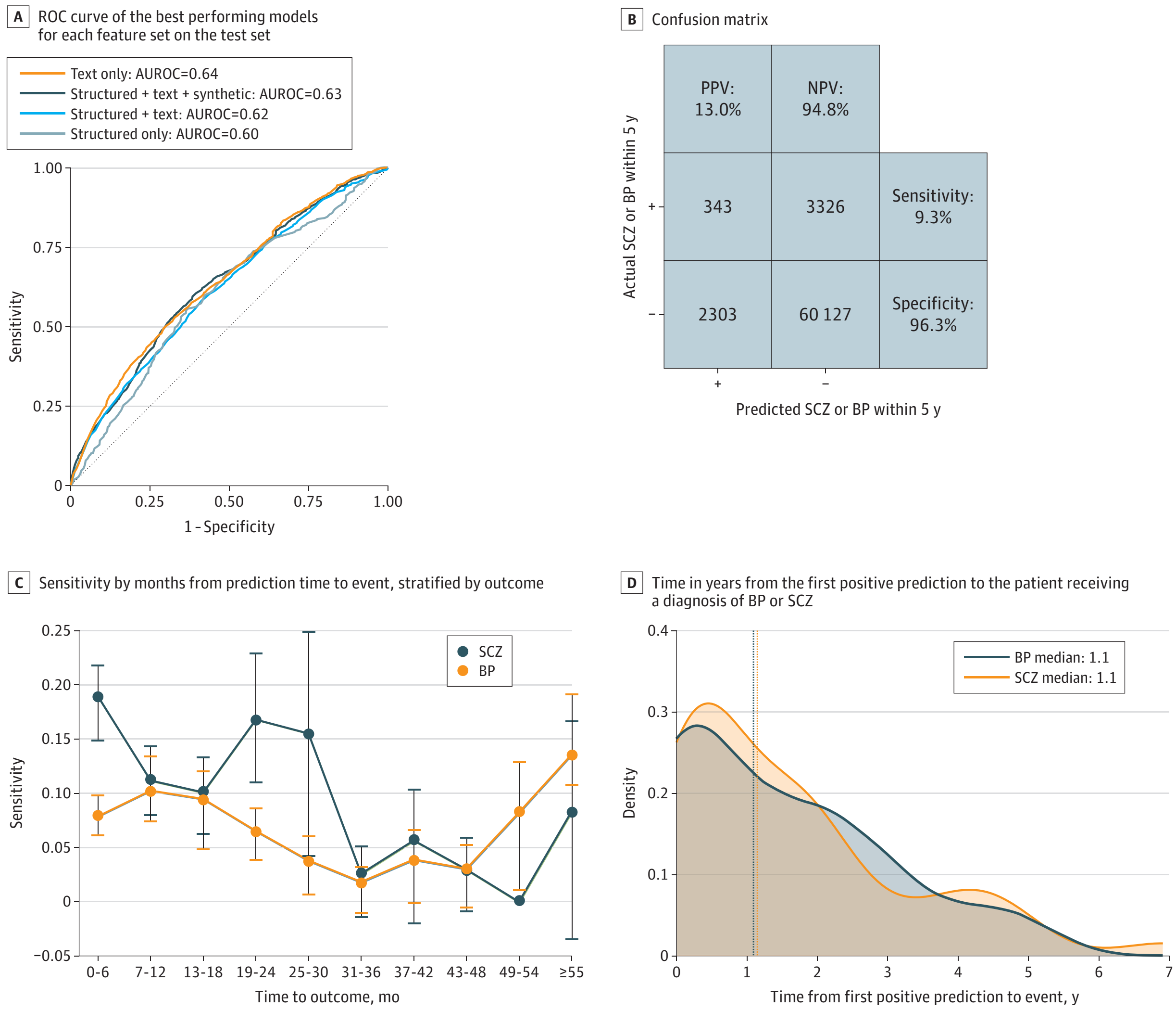
Robustness Analyses

Figure 3 shows that the performance of the joint model was stable across sex and age. The model performed slightly better on relatively young patients and older patients. Performance was quite stable across levels of time from the first contact, with some instability at the extremes, likely owing partially to lack of data. No noticeable trends were observed in the performance across calendar time. eFigure 5 in Supplement 1 shows that the model for schizophrenia was highly robust across stratifications, with slightly better performance for older patients. As shown in eFigure 6 in Supplement 1, the model for bipolar disorder was less robust, particularly across calendar time, with a noticeable dip in performance around the third quarter of 2015. The model (joint outcomes) trained on the 85% split, stratified only by patient, obtained an AUROC of 0.72 (95% CI, 0.71-0.73) on the 15% test set using the text-only feature set and the optimal hyperparameters identified previously.

