



## Impact of selection bias on polygenic risk score estimates in healthcare settings

Younga Heather Lee<sup>1,2,3</sup> , Tanayott Thaweethai<sup>3,4</sup>, Yi-Han Sheu<sup>1,2,3</sup>, Yen-Chen Anne Feng<sup>1,2,3,5,6</sup>, Elizabeth W. Karlson<sup>3,7</sup>, Tian Ge<sup>1,2,3,8</sup>, Peter Kraft<sup>9,10</sup> and Jordan W. Smoller<sup>1,2,3,8</sup> 

## Original Article

**Cite this article:** Lee YH, Thaweethai T, Sheu Y-H, Feng Y-CA, Karlson EW, Ge T, Kraft P, Smoller JW (2023). Impact of selection bias on polygenic risk score estimates in healthcare settings. *Psychological Medicine* **53**, 7435–7445. <https://doi.org/10.1017/S0033291723001186>

Received: 20 September 2022

Revised: 31 March 2023

Accepted: 11 April 2023

First published online: 25 May 2023

**Keywords:**

selection bias; polygenic risk score; biobank; inverse probability weighting; causal inference

**Corresponding author:**

Jordan W. Smoller;

Email: [jsmoller@mgh.harvard.edu](mailto:jsmoller@mgh.harvard.edu)

<sup>1</sup>Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA; <sup>2</sup>Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; <sup>3</sup>Harvard Medical School, Boston, Massachusetts, USA; <sup>4</sup>Biostatistics Center, Massachusetts General Hospital, Boston, Massachusetts, USA; <sup>5</sup>Analytic and Translational Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA; <sup>6</sup>Division of Biostatistics and Data Science, Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan; <sup>7</sup>Division of Rheumatology, Immunity, and Inflammation, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA; <sup>8</sup>Center for Precision Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA; <sup>9</sup>Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA and <sup>10</sup>Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

**Abstract**

**Background.** Hospital-based biobanks are being increasingly considered as a resource for translating polygenic risk scores (PRS) into clinical practice. However, since these biobanks originate from patient populations, there is a possibility of bias in polygenic risk estimation due to overrepresentation of patients with higher frequency of healthcare interactions.

**Methods.** PRS for schizophrenia, bipolar disorder, and depression were calculated using summary statistics from the largest available genomic studies for a sample of 24 153 European ancestry participants in the Mass General Brigham (MGB) Biobank. To correct for selection bias, we fitted logistic regression models with inverse probability (IP) weights, which were estimated using 1839 sociodemographic, clinical, and healthcare utilization features extracted from electronic health records of 1 546 440 non-Hispanic White patients eligible to participate in the Biobank study at their first visit to the MGB-affiliated hospitals.

**Results.** Case prevalence of bipolar disorder among participants in the top decile of bipolar disorder PRS was 10.0% (95% CI 8.8–11.2%) in the unweighted analysis but only 6.2% (5.0–7.5%) when selection bias was accounted for using IP weights. Similarly, case prevalence of depression among those in the top decile of depression PRS was reduced from 33.5% (31.7–35.4%) to 28.9% (25.8–31.9%) after IP weighting.

**Conclusions.** Non-random selection of participants into volunteer biobanks may induce clinically relevant selection bias that could impact implementation of PRS in research and clinical settings. As efforts to integrate PRS in medical practice expand, recognition and mitigation of these biases should be considered and may need to be optimized in a context-specific manner.

**Introduction**

In recent years, large-scale healthcare systems have contemplated integrating polygenic risk scores (PRS) into clinical practice given their potential to stratify diagnostic and therapeutic strategies in common medical conditions (e.g. diabetes, cancer, obesity) (Khera et al., 2019, 2016; Läll, Mägi, Morris, Metspalu, & Fischer, 2017; Mavaddat et al., 2019; Pashayan et al., 2015; Sharp et al., 2019) and, more recently, in psychiatric conditions (Murray et al., 2021). For example, the Electronic Medical Records and Genomics (eMERGE) Network is conducting trials evaluating the impact of returning genomic results ('return of results' or RoR) in both clinical and research venues (Electronic Medical Records and Genomics (eMERGE) Network, n.d.; Leppig et al., 2022; Madden et al., 2022; Wiesner et al., 2020). Early evidence suggests that patients are in favor of being informed of their genetic test results and receiving advice about how to interpret and act on the results (Allen et al., 2014; Karlson, Boutin, Hoffnagle, & Allen, 2016; Pet et al., 2019).

