JAMA Psychiatry | Original Investigation

Deep Learning for Cross-Diagnostic Prediction of Mental Disorder Diagnosis and Prognosis Using Danish Nationwide Register and Genetic Data

Rosa Lundbye Allesøe, MSc; Wesley K. Thompson, PhD; Jonas Bybjerg-Grauholm, PhD; David M. Hougaard, MD, PhD; Merete Nordentoft, MD, PhD; Thomas Werge, PhD; Simon Rasmussen, PhD; Michael Eriksen Benros, MD, PhD

Supplemental content

IMPORTANCE Diagnoses and treatment of mental disorders are hampered by the current lack of objective markers needed to provide a more precise diagnosis and treatment strategy.

OBJECTIVE To develop deep learning models to predict mental disorder diagnosis and severity spanning multiple diagnoses using nationwide register data, family and patient-specific diagnostic history, birth-related measurement, and genetics.

DESIGN, SETTING, AND PARTICIPANTS This study was conducted from May 1, 1981, to December 31, 2016. For the analysis, which used a Danish population-based case-cohort sample of individuals born between 1981 and 2005, genotype data and matched longitudinal health register data were taken from the longitudinal Danish population-based Integrative Psychiatric Research Consortium 2012 case-cohort study. Included were individuals with mental disorders (attention-deficit/hyperactivity disorder [ADHD]), autism spectrum disorder (ASD), major depressive disorder (MDD), bipolar disorder (BD), schizophrenia spectrum disorders (SCZ), and population controls. Data were analyzed from February 1, 2021, to January 24, 2022.

EXPOSURE At least 1 hospital contact with diagnosis of ADHD, ASD, MDD, BD, or SCZ.

MAIN OUTCOMES AND MEASURES The predictability of (1) mental disorder diagnosis and (2) severity trajectories (measured by future outpatient hospital contacts, admissions, and suicide attempts) were investigated using both a cross-diagnostic and single-disorder setup. Predictive power was measured by AUC, accuracy, and Matthews correlation coefficient (MCC), including an estimate of feature importance.

RESULTS A total of 63 535 individuals (mean [SD] age, 23 [7] years; 34 944 male [55%]; 28 591 female [45%]) were included in the model. Based on data prior to diagnosis, the specific diagnosis was predicted in a multidiagnostic prediction model including the background population with an overall area under the curve (AUC) of 0.81 and MCC of 0.28, whereas the single-disorder models gave AUCs/MCCs of 0.84/0.54 for SCZ, 0.79/0.41 for BD, 0.77/0.39 for ASD, 0.74/0.38, for ADHD, and 0.74/0.38 for MDD. The most important data sets for multidiagnostic prediction were previous mental disorders and age (11%-23% reduction in prediction accuracy when removed) followed by family diagnoses, birth-related measurements, and genetic data (3%-5% reduction in prediction accuracy when removed). Furthermore, when predicting subsequent disease trajectories of the disorder, the most severe cases were the most easily predictable, with an AUC of 0.72.

CONCLUSIONS AND RELEVANCE Results of this diagnostic study suggest the possibility of combining genetics and registry data to predict both mental disorder diagnosis and disorder progression in a clinically relevant, cross-diagnostic setting prior to clinical assessment.

JAMA Psychiatry. doi:10.1001/jamapsychiatry.2022.4076
Published online December 7, 2022.

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Authors: Michael Eriksen Benros, MD, PhD, Mental Health Centre Copenhagen, Copenhagen University Hospital, Gentofte Hospitalsvej 15, 4th Floor, 2900 Hellerup, Denmark (michael.eriksen.benros@regionh.dk); Simon Rasmussen, PhD, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, Blegdamsvej 3B, 2200 Copenhagen, Denmark (simon.rasmussen@cpr.ku.dk).

© 2022 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Sun Yat-Sen University User on 12/11/2022

The current diagnostic practice in psychiatry lacks objective markers for predicting psychiatric diagnosis and illness course to improve the consistency in the diagnostic process and facilitate more precise prevention and treatment of mental disorders.[1] In addition, studies have shown that a significant association between a disease and a feature does not necessarily mean it can be used to predict the disease clinically.[2,3] Therefore, with the increasing amounts of data available, new opportunities have emerged to apply cutting-edge methods such as deep learning (DL) to build prediction models using high-dimensional data to capture nonlinear relationships spanning multiple data sources (multimodal).[1,4,5] Currently, machine learning (ML), and specifically the DL subspeciality, have been used to establish clinically applicable prediction tools within other medical fields, such as breast cancer diagnostics, decision-making during the COVID-19 pandemic, and prediction of cardiac arrest from emergency calls.[6-13]

For mental disorders, ML models have shown great promise in predicting both diagnosis and prognosis within mental disorders using data sources such as genetics, magnetic resonance imaging (MRI), electroencephalography (EEG), and clinical and demographic data.[1,4,5,14,15] However, performance across studies and between disorders is inconsistent with relatively small sample sizes (50-3000 individuals), insufficient reporting of the model used, restriction to 1 type of data (unimodal), no validation, and has mainly been restricted to 1 diagnostic group or prognostic outcome (eg, treatment response, readmissions, or suicide). A common issue is that these do not resemble an actual clinical setting.[14-17] Therefore, prior results from both the ML field and traditional statistics have resulted in models that cannot be easily generalized or used clinically.[18]

In this study, we established cross-diagnostic prediction models, spanning multiple mental disorders, to better mimic a clinical setting. We achieved this using a large Danish population cohort from the Integrative Psychiatric Research Consortium (iPSYCH). We investigated how well we could predict both the diagnosis and the postdiagnosis severity trajectories across the mental disorders using genetic data, nationwide register data of family and patient-specific diagnostic history of mental disorders, as well as other etiological factors, such as birth-related measurements, infections, and autoimmune diseases.[19-21] The predictions were done by training DL prediction models for addressing both cross-diagnostic predictability in 1 multiclass model and the predictability of each disorder separately using binary prediction models.

Methods

Data

For this study a population-based case-cohort study, using the Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) 2012 database, data from individuals born between May 1, 1981, and December 31, 2005, and followed up until December 31, 2016, were analyzed.[22] All personal information from the registers is anonymized when used for research purposes, and the project was approved by the Danish Data Protection Agency; hence, according to Danish legislation, informed consent from participants was not required. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines (eTable 1 in Supplement 1).

