ARTICLE     OPEN

# Development and multi-site external validation of a generalizable risk prediction model for bipolar disorder

Colin G. Walsh [1✉], Michael A. Ripperger [1], Yirui Hu[2], Yi-han Sheu [3,4,5], Hyunjoon Lee[1], Drew Wilimitis[1], Amanda B. Zheutlin[3], Daniel Rocha[2], Karmel W. Choi[3], Victor M. Castro [3], H. Lester Kirchner[2], Christopher F. Chabris[2], Lea K. Davis[1] and Jordan W. Smoller[3,4,5]

Bipolar disorder is a leading contributor to disability, premature mortality, and suicide. Early identification of risk for bipolar disorder using generalizable predictive models trained on diverse cohorts around the United States could improve targeted assessment of high risk individuals, reduce misdiagnosis, and improve the allocation of limited mental health resources. This observational case-control study intended to develop and validate generalizable predictive models of bipolar disorder as part of the multisite, multinational PsycheMERGE Network across diverse and large biobanks with linked electronic health records (EHRs) from three academic medical centers: in the Northeast (Massachusetts General Brigham), the Mid-Atlantic (Geisinger) and the Mid-South (Vanderbilt University Medical Center). Predictive models were developed and valid with multiple algorithms at each study site: random forests, gradient boosting machines, penalized regression, including stacked ensemble learning algorithms combining them. Predictors were limited to widely available EHR-based features agnostic to a common data model including demographics, diagnostic codes, and medications. The main study outcome was bipolar disorder diagnosis as defined by the International Cohort Collection for Bipolar Disorder, 2015. In total, the study included records for 3,529,569 patients including 12,533 cases (0.3%) of bipolar disorder. After internal and external validation, algorithms demonstrated optimal performance in their respective development sites. The stacked ensemble achieved the best combination of overall discrimination (AUC = 0.82–0.87) and calibration performance with positive predictive values above 5% in the highest risk quantiles at all three study sites. In conclusion, generalizable predictive models of risk for bipolar disorder can be feasibly developed across diverse sites to enable precision medicine. Comparison of a range of machine learning methods indicated that an ensemble approach provides the best performance overall but required local retraining. These models will be disseminated via the PsycheMERGE Network website.

## INTRODUCTION

Bipolar disorder (BD), characterized by episodes of hypomania/mania and depression [1], is a leading cause of disability [2]. Rates of suicide among patients with BD are 20- to 30-fold higher than in the general population [3], and BD is associated with substantial premature mortality from multiple causes [4]. The diagnosis of BD can be challenging and may require a prolonged diagnostic odyssey, averaging 6–10 years [5–7]. Affected patients frequently present initially with a major depressive episode and are misdiagnosed with unipolar depression. Misdiagnosis may lead to inappropriate prescribing of antidepressants without mood stabilization and increased risk of switching into a manic state [8]. Duration of untreated bipolar disorder is associated with more severe and recurrent mood episodes and more frequent suicide attempts [9, 10].

Identifying those at risk for BD might enable targeted assessment, early intervention, and more appropriate management. A recent systematic review of clinical trials to prevent bipolar disorder showed reliance on family history for risk identification [11]. However, given BD's multifactorial nature, most affected would not have a positive family history [12]. In addition and unlike schizophrenia, no established prodrome exists for bipolar disorder. Newer methods for risk identification not reliant on existing clinical signs or symptoms might be of substantial value.

Longitudinal electronic health records (EHRs) coupled with predictive analytics might enable novel risk identification opportunities in BD. We have previously demonstrated that such data can produce valid diagnostic phenotyping of bipolar cases [13, 14]. Here, we extend this work to the domain of prognostication by leveraging the PsycheMERGE Network, a national research network of EHR-linked biobanks. Using longitudinal EHRs from three major healthcare systems, we trained and validated quantitative bipolar disorder risk prediction models based on high-dimensional structured EHR data and evaluated their performance individually and when ensembled.

