Title

Deep learning—based integration of genetics with registry data for stratification of schizophrenia and depression

Permalink

https://escholarship.org/uc/item/0nx718wc

Journal

Science Advances, 8(26)

ISSN

2375-2548

Authors

Allesøe, Rosa Lundbye
Nudel, Ron
Thompson, Wesley K
et al.

Publication Date

2022-07-01

DOI

10.1126/sciadv.abi7293

DISEASES AND DISORDERS

Deep learning–based integration of genetics with registry data for stratification of schizophrenia and depression

Rosa Lundbye Allesøe[1,2], Ron Nudel[1,3], Wesley K. Thompson[3,4,5], Yunpeng Wang[6], Merete Nordentoft[1,3,7], Anders D. Børglum[3,8,9], David M. Hougaard[3,10], Thomas Werge[3,4,7], Simon Rasmussen[2]*†, Michael Eriksen Benros[1,3,11]*†

Currently, psychiatric diagnoses are, in contrast to most other medical fields, based on subjective symptoms and observable signs and call for new and improved diagnostics to provide the most optimal care. On the basis of a deep learning approach, we performed unsupervised patient stratification of 19,636 patients with depression [major depressive disorder (MDD)] and/or schizophrenia (SCZ) and 22,467 population controls from the iPSYCH2012 case cohort. We integrated data of disorder severity, history of mental disorders and disease comorbidities, genetics, and medical birth data. From this, we stratified the individuals in six and seven unique clusters for MDD and SCZ, respectively. When censoring data until diagnosis, we could predict MDD clusters with areas under the curve (AUCs) of 0.54 to 0.80 and SCZ clusters with AUCs of 0.71 to 0.86. Overall cases and controls could be predicted with an AUC of 0.81, illustrating the utility of data-driven subgrouping in psychiatry.

INTRODUCTION

The current psychiatric diagnostic categories remain restricted to subjective symptoms and subjectively observable signs, in contrast to most other medical fields, where diagnoses are often made based on quantifiable biomarkers. Schizophrenia (SCZ) and major depressive disorder (MDD) are both severe mental disorders with a large impact on the individual's well-being and are among the biggest societal health burdens (*1*, *2*). Within the current diagnostic scheme of both SCZ and MDD, large variation exists in terms of treatment response, clinical presentation at onset, and disease progression (*3*). Furthermore, both SCZ and MDD have been shown to have polygenic architectures, associations with birth-related factors, as well as complex overlaps with other mental disorder diagnoses and general medical conditions, particularly diseases with immunological pathophysiological mechanisms, which could be used for stratification and prediction models (*4*–*6*). However, most prior studies have focused on the overall diagnostic categories and included a limited number of features, which might not be adequate to add the desired additional clinical value for these highly diverse disorders (*6*, *7*). Together, this calls for new and improved diagnoses and evaluations in psychiatry that can help stratify patients more efficiently and guide clinicians in providing optimal care (*6*).

For stratification of MDD and SCZ, a limited number of studies have applied machine learning for clustering, and no prior studies have used a deep learning (DL) framework (*8*–*10*). DL methods are able to process high-dimensional data and capture nonlinear structures (*6*, *11*). Specifically, variational autoencoders (VAEs), which are based on deep neural networks (NNs), have been highly useful for unsupervised learning of structures in large datasets (*12*). VAEs work by compressing high-dimensional data into lower-dimensional latent representations through the training of NNs, thereby capturing nonlinear correlations in the data. DL has shown promising results in identifying biologically relevant low-dimensional information from highly heterogeneous data such as transcriptomics, single-cell sequencing data, and integration of multiple datasets on human microbiome data (*13*–*16*). Currently, DL models have mainly been used to predict SCZ or MDD from data such as brain imaging via magnetic resonance imaging (MRI), genotype data, electroencephalographic, or social media data (*7*, *17*–*21*). However, these studies generally lack sufficient sample sizes (majority with no more than 400 cases with up to ~5500 for genetic models), have mainly looked at one diagnostic group, and lack true population controls to conclude on the generalization of the predictive power. To fully use the potential of the DL approach in gaining a deeper understanding of the complexity of the mental disorders, integration of more data from larger samples sizes is needed.

In this study, we used a large Danish population cohort from the Integrative Psychiatric Research Consortium (iPSYCH) of 19,636 individuals with MDD and/or SCZ including a population control group of 22,467 individuals, all genotyped and linked with the Danish nationwide registers (table S1). DL models were applied for efficient data integration to establish a clinically predictable stratification of individuals with MDD and SCZ. Individuals were stratified by all available data pertaining to them and their family medical history of mental disorders and other medical conditions, as well as birth-related

[1]Copenhagen Research Centre for Mental Health, Mental Health Centre Copenhagen, Copenhagen University Hospital, Copenhagen, Denmark. [2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [3]iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus, Denmark. [4]Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Mental Health Services Copenhagen, Roskilde, Denmark. [5]Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, San Diego, CA, USA. [6]Lifespan Changes in Brain and Cognition (LCBC), Department of Psychology, University of Oslo, Forskningsveien 3A, 0317 Oslo, Norway. [7]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [8]Department of Biomedicine, Aarhus University and Centre for Integrative Sequencing, iSEQ, Aarhus, Denmark. [9]Aarhus Genome Center, Aarhus, Denmark. [10]Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark. [11]Department of Immunology and Microbiology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
*Corresponding author. Email: simon.rasmussen@cpr.ku.dk (S.R.); michael.eriksen.benros@regionh.dk (M.E.B.)
†These authors contributed equally to this work.

variables, previously identified genetic markers from the literature, and the severity of their mental disorders presented by hospital contacts. Using the integrated data, we aimed to achieve a better stratification of possible disorder trajectories to get a more complete picture of the complexity between and within SCZ and MDD compared to a background population. Furthermore, by establishing new DL prediction models, we investigate the clinical predictability of the identified subgroups only including data up until their initial diagnosis. Given the size of the cohort and the fact that all individuals were followed from birth, we have increased power to gain new insights into the etiologies of these disorders using state-of-the-art DL methods.

RESULTS

Data-driven stratification of population controls and the severe mental disorders

We applied our DL VAE framework to integrate all the register-based and genetic data available (see Materials and Methods and data file S1) into a common latent representation to be used in the cluster analysis. Before clustering, the optimal model hyperparameters were determined on a random subset of 6000 individuals from the full cohort divided into training (5000) and test (1000) datasets to make it feasible for an exhaustive grid search (see Materials and Methods and fig. S1). In the cluster analysis of the full cohort of all 42,103 individuals, referred to as cluster analysis A, we identified an optimum of six clusters with an expected clear overall separation with three clusters characterized as background population (A-Back_pop1, A-Back_pop2, and A-Back_pop3), two clusters as MDD (A-MDD1 and A-MDD2), and one cluster in which 99% were diagnosed with SCZ (A-SCZ1) (Fig. 1, A and B, and fig. S2). Compared to using a standard principal components analysis (PCA), a sparse PCA (SPCA), a truncated singular value decomposition, or a Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) for reduction, the VAE showed a better separation of population control from mental disorders (fig. S3 and table S2). In the feature importance analysis, we found that psychiatric disorders and the severity of these had the largest impact on the clustering measured in the change of adjusted Rand index between subgroups (0.76 and 0.80), followed by family history and genetics (0.50 and 0.58) (table S3).

Disorder clusters resemble the overall diagnostic categories

When investigating the signatures of the clusters, we found that two of the background population clusters, A-Back_pop1 and A-Back_pop2, did not have any individuals diagnosed with MDD, SCZ, or other mental disorders [Fig. 1B and data file S2 for all *P* values and confidence intervals (CIs)]. However, A-Back_pop3 had a higher similarity to the three primary mental disorder clusters and included individuals with infections [adjusted *P* value of <0.001 – average 1.38 per individual (pi)] and family history of mental disorders and infections; however, the average number of mental disorders (0.4 pi) was far lower compared to the disease clusters (2.7 to 4.4 pi). Furthermore, this cluster had a higher average polygenic risk score (PRS) for both MDD and SCZ than the two other background clusters (data file S2). The two MDD clusters were mainly separated by severity with A-MDD2 having more inpatient hospital contacts, suicide attempts, infections, and a higher PRS for MDD compared to A-MDD1 (all adjusted *P* values of <0.001) (data file S2). This highlights that, in broad terms, the overall diagnostic categories more or less form as separate distinct subgroups when compared to a background population, with the only subdivision being severe or less severe MDD.

Excluding population controls gave a more detailed stratification of depression

To get a deeper picture of possible subgroups within MDD and SCZ, we repeated the VAE integration and clustering of mental disorder cases only, referred to as cluster analysis B. Here, we identified seven clusters (fig. S4A), of which six were mainly MDD clusters (91 to 100%) and one resembled the A-SCZ1 cluster from cluster analysis A (B-SCZ1) (Fig. 2A and data file S3 for all *P* values and CI). Cluster B-MDD1 and B-MDD2 were the most similar of all clusters with a cluster distance of only 0.015 (fig. S5A) and were mainly distinguished by B-MDD1 having significantly more infections and other medical conditions and B-MDD2 having a slightly higher average of outpatient hospital contacts (Fig. 2A). Overall, the severity in terms of hospital contacts, suicide attempts, and mental comorbidities increased going from B-MDD1/B-MDD2 to B-MDD6 and to B-SCZ1. In cluster B-MDD3, 99.7% also had anxiety disorders; cluster B-BDD4 had more hospital contacts for MDD, suicide attempts, housing days, and infections; in B-MDD5, 99.9% had behavioral disorders with onset in childhood and adolescence; and in B-MDD6, 94.0% had developmental disorders. Furthermore, both B-MDD5 and B-SCZ1 were defined by more parental history of infections and mental disorders as well as more infections in B-MDD5. Both clusters were also highly severe in terms of suicide attempts and many inpatient hospital contacts for their MDD. Overall, the four clusters (B-MDD3, B-MDD5, B-MDD6, and B-SCZ1) with one other major mental comorbidity (>90%) had more additional mental comorbidities compared to the three other clusters (average of 4 to 6 pi compared to 1 to 2 pi).