The genotype data from dry blood spots taken at birth were processed as described in detail by Schork et al,[23] and quality control included the removal of individuals of non-European descent and related individuals with the remaining individuals serving as our full study cohort. From this cohort were the individuals in our case group, consisting of all individuals in Denmark during the study period who had at least 1 hospital contact with diagnosis of the following mental disorders: autism spectrum disorder (ASD) (*International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* [*ICD-10*] codes F84.0-1, F84.5, F84.8-9), attention-deficit/hyperactivity disorder (ADHD; code F90.0), schizophrenia spectrum disorders (SCZ; codes F20/F20-F29), bipolar disorder (BD; code F31), and major depressive disorder (MDD; codes F32-F33). The remaining individuals were a population control group, ie, randomly sampled from the population and with no hospital contacts for any of the previously mentioned psychiatric diagnoses. All data from the registers were included up until 2016 for mental disorders and 2012 for infections and autoimmune diseases taking any revised diagnostic changes into account.[24] We included all psychiatric diagnoses from the Danish registers based on *ICD-10* or *ICD Eighth Revision* codes, including information on the age of onset, all psychiatric-related hospital contacts for the patients, and all diagnoses of their parents (family history data set).[25] Additionally, we included a count of how many siblings were diagnosed with MDD, SCZ, or any mental disorder diagnosis. For both patients and parents, registered infections, autoimmune diseases, and other medical conditions (diabetes, migraine, and epilepsy) were included as well as information from the medical birth registry (MBR). Age was recorded at the end of 2016 or at the time of the first diagnosis with any of the 5 main mental disorders in the study (ADHD, ASD, MDD, BD, or

Key Points

Question Prior to the diagnosis, can it be predicted if an individual will be diagnosed with a severe mental disorder and the subsequent severity trajectories using the national Danish health registry and genetic data?

Findings In this diagnostic study including 63 535 individuals, the specific diagnostic category within the mental disorder group could be predicted in a multidiagnostic model including a randomly sampled population control group. The most predictable group was the most severe group.

Meaning Results suggest that the multidiagnostic model resembling a clinical setting prior to the examination can predict the mental disorder diagnosis with high accuracy based only on registry data and genetic information; prediction of the subsequent severity trajectory progression of the disorder based on information up to the diagnosis only performed with lower accuracy.

© 2022 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Sun Yat-Sen University User on 12/11/2022

SCZ). A total of 1181 variables spanning both register and genetic data were included (eTable 6 in Supplement 2).

Preprocessing of Genetic Data

Genotype data were included as 516 genome-wide significant single-nucleotide variants of risk alleles from the genome-wide association study catalog for mental disorders and for general medical conditions that have previously been associated with mental disorders, such as autoimmune disease and infections (as described in Allesøe et al[26]). Briefly, from the search of genome-wide association studies of the diseases, we only excluded studies based on missing information on the risk allele or on the study cohort, such as ethnicity. Additionally, we included imputed human leukocyte antigen (HLA)[19] alleles. Both were included as being either homozygotic for the allele, heterozygotic, or not having the allele. In addition, we included polygenic risk scores (PRS) for 142 traits related to mental disorders and state (loneliness, mood swings, etc), autoimmune disease, infections, and educational level/intelligence (eTable 6 in Supplement 2). The PRSs were calculated from polygenic scores using PRSice2, version 2.3.3.[27] We used a P value threshold of 0.10 and a clumping r^2 threshold of 0.1 in a window of 250 kilobase pairs and the sum method to calculate the scores (eMethods in the Supplement 1).

Data Encoding

The continuous data variables, eg, age of comorbid diagnoses, the number of hospitalizations, hospital contacts, age, and birth measures, were all normalized by scaling them between 0 and 1 per feature for optimal training of the models. Missing continuous data were encoded as the mean value of the feature. Categorical data with more than 2 categories, MBR data, genotype, and HLA data were encoded through a feature embedding layer as implemented in Pytorch,[28] version 1.4.0. Family history data were encoded as binary for the diagnoses.

Data Processing for Diagnostic Prediction Models

To resemble the data available in a clinical setting at the time of diagnosis, we removed all data occurring after the main diagnosis of either ADHD, ASD, MDD, BD, or SCZ using a ranking system in cases of multiple diagnoses (rank order: ADHD < ASD < MDD < BD < SCZ). All prior diagnostic data, such as the age at diagnosis, family history, medical birth data, and genetics, were included in the model. In the model of the combined case group, individuals were included with data up until the first diagnosis with any of the 5 disorders and therefore have less information available compared with the multidiagnostic models. In single-disorder models (models 4-8), all individuals with that given diagnosis were included with data up until the diagnosis of the given disorder (**Figure 1**A). Similar to the masking of data in the mental disorder group, we masked data in the control group to get the same age distributions (eMethods in Supplement 1).

Data Processing for Prognostic Prediction Models

The severity features were collected and summed for all 5 included mental disorders and included the following: (1) the number of outpatient contacts, (2) hospital admission, (3) length of admission, (4) suicide attempt, (5) housing facility days, and (6) suicide. Each severity metric was divided into 3 to 4 categories of increasing severity, whereas suicide and housing facility days were included as binary (eMethods and eTable 3 in Supplement 1) (Figure 1C). The combined severity score was calculated as a weighted sum of the groups with the highest number being more severe as defined in eTable 3 in Supplement 1. All counts were normalized for years from the time of diagnosis until the end of 2016 before creating the severity groups. In these models, in addition to the data in the diagnostic model, we also included the age of their main mental disorder diagnosis of ADHD, ASD, MDD, BD, or SCZ and all new diagnoses of both mental disorders and infections given within the first half-year after the diagnosis.

Statistical Analysis

All prediction models used a feed-forward neural network architecture and were implemented using Python PyTorch,[28] version 1.4.0 (Figure 1D) (eMethods in Supplement 1). The output of the model was probability scores for each of the N classes, and the final prediction was the class with the highest probability. A weighted sampler based on the observed class prevalence was used as implemented in PyTorch to ensure that each batch contained a proportional number of all classes to account for class imbalance.

Hyperparameter Selection and Model Validation

We used a stratified 3-fold cross-validation evaluation scheme using Python scikit-learn, version 0.21.3[29] stratified by the label to ensure equal splits across the mental disorders categories and the background population group. Within each training data set, 10% of the data was used as a validation set for the hyperparameter optimization and the rest was used to train the model (eFigure 1 in Supplement 1). The best-performing model (eMethods in Supplement 1 contains details on hyperparameter optimization setup) was evaluated on the test set by calculating the receiver operating characteristic area under the curve (AUC) of the overall multiclass model calculated as the microaverage and for each class individually by binarizing the output.[30] Furthermore, we calculated the overall accuracy and the Matthews correlation coefficient (MCC) combined across all models, which have shown to be a superior performance measurement in both multiclass and binary predictions.[31,32] The MCC ranges from −1 to 1 with 1 being a perfect prediction and 0 being a random prediction. For accuracy, the lift from the by-chance accuracy was calculated due to the class imbalance. All performance calculations were done using the scikit-learn implementations and compared with multiple linear regression and random forest.