Discussion

This cohort study investigated the feasibility of predicting diagnostic progression to schizophrenia or bipolar disorder within 5 years among patients with pre-existing mental illness. A model predicting the progression to either of the 2 disorders achieved an AUROC of 0.64 on the test set, with notable disparity in predictive performance between the 2 disorders. While the best model for predicting progression to schizophrenia achieved an AUROC of 0.80 on the test set, predicting bipolar disorder only achieved an AUROC of 0.62. This discrepancy may be attributed to the relatively larger heterogeneity within bipolar disorder compared with schizophre-

Figure 2. Performance of the Joint Model on the Test Set



A, Receiver operating characteristics (ROC) curve of the best performing models for each feature set on the test set. The model with the highest area under the ROC curve (AUROC) on the test (text only) was used in panels B to D, with a classification threshold corresponding to 4% positives. B, Confusion matrix showing positive predictive value (PPV) and negative predictive value (NPV). C, Sensitivity by months from prediction time to event, stratified by outcome (bipolar disorder [BD] or schizophrenia [SCZ]). D), Time (years) from the first positive prediction to the patient receiving a diagnosis of BP or SCZ. The dotted lines indicate the median time for each group.

nia and the distinctiveness of the psychotic symptoms of schizophrenia. Bipolar disorder covers a wide spectrum of illness manifestations, with some individuals initially presenting with mania and others with depression. In contrast, most individuals with schizophrenia have the paranoid schizophrenia subtype (*ICD-10* code F20), accounting for approximately 72% in the CDR,[29] which is relatively more homogenous in its presentation.

A substantial drop in performance was observed when moving from the training set to the test set, particularly for the joint model and the bipolar disorder model. The training and test sets contained data from different hospitals in the CDR; this indicates that substantial distribution shifts can occur even in a relatively homogenous population in close geographical proximity using the same health care system and clinical guidelines. These shifts might be caused by slightly different patient populations and/or variations in diagnostic practices. Indeed, a model trained and evaluated on an 85% to 15% patient-level dataset split achieved an AUROC of 0.72. This performance closely mirrors the out-of-fold performance observed in the training set of the models presented in the results section, which suggests that the performance decline across sites was not attributable to hyperparameter overfitting, but instead results from distribution shifts between sites. This change in performance across sites supports the argument that external validation should not be a strict requirement for scientific publication or model evaluation.[30] Rather, models should be tested in the specific context where they will be applied.

Model performance seems mainly to be driven by the inclusion of text-based predictors extracted from clinical notes.

Table 2. Top 10 Most Important Predictors by Information Gain in the Best-Performing (Text-Only) Joint Model

Danish token[a]	English translation	Information gain	Interpretation[b]
Udskrivelse	Discharge	0.003781	Discharge from an inpatient stay.
Spillet	The game	0.002872	Used when describing the activities of inpatients at wards (eg, playing board games with staff or other inpatients); likely a proxy for inpatient stays (of long duration)
Veninder	Female friends	0.002719	Appears in both positive and negative contexts (eg, has been visited by female friends; no contact with female friends)
Stemmerne	The voices	0.002644	Primarily used when referring to auditory hallucinations
Udskrives	Is being discharged	0.002577	Used when a patient has recovered and is about to be discharged from an inpatient stay
Gav	Gave, made, or resulted in	0.002393	Ambiguous, but often used to describe psychopathology (eg, gave him the urge to...; the pro re nata medication made her calm)
Indlæggelsen	The admission or inpatient stay	0.002263	Used to describe various aspects of inpatient stays or treatment
Spille	To play	0.002237	Similarly to spillet (the game), but as a verb
Morgenstunden	Early morning	0.002180	The first part of many notes describing patients during inpatient stays
Forklare	Explain	0.002103	Often used when patients are unable to describe why they acted in a certain way; possibly an indicator of delusions or derealization

[a] All predictors are term frequency-inverse document frequency with a 2-year look-behind and mean aggregation function.