[1]Vanderbilt University Medical Center Health System, Nashville, TN, USA. [2]Geisinger Health System, Danville, PA, USA. [3]Massachusetts General-Brigham Health System, Boston, MA, USA. [4]Center for Precision Psychiatry, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [5]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. ✉email: Colin.walsh@vumc.org

**Table 1.** Baseline study participant characteristics.

| Site | VUMC (%) | GHS (%) | MGB (%) |
|---|---|---|---|
| Total patients in study, N | 932,784 | 934,749 | 1,662,036 |
| Total patients, bipolar coded-broad, N (%) | 3,357 (0.36%) | 3,101 (0.33%) | 6,075(0.36%) |
| Internal validation training set, N (%) | 746,226 (80.0%) | 701,147 (75.0%) | 997,687 (60.0%) |
| Internal validation testing set, N (%) | 186,558 (20.0%) | 232,942 (25.0%) | 664,349 (40.0%) |
| Sex, women | 535,273 (57.4%) | 524,479 (56%) | 980,586 (59%) |
| Sex, men | 392,477 (42.6%) | 410,270 (44%) | 681,372 (41%) |
| Sex, unknown | 34 (0.0036%) | 0 (0%) | 78 (0.0047%) |
| Coded Race, White | 701,525 (75.2%) | 881,433 (94.3%) | 1,282,679 (77%) |
| Coded Race, Black | 100,462(10.8%) | 33,025 (3.6%) | 106,362 (6.4%) |
| Coded Race, Asian | 16,055 (1.72%) | 9,129 (1.0%) | 69,651 (4.1%) |
| Coded Race, Other | 32,063 (3.43%) | 5,786 (0.6%) | 203,315 (12%) |
| Coded Race, Unknown | 82,679 (8.86%) | 5,490 (0.6%) | 29 (0%) |
| Age, mean (Range) | 49.5 (18–89) | 52.2 (18–89) | 52.82 (18–89) |
| EHR length in years, median (Q1-Q3) | 7 (3–13) | 8 (3–15) | 9 (4–15) |

## METHODS

### Study settings

Participating study sites included three major academic medical centers in the United States: the Northeast (Mass General Brigham [MGB]), the Mid-South (Vanderbilt University Medical Center [VUMC]), the Mid-Atlantic (Geisinger [GHS]). Each site participates in the PsycheMERGE Network and has an extensive EHR repository linked to a biobank. On average, these sites each serve 1.4 M patients per year and have EHR repositories of ~2.7 M patients linked via EHRs.

The methods were performed in accordance with relevant guidelines and regulations and approved by Institutional Review Boards at each participating study site: Vanderbilt University Medical Center, Geisinger Health System, Mass General Brigham.

### Outcome definition

Cases of BD were defined by the published "Bipolar Coded-Broad" definition per Castro et al. [14]. This rule-based algorithm demonstrated high positive predictive value (0.80) using a gold-standard of semi-structured diagnostic interviews (SCID-IV) by experienced doctoral-level clinicians blind to algorithmic results. To meet "Bipolar Coded-Broad," cases must have at least two BD diagnostic codes versions nine or ten of the International Classification of Diseases (ICD) schema with a minimum of four weeks between each code and at least two documented medications used to treat BD (e.g., lithium or valproic acid) within one year of the index BD diagnosis. To rule out patients with related disorders, we required the number of diagnostic codes for major depressive disorder (MDD), schizophrenia (SCZ), schizoaffective disorder, or organic affective syndrome (OAS) to total fewer than half the number of BD codes. Only adult patients aged 18 years and older at the time of their index BD diagnosis were included in this analysis.

All adult patients were included if they had a minimum of three documented healthcare encounters over a minimum of six months, regardless of case status.

### Predictive modeling approach

We compared three predictive modeling approaches that together span a range of model architectures and strategies for handling feature relationships (see "Algorithmic Details", below). Because of prior algorithmic experience at each study site, each team validated internally one of three types of models: L2-penalized regression (abbreviated here as "Ridge") at MGB; random forests (RF) at VUMC; gradient boosting machines (GBM) at GHS. In external validation, the remaining two of three models were tested across sites, e.g., MGB validated externally RF from VUMC and GBM from GHS.

Internal validation was conducted at each site with a randomly selected hold-out test and the best internally performing algorithms were shared for external validation. This reciprocal validation strategy tested generalizability of each algorithm without the need for each site to train three separate algorithms in parallel.