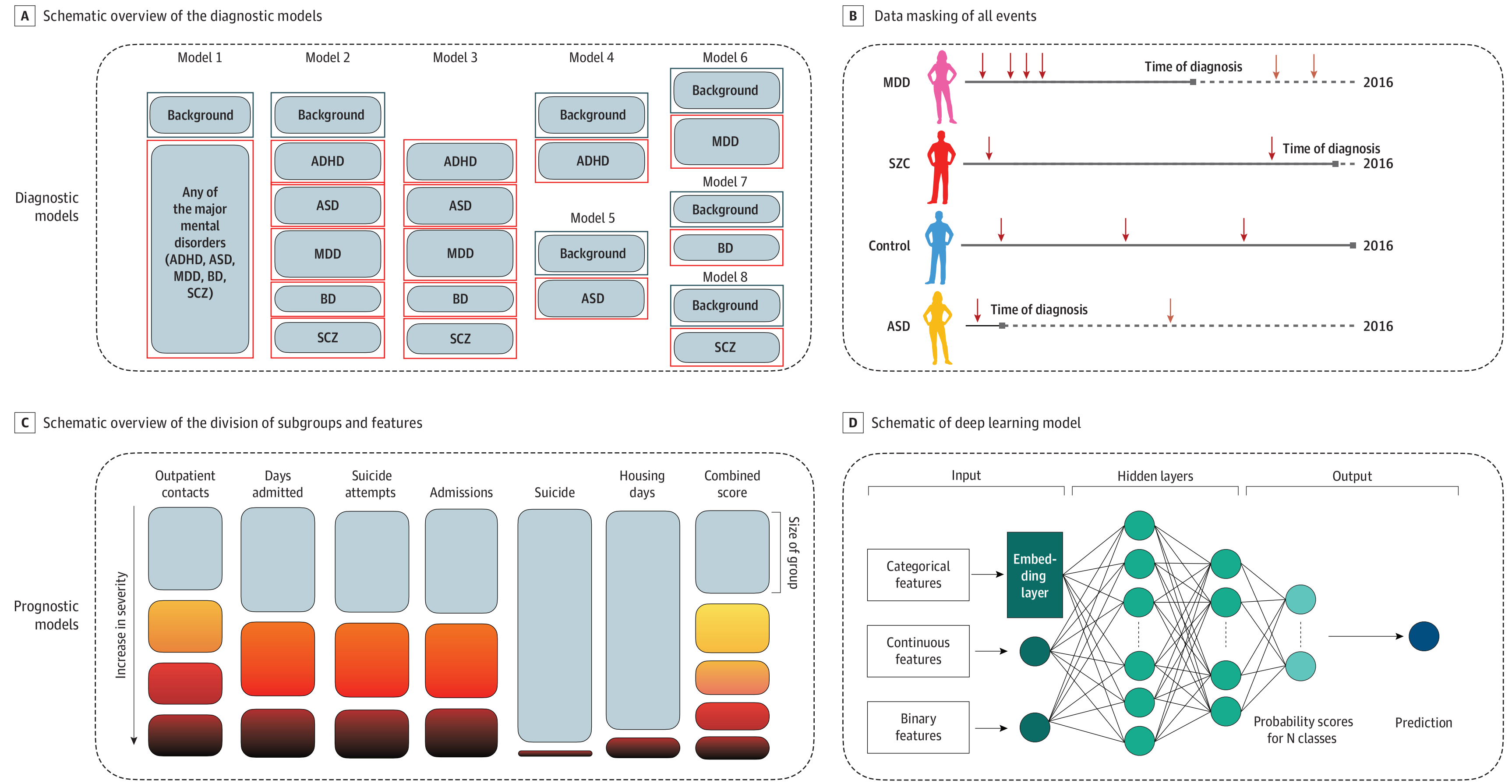
Feature Importance Calculation

Feature importance was calculated separately for each data type: family diagnosis history, age, individual genotype data, PRS, mental disorder diagnosis, and MBR. We assessed the importance in each of the 3 models in the 3-fold cross-validation by removing all variables in each data set one data set at a time and retrain the model. The final feature impor-

© 2022 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Sun Yat-Sen University User on 12/11/2022

Figure 1. Cohort and Model Design



A, Schematic overview of the diagnostic models with models 1 to 3 being cross-diagnostic models including all individuals with attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), major depressive disorder (MDD), bipolar disorder (BD), and schizophrenia or related disorders (SCZ). Models 1 to 2 also include the background population. Models 4 to 8 are single-disorder vs background population models. B, Data masking of all events from the time of diagnosis to resemble the clinical assessment. Each arrow represents an event such as a mental disorder diagnosis or infection. From the time of diagnosis, all future events are removed from the data and not presented to the model for prediction. C, Schematic overview of the division of the subgroups and features predicted in the prognostic prediction models. The groups are divided by increased severity for the specific feature such as more or longer admissions to the hospital. The colors from blue to red represent the severity of the group (low to high severity). Note the size of the boxes are approximations and not exact representations of the size of the groups. D, Schematic of the deep learning model used in the prediction.

© 2022 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Sun Yat-Sen University User on 12/11/2022

tance measure was the average relative drop in model accuracy across the 3 models. Data were analyzed from February 1, 2021, to January 24, 2022.

Results

Prediction of Severe Mental Disorders From a Background Population

A total of 63 535 individuals (mean [SD] age, 23 [7] years; 34 944 male [55%]; 28 591 female [45%]) were included in the model (eTable 2 in Supplement 1). These included individuals diagnosed with ASD (n = 12 878), ADHD (n = 15 969), SCZ spectrum disorders (n = 5120), BD (n = 1719), and MDD (n = 19 159) as well as 20 681 controls. When predicting the combined case group vs the population control group in a binary prediction model (model 1) (Figure 1A and eFigure 4 in Supplement 1), we achieved an AUC of 0.72, an accuracy of 66% (1.2-fold lift from by-chance accuracy of 56%), and an MCC of 0.27 (eFigure 2A and 2B in Supplement 1). Our multidiagnostic model could predict the specific mental disorder diagnoses from the background population (Figure 1A model 2) with an AUC of 0.81, with an accuracy (equal to the positive predictive value for multiclass models) of 44% (1.9-fold lift from by-chance accuracy of 23%), and an MCC of 0.28 (**Figure 2**A). This was similar to multiple linear regression and random forest performance (eResults in Supplement 1).