For all clusters, we only observed some single-nucleotide polymorphisms (SNPs) with differences in homozygote/heterozygote distributions but not a single SNP or HLA signal that characterized the clusters (data file S3). In the SNP distribution and MDD-PRS, B-MDD2 was the most genetically different to the other subgroups and had a lower genetic load (PRS 3.9×10^{-5}, *P* value of 0.0014), whereas A-MDD5 and A-MDD6 had the highest genetic load (PRS 4.1×10^{-5}, *P* value of 0.0019 and 0.11). When looking at specific human leukocyte antigen (HLA) alleles that were at least nominally significantly associated with SCZ or MDD in previous studies (*22*, *23*), we observed that HLA*A0101 was significant in B-MDD2. Together, we identified clear subgroups within MDD with distinct disorder prognostics and comorbid signatures that highlight the high heterogeneity and potential clinical use to move toward a more personalized treatment strategy.

Clear stratification of SCZ subgroups in terms of severity and comorbidities

For a more detailed stratification of the 3896 patients with SCZ, we conducted a subsequent analysis only including the patients diagnosed with SCZ referred to as cluster analysis C. We identified seven stable SCZ clusters from the latent representation with unique signatures across the included datasets (Fig. 2B and fig. S6A). Here, the feature importance showed a higher impact on each data modality, measured by the average change in the identified clusters using adjusted Rand index, compared to analyses A and B (0.47 compared to

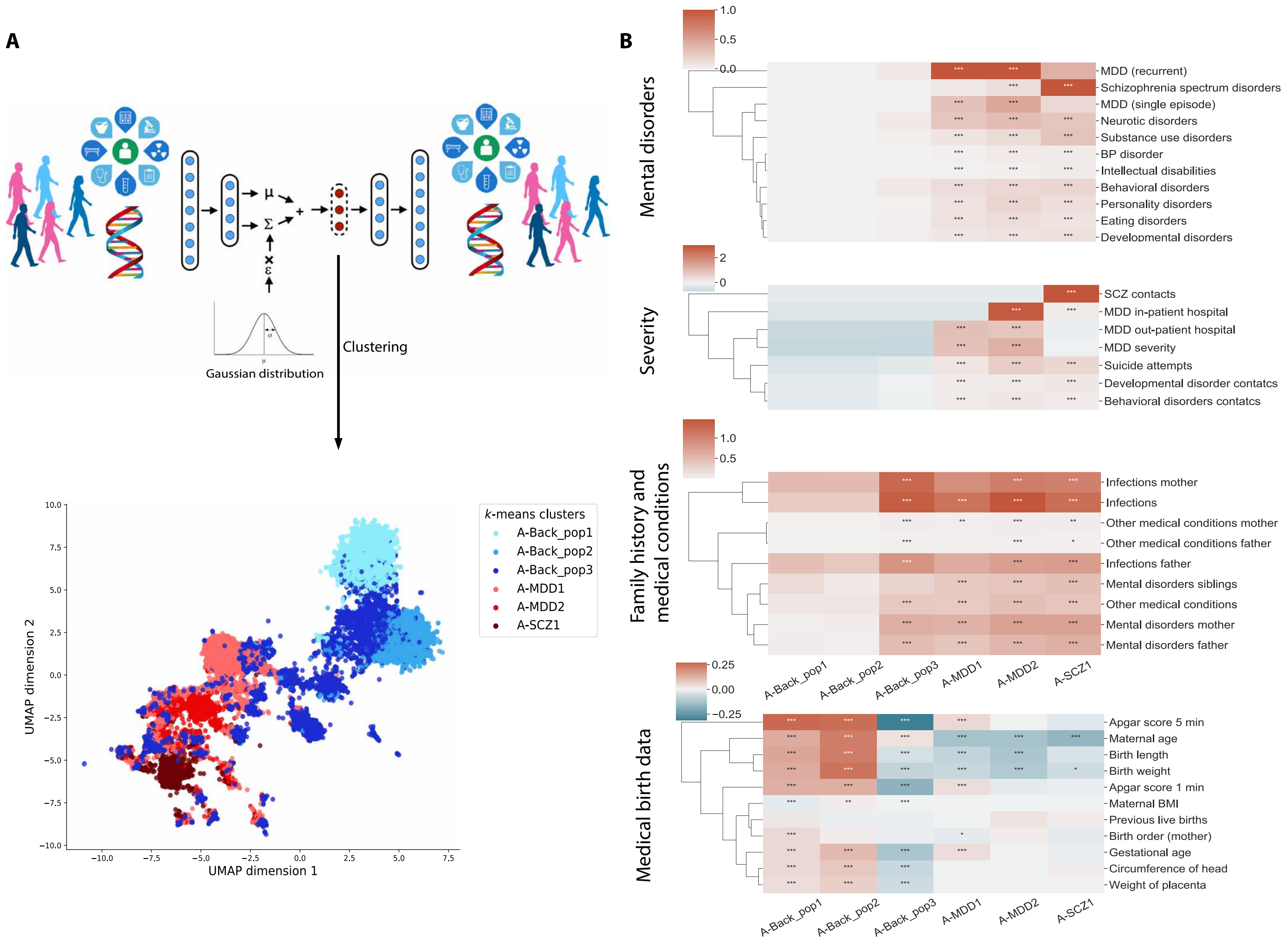

Fig. 1. Clear stratification of depression and SCZ from background population subgroups using data-driven DL integration and clustering. (**A**) Overview of the VAE framework for integrating genetic data with all the available register data including their own and family history of diagnostic data for both mental disorders and immune-related disease, as well as birth-related measurements, previously identified genetic markers, and the severity of the mental disorders presented by hospital contacts, suicide attempts, and housing days. The overview illustrates how the VAE is used to learn the lower-dimensional latent representation for clustering of the individuals into distinct subgroups. The latent representation of the individuals is additionally used for a two-dimensional UMAP visualization to illustrate the clusters. The UMAP visualization is of all the individuals and illustrates the six identified clusters of background population, depression (MDD), and SCZ. Each dot represents a patient, and for GDPR (General Data Protection Regulation) purposes, we have masked all single occurrences not in close proximity with other individuals to ensure privacy (less than 30 individuals were masked in total). (**B**) Heatmap visualization of the cluster signatures identified in cluster analysis A. Here, we have for mental disorders grouped into the ICD-10 block within the F chapter, except for mood disorders that are divided into single-episode MDD, recurrent or single-episode MDD, and BD. The scale illustrates the fraction of individuals with at least one diagnosis within the ICD-10 block for that cluster. Both severity and medical birth data are *z*-score–normalized showing from low to high values. For severity, hospital contacts are a combined average of both days admitted and number of admissions. Family history and medical conditions are a combined count of the average number of occurrences across all included diagnoses per individual in the cluster. The significance levels are defined as **p* < 0.05, ***p* < 0.001, and ****p* < 0.0001.

0.44 and 0.38) (table S3). Of these, 62% had been diagnosed with MDD as well. In the two clusters, C-SCZ2 and C-SCZ5, all individuals were also diagnosed with MDD (single episode or recurrent) with C-SCZ5 having more severe MDD in terms of more inpatient and outpatient hospital contacts as well as more housing days and suicide attempts. C-SCZ5 had, in addition to the MDD diagnosis, a significantly higher average number of infections and other immune-related diseases per individual.

In general, we observed an increase in disorder severity with significantly more comorbidities, hospital contacts, suicide attempts, and housing days going from C-SCZ1 to C-SCZ7 (all adjusted *P* values <0.05; Fig. 2B and data file S4 for all *P* values and CI). The least severe cluster, C-SCZ1, had no significant fraction of other mental comorbidities and was defined by no significant increase in any severity measurements. C-SCZ3 was defined by having the most hospital contacts for their SCZ spectrum disorder and a higher amount of comorbid substance use disorder (21%). C-SCZ4 was characterized by a high fraction of multiple different mental disorder comorbidities (average of six mental disorder diagnosis pi) and a higher fraction of a maternal family history of mental disorders. We found that C-SCZ6 had a similar characteristic with B-MDD6 by having developmental disorders (89%) and an early average age of