### Feature engineering

Variables for prediction included demographics: age (continuous), coded sex (categorical: Male, Female, and Unknown), coded race (categorical: White, Black, Asian, Other and Unknown); inpatient-administered and outpatient-prescribed medications (log-transformed counts); and diagnostic codes (log-transformed counts). For log-transformed counts, we identified each diagnostic code and medication. We grouped medications by RxNorm ingredient and each diagnostic code by Clinical Classification Software, Level 2, codes. We then counted occurrences of each group per person in the time before the prediction. The "log-transform" means the counts all were incremented by 1 (to avoid log of zero) and then the natural log of that count +1 value was calculated for all count-based predictors. This common transform has been shown to be helpful in sparse and skewed predictors, i.e., many zero values with few very high counts. Such predictors are common in EHR-based datasets.

Dimensionality reduction included grouping medications by their RxNorm ingredients [15] and diagnostic codes mapped from ICD-9-CM and ICD-10-CM to Clinical Classification Software (CCS) Level 2 codes [16]. The final feature list numbered up to ~2500. Missing data for count variables were imputed as zeroes and categorized as unknown for race and sex (see Table 1).

Records meeting "Bipolar Coded Broad" were right-censored until the day before index diagnosis to represent a useful prediction target and to prevent leakage of bipolar-related data from driving model predictions. For those not meeting bipolar disorder criteria, right-censoring occurred at the last date of a visit or first occurrence of an ICD code for BD in the EHR.

### Algorithmic details

*Ridge regression.* Ridge Regression [17] is a regularized regression model that imposes shrinkage of regression coefficients to reduce multi-collinearity and model variance, and thereby increasing prediction performance. Despite the shrinkage, the regression coefficients are never shrunk to zero, and therefore all features remain in the final model. We used the widely adopted *glmnet* [18–20] package in R for model training, using the main (first order) effects of all available features. The model was developed with 10-fold cross-validation using 60% of all data to find the best Lambda value (i.e., strength of regularization) and estimate model parameters, while the remaining 40% data were used as a hold-out test set. The 60–40 split was chosen due to a larger sample size at MGB, and the 60% training/validation split approached the limits in input data size for the *glmnet* package. Preliminary analyses showed minimal performance differences by varying the train/test ratio.

*RF.* VUMC implemented the decision-tree based RF. A commonly employed nonparametric algorithm, RF permits nonlinear relationships between predictors, samples with replacement for feature inclusion and model training, and it tolerates collinearity likely to be present in EHR data. After preliminary analyses varying the following, parameters of 200 trees, minimum node size of five, and purity for importance were used. RF were trained with an 80-20% train-test split. RF were implemented using the ranger package in R [21].

**Table 2.** Model performance by site.

|  | AUROC | AUPR | Brier score | Calibration slope | Calibration intercept |
|---|---|---|---|---|---|
| **VUMC** |  |  |  |  |  |
| RIDGE | 0.796 | 0.020 | 0.004 | 2.108 | 6.085 |
| RF | *0.836* | *0.046* | *0.004* | *0.946* | *−0.303* |
| GBM | 0.808 | 0.025 | 0.004 | 1.011 | 0.092 |
| ENSEMBLE | 0.837 | 0.049 | 0.004 | 0.996 | −0.021 |
| **GHS** |  |  |  |  |  |
| RIDGE | 0.775 | 0.015 | 0.003 | 1.987 | 5.310 |
| RF | 0.775 | 0.015 | 0.003 | 0.625 | −2.108 |
| GBM | *0.873* | *0.054* | *0.003* | *1.085* | *0.387* |
| ENSEMBLE | 0.825 | 0.032 | 0.003 | 1.027 | 0.224 |
| **MGB** |  |  |  |  |  |
| RIDGE | *0.865* | *0.026* | *0.004* | *1.502* | *2.674* |
| RF | 0.802 | 0.019 | 0.004 | 0.726 | −1.626 |
| GBM | 0.852 | 0.033 | 0.004 | 1.000 | −0.002 |
| ENSEMBLE | 0.822 | 0.026 | 0.004 | 0.981 | −0.111 |

Internal validation italicized.
*AUROC* Area Under the Receiver Operating Characteristic, *AUPR* Area Under the Precision-Recall Curve.