Misclassifications in the population control group were lowest for BD and SCZ (11.3% and 14.3% of the predictions) with the majority of misclassifications being for one of the other mental disorders with 67.8% and 44.3% of the predictions, respectively. ADHD, ASD, and MDD all had misclassifications as the population control group on 25% to 30% of the predictions (Figure 2B and eFigure 3 in Supplement 1). When training separate binary prediction models for each of the 5 disorders against the background population (models 4-8) (Figure 1A), the best performances were observed for SCZ and BD with AUCs of 0.84 and 0.79 (MCC, 0.54 and 0.41), respectively (Figure 1A and eTable 4 in Supplement 1).

Prediction of Mental Disorders in a Multidiagnostic Model Among Cases Only

The prediction of the diagnosis in a multidiagnostic model only considering cases (model 3) yielded an overall AUC of 0.82, MCC of 0.36, and accuracy of 53% (2-fold lift from by-chance accuracy of 27%) (Figure 1A and Figure 2C). The positive predictive value (precision) was highest for MDD and ASD at values of 0.66 and 0.56, respectively, whereas the lowest was for BD at 0.18 (model 3) with most misclassifications as SCZ or MDD (each representing 35% of the classifications) (**Table** and Figure 2D). We also observed that both ADHD and SCZ were often predicted as MDD (21% and 33%) as well as difficulties separating ADHD from ASD (34% of ADHD predicted as ASD).

Feature Importance for the Diagnostic Prediction Models

Our feature importance analysis showed that in general, no single data type had an extreme influence on the performance across the models (**Figure 3**A and eTable 5 in Supplement 1). Specifically, for model 1, combined case group vs background, the highest observed influence was a 5% relative reduction in prediction accuracy. For all the multidiagnostic models (models 1-3), the highest impact was from previous mental disorder diagnoses (5%-23%). In all 3 models, the total genetic impact (PRS, individual genotypes, and HLA) ranged from 2% to 5%, with model 3 having the highest genetic impact (5%).

Similarly, for the single-disorder models vs background population (models 4-8), the highest observed impact was for previous mental disorder diagnoses and ranged from 7% to 8% for ADHD, ASD, and MDD to 20% to 21% for SCZ and BD. The single-disorder models gave AUCs/MCCs of 0.84/0.54 for SCZ, 0.79/0.41 for BD, 0.77/0.39 for ASD, 0.74/0.38, for ADHD, and 0.74/0.38 for MDD. In the single-disorder models, the individual genotypes had a higher impact (2%-5%) compared with the PRS score (0.5%-1%). The highest total genomic impact was when predicting ASD and BD (6% and 7%). Interestingly, the categorical MBR data set, consisting of the level of malformations at birth and smoking during pregnancy, was only of major importance for the ASD-specific prediction model (11%) compared with less than 2% for the remaining single-disorder prediction models.

Prediction of Post-Diagnosis Severity Levels

When predicting cross-disorder level of severity by dividing the cases into 5 groups of increased severity across multiple measurements (Methods in Supplement 1 and Figure 1C), we achieved an AUC of 0.65, MCC of 0.13, and accuracy of 32% (lift of 1.3 from by-chance accuracy 25%). Of all the groups, the most severe group was the easiest to predict (AUC of 0.72) (Figure 3B and C). When predicting the combined severity score for each disorder separately, we found that the overall performance dropped with AUCs between 0.58 to 0.62 (eTable 4 in Supplement 1). The feature importance for the combined severity score on all patients showed that comorbid psychiatric diagnoses were the most important with 23% followed by total birth-related measurements, family diagnoses, and total genetic data (all approximately 4%) (Figure 3D and eTable 5 in Supplement 1). For the within-disorder severity, mental disorder diagnoses had a high influence for all models (16%-28%). However, prediction of the individual severity measurement (eg, suicide attempts and admissions) had an equally high influence, between 9% and 11%, for psychiatric diagnoses and individual genetics across all models (eTable 5 in Supplement 1).