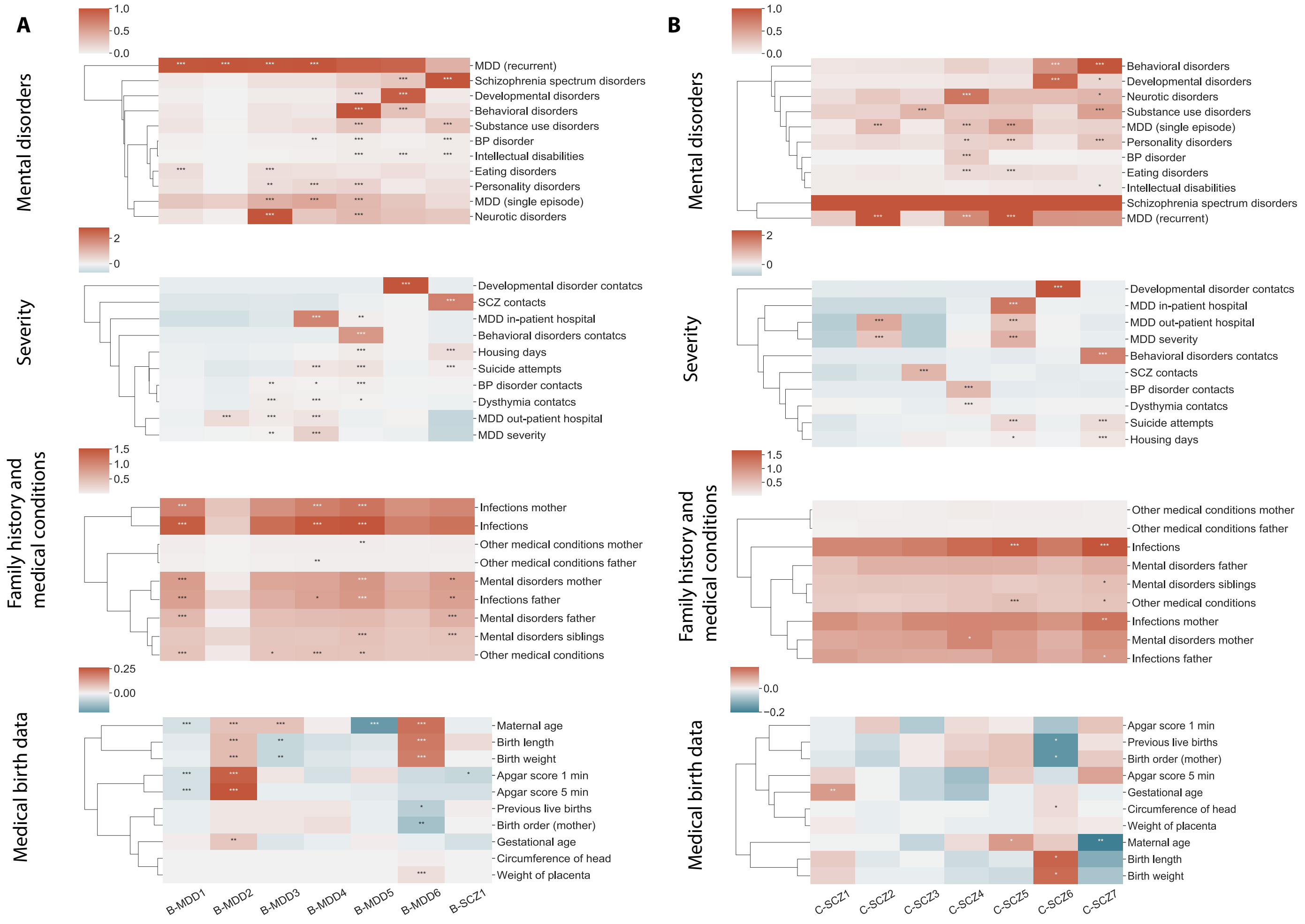

Fig. 2. Distinct clinical signatures in both comorbidities and severity measurements for depression and SCZ subgroups. (**A**) Heatmap visualization of the cluster signatures identified in cluster analysis B of depression (MDD) and SCZ. Here, we have mental disorders grouped into the ICD-10 block within the F chapter, except for mood disorders that are divided into single-episode MDD, recurrent or single-episode MDD, and BD. The scale illustrates the fraction of individuals with at least one diagnosis within the ICD-10 block for that cluster. Both severity and medical birth data are *z*-score–normalized showing from low to high values. For severity, hospital contacts are a combined average of both days admitted and number of admissions. Family history and medical conditions are a combined count of the average number of occurrences across all included diagnoses per individual in the cluster. The significance levels are defined as $^*p < 0.05$, $^{**}p < 0.001$, and $^{***}p < 0.0001$. (**B**) Heatmap visualization of the cluster signatures identified in cluster analysis C of SCZ of the same data as described for the heatmaps in (A).

diagnosis for most of the comorbid mental disorders as well an early SCZ diagnosis of approximately 18 years (figs. S4B and S6B). The most severe cluster, C-SCZ7, had the highest average of housing days and suicide attempts as well as many comorbidities, within both mental disorders and infections. In this analysis, we found no significant signal for individual SNPs in the genetics data and only a few SNPs with significantly different homozygote/heterozygote distributions (data file S4).

Prediction of cluster membership and MDD or SCZ diagnostics

To get an estimate of the clinical predictability of the clusters, we trained three feed-forward neural network (FFNN) prediction models for predictions of cluster membership on data up until their main mental disorder diagnosis of MDD or SCZ (prediction models A to C). To mimic a clinical situation at the time of diagnosis, we removed all disease severity measurements as well as events after the diagnosis of either MDD or SCZ including the diagnosis itself from the data used as input to the VAE. The time masked data were then used directly as input to the FFNN model for evaluating the prediagnostic predictability of clusters identified on the full-disorder trajectory. Other previous mental diagnoses were therefore still included; however, a diagnosis with MDD was only included in the prediction of the clusters in analysis C, to avoid introducing a biased increase in prediction accuracy for the patients with SCZ because of the nature of the cohort and the study setup. In addition, individuals without genotype data were removed from the prediction analysis to not bias the prediction due to imbalanced distribution of missingness between cases and controls. Furthermore, we, for comparison, trained a prediction model to predict the overall diagnostics groups of background, MDD, or SCZ using the same data (overall prediction model). All performance evaluations were done on a test set excluded from training and hyperparameter optimization (see Materials and Methods).

The first overall prediction model of the current broad clinical diagnosis of MDD and SCZ showed good performance with an overall AUC of 0.81 (AUC 0.74 for MDD and AUC 0.65 for SCZ), a Matthew correlation coefficient (MCC) of 0.39, and an accuracy of 65.0% (multiclass by-chance accuracy 41.6%) on the test set (Fig. 3A and table S4). We identified misclassifications between all three categories suggesting similarities in the groups when only considering data before a possible diagnosis of MDD or SCZ (fig. S8). In prediction model A (Fig. 3C and table S4), we observed a comparable performance with an AUC of 0.83, an MCC of 0.37, and an accuracy of 49.1% (multiclass by-chance accuracy 21.3%). The highest individual cluster AUCs were for the background population clusters with AUCs of 0.81 to 0.97 most likely caused by the postdisorder data censoring not changing their input data. However, the results further support A-Back_pop3 (AUC of 0.81) being a high-risk group as most of the misclassifications between the mental disorder clusters and background population were predicted to be A-Back_pop3 (fig. S2C).

Differences in the predictability of MDD and SCZ subgroups before diagnosis

The SCZ subgroups in cluster analysis C were easier to predict with an AUC of 0.79, an MCC 0.29, and 40.5% accuracy (cluster AUCs ranging from 0.70 for C-SCZ1 to 0.86 for C-SCZ6) than the MDD subgroups in cluster analysis B with an AUC of 0.72, an MCC of 0.17, and 31.4% accuracy (cluster AUCs ranging from 0.54 for B-MDD4 to 0.80 for B-MDD2) (Fig. 4, A and C, and table S4). In general, we found that most clusters in both analyses B and C were highly similar in their data up until diagnosis, but specifically for cluster analysis B, some clusters were more or less identical, making prediction a complicated task (figs. S5 and S7). With the decrease in