*GBM.* GHS implemented GBM in for internal validation. Boosting is an ensemble technique based on using multiple "weak learner" algorithms to train a strong one through sequential training to iteratively improve prediction accuracy. GBM is a high-performance gradient boosting framework based on decision trees capable of handling imbalanced datasets, as the boosting can strengthen the impact of the positive class (here, cases of BD). Tuning parameters included the ratio of features used, the ratio of training instances, maximum depth of trees and the learning rate. In preliminary analyses, dimensionality impacted model performance, so we selected the most prevalent medications across the GHS EHR by including those accounting for 95% of cumulative medication counts by Pareto analysis. GBM were trained with a 75-25% train-test split. Here, GBM were implemented in Python (package 'lightgbm') [22].

### Ensembling
Ensembling is designed to improve prediction accuracy by aggregating the strengths of diverse machine learning models into a single predictive model. Here, we ensembled the three algorithms via stacking of all three algorithms at each site and evaluated performance. We combined Ridge, GBM, and RFs with a stacked ensemble trained with ten-fold cross validation and logistic regression using the three individual predictions as multivariate predictors on each site's training set to avoid leakage of test data for the internally valid model into the ensemble.

### Model evaluation
Evaluation metrics included Area Under the Receiver Operating Characteristic (AUROC), Precision-Recall Curves and Area Under the P-R Curve (AUPR). Metrics at specific risk thresholds included sensitivity/recall, specificity, positive predictive value (PPV), and number needed to screen (NNS [23], the reciprocal of PPV for a predictive model) Calibration was measured with Brier score and calibration slope/intercept [24].

### Feature importance by site
Because each algorithm, GBM/RF/Ridge, have different inherent metrics of importance, we applied L1-penalized regression (LASSO) to the training datasets at each site. L1-penalized regression importance was defined as the magnitude of regression coefficients by predictor. The top ten L1-penalized regression selected features for each site are shown here, ranked (Table 4).

### RESULTS
Study site data are shown (Table 1).

### Individual model performance by site
Discrimination performance is shown for each algorithm by training site with internal validation (testing within site) denoted visually for ready comparison to external validation (testing across sites) (Table 2).

As shown in Table 2, models performed comparably within and across sites with a tendency for better discrimination at internal validation sites for locally trained algorithms and better calibration for GBM and Ensembles of GBM, RF and Ridge.

### Optimal thresholds and performance metrics by algorithm and site
Varying risk percentile thresholds by algorithm and by site showed specificity was closely linked to the thresholds themselves, while sensitivity (recall) and PPV tended to decrease and increase, respectively, as thresholds increased (Table 3). NPV for all algorithms above these thresholds (90% and above) was over 99%, largely because of the rarity of the outcomes in question.

### Predictor importance
The most important predictors to each model are listed in Table 4 (top ten by site) and in the eSupplement (top fifty by site). Features were driven by medications and contrast agents, e.g., those used in imaging studies, at the three sites. We caution overinterpretation of such predictor weights and underline these statistics are correlative, not causal.

### DISCUSSION
Early identification of individuals at risk for BD offers opportunities for targeted assessment and prevention. Although a number of risk factors for BD have been established including family history [25] and stressful life events [26], quantitative, scalable prediction of risk is challenging. Prior studies have largely focused on individuals with a history of depression and/or have included relatively small samples [27, 28]. Here, we validated multiple algorithmic approaches across multiple well-powered longitudinal EHR sites in the absence of a common data model to generate a novel suite of prediction algorithms for BD. These models performed well across diverse geography and broad,