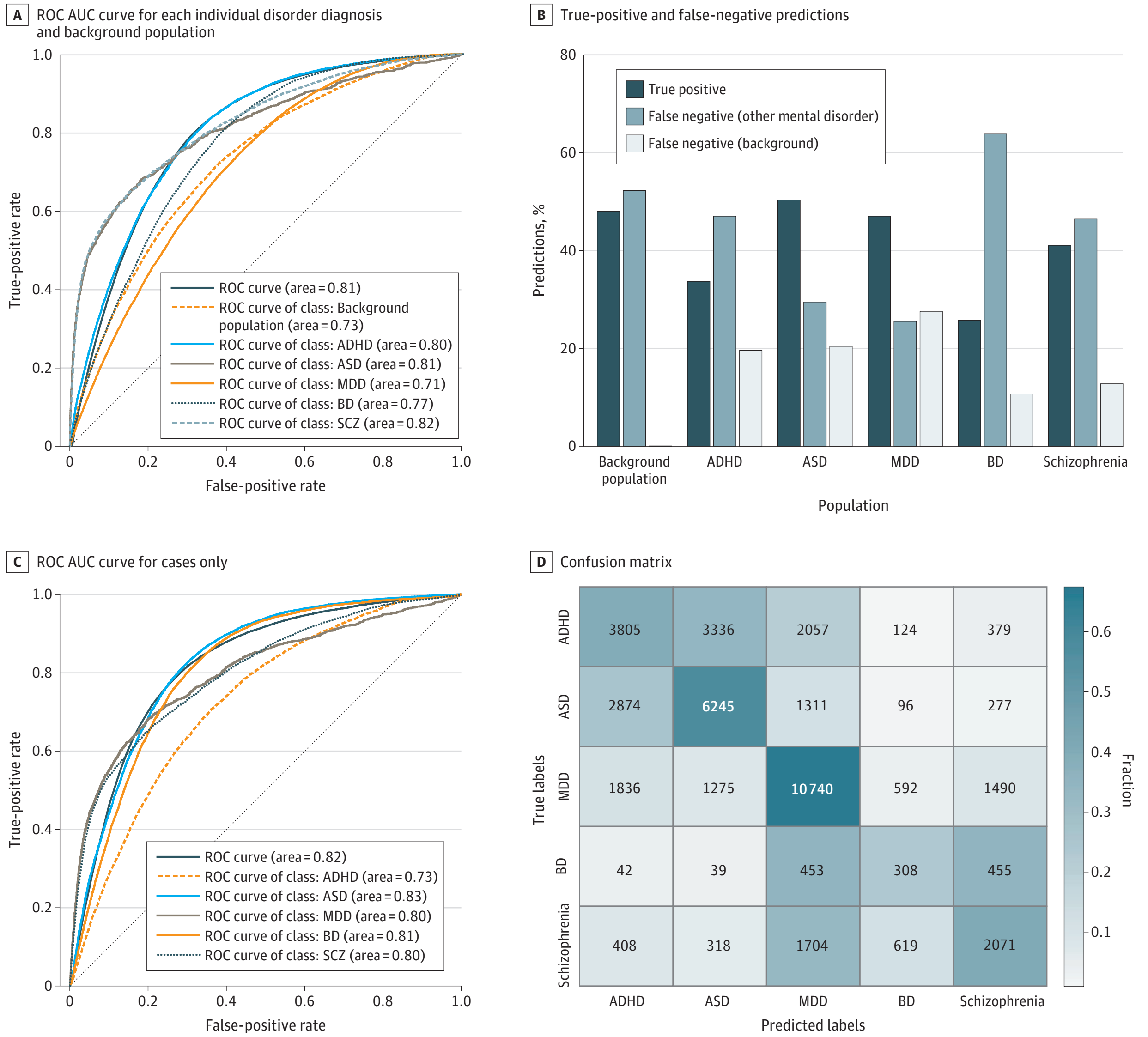
Discussion

In this diagnostic study, results suggest that a cross-diagnostic model using features collected prior to clinical assessment can have high predictive power on diagnostics (AUC between 0.71 and 0.82). We found that combining clinical register data with genetics had a positive association with prediction accuracy. This was indicated by a relative decrease of up to 7% in the accuracy for predicting diagnosis and 11% for

© 2022 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Sun Yat-Sen University User on 12/11/2022

Figure 2. Diagnostic Model Performance



A. Receiver operating characteristic (ROC) area under the curve (AUC) for the multidiagnostic prediction model (model 2) of each individual disorder diagnosis and background population. Included are the overall multilabel performance and performance on each of the individual diagnostic labels in the model. The dotted line shows the curve for a random performance. B, The percentage of true-positive and false-negative predictions are divided into the background population or a combined percentage of any other mental disorders in the multidiagnostic prediction model (model 2). C, ROC-AUC curve for the multidiagnostic prediction model (model 3) for cases only. Included are the overall multilabel performance and performance on each of the individual diagnostic labels in the model. The dotted line shows the curve for a random performance. D, Confusion matrix showing the distribution of predicted classes in the multidiagnostic prediction model on cases only (model 3) with the color intensity showing the fraction across the rows (true label). The numbers in the table are the number of predictions as a given class for each of the true classes. Therefore, each row sums up the total number of individuals in the class, and the sum of each column is the total number of predictions for the class. ADHD indicates attention-deficit/hyperactivity disorder; ASD, autism spectrum disorder; BD, bipolar disorder; MDD, major depressive disorder; SCZ, schizophrenia or related disorders.

prognostic scores when not included in the model. Furthermore, we could predict the overall cross-diagnostic severity group with an AUC of 0.65, with the most predictable group being the most severe group with an AUC of 0.72.

Contrary to most previous prediction studies in psychiatry,[14] this study combines data for multiclass predictions of several mental disorders and provides insights into how the model uses the data sets across the different prediction tasks. Even though previous work has shown promising performance (ranging between 0.48-0.95 AUC or 45%-100% accuracy),[14-16] we found it to be particularly important to use cross-diagnostic models when evaluating model performances for a clinical setting due to the high misclassification rates between the disorders.[33] Specifically, we observed overlaps in predictions of ADHD and ASD as well as BD, MDD, and SCZ, thereby highlighting the known overlaps in preassessment predictors such as earlier diagnoses, genetics, and family history between the disorders.[34,35]

© 2022 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Sun Yat-Sen University User on 12/11/2022

Table. Performance Measurements in the Multiclass Prediction Model[a]

Class	Accuracy	Precision	Sensitivity	Specificity	AUC
Model 2 (cross-diagnostic prediction including population control group)					
Back_pop	0.68	0.52	0.48	0.78	0.73
ADHD	0.80	0.35	0.34	0.89	0.80
ASD	0.80	0.42	0.50	0.86	0.81
MDD	0.75	0.50	0.47	0.84	0.71
BD	0.96	0.18	0.26	0.98	0.77
SCZ	0.90	0.40	0.41	0.95	0.82
Model 3 (cross-diagnostic prediction of cases only)					
ADHD	0.74	0.42	0.39	0.84	0.73
ASD	0.78	0.56	0.58	0.84	0.83
MDD	0.75	0.66	0.67	0.79	0.80
BD	0.94	0.18	0.24	0.97	0.81
SCZ	0.87	0.44	0.40	0.93	0.80

Abbreviations: ADHD, attention-deficit/hyperactivity disorder; ASD, autism spectrum disorder; AUC, area under the curve; Back_pop, background population control; BD, bipolar disorder; FN, false negative; FP, false positive; MDD, major depressive disorder; SCZ, schizophrenia spectrum disorders; TN, true negative; TP, true positive.

[a] Performance as accuracy, precision (positive predictive value), sensitivity (recall), specificity, and AUC for each diagnostic category in the multiclass models. The evaluations are done by considering each class/diagnosis separately in the model and collapsing all other classes into one. In the table, accuracy is calculated as TP + TN / TP + TN + FP + FN; precision as TP / TP + FP; sensitivity as TP / TP + FN; and specificity as TN / TN + FP.

Combining high-dimensional genetic data with clinical records increased prediction accuracy by 2% to 11% in our study. However, the overall low decrease in accuracy for each data set shows that the different data modalities have a high overlap in predictive information. In the prediction of the specific diagnosis (models 2 and 3), the overall importance of each data set was higher compared with predicting any case vs background (model 1). This clearly shows that more predictive information is needed to separate between the diagnoses than only separating any case from a background population. Additionally, the individual genotype data contributed to the predictive power, indicating that the exact genetic composition holds information that is predictive for both diagnosis and prognosis in mental disorders. Therefore, multimodal models can both increase prediction power and add to the understanding of the disorder complexity with the potential to support clinical decision-making by including objective markers. Nevertheless, our findings still support a low added predictive value of PRS scores as observed previously for SCZ.[36] Therefore, further analysis, including results from other cohorts, is needed to determine if including genetic data would be beneficial for clinical applications in psychiatry in settings where detailed family diagnostics are available.