[b] The interpretation column is based on reading of a subset of clinical notes containing the token in question. Descriptions in parentheses are hypothetical examples of the content that could be found in the electronic health record.

Indeed, models trained with both structured and text-based predictors performed practically equivalent to models trained with only text-based features; this underscores the importance of text in clinical prediction modeling within psychiatry.[12,31] Inspection of the most important words by feature importance and the context they appeared in within the EHRs revealed that many were related to hospital admission or psychiatric symptoms. Specifically, *admission* and *discharge* directly pertained to hospitalization. The terms *play* and *the game* often described patients' interactions with staff or other patients (eg, playing board games) during their inpatient stay. The phrase *early morning* frequently appeared at the beginning of notes documenting daily activities of inpatients. Symptom-related terms included *the voices*, typically referring to auditory hallucinations, while *female friends* was often used to describe social interactions or lack thereof (ie, social withdrawal). The term *explain* commonly appeared when patients struggled to articulate the reasons for their actions or experiences, potentially reflecting delusions or derealization.

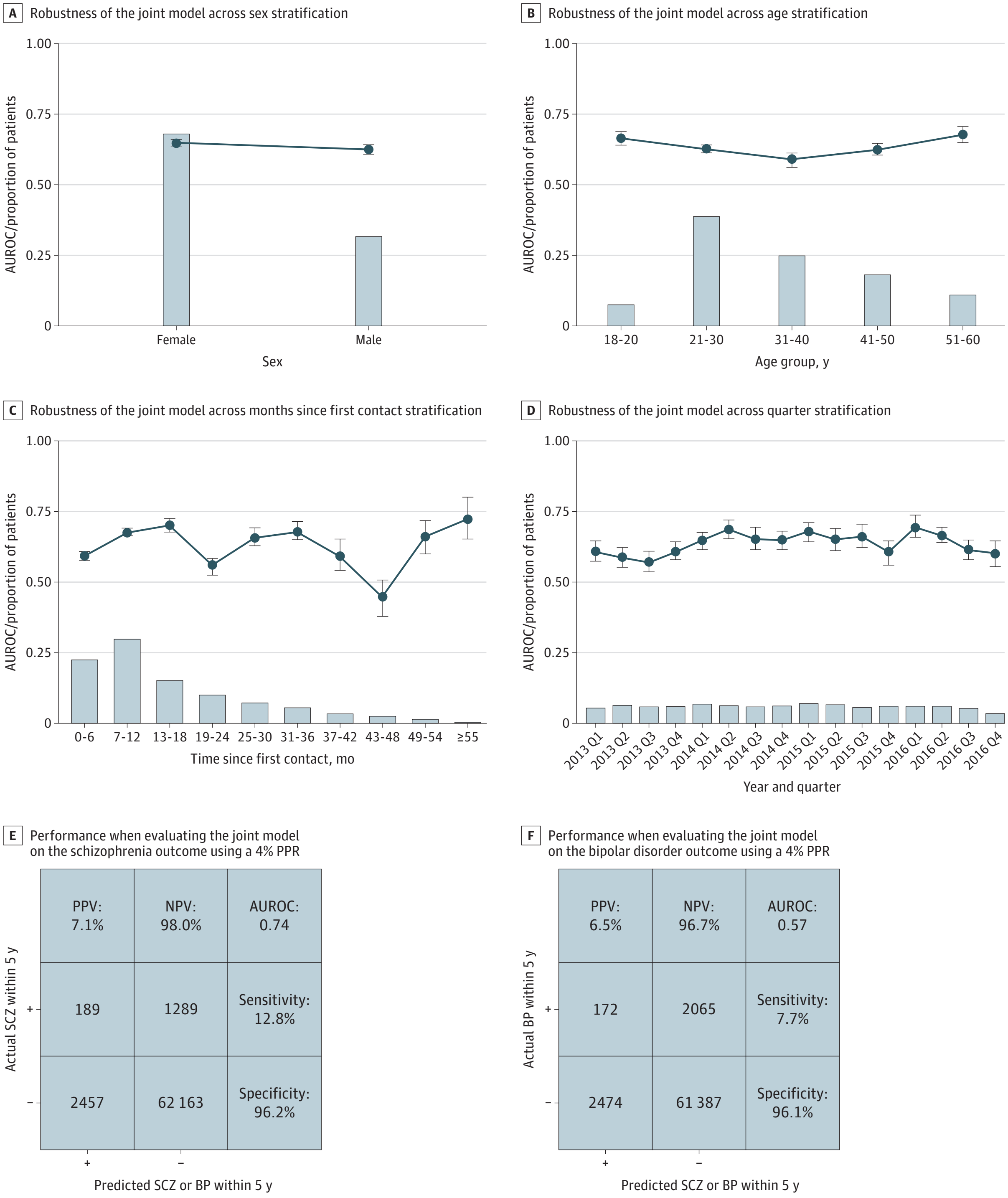
Performance from our models predicting schizophrenia is in line with the literature, for example with Irving et al,[12] which achieved a Harrell C of 0.79 to 0.86 for 10-year survival prediction of onset of psychosis from an index date. Harrell C and AUROC are equivalent in binary outcomes, but direct comparisons cannot be made with censored data such as those used by Irving and colleagues.[12] Irving et al[12] made a single prediction at an index date and only provide aggregate performance metrics such as Harrell C, Brier score, and the calibration slope. Wang et al[13] achieved an AUROC of 0.80 in predicting early onset bipolar disorder 3 years into the future from a single randomly sampled time between age 10 to 25 years per patient. The performance discrepancy in bipolar disorder prediction likely stems from the more age-restricted cohort in Wang et al,[13] and major differences in the definition of the outcome with Wang et al requiring (1) at least 2 *International Classification of Diseases, Ninth Revision (ICD-9)* or *ICD-10* codes for bipolar disorder, (2) predominance of bipolar disorder diagnoses, and (3) treatment with at least 2 medications commonly used for bipolar disorder. In summary, direct comparison is difficult because most studies only make a single prediction at an index date or only report aggregated measures such as AUROC or the C-index. In contrast, we issued predictions dynamically at clinically relevant times (before an outpatient contact) and reported performance at multiple decision thresholds to facilitate maximal clinical utility and critical scrutiny.