**Table 3.** Model performance by risk percentile threshold.

| | Risk percentile cutoff | Specificity (%) | Sensitivity (%) | PPV (%) | NPV (%) |
|---|---|---|---|---|---|
| **Random forest** | | | | | |
| VUMC | *90* | *90.2* | *58.6* | *2.1* | *99.8* |
| | *95* | *95.2* | *45.5* | *3.3* | *99.8* |
| | *99* | *99.1* | *19.2* | *6.9* | *99.7* |
| MGB | 90 | 90.2 | 51.7 | 1.9 | 99.8 |
| | 95 | 95.1 | 34.9 | 2.5 | 99.7 |
| | 99 | 99 | 9.1 | 3.3 | 99.7 |
| GHS | 90 | 90.1 | 44.8 | 1.5 | 99.8 |
| | 95 | 95.1 | 30.1 | 2 | 99.8 |
| | 99 | 99 | 7.6 | 2.5 | 99.7 |
| **GBM** | | | | | |
| VUMC | 90 | 90.2 | 52.7 | 1.9 | 99.8 |
| | 95 | 95.1 | 40.5 | 2.9 | 99.8 |
| | 99 | 99 | 13.1 | 4.7 | 99.7 |
| MGB | *90* | *90.2* | *63.1* | *2.3* | *99.9* |
| | *95* | *95.1* | *45.3* | *3.3* | *99.8* |
| | *99* | *99.1* | *15.1* | *5.5* | *99.7* |
| GHS | 90 | 90.2 | 64.2 | 2.1 | 99.9 |
| | 95 | 95.1 | 48.8 | 3.2 | 99.8 |
| | 99 | 99.1 | 20.4 | 6.8 | 99.7 |
| **Ridge** | | | | | |
| VUMC | 90 | 90.1 | 50 | 1.8 | 99.8 |
| | 95 | 95.1 | 34.8 | 2.5 | 99.8 |
| | 99 | 99 | 9.2 | 3.3 | 99.7 |
| MGB | 90 | 90.2 | 63.6 | 2.3 | 99.9 |
| | 95 | 95.1 | 43.9 | 3.2 | 99.8 |
| | 99 | 99 | 10.4 | 3.8 | 99.7 |
| GHS | *90* | *90.1* | *48.8* | *1.6* | *99.8* |
| | *95* | *95.1* | *35.1* | *2.3* | *99.8* |
| | *99* | *99* | *10.5* | *3.5* | *99.7* |
| **Ensemble** | | | | | |
| VUMC | 90 | 90.2 | 58.2 | 2.1 | 99.8 |
| | 95 | 95.2 | 46.9 | 3.4 | 99.8 |
| | 99 | 99.1 | 18.9 | 6.8 | 99.7 |
| MGB | 90 | 90.2 | 56.4 | 2.1 | 99.8 |
| | 95 | 95.1 | 39.7 | 2.9 | 99.8 |
| | 99 | 99 | 12.4 | 4.5 | 99.7 |
| GHS | 90 | 90.1 | 52.5 | 1.8 | 99.8 |
| | 95 | 95.1 | 36.3 | 2.5 | 99.8 |
| | 99 | 99 | 14 | 4.9 | 99.7 |

Internal validation italicized.

heterogeneous patient populations. However, difficulty in portability and transferring algorithms across sites remains a primary barrier to replicative and implementation studies.

Our results demonstrate the feasibility and comparative performance of prediction algorithms using federated analyses of EHR data across the PsycheMERGE network. We compared three different machine learning approaches, each reliant on different assumptions and means of handling noisy, high dimensional data. Finally, we tested ensembles of these methods via stacking.

We highlight several noteworthy findings. First, we found that, regardless of method, performance was optimal at the site at which the model was developed, supporting the inference that portability of models may be limited by site-specific features - e.g., a local care practice common in one setting or region and uncommon in another. It also suggests potential for overly optimistic performance estimates with internal validation - underlining again that no substitute exists for external validation in model evaluation. We also note little overlap to the most important predictive features at each site, which likely relates to both site-specific differences and

**Table 4.** Ten most important predictors by algorithm using L1-penalized regression for feature selection and weighting.

| MGB | GHS | VUMC |
|---|---|---|
| Clemastine/phenylpropanolamine | Varicella zoster virus glycoprotein E | Glycolate |
| Carboxymethylcellulose/glycerin | Mood Disorders | Carbetapentane |
| Gentian violet | Ioversol | Hydrogen Peroxide |
| Erythromycin/sulfisoxazole | Niacinamide | Telithromycin |
| Flaxseed extract | Dobutamine | Colloidal oatmeal |
| Chlorpheniramine/ibuprofen/pseudoephedrine | Iopamidol | Tipranavir |
| Droperidol | Gadobutrol | Avibactam |
| Amyl nitrite | Race, Unknown | Ceftibuten |
| Manganese | Zonisamide | Ticarcillin |
| Cisapride | Witch Hazel | Clove oil |