Strengths and Limitations

The main strength of this study is the large-scale population-based data with high diagnostic validity[37-40] that span multiple mental disorders and therefore provide a more accurate representation of a real clinical setting.[14,41,42]Additionally, access to large-scale genetic data provides a unique possibility to investigate the added value of genetics in diagnostic practice. However, the study and the possible future implementations from these results are limited to Danish individuals and a setting with access to nationwide register data[43]; thus, implementation in other populations might require adjustment of variables and further testing. Furthermore, in future models with full access to richer data from the Danish registers, more advanced models such as the LSTM, or long-short-term memory, deep learning architecture could be beneficial.[43]

A limitation of all ML-based prediction models is that they are restricted to learning from historical data. Therefore, possible previous misdiagnoses of the patients or changes in diagnostic practices would lead to a wrongful profile of the diagnoses, and as a result, the prediction model will conserve these biases in future predictions. Consequently, these models should be continuously assessed and evaluated by trained clinicians and never stand alone in the decision-making. Finally, the oversampling of mental disorders compared with the population controls resulted in a skewed population distribution where approximately 100% of all cases (n = 42 854), and only approximately 2% of the background population (n = 20 681) were sampled. However, the models were designed to be applied in a clinical setting at the time of assessment, in which an overrepresentation of cases would be expected. Therefore, these models are expected to perform as intended given the true distribution within the disorders.

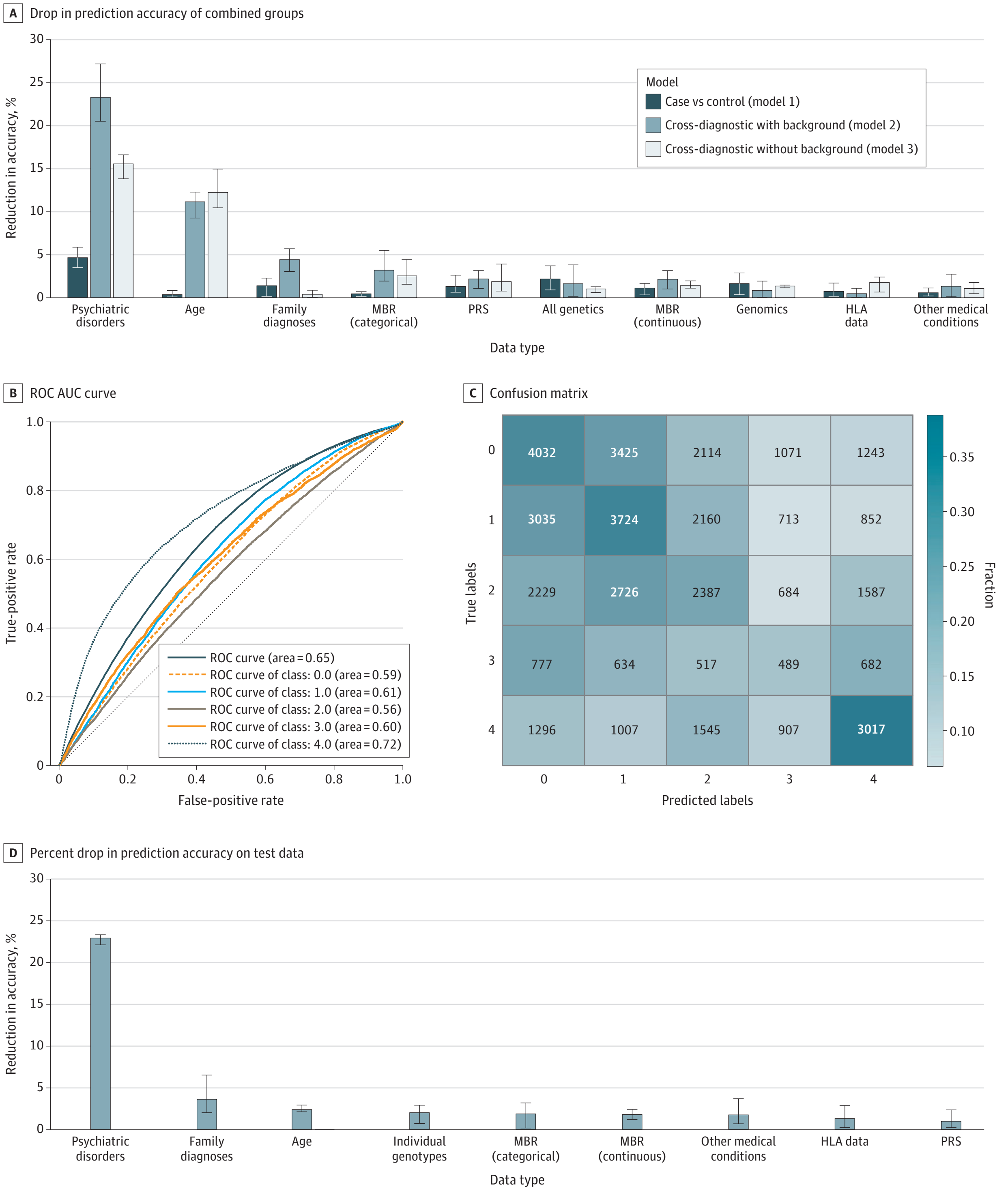
Conclusions

Results of this diagnostic study suggest the potential and challenges in multiclass prediction models based on multimodal data. To further expand the clinical applicability of the models, more detailed phenotypic data could be added. We envision that the predicted distribution score of each mental disorder category or control from the multiclass model (model 2) can be used as part of a decision-support tool at the time of the clinical assessment. In case of no clear prediction, results

© 2022 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Sun Yat-Sen University User on 12/11/2022

Figure 3. Feature Importance of Diagnostic Models and Prognostic Model Performance



True labels \ Predicted labels	0	1	2	3	4
0	4032	3425	2114	1071	1243
1	3035	3724	2160	713	852
2	2229	2726	2387	684	1587
3	777	634	517	489	682
4	1296	1007	1545	907	3017

A, Percentage drop in prediction accuracy on the test data when masking out each of the included data modalities in evaluating the cross-diagnostic prediction models (models 1-3). B, Receiver operating characteristic (ROC) area under the curve (AUC) curve for prediction of the combined cross-diagnostic prognostic groups of increased severity with class 0 being the least severe and class 4 the most severe. Included are the overall multilabel performance and performance on each of the individual severity group labels in the model. C, Confusion matrix showing the distribution of predicted classes for the cross-diagnostic prediction of the severity groups 0 to 4 with the color intensity showing the fraction across the rows (true label). The numbers in the table are the number of predictions as a given class for each of the true classes. Therefore, each row sums up the total number of individuals in the class, and the sum of each column is the total number of predictions for the class. D, Percentage drop in prediction accuracy on the test data when masking out the data in evaluating the prognostic model of the combined cross-diagnostic severity groups. HLA indicates human leukocyte antigen; MBR, medical birth registry; PRS, polygenic risk scores.