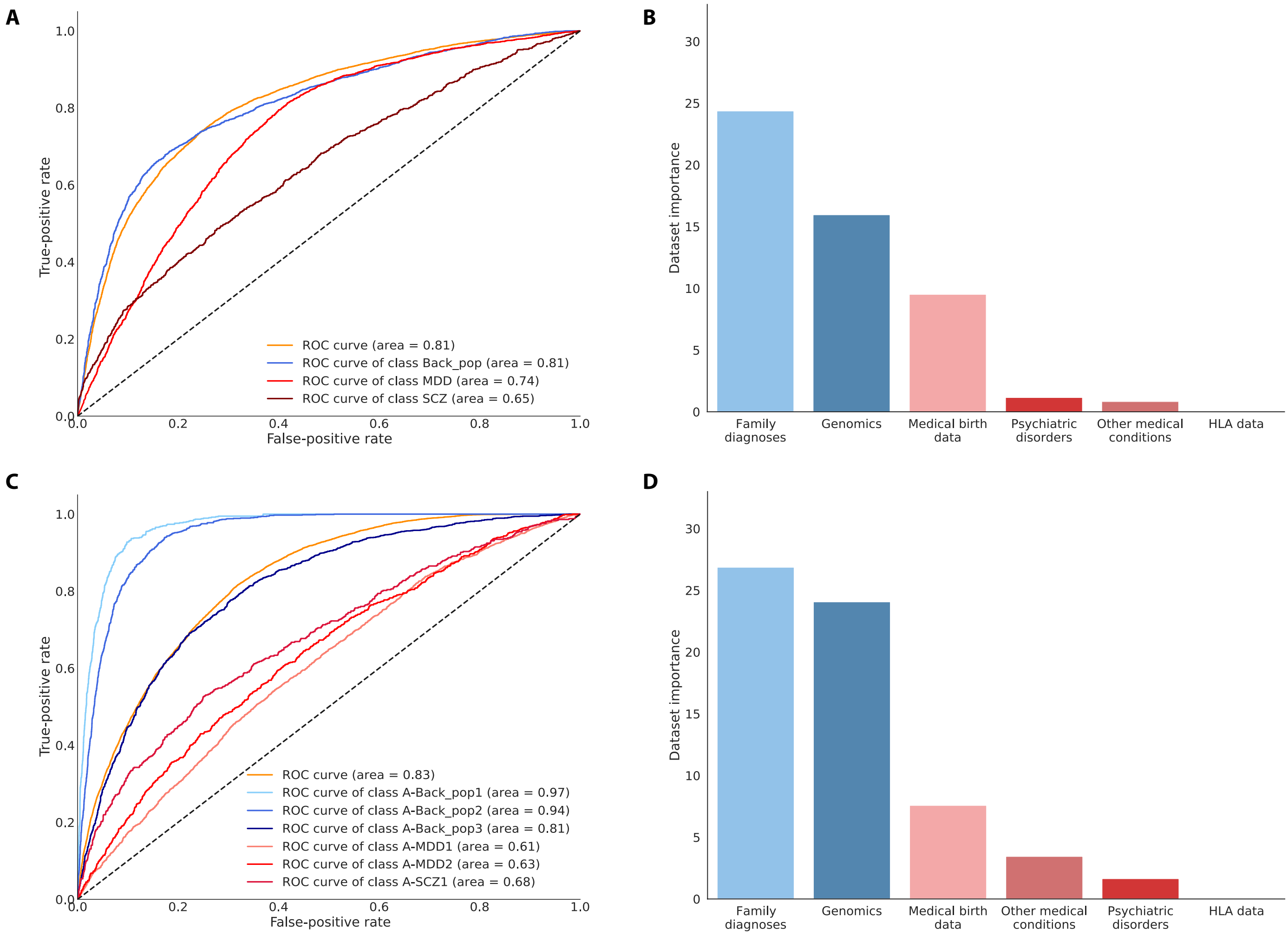

Fig. 3. High predictability of subgroups and the broad diagnoses of depression and SCZ from a background population with clear impact from family history and genetics. (**A**) ROC-AUC or AUC for the performance of the prediction model and specifically for each of the diagnostic categories of control, depression (MDD), and SCZ. The model performance was calculated on the basis of the prediction on the randomly sampled test set. (**B**) The reduction in prediction accuracy across all categories in the prediction model when removing each of the datasets completely from the prediction on the test set. The dataset was removed by setting all features to the mean for continuous data (zero as the data are *z*-score–normalized) and to missing for categorical features (a zero vector in the one-hot encoding) and passing the test data through the pretrained prediction model. (**C**) ROC curves for prediction model A of the six clusters identified in cluster analysis A of all individuals with MDD, SCZ, and the background population, again showing the overall model performance across all clusters and the individual cluster performances. (**D**) The dataset-specific accuracy reduction for prediction model A calculated the same way as in (B). The colors are linked to the dataset.

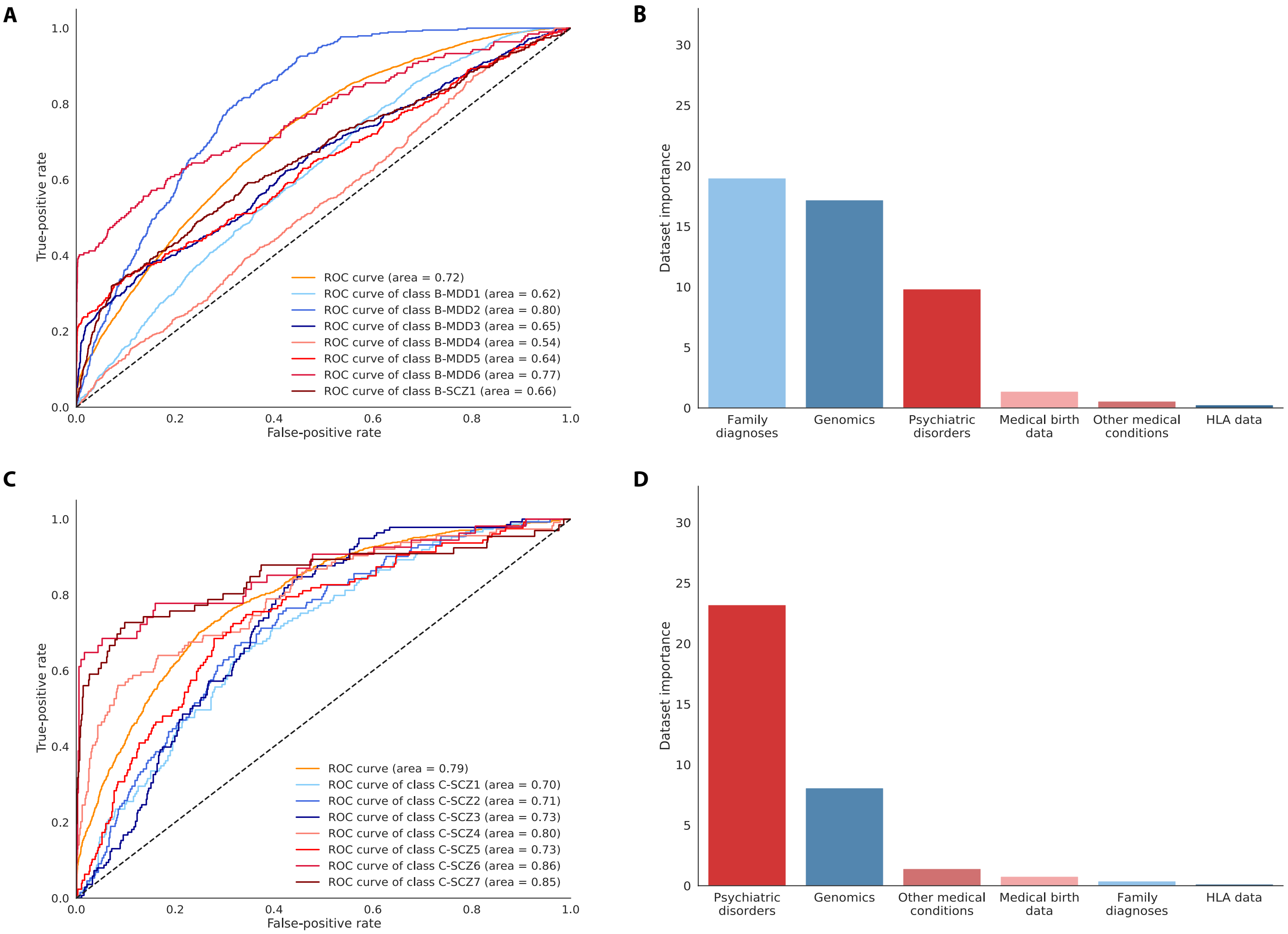

Fig. 4. Subgroups within SCZ are easier to predict than within depression when only considering data before diagnosis. (**A**) ROC-AUC or AUC for the performance of the prediction model of the clusters identified in cluster analysis B of depression (MDD) and SCZ and the performance for each of the seven clusters. The model performance was calculated on the basis of the prediction on the randomly sampled test set. (**B**) The reduction in prediction accuracy across all categories in the prediction model when removing each of the datasets completely from the prediction on the test set. The dataset was removed by setting all features to the mean for continuous data (zero as the data are *z*-score–normalized) and to missing for categorical features (a zero vector in the one-hot encoding) and passing the test data through the pretrained prediction model. (**C**) ROC curves for prediction model C of the seven clusters identified in cluster analysis C of all individuals with MDD, SCZ, and the background population, again showing the overall model performance across all clusters and the individual cluster performances. (**D**) The dataset-specific accuracy reduction for prediction model B calculated the same way as in (B). The colors are linked to the dataset.

overall model performance, these results highlight that, in particular, individuals diagnosed with MDD were more homogeneous at the time of diagnosis, making the prediction of subgroup trajectories more difficult compared to the easier prediction task of diseased versus healthy.

Feature importance for the prediction models

In both prediction models, which included the control population, the family history of diagnoses and genetics had the highest impact on the model accuracy; when removed, we observe a reduction in accuracy of 24.4 and 26.9% for the overall prediction model and prediction model A, respectively. Similarly, both models had genomics as the second most important feature with a reduction of 16.0 and 24.1%, followed by medical birth data with 9.5 and 7.6% (Fig. 3, B and D, and table S5). This pattern was also observed for prediction model B, with a 19% reduction for family history and a 17.2% reduction for genotype data. However, the patient's own prior diagnoses of mental disorders and other medical comorbidities proved to be the third most important feature, with a 9.8% reduction (Fig. 4B and table S5). For prediction model C, we found that, in contrast to the other prediction models, the family diagnoses were among the least important (only 0.38% reduction), whereas the most important was the mental disorder diagnoses with a 23.2% reduction, followed by the genotype data with an 8.1% reduction (Fig. 4D and table S5). This shows that the prediction of SCZ subgroups is mainly based on earlier mental diagnoses. The general increase in the importance of each dataset in the prediction of the more detailed stratification suggests, as expected, that the clusters have a more complex structure that requires more variables to predict and thus removing a dataset causes a larger decrease in predictive power.