Importance refers to impact on model performance and confers insight into correlation, not causation.

algorithmic differences, e.g., parametric (Ridge) and nonparametric (tree-based) approaches. The most generalizable algorithm, the stacked ensemble, matched internally valid algorithms in discrimination and was the only well-calibrated model, in part owing to its recalibration via regression at each site during the stacking process. Vigilance for drift and miscalibration over time would be necessary in planning implementation downstream.

Race was included as a predictor in these models, a feature being reconsidered for a number of clinical predictive applications. We opted not to blind our algorithms in this case as this process has been shown not to prevent algorithmic bias and might, in fact, introduce it. We emphasize race is a social construct that does not itself cause mental illness but can be a marker of inequitable healthcare access, experiences of adversity, and systemic inequity of opportunity. As such, it might be predictive of a coded diagnosis despite not being in the causal path for the outcome. Prior to implementation of models like these, close attention to algorithmic bias and potential for disparities should be considered using variables like coded race for prediction [29, 30]. As an additional check that race as a predictor did not bias our model unfairly, we retrained the RF at VUMC and compared validation set performance by coded race across 1,000 bootstrap replicates. Performance distributions did not differ across or within race: coded White race AUC 0.8 [0.79, 0.82] and 0.79 [0.77, 0.8] for model with and without race, respectively; coded Black race AUC 0.79 [0.76, 0.83] and 0.78 [0.75, 0.82] for model with and without race, respectively.

The clinical utility of predictive models for rare events like BD (<1% at each site) merits consideration and attention to the importance of PPV. Here, the stacked ensembles achieved the best threshold-specific PPVs across all three sites (a ~forty-fold increase compared to case prevalence). A resource-limited clinical environment prioritizing identifying those most likely to have undiagnosed BD or predicting onset of BD might benefit from models that provide such PPVs in the setting of such rare outcomes. Of note, our models achieved a NNS as low as 20 or fewer at each site, meaning that fewer than 20 patients would be identified as high risk for every true case detected.

These models rely on EHR data common to any modern hospital agnostic to a common data model: demographics, diagnostic codes in a universally accepted schema (ICD), medications mapped to a public ontology (RXNORM). For those who wish to leverage these trained and tested algorithms, a library of the individual models and the stacked ensemble will be made available on the PsycheMERGE Network website (psychemerge.com) [31].

## Strengths
This study leveraged three large health systems with a validated definition of bipolar disorder as the prediction target. We applied three accepted algorithms (RF, Ridge Regression, GBM) to large real-world cohorts and assessed generalizability and model fit across partner sites. We ensembled these algorithms on over 3 million patient lives across three major biobanks - the largest modeling study of this kind in BD, to our knowledge. We relied on readily available structured EHR data for feature engineering. Finally, we disseminate these tools via the PsycheMERGE Network to facilitate replication studies and local deployment.

## Limitations
Our results should also be interpreted in light of several limitations. First, while we explored performance of multiple different modeling approaches, there are others (including deep learning approaches) that were not tested. Second, our study relied on structured longitudinal EHR data, a decision we made to facilitate ready implementation across sites. However, natural language processing of narrative text might offer performance advantages longer term. Third, covariate shift in real-world data like these mean the joint distribution of model inputs and outputs may differ between training and testing across sites [32]. Methods of covariate shift detection and adaption might be investigated using importance re-weighting or feature dropping methods in future studies to improve model performance. Finally, class imbalance remains a notable challenge in this study and studies like it, and creates the potential for overfitting and spuriously high model performance metrics (e.g., high AUROCs simply because of identification of the majority class, here non-BD).

## CONCLUSIONS
Generalizable predictive models of bipolar disorder trained and validated across health systems are feasible targets of clinical and precision medicine focused initiatives, even in the absence of common data models across sites. Implications of these models include BD risk research acceleration, catalyzing pharmacoepidemiologic studies, and potential for similar models to serve as probabilistic phenotypes in precision medicine research. Future work should assess their clinical utility and potential to phenotype quantitatively this serious mental illness.