© 2022 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Sun Yat-Sen University User on 12/11/2022

suggest that the clinicians should not rely on the model for the final diagnostic decision. The model could eventually be implemented to summarize all historical data into an underlying probability for each mental disorder diagnosis that the clinician can use together with the regular assessment to provide more standardized care across clinics.

ARTICLE INFORMATION

Accepted for Publication: October 12, 2022.

Published Online: December 7, 2022. doi:10.1001/jamapsychiatry.2022.4076

Author Affiliations: Copenhagen Research Centre for Mental Health, Mental Health Centre Copenhagen, Copenhagen University Hospital, Copenhagen, Denmark (Allesøe, Nordentoft, Benros); Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark (Allesøe, Rasmussen); Division of Biostatistics and Department of Radiology, Population Neuroscience and Genetics Lab, University of California, San Diego, La Jolla (Thompson); iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Department of Immunology and Microbiology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark (Bybjerg-Grauholm, Hougaard, Nordentoft, Werge); Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark (Bybjerg-Grauholm, Hougaard); Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark (Nordentoft, Werge); Institute of Biological Psychiatry, Mental Health Centre Sct Hans, Mental Health Services Copenhagen, Roskilde, Denmark (Werge); Department of Immunology and Microbiology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark (Benros).

Author Contributions: Dr Benros 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. Drs Rasmussen and Benros are considered co–senior authors.
Concept and design: Allesøe, Thompson, Hougaard, Nordentoft, Werge, Rasmussen, Benros.
Acquisition, analysis, or interpretation of data: Allesøe, Bybjerg-Grauholm, Nordentoft, Werge, Rasmussen, Benros.
Drafting of the manuscript: Allesøe, Thompson, Rasmussen.
Critical revision of the manuscript for important intellectual content: Allesøe, Bybjerg-Grauholm, Hougaard, Nordentoft, Werge, Rasmussen, Benros.
Statistical analysis: Allesøe, Thompson, Hougaard, Rasmussen.
Obtained funding: Nordentoft, Werge, Rasmussen, Benros.
Administrative, technical, or material support: Bybjerg-Grauholm, Nordentoft, Benros.
Supervision: Nordentoft, Werge, Rasmussen, Benros.

Conflict of Interest Disclosures: Dr Werge reported that he acted as a lecturer and scientific counselor to H. Lundbeck A/S. No other disclosures were reported.

Funding/Support: This specific study was funded by grant R268-2016-3925 from The Lundbeck Foundation, Denmark, and independent research grants R278-2018-1411 and NNF14CC0001 from the Novo Nordisk Foundation. The genotyping of the iPSYCH samples was supported by grants R102-A9118 and R155-2014-1724 from The Lundbeck Foundation, Denmark, the Stanley Foundation, grant SFARI 311789 from the Simons Foundation, and grant 5U01MH094432-02 from the National Institute of Mental Health.

Role of the Funder/Sponsor: The funders 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.

Additional Information: This research has been conducted using the Danish National Biobank resource supported by the Novo Nordisk Foundation. IPSYCH data were stored and analyzed at the Computerome HPC Facility (http://www.computerome.dtu.dk/). Preben B. Mortensen, DrMedSc, Anders D. Børglum, PhD, David M. Hougaard, MD, PhD, Ole Mors, MD, PhD, Thomas Werge, PhD, and Merete Nordentoft, MD, PhD, are the principal investigators of iPSYCH. iPSYCH data is stored in a national HPC facility in Denmark. The iPSYCH initiative is committed to providing access to these data to the scientific community, in accordance with Danish law. All code is available on GitHub at https://github.com/RasmussenLab/PSYCH-Pred/ and is freely available under the permissive MIT License. All analyses were run using PSYCH-Pred v1.0.

REFERENCES

1. Chekroud AM, Bondar J, Delgadillo J, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*. 2021;20(2):154-170. doi:10.1002/wps.20882

2. Varga TV, Niss K, Estampador AC, Collin CB, Moseley PL. Association is not prediction: a landscape of confused reporting in diabetes—a systematic review. *Diabetes Res Clin Pract*. 2020;170:108497. doi:10.1016/j.diabres.2020.108497

3. Bzdok D, Varoquaux G, Steyerberg EW. Prediction, not association, paves the road to precision medicine. *JAMA Psychiatry*. 2021;78(2):127-128. doi:10.1001/jamapsychiatry.2020.2549

4. Koppe G, Meyer-Lindenberg A, Durstewitz D. Deep learning for small and big data in psychiatry. *Neuropsychopharmacology*. 2021;46(1):176-190. doi:10.1038/s41386-020-0767-z

5. Squarcina L, Villa FM, Nobile M, Grisan E, Brambilla P. Deep learning for the prediction of treatment response in depression. *J Affect Disord*. 2021;281:618-622. doi:10.1016/j.jad.2020.11.104

6. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6

7. An C, Lim H, Kim DW, Chang JH, Choi YJ, Kim SW. Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. *Sci Rep*. 2020;10(1):18716. doi:10.1038/s41598-020-75767-2

8. Blomberg SN, Folke F, Ersbøll AK, et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation*. 2019;138:322-329. doi:10.1016/j.resuscitation.2019.01.015

9. Santos A, Colaço AR, Nielsen AB, et al. Clinical knowledge graph integrates proteomics data into clinical decision-making. *bioRxiv*. Preprint posted online May 10, 2020. doi:10.1101/2020.05.09.084897

10. Choi GH, Yun J, Choi J, et al. Development of machine learning-based clinical decision support system for hepatocellular carcinoma. *Sci Rep*. 2020;10(1):14855. doi:10.1038/s41598-020-71796-z