DISCUSSION

In this first large-scale case-cohort study of 19,636 individuals with severe mental disorders and 22,467 population controls using DL for integration of genetics and registry data, we identified distinct subgroups with unique disorder severities and comorbidity signatures. Furthermore, we showed different predictability of the subgroups when only using data up until diagnosis ranging from AUC 0.55 to 0.97. Our DL data-driven stratification analysis clearly demonstrates the heterogeneity and overlap within and between mental disorder diagnostic categories of MDD and SCZ. Furthermore, several subgroups show high clinical predictability with an AUC above 0.8 for B-MDD2, C-SCZ4, C-SCZ6, and C-SCZ7. The important features defining our subgroups and used to predict them span both the pragmatic (based on disease-relevant features) and biological (based on causal factors) classifications, as described in the disease subtype framework recently proposed by Dahl and Zaitlen (*24*). This emphasizes the importance of our approach in the translational setting.

In previous studies, the predictive power of the model for a diagnosis of MDD and SCZ ranges between 0.54 and 0.95 in AUC or 45 to 99% accuracy (*7*, *17*, *19*). However, it was common for most models to have a questionable model evaluation and selection biases (*7*) and a low sample size of approximately 10 to 400 cases (up to ~5500 for genetic models) compared to ~4000 and ~16,000 cases of SCZ and MDD in our study, respectively. Furthermore, a direct comparison of the performances is not possible because our models are cross-diagnostic and the data types used in the studies are mostly different to ours, i.e., MRI and EGG (Electrogastrogram) data or models only relying on genetic data. In a real clinical setting, models that only predict one diagnosis compared to healthy individuals are not as clinically useful and would be comparable to our distinction of cases from population controls with an AUC of 0.81. However, our model has the improvement of a realistic control group as this also includes individuals with mental disorders, other than MDD and SCZ, thus being closer to a real clinical setting. As we also show here, cross-diagnostic signals exist between the mental disorders and a true population control group, which our prediction model encompasses. Therefore, there is a need for a better understanding of the cross-diagnosis similarities and differences to build a model that resembles a real clinical setting as many patients will likely present a continuum of symptoms within multiple mental disorders.

We identified an overlap in the characteristics of the subgroups identified within MDD to those within SCZ. For instance, we identified that both the MDD and SCZ cluster with high fraction of behavioral disorders, where B-MDD5 and C-SCZ7 had significantly more infections and other medical conditions compared to most other clusters as well as a lower maternal age. Similarly, we observed a high similarity between the two clusters with developmental disorders, B-MDD6 and C-SCZ6, with an earlier age at diagnosis across all comorbidities and a high predictability before diagnosis compared to the other clusters (AUC 0.77 and 0.86, respectively). This suggests that the subgroups of individuals where the majority is diagnosed with developmental disorders, mainly autism, tend to get their MDD or SCZ diagnosis earlier compared to the other subgroups. Consequently, the majority of the individuals in these clusters had a diagnosis that was given before MDD or SCZ diagnosis, which could improve the prediction.

For the prediction of clusters based on prediagnostic data, we do note that this will include partially overlapping data used for both clustering and prediction. While this is similar to a clinical application of such a prediction model of full-disorder trajectories, in a strict machine learning sense, this can result in an overestimation of the predictive performance. This potential effect on the performance measures is likely minor as most of the clusters were strongly driven by postdiagnostic features, such as hospital contacts, that were not included in the prediction models. Nonetheless, this should be considered when interpreting our performances compared with studies not using prediction for prediagnostic subgroup evaluation.

In all three cluster analyses, we did not find individual SNPs that were significantly overrepresented in the mental disorder subgroups or an overall separation in the genetic load of all SNPs. However, we did identify some significant differences in the distributions of the number of homozygote and heterozygote alleles for each SNP. Furthermore, the SNP distributions had predictive power in all the prediction models. This could be due to the fact that the presence or absence of a single SNP is not a strong enough signal to define the path of an entire subpopulation, and the underlying genetic signal is more likely to be a more complex combination of different genetic signatures (*25*, *26*). These were likely captured by the prediction model and resulted in the observed large impact of genomics data on the predictive power. This highlights the importance of the underlying genetics and their role in the prognostics of mental disorders. Overall, the signal from all combined HLA alleles was not strong.

One of the strengths of this study is the size of the cohort and the fact that it is a nationwide population-based cohort, which eliminates the majority of selection biases. However, the MDD cases are defined by having at least one visit to a hospital and, thus, only include the most severe MDD cases. Those with a less severe depression diagnosis will only have visits to general practitioners (GPs) and will not be registered in the hospital-based registers. This might explain some of the homogeneity we observe in the data before diagnosis and why many of these individuals might resemble the individuals diagnosed with SCZ. It is also likely that the background population individuals in the high-risk A-Back_pop3 cluster have been diagnosed with depression by their GP without being severely depressed, at least not within the time constrained of the study. From the data available in this analysis, we also have only a snapshot of the diagnosis data and an age of the first diagnosis. Some important differences might be observed in the number of repeated infections or severity of the infections for which data were not available to us.

Using DL approaches, we identified clinically distinct subgroups within both MDD and SCZ underlining the potential added values of a deeper stratification rather than only using the broad diagnostic categories. With this study being conducted on one of the largest cohorts in the field, we have shown that we can predict the most likely subgroup of either background population or MDD and SCZ as well as predict the specific subgroup within MDD and SCZ. With more detailed clinical information, such as omics data, EGG and MRI data, detailed clinical health records, and clinical assessments data with more information on treatment responses from clinical cohorts, the potential response differences within the subgroups can be further addressed. Thus, integrating the models presented in this study with further large-scale studies with more detailed molecular profiles and psychopathology would allow us to investigate the exact mechanisms involved and gain further insights into this proof-of-concept study. With further insights into the subgroups, these clinical prediction models can assist clinicians in the diagnosis and prognostics of patients with mental disorders providing more standardized and personalized care to the benefit of the individual patient.

MATERIALS AND METHODS

Cohort and experimental setup

The data in this study are from the large Danish population-based case-cohort sample (iPSYCH2012) from the iPSYCH (*27*). The iPSYCH2012 case-cohort sample consists of 87,764 Danish individuals who were born in Denmark between 1981 and 2005 and had dry blood spots from samples taken at birth available from the Danish Neonatal Screening Biobank (*28*). In this cohort, there are 57,764 people with at least one major mental disorder and 30,000 individuals who were randomly sampled from the population without regard for the psychiatric disorders. The mental disorders included in the iPSYCH cohort are autism spectrum disorder (F84.0-1, F84.5, and F84.8-9), attention-deficit/hyperactivity disorder (ADHD) (F90.0), SCZ (F20), bipolar disorder (BD) (F31), and MDD (F32 and F33). Genotyping, data cleaning, and imputation were described in detail by Schork *et al.* (*29*). For this study, we only included the individuals diagnosed with SCZ and/or MDD based on the International Classification of Diseases (ICD) 10 F20 and F32-F33 [ICD-8 codes 295.x9 (excl 295.79) for SCZ and 296.09, 296.29, 298.09, and 300.49 for MDD]. This included 19,636 individuals with either MDD or SCZ, of whom 15,740 were diagnosed with MDD and 3896 were diagnosed with SCZ. Furthermore, a population control group of approximately the same size, 22,467, was randomly selected from the control population in the iPSYCH cohort, not including individuals with either of the two focus diagnoses. From the Danish Psychiatric Central Research Register (*30*), we included all psychiatric diagnoses in ICD-10/ICD-8 codes including information on the age of diagnosis and psychiatric-related hospital contacts. In addition, registered infections, autoimmune diseases, the medical birth registry including complications and birth weight, and a selected number of other diseases were included. For the family history of the included individual, we only included parental psychiatric diagnoses, and siblings were included as a combined count of how many with the specific diagnoses. All data from the registers were included up until 2016 for mental disorders and 2012 for infections and autoimmune diseases. A total of 1185 variables were included before preprocessing of the data.

Preprocessing of data

The genotype data were included as the genotypes of risk alleles toward MDD, SCZ, bipolar spectrum disorder (BD), autism, ADHD, suicide ideation, autoimmune diseases, or infections such as influenza, human immunodeficiency virus, and hepatitis. The alleles were identified from the genome-wide association study (GWAS) catalog (*31*) and only included if the risk allele was specified. Note that, with regard to the genetic data, individuals who failed the genetic quality control as described in Schork *et al.* (*29*) were treated as having missing data. The quality control included removal of individuals with non-European descent as well as related individuals. We additionally ran a PCA using SmartPCA (*32*) on the 31,863 individuals with genotype data together with part of the HapMap sample [Japanese (JPT), Chinese (CHB), Yoruba (YRI), and European (CEU) descent] (*33*) to test for any major ancestral structure among the individuals in the 516 included SNPs. The cohort clustered with the European individuals, as expected, and showed no clear indication of ancestral structures (fig. S9). Furthermore, we found that none of the 10 first PCs were strongly correlated with the clusters (all PCC < 0.1). The genotype and HLA alleles were both included as being either homozygote for the allele, heterozygote, or not having the allele. The continuous data variables such as age of diagnosis, the number of hospitalizations, patient contacts, and birth measures were all normalized and centered around zero by *z*-score normalization per feature. Missing continuous data were encoded as the mean, which, for z-score–normalized data, is zero. The categorical data, genotype, HLA, and family history, were one-hot–encoded and flattened for input to the model. All missing categorical data were encoded as a zero vector. We only included data that were observed in at least 1% of the individuals in the analysis, and therefore, we had a small difference in the number of features including analyses A to C with 668, 692, and 711 features in A to C, respectively (see data file S1 for the full list). For the prediction models, we applied an individual-based age masking of the input data to resemble a clinical setting at the time of the diagnosis with either MDD or SCZ. This included masking of all diagnosis given after the exact age of MDD and SCZ including the diagnosis itself. Therefore, individuals with previous mental disorder diagnoses would still have these available for the prediction. Previous diagnoses of MDD were excluded from all prediction models that included both patients with MDD and SCZ as this would create a bias because only patients with SCZ could have this diagnosis after the masking. All severity measurements (hospital contacts and suicide attempts) were removed as these were available as a total sum of contacts and not individual events.