## DATA AVAILABILITY
Study data including de-identified electronic health records linked to biobanks. However, complete anonymization to prevent inadvertent or intentional reidentification is not possible with granular healthcare data as those used here. Study-related analytic code and trained algorithms will be made available with publication as per the manuscript text.

6

## REFERENCES

1. Carvalho AF, Firth J, Vieta E. Bipolar disorder. N Engl J Med. 2020. https://doi.org/10.1056/NEJMra1906193.
2. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry. 2016;3:171–8. https://doi.org/10.1016/S2215-0366(15)00505-2.
3. Plans L, Barrot C, Nieto E, Rios J, Schulze TG, Papiol S, et al. Association between completed suicide and bipolar disorder: a systematic review of the literature. J Affect Disord. 2019;242:111–22. https://doi.org/10.1016/j.jad.2018.08.054.
4. Weye N, Momen NC, Christensen MK, Iburg KM, Dalsgaard S, Laursen TM, et al. Association of specific mental disorders with premature mortality in the Danish population using alternative measurement methods. JAMA Netw Open. 2020;3:e206646. https://doi.org/10.1001/jamanetworkopen.2020.6646.
5. Drancourt N, Etain B, Lajnef M, Henry C, Raust A, Cochet B, et al. Duration of untreated bipolar disorder: missed opportunities on the long road to optimal treatment. Acta Psychiatr Scand. 2013;127:136–44. https://doi.org/10.1111/j.1600-0447.2012.01917.x.
6. Fritz K, Russell AMT, Allwang C, Kuiper S, Lampe L, Malhi GS. Is a delay in the diagnosis of bipolar disorder inevitable? Bipolar Disord. 2017;19:396–400. https://doi.org/10.1111/bdi.12499.
7. Dagani J, Signorini G, Nielssen O, Bani M, de Pastore A, et al. Meta-analysis of the interval between the onset and management of bipolar disorder. Can J Psychiatry. 2017;62:247–58. https://doi.org/10.1177/0706743716656607.
8. Bowden CL. Strategies to reduce misdiagnosis of bipolar depression. Psychiatr Serv. 2001;52:51–55. https://doi.org/10.1176/appi.ps.52.1.51.
9. Altamura AC, Dell'Osso B, Berlin HA, Buoli M, Bassetti R, Mundo E. Duration of untreated illness and suicide in bipolar disorder: a naturalistic study. Eur Arch Psychiatry Clin Neurosci. 2010;260:385–91. https://doi.org/10.1007/s00406-009-0085-2.
10. Altamura AC, Buoli M, Caldiroli A, Caron L, Melter CC, Dobrea C, et al. Misdiagnosis, duration of untreated illness (DUI) and outcome in bipolar patients with psychotic symptoms: a naturalistic study. J Affect Disord. 2015;182:70–75. https://doi.org/10.1016/j.jad.2015.04.024.
11. Saraf G, Moazen-Zadeh E, Pinto JV, Ziafat K, Torres IJ, Kesavan M, et al. Early intervention for people at high risk of developing bipolar disorder: a systematic review of clinical trials. Lancet Psychiatry. 2020. https://doi.org/10.1016/S2215-0366(20)30188-7.
12. Yang J, Visscher PM, Wray NR. Sporadic cases are the norm for complex disease. Eur J Hum Genet. 2010;18:1039–43. https://doi.org/10.1038/ejhg.2009.177.
13. Chen C-Y, Lee PH, Castro VM, Minnier J, Charney AW, Stahl EA, et al. Genetic validation of bipolar disorder identified by automated phenotyping using electronic health records. Transl Psychiatry. 2018;8:86 https://doi.org/10.1038/s41398-018-0133-7.
14. Castro VM, Minnier J, Murphy SN, Kohane I, Churchill SE, Gainer V, et al. Validation of electronic health record phenotyping of bipolar disorder cases and controls. Am J Psychiatry. 2015;172:363–72. https://doi.org/10.1176/appi.ajp.2014.14030423.
15. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6years. J Am Med Inf Assoc. 2011;18:441–8. https://doi.org/10.1136/amiajnl-2011-000116.
16. Healthcare Cost and Utilization Project (HCUP). HCUP Clinical Classifications Software (CCS) for ICD-9-CM. Agency for Healthcare Research and Quality; 2006. www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed 1 May 2017.
17. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12:55–67. https://doi.org/10.1080/00401706.1970.10488634.
18. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw. 2011;39:1–13. https://doi.org/10.18637/jss.v039.i05.
19. Friedman AJ, Hastie T, Simon N, Tibshirani R, Hastie MT. Package ' glmnet '. 2015. https://glmnet.stanford.edu/.
20. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1–22.
21. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. 2015. http://arxiv.org/abs/1508.04409.
22. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146–54.
23. Rembold CM. Number needed to screen: development of a statistic for disease screening. BMJ. 1998;317:307–312.
24. Steyerberg E. Clinical prediction models: a practical approach to development, validation, and updating. Springer-Verlag; 2009. www.springer.com/us/book/9780387772431. Accessed October 30, 2018.
25. Smoller JW, Finn CT. Family, twin, and adoption studies of bipolar disorder. Am J Med Genet C Semin Med Genet. 2003;123C:48–58. https://doi.org/10.1002/ajmg.c.20013.
26. Gilman SE, Ni MY, Dunn EC, Breslau J, McLaughlin KA, Smoller JW, et al. Contributions of the social environment to first-onset and recurrent mania. Mol Psychiatry. 2015;20:329–336. https://doi.org/10.1038/mp.2014.36.
27. Pradier MF, Hughes MC, McCoy TH, Barroilhet SA, Doshi-Velez F, Perlis RH. Predicting change in diagnosis from major depression to bipolar disorder after antidepressant initiation. Neuropsychopharmacology. 2021;46:455–461. https://doi.org/10.1038/s41386-020-00838-x.
28. Rabelo-da-Ponte FD, Feiten JG, Mwangi B, Barros FC, Wehrmeister FC, Menezes AM, et al. Early identification of bipolar disorder among young adults – a 22-year community birth cohort. Acta Psychiatr Scand. 2020;142:476–485. https://doi.org/10.1111/acps.13233.
29. Walsh CG, Chaudhry B, Dua P, Goodman KW, Kaplan B, Kavuluru R, et al. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. Jamia Open. 2020;3:9–15. https://doi.org/10.1093/jamiaopen/ooz054.
30. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447–453. https://doi.org/10.1126/science.aax2342.
31. Makadia R, Ryan PB. Transforming the premier perspective hospital database into the observational medical outcomes partnership (OMOP) common data model. EGEMS. 2014;2:1110 https://doi.org/10.13063/2327-9214.1110.
32. Sugiyama M, Kawanabe M. Machine learning in non-stationary environments: Introduction to Covariate Shift Adaptation. The MIT Press; 2012. https://doi.org/10.7551/mitpress/9780262017091.001.0001.

## AUTHOR CONTRIBUTIONS
Design and conduct of the study (all authors) Data collection, management, analysis (CGW, YH, MAR, DR, DW, YS, KWC, ABZ). Interpretation of the data (all authors) Preparation, review, or approval of the manuscript (all authors).

## COMPETING INTERESTS
Authors (MAR, YH, YS, DW, DR, KWC, VMC, HLK, CFC, LKD) declare no Competing Financial or Non-Financial Interests. JWS declares no Competing Non-Financial Interests but the following Competing Financial Interests: PI of a collaborative study of the genetics of depression and bipolar disorder sponsored by 23andMe for which 23andMe provides analysis time as in-kind support but no payments. He is also a member of the Scientific Advisory Board of Sensorium Therapeutics (with equity). ABZ declares no Competing Non-Financial Interests but the following Competing Financial Interests: currently employment by Janssen Pharmaceuticals. Other than manuscript revision, she contributed to this research only during her time as a postdoctoral fellow at MGB. CGW declares no Competing Non-Financial Interests but the following Competing Financial Interests: equity interest in Sage Al, LLC (unrelated to healthcare).

## ADDITIONAL INFORMATION
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41398-023-02720-y.

**Correspondence** and requests for materials should be addressed to Colin G. Walsh.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.