11. Nielsen AB, Thorsen-Meyer HC, Belling K, et al. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. *Lancet Digit Health*. 2019;1(2):e78-e89. doi:10.1016/S2589-7500(19)30024-X

12. Bachtiger P, Peters NS, Walsh SL. Machine learning for COVID-19-asking the right questions. *Lancet Digit Health*. 2020;2(8):e391-e392. doi:10.1016/S2589-7500(20)30162-X

13. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342-1350. doi:10.1038/s41591-018-0107-6

14. Salazar de Pablo G, Studerus E, Vaquerizo-Serrano J, et al. Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. *Schizophr Bull*. 2021;47(2):284-297. doi:10.1093/schbul/sbaa120

15. Gao S, Calhoun VD, Sui J. Machine learning in major depression: from classification to treatment outcome prediction. *CNS Neurosci Ther*. 2018;24(11):1037-1052. doi:10.1111/cns.13048

16. Bracher-Smith M, Crawford K, Escott-Price V. Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Mol Psychiatry*. 2021;26(1):70-79. doi:10.1038/s41380-020-0825-2

17. Mumtaz W, Qayyum A. A deep learning framework for automatic diagnosis of unipolar depression. *Int J Med Inform*. 2019;132:103983. doi:10.1016/j.ijmedinf.2019.103983

18. Meehan AJ, Lewis SJ, Fazel S, et al. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Mol Psychiatry*. 2022;27(6):2700-2708. doi:10.1038/s41380-022-01528-4

19. Nudel R, Benros ME, Krebs MD, et al. Immunity and mental illness: findings from a Danish population-based immunogenetic study of seven psychiatric and neurodevelopmental disorders. *Eur J Hum Genet*. 2019;27(9):1445-1455. doi:10.1038/s41431-019-0402-9

20. Benros ME, Waltoft BL, Nordentoft M, et al. Autoimmune diseases and severe infections as risk factors for mood disorders: a nationwide study. *JAMA Psychiatry*. 2013;70(8):812-820. doi:10.1001/jamapsychiatry.2013.1111

21. Benros ME, Mortensen PB, Eaton WW. Autoimmune diseases and infections as risk factors

© 2022 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Sun Yat-Sen University User on 12/11/2022

for schizophrenia. *Ann N Y Acad Sci*. 2012;1262:56-66. doi:10.1111/j.1749-6632.2012.06638.x

22. Pedersen CB, Bybjerg-Grauholm J, Pedersen MG, et al. The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol Psychiatry*. 2018;23(1):6-14. doi:10.1038/mp.2017.196

23. Schork AJ, Won H, Appadurai V, et al. A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nat Neurosci*. 2019;22(3):353-361. doi:10.1038/s41593-018-0320-0

24. Pedersen CB, Gøtzsche H, Møller JO, Mortensen PB. The Danish Civil Registration System: a cohort of 8 million persons. *Dan Med Bull*. 2006;53(4):441-449.

25. Mors O, Perto GP, Mortensen PB. The Danish Psychiatric Central Research Register. *Scand J Public Health*. 2011;39(7)(suppl):54-57. doi:10.1177/1403494810395825

26. Allesøe RL, Nudel R, Thompson WK, et al. Deep learning-based integration of genetics with registry data for stratification of schizophrenia and depression. *Sci Adv*. 2022;8(26):eabi7293. doi:10.1126/sciadv.abi7293

27. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience*. 2019;8(7):giz082. doi:10.1093/gigascience/giz082

28. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. Accessed June 21, 2019. https://openreview.net/pdf?id=BJJsrmfCZ

29. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. Preprint posted online January 2, 2012. doi:10.48550/arXiv.1201.0490

30. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach Learn*. 2001;45(2):171-186. doi:10.1023/A:1010920819831

31. Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. *PLoS One*. 2012;7(8):e41882. doi:10.1371/journal.pone.0041882

32. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1):6. doi:10.1186/s12864-019-6413-7

33. Shah JL, Scott J, McGorry PD, et al; International Working Group on Transdiagnostic Clinical Staging in Youth Mental Health. Transdiagnostic clinical staging in youth mental health: a first international consensus statement. *World Psychiatry*. 2020;19(2):233-242. doi:10.1002/wps.20745

34. Perlis RH, Brown E, Baker RW, Nierenberg AA. Clinical features of bipolar depression versus major depressive disorder in large multicenter trials. *Am J Psychiatry*. 2006;163(2):225-231. doi:10.1176/appi.ajp.163.2.225

35. Purcell SM, Wray NR, Stone JL, et al; International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009;460(7256):748-752. doi:10.1038/nature08185

36. Landi I, Kaji DA, Cotter L, et al. Prognostic value of polygenic risk scores for adults with psychosis. *Nat Med*. 2021;27(9):1576-1581. doi:10.1038/s41591-021-01475-7

37. Lauritsen MB, Jørgensen M, Madsen KM, et al. Validity of childhood autism in the Danish Psychiatric Central Register: findings from a cohort sample born 1990-1999. *J Autism Dev Disord*. 2010;40(2):139-148. doi:10.1007/s10803-009-0818-0

38. Svensson E, Voldsgaard I, Haller LG, Baandrup L. Validation study of the population included in the Danish Schizophrenia Registry. *Dan Med J*. 2019;66(10):A5571.

39. Frederiksen LH, Bilenberg N, Andersen L, et al. The validity of child and adolescent depression diagnoses in the Danish psychiatric central research register. *Acta Psychiatr Scand*. 2021;143(3):264-274. doi:10.1111/acps.13258

40. Bock C, Bukh JD, Vinberg M, Gether U, Kessing LV. Validity of the diagnosis of a single depressive episode in a case register. *Clin Pract Epidemiol Ment Health*. 2009;5:4. doi:10.1186/1745-0179-5-4

41. Fusar-Poli P, Solmi M, Brondino N, et al. Transdiagnostic psychiatry: a systematic review. *World Psychiatry*. 2019;18(2):192-207. doi:10.1002/wps.20631

42. McGorry P, van Os J. Redeeming diagnosis in psychiatry: timing versus specificity. *Lancet*. 2013;381(9863):343-345. doi:10.1016/S0140-6736(12)61268-9

43. Caspi A, Houts RM, Ambler A, et al. Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the Dunedin Birth Cohort Study. *JAMA Netw Open*. 2020;3(4):e203221. doi:10.1001/jamanetworkopen.2020.3221

© 2022 American Medical Association. All rights reserved.

Downloaded From: https://jamanetwork.com/ by a Sun Yat-Sen University User on 12/11/2022