DL model

The DL model used to integrate the data was a VAE (*12*). A VAE consists of an encoder network, followed by a latent layer of size NL that is passed on to a decoder of the same sizes as the encoder layers arranged in reversed order. The model framework was built to account for a variable number of fully connected hidden layers in both the encoder and decoder. Each hidden layer included both batch normalization (*34*) and dropout ($P = 0.1$) (*35*) with Leaky Rectified Linear Units (*36*) as activation function. The latent layer was built from sampling from a Gaussian distribution N(0,I) of two fully connected layers of the means (μ) and SDs (σ), both of size NL.

When training the model, each dataset was merged to one input layer including both categorical and continuous variables and passed through the network. The reconstruction error was calculated separately for categorical and continuous datasets by splitting up the reconstructed output vector. The loss functions applied were cross-entropy for categorical data and mean-squared error for continuous data as implemented in PyTorch. For the categorical variables, we avoided back propagating of missing values using the ignore index implementation in PyTorch. For continuous variables, we set the reconstructions for all missing values to zero to match the input, so these would not contribute to the loss. In addition, the sampling of the latent layer was constrained to the Gaussian distribution by penalizing the deviance by adding the Kullback-Leibler divergence (KLD) to the loss. The final loss was defined as

$$L = E_{\mathrm{cat}} + E_{\mathrm{con}} + W_{\mathrm{KLD}} \times \mathrm{KLD}$$

Here, E_{cat} and E_{con} are the normalized reconstruction error for the continuous and categorical data. W_{KLD} is a weight put on the KLD defined as $W_{\text{KLD}} = \beta \times N_L^{-1}$, for which we used a β of 0.0001 for the final model. The KLD was defined as

$$\text{KLD} = -\frac{1}{2} \times \sum 1 + \ln(\sigma) - \mu - \sigma$$

The VAE model was trained with the Adam optimizer (*37*) and used the same settings to train three models for each analysis referred to as models A, B, and C. Between the models, the only changes in the training parameters were the batch size due to the differences in the sizes of the subsets of the data. Here, we used a batch size of 30, 15, and 10 for models A to C, respectively. We increased the batch size by a factor of 1.5 during training after every 50 epochs and used KLD warm-up by slowly increasing WKLD from zero to β × NL-1 at epochs 4, 6, 8, and 10 (*38*). The number of training epochs was determined to be 250 based on early stopping on the test set described under the hyperparameter testing. Hereafter, the latent representation of each individual was obtained by passing them through the trained VAE and extracting the μ layer. The VAE was implemented using PyTorch (v.1.4.0) (*39*).

Hyperparameter selection and stability testing

To identify the optimal parameters in terms of reconstruction accuracy, cluster capability, and model stability, different combinations of size of hidden layers, number of hidden layers, and size of latent space were evaluated. For the number of hidden neurons, we tested the sizes 400, 600, and 800, with the number of layers ranging between 1 and 3 with the same number of hidden neurons in each layer. The number of latent neurons used was 10, 20, or 40. We did an exhaustive search of all combinations on a randomly selected sample of 5000 individuals from the whole dataset for training and another 1000 individuals as a test set. The model performance was then evaluated in terms of test log-likelihood and reconstruction accuracy as well as reconstruction accuracy on the training data. An accurate reconstruction was for categorical variables defined as the class with the highest probability corresponding to the class given by the input. For continuous variables, the accuracy was assessed by comparing the reconstructed array with the input array using cosine similarity for each feature. Only nonmissing values were used when calculating the accuracy in the reconstruction. The performance of the model was also assessed by comparing the ability to cluster the data in terms of the intercluster separation calculated as the sum of squared distances (see more detail under the "Clustering" section). We chose the number of training epochs based on the lowest test error during training and rounded up to the nearest 100 epochs to ensure sufficient training to learn the complexity of the data. The stability of the model was evaluated by repeating training three times with the same hyperparameters and calculating the difference in cosine similarity to all other individuals for each individual in the dataset. If the model produced the same result, the average change in cosine similarity should be zero. The model with the average change closest to zero is then considered the most stable. The final model was then selected on the basis of all the above performance measures by selecting the model with the highest total rank across all tests. From this analysis, the optimal network architecture was found to be two hidden layers of 800 neurons and a latent space size of 40 dimensions (fig. S1).

Clustering

The clustering of the data was done using *k*-means, and all clustering and testing were done using Python scikit-learn (v.0.21.3) (*40*). The optimal number of clusters for the *k*-means approach was determined using the silhouette score to measure the spread of the clusters as well as the elbow method using the sum of squared distances to measure the intercluster separation as implemented in the scikit-learn package. From the latter, the optimal number of clusters was determined by first calculating the sum of squared distances (SSDs) of every number of clusters. From the SSD for the smallest and largest number of clusters, we then fitted a linear line and selected the optimal cluster number as the one with the largest distance to this line. The final clusters were selected on the basis of *k*-means clustering on a consensus matrix of repeated clustering of the latent space using the same parameters to ensure stable clusters.

Comparing PCA and UMAP-based reduction for clustering

We tested whether performing the dimensionality reduction before the cluster analysis using other methods would give similar results. Here, we used a PCA, UMAP (*41*), SPCA, and truncated singular value decomposition (TSVD) with the same reduction size (40 dimensions) (fig. S3). Here, we identified some overlap in cluster signatures when identifying the same number of clusters with adjusted Rand indices of 0.35, 0.23, 0.40, and 0.42 for PCA, UMAP, SPCA, and TSVD, respectively, using Python SciPy (v.1.3.1) (*42*). However, these clusters had a worse separation of population control from mental disorders. Using PCA and UMAP, we only found one clear background population (98 to 100%) cluster and two mixed (17 to 85% in background population clusters for PCA and 18 to 90% for UMAP) (fig. S3 and table S2). When based on UMAP reduction, no clusters were identified without any background population (3 to 13% in the remaining three clusters). SPCA and TSVD were both closer to the VAE clustering in the distribution of diagnoses in the main disorder groups (two 100% MDD with between 6 and 8% or between 12 and 14% SCZ and one SCZ cluster with 44 to 47% MDD). However, the background clusters had a higher percent of the cases with only 83 to 86% as background. This highlights the added value of using the VAE for dataset reduction with the two sparse methods, SPCA and TSVD, showing better performance compared with PCA and UMAP.

Assessment of impact of genotype missingness on clustering

The VAE integration can handle missing data by not using that information in the generation of the latent space. In our analysis, we had 10% of the individuals not having genotype data available because of failed QC (quality control). Of these, the majority was within the background population group, which had 18% missing data. We tested our VAE integration and cluster analysis without including these individuals and were able to identify the same six groups of three background groups, two MDD subgroups, and one SCZ subgroup (fig. S10). Therefore, we concluded that our method is stable and can handle a high percent missingness that is unevenly distributed between the labels included in the analysis.

Feature importance on clustering

We calculated the impact of each data modality on the clustering by setting each of the datasets to missing before VAE integration. We then recalculated the clusters by using the same *k*-means approach. The impact was then calculated from the adjusted Rand index between the true labels and the labels identified with the dataset set to missing. The reported impact is the change from a perfect overlap of 1.

Polygenic risk score calculation

PRSs for MDD and SCZ were calculated for iPSYCH individuals using PRSice2 v2.3.3 (*43*) and summary statistics from a meta-analysis of GWAS of MDD and SCZ. The summary statistics were based on Psychiatric Genomics Consortium samples from which Danish samples were excluded before meta-analysis. The MDD sample included 49,982 cases and 136,871 controls, and the SCZ sample included 34,129 cases and 45,512 controls (*44*, *45*). We included all SNPs in the calculation using a *P* value threshold of 1 and used a clumping r2 threshold of 0.1 in a window of 250 kilo–base pairs. The "--score" sum method was used to calculate the scores. Otherwise, the default parameters were used.

Statistical analysis

The identification of cluster-defining variables was done using a chi-square test for categorical variables using Python SciPy (v.1.3.1) (*42*). For the mental diagnoses, calculation was done by combining all diagnoses within the same block in the ICD-10 system and ICD-8 codes included in the corresponding ICD-10 block, and calculating the fraction of the individuals in a given cluster with at least one diagnosis within each block in the chapter. For the other medical conditions, these were combined into infections and others (migraine, diabetes, etc.) by counting the average number of each type of the two conditions each individual had within each cluster. The same calculation was done for the family history separated both into mental diagnoses and other medical conditions and infections as well as parental and maternal. For siblings, only mental diagnosis for MDD, BPD, SCZ, or any was available and included as a fraction of individuals with siblings with a diagnosis within each cluster. Calculation of genetic signals for both disorder-related phenotype SNPs and HLA alleles was done as the number of alleles out of all possible alleles within each cluster. All statistical tests for these were calculated as categorical with a one-sided chi-square test for identification of overrepresentation and cluster defining of diagnoses for each cluster. Furthermore, we also calculated whether there was a difference in the distribution of being a homozygote and being a heterozygote for each SNP with a two-sided chi-square test. For the medical birth data and severity variables, the numbers were included as the *z*-score–normalized values. For the medical birth data, a two-sided Student's *t* test was used to calculate whether a significant higher or lower, e.g., birth weight was defining the cluster. The hospital contacts for the severity measurements were combined as an average across all variables for the disorder (such as number of contacts and days admitted). For MDD, these were separated into outpatient and inpatient hospital contacts, and the statistics were done with a one-sided Student's *t* test to test whether a higher severity was defining the cluster. All tests were corrected for multiple testing using Bonferroni correction, and all the *P* values reported are corrected.