If applied within the Psychiatric Services of the CDR, the model's positive predictions should be automatically presented to the staff through the EHR system, enabling intervention at the level of the individual patient. Specifically, increased focus on symptoms compatible with schizophrenia or bipolar disorder (eg, via a focused diagnostic interview at the next outpatient consultation) would seem reasonable. Models predicting schizophrenia might be more suitable for implementation than those predicting bipolar disorder due to substantially better predictive performance.

Limitations

The study should be interpreted considering the following limitations. First, the data were restricted to patients receiving psychiatric treatment and did not contain information from primary care. Consequently, the prediction models are primarily useful for patients who are progressing from less severe mental disorder to schizophrenia or bipolar disorder. Patients whose initial contact to the psychiatric services is due to clinical suspicion of schizophrenia or bipolar disorder will not see additional benefits from the model. The study's focus on a high-risk population likely limits the generalizability of the findings to broader, low-risk populations such as those attending primary care. However, we believe this targeted approach enhances the practical utility of the developed models in real-world clinical settings, where interventions must be allocated carefully. Second, text models might be at risk of fitting to al-

Figure 3. Robustness of the Joint Model Across Stratifications on the Test Set



	PPV: 7.1%	NPV: 98.0%	AUROC: 0.74
+	189	1289	Sensitivity: 12.8%
–	2457	62 163	Specificity: 96.2%
	+	–	

	PPV: 6.5%	NPV: 96.7%	AUROC: 0.57
+	172	2065	Sensitivity: 7.7%
–	2474	61 387	Specificity: 96.1%
	+	–	