Cluster visualization and distances

We applied UMAP (*41*) on the latent representations for visualizing the clusters in two dimensions to capture potential nonlinear structures. The reduction was only applied for visualization purposes and not used to define the clusters or infer any relationships or distances between clusters. Cluster distances are computed as the correlation distance between the mean values of each cluster pair using the "distance.correlation" from the SciPy python package (v.1.3.1) (*42*). The distance between clusters ranges from 0 to 1, with 1 being the maximum distance.

Prediction models

The prediction models for the subsequent prediction of cluster labels and main disorder group (MDD, SZC, or background population control) were performed using an FFNN implemented using Python PyTorch (v.1.4.0) (*39*). We initially also tested the performance using a random forest model, using scikit-learn (v.0.21.3) (*40*); however, it was unable to handle the class imbalance compared to the performance of the FFNN with overprediction of the largest class (MCC of 0.26 in analysis A). We trained the model on 80% of the data and used the remaining 20% as a test set to evaluate the model. In the training, we used 10% of the data as a validation set for the hyperparameter optimization. The data were split using the train_test_split function implemented in Python scikit-learn (v.0.21.3) (*40*) to ensure equal splits across classes. Furthermore, we used the WeightedRandomSampler in PyTorch to account for class imbalance. We used the same implementations of activation function, dropout, and batch normalization as in the VAE and cross-entropy for loss calculation. The hyperparameter optimization was done with a full grid search of all combinations, and the model with the lowest prediction error on the validation set was used to evaluate the model. Early stopping was applied by saving the model with the lowest validation error during the 200 epochs of model training. On the basis of the performance of the VAE, we only tested prediction models with two hidden layers with the sizes [64, 32], [128, 64], or [256, 128]. Furthermore, we tested the learning rate (0.001 or 0.0001) and the batch size with different ranges depending on the analysis ([40, 50, 100] for analysis A, [20, 30, 40, 50] for analysis B, and [5, 10, 20, 25] for analysis C). The best-performing model was evaluated on the test set by calculating the receiver operating characteristic area under the curve (ROC-AUC or AUC) of the predictive power of each of the clusters individually and across all clusters. Furthermore, we calculated the overall accuracy and the MCC for each model. All performance calculations were done using the scikit-learn implementations. For all randomizations, the seed was set to 42. In the prediction model, all individuals with missing genetic data were removed in analysis A due to the large bias in the distribution between background and cases that only our VAE integration and clustering could account for. This removed around 10% of the individuals from the model, leaving 31,863 individuals in total. For analysis A, the optimal hyperparameter for the best-performing model on the validation set was a batch size of 100, a learning rate of 0.0001, and two layers of 256 and 128 neurons. For the second model in analysis A on predicting MDD, SCZ, or background, we used the same hyperparameters identified for prediction of the cluster membership. For analysis B, we found the optimal hyperparameters to have a batch size of 50, a learning rate of 0.0001, and two layers of 256 and 128 neurons cluster membership, which were also used in the MDD or SCZ prediction. Last, the parameters for the prediction model in analysis C were a batch size of 10, a learning rate of 0.0001, and two layers of 256 and 128.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abi7293

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. J.-P. Lépine, M. Briley, The increasing burden of depression. *Neuropsychiatr. Dis. Treat.* **7**, 3–7 (2011).

2. R. Mangalore, M. Knapp, Cost of schizophrenia in England. *J. Ment. Health Policy Econ.* **10**, 23–41 (2007).
3. S. F. Austin, O. Mors, E. Budtz-Jørgensen, R. G. Secher, C. R. Hjorthøj, M. Bertelsen, P. Jeppesen, L. Petersen, A. Thorup, M. Nordentoft, Long-term trajectories of positive and negative symptoms in first episode psychosis: A 10-year follow-up study in the OPUS cohort. *Schizophr. Res.* **168**, 84–91 (2015).
4. P. Lichtenstein, B. H. Yip, C. Björk, Y. Pawitan, T. D. Cannon, P. F. Sullivan, C. M. Hultman, Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: A population-based study. *Lancet* **373**, 234–239 (2009).
5. M. E. Benros, P. B. Mortensen, W. W. Eaton, Autoimmune diseases and infections as risk factors for schizophrenia. *Ann. N. Y. Acad. Sci.* **1262**, 56–66 (2012).
6. G. Koppe, A. Meyer-Lindenberg, D. Durstewitz, Deep learning for small and big data in psychiatry. *Neuropsychopharmacology* **46**, 176–190 (2021).
7. M. Bracher-Smith, K. Crawford, V. Escott-Price, Machine learning for genetic prediction of psychiatric disorders: A systematic review. *Mol. Psychiatry* **26**, 70–79 (2021).
8. E. Helmes, J. Landmark, Subtypes of schizophrenia: A cluster analytic approach. *Can. J. Psychiatry* **48**, 702–708 (2003).
9. C. F. Hybels, D. G. Blazer, C. F. Pieper, L. R. Landerman, D. C. Steffens, Profiles of depressive symptoms in older adults diagnosed with major depression: Latent cluster analysis. *Am. J. Geriatr. Psychiatry* **17**, 387–396 (2009).
10. A. Talpalaru, N. Bhagwat, G. A. Devenyi, M. Lepage, M. M. Chakravarty, Identifying schizophrenia subgroups using clustering and supervised learning. *Schizophr. Res.* **214**, 51–59 (2019).
11. L. Squarcina, F. M. Villa, M. Nobile, E. Grisan, P. Brambilla, Deep learning for the prediction of treatment response in depression. *J. Affect. Disord.* **281**, 618–622 (2021).
12. D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes. arXiv: 1312.6114v10 [stat.ML] (20 December 2013).
13. G. P. Way, C. S. Greene, Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput.* **23**, 80–91 (2018).
14. C. H. Grønbech, M. F. Vording, P. N. Timshel, C. K. Sønderby, T. H. Pers, O. Winther, scVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics* **36**, 4415–4422 (2020).
15. L. Rampášek, D. Hidru, P. Smirnov, B. Haibe-Kains, A. Goldenberg, Dr.VAE: Improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics* **35**, 3743–3751 (2019).
16. J. N. Nissen, J. Johansen, R. L. Allesøe, C. K. Sønderby, J. J. A. Armenteros, C. H. Grønbech, L. J. Jensen, H. B. Nielsen, T. N. Petersen, O. Winther, S. Rasmussen, Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
17. L.-L. Zeng, H. Wang, P. Hu, B. Yang, W. Pu, H. Shen, X. Chen, Z. Liu, H. Yin, Q. Tan, K. Wang, D. Hu, Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *EBioMedicine* **30**, 74–85 (2018).
18. J. Kim, J. Lee, E. Park, J. Han, A deep learning model for detecting mental illness from user content on social media. *Sci. Rep.* **10**, 11846 (2020).
19. S. Gao, V. D. Calhoun, J. Sui, Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neurosci. Ther.* **24**, 1037–1052 (2018).
20. L. A. Maglanoc, N. I. Landrø, R. Jonassen, T. Kaufmann, A. Córdova-Palomera, E. Hilland, L. T. Westlye, Data-driven clustering reveals a link between symptoms and functional brain connectivity in depression. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging.* **4**, 16–26 (2019).
21. W. Mumtaz, A. Qayyum, A deep learning framework for automatic diagnosis of unipolar depression. *Int. J. Med. Inform.* **132**, 103983 (2019).
22. K. P. Glanville, J. R. I. Coleman, K. B. Hanscombe, J. Euesden, S. W. Choi, K. L. Purves, G. Breen, T. M. Air, T. F. M. Andlauer, B. T. Baune, E. B. Binder, D. H. R. Blackwood, D. I. Boomsma, H. N. Buttenschøn, L. Colodro-Conde, U. Dannlowski, N. Direk, E. C. Dunn, A. J. Forstner, E. J. C. de Geus, H. J. Grabe, S. P. Hamilton, I. Jones, L. A. Jones, J. A. Knowles, Z. Kutalik, D. F. Levinson, G. Lewis, P. A. Lind, S. Lucae, P. K. Magnusson, P. McGuffin, A. M. McIntosh, Y. Milaneschi, O. Mors, S. Mostafavi, B. Müller-Myhsok, N. L. Pedersen, B. W. J. H. Penninx, J. B. Potash, M. Preisig, S. Ripke, J. Shi, S. I. Shyn, J. W. Smoller, F. Streit, P. F. Sullivan, H. Tiemeier, R. Uher, S. Van der Auwera, M. M. Weissman; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, P. F. O'Reilly, C. M. Lewis, Classical human leukocyte antigen alleles and C4 haplotypes are not significantly associated with depression. *Biol. Psychiatry* **87**, 419–430 (2020).
23. R. Nudel, M. E. Benros, M. D. Krebs, R. L. Allesøe, C. K. Lemvigh, J. Bybjerg-Grauholm, A. D. Børglum, M. J. Daly, M. Nordentoft, O. Mors, D. M. Hougaard, P. B. Mortensen, A. Buil, T. Werge, S. Rasmussen, W. K. Thompson, Immunity and mental illness: Findings from a Danish population-based immunogenetic study of seven psychiatric and neurodevelopmental disorders. *Eur. J. Hum. Genet.* **27**, 1445–1455 (2019).
24. A. Dahl, N. Zaitlen, Genetic influences on disease subtypes. *Annu. Rev. Genomics Hum. Genet.* **21**, 413–435 (2020).
25. N. R. Wray, M. E. Goddard, P. M. Visscher, Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).
26. A. L. Collins, P. F. Sullivan, Genome-wide association studies in psychiatry: What have we learned? *Br. J. Psychiatry* **202**, 1–4 (2013).
27. C. B. Pedersen, J. Bybjerg-Grauholm, M. G. Pedersen, J. Grove, E. Agerbo, M. Bækvad-Hansen, J. B. Poulsen, C. S. Hansen, J. J. McGrath, T. D. Als, J. I. Goldstein, B. M. Neale, M. J. Daly, D. M. Hougaard, O. Mors, M. Nordentoft, A. D. Børglum, T. Werge, P. B. Mortensen, The iPSYCH2012 case-cohort sample: New directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).
28. B. Nørgaard-Pedersen, D. M. Hougaard, Storage policies and use of the Danish Newborn Screening Biobank. *J. Inherit. Metab. Dis.* **30**, 530–536 (2007).
29. A. J. Schork, H. Won, V. Appadurai, R. Nudel, M. Gandal, O. Delaneau, M. R. Christiansen, D. M. Hougaard, M. Bækved-Hansen, J. Bybjerg-Grauholm, M. G. Pedersen, E. Agerbo, C. B. Pedersen, B. M. Neale, M. J. Daly, N. R. Wray, M. Nordentoft, O. Mors, A. D. Børglum, P. B. Mortensen, A. Buil, W. K. Thompson, D. H. Geschwind, T. Werge, A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nat. Neurosci.* **22**, 353–361 (2019).
30. O. Mors, G. P. Perto, P. B. Mortensen, The Danish psychiatric central research register. *Scand. J. Public Health* **39**, 54–57 (2011).
31. A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, H. Parkinson, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
32. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLOS Genet.* **2**, e190 (2006).
33. International HapMap Consortium, The International HapMap Project. *Nature* **426**, 789–796 (2003).
34. S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv: 1502.03167 [cs.LG] (1 February 2015).
35. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580 [cs.NE] (3 July 2012).
36. A. L. Maas, A. Y. Hannun, A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models (2013); available at https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf.
37. D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization. arXiv: 1412.6980 [cs.LG] (22 December 2014).
38. C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, O. Winther, Ladder Variational Autoencoders. arXiv: 1602.02282 [stat.ML] (6 February 2016).
39. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch (2017); available at https://openreview.net/pdf?id=BJJsrmfCZ.
40. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
41. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv: 1802.03426 [stat.ML] (9 February 2018).
42. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1. 0 Contributors, SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python. arXiv: 1907.10121 [cs.MS] (23 July 2019).
43. S. W. Choi, T. S.-H. Mak, P. F. O'Reilly, Tutorial: A guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
44. Schizophrenia Working Group of the Psychiatric Genomics Consortium, Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
45. N. R. Wray, S. Ripke, M. Mattheisen, M. Trzaskowski, E. M. Byrne, A. Abdellaoui, M. J. Adams, E. Agerbo, T. M. Air, T. M. F. Andlauer, S.-A. Bacanu, M. Bækvad-Hansen, A. F. T. Beekman, T. B. Bigdeli, E. B. Binder, D. R. H. Blackwood, J. Bryois, H. N. Buttenschøn, J. Bybjerg-Grauholm, N. Cai, E. Castelao, J. H. Christensen, T.-K. Clarke, J. I. R. Coleman, L. Colodro-Conde, B. Couvy-Duchesne, N. Craddock, G. E. Crawford, C. A. Crowley, H. S. Dashti, G. Davies, I. J. Deary, F. Degenhardt, E. M. Derks, N. Direk, C. V. Dolan, E. C. Dunn, T. C. Eley, N. Eriksson, V. Escott-Price, F. H. F. Kiadeh, H. K. Finucane, A. J. Forstner, J. Frank, H. A. Gaspar, M. Gill, P. Giusti-Rodríguez, F. S. Goes, S. D. Gordon, J. Grove, L. S. Hall, E. Hannon, C. S. Hansen, T. F. Hansen, S. Herms, I. B. Hickie, P. Hoffmann, G. Homuth, C. Horn, J.-J. Hottenga, D. M. Hougaard, M. Hu, C. L. Hyde, M. Ising, R. Jansen, F. Jin, E. Jorgenson, J. A. Knowles, I. S. Kohane, J. Kraft, W. W. Kretzschmar, J. Krogh, Z. Kutalik, J. M. Lane, Y. Li, Y. Li, P. A. Lind, X. Liu, L. Lu,