Panels A to D show robustness of the joint model across stratifications on the test set. The gray line is the area under the receiver operating characteristics curve (AUROC). Grey bars represent the proportion of prediction times in each bin. Error bars are 95% CIs from 100-fold bootstrap. Panels E and F show the performance when evaluating the joint model on the schizophrenia and bipolar disorder outcome, respectively, using a 4% predicted positive rate (PPR). NPV indicates negative predictive value; PPV indicates positive predictive value. BP indicates bipolar disorder; SCZ, schizophrenia.

ready present clinical suspicion and thereby provide less value. As shown in eTable 6 in Supplement 1, the median time from the first positive prediction to the outcome was highest for the model using only structured predictors, despite having the lowest overall performance in terms of AUROC. This finding indicates that while text models might lead to more cases being identified correctly, this might be at the cost of less lead time. Third, the test performance of our models was evaluated on data from 2 hospital sites. More robust estimates of external performance and performance heterogeneity could be obtained by performing internal-external cross-validation.[32] Computational limitations, however, precluded this option because we were restricted to run our analyses on in-house hardware (on the CDR server) due to the highly sensitive nature of the EHR data. Fourth, the generalizability of the prediction models outside the Psychiatric Services in the Central Denmark Region is unknown. Likely, retraining would be required for successful transfer, especially if testing the model where a different language is used. Fifth, our dataset only included data on medication administered during inpatient stays, not prescribed medication. Because most antipsychotic medication is prescribed (eg, to outpatients) rather than administered, we were unable to estimate the number of patients already receiving antipsychotic medication. Sixth, further research would benefit from comparing the predictive performance of prediction models such as the one developed in the present study with traditional instruments such as the Structured Interview for Psychosis-Risk Syndromes[33] or the Comprehensive Assessment of At Risk Mental States.[34] It bears noting that the prediction models developed in our study do not require interviewing or collection of other data, but rather operate on data stemming from documentation of routine clinical practice. Seventh, before implementation, investigations of the temporal stability of the model should be conducted.

Conclusions

This cohort study developed and validated models for predicting progression to schizophrenia or bipolar disorder using EHR data. The model predicting schizophrenia performed substantially better than the model predicting bipolar disorder, likely due to heterogenic clinical manifestations of the latter. These findings suggest that text-based features from clinical notes show great promise for improving prediction of psychiatric outcomes.

ARTICLE INFORMATION

Accepted for Publication: December 1, 2024.

Published Online: February 19, 2025. doi:10.1001/jamapsychiatry.2024.4702

Author Contributions: Dr Hansen had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: All authors.
Acquisition, analysis, or interpretation of data: Hansen, Bernstorff, Enevoldsen, Kolding, Damgaard, Perfalk, Danielsen, Østergaard.
Drafting of the manuscript: Hansen.
Critical review of the manuscript for important intellectual content: Bernstorff, Enevoldsen, Kolding, Damgaard, Perfalk, Nielbo, Danielsen, Østergaard.
Statistical analysis: Hansen, Bernstorff, Enevoldsen.
Obtained funding: Østergaard.
Supervision: Danielsen, Østergaard.

Conflict of Interest Disclosures: Dr Danielsen reported receiving personal fees from Otsuka Pharma Scandinavia AB outside the submitted work. Dr Østergaard reported receiving grants from The Novo Nordisk Foundation (grant No. NNF20SA0062874), The Lundbeck Foundation (grant No. R358-2020-2341), and Independent Research Fund Denmark (grant Nos. 7016-00048B and 2096-00055A); receiving the 2020 Lundbeck Foundation Young Investigator Prize; owning or having owned units of mutual funds with stock tickers DKIGI, IAIMWC, SPIC25KL and WEKAFKI; and owning or having owned units of exchange traded funds with stock tickers BATE, TRET, QDV5, QDVH, QDVE, SADM, IQQH, USPY, EXH2, 2B76, IS4S, OM3X and EUNL outside the submitted work. No other disclosures were reported.