D. J. M. Intyre, D. F. M. Kinnon, R. M. Maier, W. Maier, J. Marchini, H. Mbarek, P. M. Grath, P. M. Guffin, S. E. Medland, D. Mehta, C. M. Middeldorp, E. Mihailov, Y. Milaneschi, L. Milani, J. Mill, F. M. Mondimore, G. W. Montgomery, S. Mostafavi, N. Mullins, M. Nauck, B. Ng, M. G. Nivard, D. R. Nyholt, P. F. O'Reilly, H. Oskarsson, M. J. Owen, J. N. Painter, C. B. Pedersen, M. G. Pedersen, R. E. Peterson, E. Pettersson, W. J. Peyrot, G. Pistis, D. Posthuma, S. M. Purcell, J. A. Quiroz, P. Qvist, J. P. Rice, B. P. Riley, M. Rivera, S. S. Mirza, R. Saxena, R. Schoevers, E. C. Schulte, L. Shen, J. Shi, S. I. Shyn, E. Sigurdsson, G. B. C. Sinnamon, J. H. Smit, D. J. Smith, H. Stefansson, S. Steinberg, C. A. Stockmeier, F. Streit, J. Strohmaier, K. E. Tansey, H. Teismann, A. Teumer, W. Thompson, P. A. Thomson, T. E. Thorgeirsson, C. Tian, M. Traylor, J. Treutlein, V. Trubetskoy, A. G. Uitterlinden, D. Umbricht, S. Van der Auwera, A. M. van Hemert, A. Viktorin, P. M. Visscher, Y. Wang, B. T. Webb, S. M. Weinsheimer, J. Wellmann, G. Willemsen, S. H. Witt, Y. Wu, H. S. Xi, J. Yang, F. Zhang; eQTLGen; 23andMe, V. Arolt, B. T. Baune, K. Berger, D. I. Boomsma, S. Cichon, U. Dannlowski, E. C. J. de Geus, J. R. De Paulo, E. Domenici, K. Domschke, T. Esko, H. J. Grabe, S. P. Hamilton, C. Hayward, A. C. Heath, D. A. Hinds, K. S. Kendler, S. Kloiber, G. Lewis, Q. S. Li, S. Lucae, P. F. A. Madden, P. K. Magnusson, N. G. Martin, A. M. McIntosh, A. Metspalu, O. Mors, P. B. Mortensen, B. Müller-Myhsok, M. Nordentoft, M. M. Nöthen, M. C. O'Donovan, S. A. Paciga, N. L. Pedersen, B. W. J. H. Penninx, R. H. Perlis, D. J. Porteous, J. B. Potash, M. Preisig, M. Rietschel, C. Schaefer, T. G. Schulze, J. W. Smoller, K. Stefansson, H. Tiemeier, R. Uher, H. Völzke, M. M. Weissman, T. Werge, A. R. Winslow, C. M. Lewis, D. F. Levinson, G. Breen, A. D. Børglum, P. F. Sullivan; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).

Acknowledgments: 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 High Performance Computing (HPC) Facility (http://computerome.dtu.dk/). A.D.B., D.M.H., O. Mors, T.W., P. B. Mortensen, and M.N. are the principal investigators of iPSYCH. **Funding:** This specific study was funded by an independent research grant by the Lundbeck Foundation, Denmark (grant numbers R278-2018-1411 and R268-2016-3925). S.R. was supported by a grant from the Novo Nordisk Foundation (NNF14CC0001). The genotyping of the iPSYCH samples was supported by grants from The Lundbeck Foundation, Denmark (grants R102-A9118 and R155-2014-1724); the Stanley Foundation; the Simons Foundation (SFARI 311789); and NIMH (5U01MH094432-02). **Author contributions:** R.L.A. wrote the code, carried out the analyses, interpreted the results, and wrote the manuscript. S.R. and M.E.B. designed and supervised the analyses, interpreted the results, and wrote the manuscript. R.N. and W.K.T. interpreted the results and revised the manuscript. Y.W. provided summary statistics for PRS calculations. A.D.B., D.M.H., T.W., and M.N. are the principal investigators of iPSYCH and conceptualized the consortium, and contributed to the acquisition and processing of the Danish registry data and/or to the generation of the genetic data. All authors have read and approved the manuscript. **Competing interests:** T.W. states that he has acted as a lecturer and scientific counselor to H. Lundbeck A/S. All other authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. iPSYCH data are 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 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. All code is available at GitHub at https://github.com/RasmussenLab/PSYCH-VAE/ and is freely available under the permissive MIT License and at Zenodo (DOI 10.5281/zenodo.5910302). All analyses were run using PSYCH-VAE v1.0.

Submitted 26 April 2021
Accepted 11 May 2022
Published 29 June 2022
10.1126/sciadv.abi7293