Funding/Support: The study is supported by grants from the Lundbeck Foundation (grant No. R344-2020-1073 to Dr Østergaard), the Danish Cancer Society (grant No. R283-A16461 to Dr Østergaard), the Central Denmark Region Fund for Strengthening of Health Science (grant No. 1-36-72-4-20 to Dr Østergaard), the Danish Agency for Digitisation Investment Fund for New Technologies (grant No. 2020-6720 to Dr Østergaard) and by a private donation to the Department of Affective Disorders at Aarhus University Hospital—Psychiatry to support research into bipolar disorder.

Role of the Funder/Sponsor: The sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. Goldman ML, Pincus HA, Mangurian C. Schizophrenia. *N Engl J Med*. 2020;382(6):583-584. doi:10.1056/NEJMc1915943

2. Vieta E, Berk M, Schulze TG, et al. Bipolar disorders. *Nat Rev Dis Primers*. 2018;4:18008. doi:10.1038/nrdp.2018.8

3. Laursen TM, Wahlbeck K, Hällgren J, et al. Life expectancy and death by diseases of the circulatory system in patients with bipolar disorder or schizophrenia in the Nordic countries. *PLoS One*. 2013;8(6):e67133. doi:10.1371/journal.pone.0067133

4. Altamura AC, Buoli M, Caldiroli A, et al. Misdiagnosis, duration of untreated illness (DUI) and outcome in bipolar patients with psychotic symptoms: a naturalistic study. *J Affect Disord*. 2015;182:70-75. doi:10.1016/j.jad.2015.04.024

5. Penttilä M, Jääskeläinen E, Hirvonen N, Isohanni M, Miettunen J. Duration of untreated psychosis as predictor of long-term outcome in schizophrenia: systematic review and meta-analysis. *Br J Psychiatry*. 2014;205(2):88-94. doi:10.1192/bjp.bp.113.127753

6. Hafeman DM, Merranko J, Axelson D, et al. Toward the definition of a bipolar prodrome: dimensional predictors of bipolar spectrum disorders in at-risk youths. *Am J Psychiatry*. 2016; 173(7):695-704. doi:10.1176/appi.ajp.2015.15040414

7. Fusar-Poli P, Oliver D, Spada G, Estrade A, McGuire P. The case for improved transdiagnostic detection of first-episode psychosis: electronic health record cohort study. *Schizophr Res*. 2021; 228:547-554. doi:10.1016/j.schres.2020.11.031

8. Musliner KL, Østergaard SD. Patterns and predictors of conversion to bipolar disorder in 91 587 individuals diagnosed with unipolar depression. *Acta Psychiatr Scand*. 2018;137(5): 422-432. doi:10.1111/acps.12869

9. Musliner KL, Munk-Olsen T, Mors O, Østergaard SD. Progression from unipolar depression to schizophrenia. *Acta Psychiatr Scand*. 2017;135(1): 42-50. doi:10.1111/acps.12663

10. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2018;3(3):223-230. doi:10.1016/j.bpsc.2017.11.007

11. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. *NPJ Digit Med*. 2022;5 (1):46. doi:10.1038/s41746-022-00589-7

12. Irving J, Patel R, Oliver D, et al. Using natural language processing on electronic health records to enhance detection and prediction of psychosis risk. *Schizophr Bull*. 2021;47(2):405-414. doi:10.1093/schbul/sbaa126

13. Wang B, Sheu YH, Lee H, Mealer RG, Castro VM, Smoller JW. Machine Learning models for the prediction of early-onset bipolar using electronic health records. *MedRxiv*. Preprint posted online February 21, 2024. doi:10.1101/2024.02.19.24302919 doi:10.1101/2024.02.19.24302919

14. Raket LL, Jaskolowski J, Kinon BJ, et al. Dynamic electronic health record detection (DETECT) of individuals at risk of a first episode of psychosis: a case-control development and validation study. *Lancet Digit Health*. 2020;2(5): e229-e239. doi:10.1016/S2589-7500(20)30024-8

15. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378

16. Bernstorff M, Hansen L, Enevoldsen K, et al. Development and validation of a machine learning model for prediction of type 2 diabetes in patients with mental illness. *Acta Psychiatr Scand*. 2024. Published online April 4, 2024. doi:10.1111/acps.13687

17. Hansen L, Enevoldsen KC, Bernstorff M, Nielbo KL, Danielsen AA, Østergaard SD. The psychiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of electronic health records in the treatment of mental disorders. *Acta Neuropsychiatr*. 2021;33(6):323-330. doi:10.1017/neu.2021.22

18. Bernstorff M, Hansen L, Perfalk E, Danielsen AA, Østergaard SD. Stability of diagnostic coding of psychiatric outpatient visits across the transition from the second to the third version of the Danish National Patient Registry. *Acta Psychiatr Scand*. 2022;146(3):272-283. doi:10.1111/acps.13463

19. Hansen L, Enevoldsen K, Bernstorff M, et al. Lexical stability of psychiatric clinical notes from electronic health records over a decade. *Acta Neuropsychiatr*. 2023;1-11. Published online August 25, 2023 doi:10.1017/neu.2023.46. doi:10.1017/neu.2023.46

20. Howard R, Rabins PV, Seeman MV, Jeste DV; The International Late-Onset Schizophrenia Group. Late-onset schizophrenia and very-late-onset schizophrenia-like psychosis: an international consensus. *Am J Psychiatry*. 2000;157(2):172-178. doi:10.1176/appi.ajp.157.2.172

21. Schürhoff F, Bellivier F, Jouvent R, et al. Early and late onset bipolar disorders: two different forms of manic-depressive illness? *J Affect Disord*. 2000;58(3):215-221. doi:10.1016/S0165-0327(99)00111-1

22. Bernstorff M, Enevoldsen K, Damgaard J, Danielsen A, Hansen L. Timeseriesflattener: A Python package for summarizing features from (medical) time series. *J Open Source Softw*. 2023;8(83):5197. doi:10.21105/joss.05197

23. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-Networks. In: Inui K, Jiang J, Ng V, Wan X, eds. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics; 2019:3982-3992. doi:10.18653/v1/D19-1410

24. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. ACM Digital Library. Published August 13, 2016. Accessed January 15, 2025. https://dl.acm.org/doi/10.1145/2939672.2939785

25. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. Published July 25, 2019. ACM Digital Library. Accessed January 15, 2025. https://dl.acm.org/doi/10.1145/3292500.3330701

26. Le H, Eng-Jon O, Miroslaw B. SurvTimeSurvival: survival analysis on the patient with multiple visits/records. *ArXiv*. Published online November 16, 2023. doi:10.48550/arXiv.2311.09854

27. Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H. Synthetic data augmentation using GAN for improved liver lesion classification. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE; 2018:289-293. doi:10.1109/ISBI.2018.8363576

28. Perfalk E. Aarhus-Psychiatry-Research. Github. Updated January 2025. Accessed January 28, 2024. https://github.com/Aarhus-Psychiatry-Research/psycop-common

29. Köhler-Forsberg O, Antonsen S, Pedersen CB, Mortensen PB, McGrath JJ, Mors O. Schizophrenia spectrum disorders in Denmark between 2000 and 2018: Incidence and early diagnostic transition. *Acta Psychiatr Scand*. 2023;148(2):190-198. doi:10.1111/acps.13565

30. Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*. 2024;384:e074819. doi:10.1136/bmj-2023-074819

31. Rumshisky A, Ghassemi M, Naumann T, et al. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl Psychiatry*. 2016;6(10):e921-e921. doi:10.1038/tp.2015.182

32. Debray TPA, Collins GS, Riley RD, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster): explanation and elaboration. *BMJ*. 2023;380:e071058. doi:10.1136/bmj-2022-071058. doi:10.1136/bmj-2022-071058

33. McGlashan TH, Walsh BC, Woods SW, et al. Structured interview for psychosis-risk syndromes. Prime Research Clinic, Yale School of Medicine. Updated March 5, 2014. Accessed August 19, 2024. https://easacommunity.org/wp-content/uploads/2023/09/SIPS_5-5_0325141-correct.pdf

34. Yung A, Phillips L, McGorry P, Ward J, Donovan K, Thompson K. Comprehensive assessment of at-risk mental states (CAARMS). PACE Clinic, University of Melbourne, Department of Psychiatry. Published online 2002. Updated 2015. Accessed August 19, 2024. https://vizdom.cz/wp-content/uploads/2021/12/CAARMS_FINAL.pdf