With the prospect of using PRS to guide clinical decision making, optimizing the accuracy of the risk estimates they provide becomes especially important (Polygenic Risk Score Task Force of the International Common Disease Alliance, 2021). In research settings, including biobank-based studies, genetic analyses are usually restricted to individuals who have volunteered to provide biospecimens for research investigations. More specifically, application of PRS in a biobank or other research cohort typically entails a sequence of sampling procedures

(see Fig. 1a). First, the cohort is limited to participants who provided consent, and had blood samples drawn and genotyped prior to the time of analysis. Next, this subsample is further restricted to those who have passed a genomic quality control (QC) process. However, restricting analyses without considering the complexity of selection mechanism can change or induce spurious associations between factors directly or indirectly related to selection into the PRS analysis.

Inverse probability (IP) weighting is an established method for correcting such bias in which the contribution of each sampled individual is weighted by the inverse of their probability of being sampled (Seaman & White, 2013). In most volunteer-based studies, information about those who were not enrolled is typically limited, precluding in-depth exploration of selection bias that can result from non-random sampling. However, biobanks nested within healthcare systems where demographic and clinical data are available for the full healthcare system population provide a unique opportunity to evaluate factors that may influence the probability of being selected into an analytic sample. In these settings, one can use IP weighting to construct a hypothetical population in which participants are weighted such that they represent the entire population of participants and non-participants with respect to the predictors of selection and conduct analyses that account for non-random sampling.

A key assumption of IP weighting, however, is that one has correctly identified and weighted the predictors of sampling; violation of this assumption may lead to residual or even greater bias (Cole & Hernán, 2008). Meeting this requirement could be particularly challenging in the case of hospital-based biobanks, since selection may be dynamic and reflect a large number of poorly understood factors – including patient comorbidity profiles and the diversity of clinical settings in which recruitment was conducted. Instead of solely relying on expert knowledge to specify the weight model, Haneuse and Daniels suggest combining clinical knowledge with data-driven strategies for covariate selection (Meier, Van De Geer, & Bühlmann, 2008; Tibshirani, 1997; Zou, 2006), especially when working with high-dimensional electronic health records (EHRs) (Haneuse & Daniels, 2016). Accordingly, we use a two-step approach to correct for non-random sampling in PRS analyses. First, we apply a machine learning approach to examine the relative contribution of sociodemographic, clinical, and healthcare utilization characteristics (captured in the longitudinal EHRs) and estimate IP weights for selection into the nested biobank study. Next, we estimate the association between PRS and the target conditions in an IP-weighted sample in which selection into the biobank study occurred at random. Using this two-step approach, we find that standard PRS analyses that do not account for the non-random sampling of biobank samples may lead to biased estimation of polygenic risk in the context of psychiatric conditions.

Finally, we address the fact that selection into biobank-based studies typically involves *multiple* steps – such as recruitment, consent, biospecimen collection, genotyping, and genomic QC – each of which may be influenced by a unique set of determinants. Haneuse and Daniels proposed a general statistical framework for EHR-based research that explicitly models the chain of interactions and decisions made by patients and healthcare providers which ultimately shape the underlying mechanism for study participation and data availability in a given health system (Haneuse & Daniels, 2016). In practical terms, they recommend *modularizing* the complex selection mechanism into a series of sub-mechanisms that are relatively easier to characterize and model

(Haneuse, Arterburn, & Daniels, 2021). Applying this modular IP weighting framework, we evaluate the discrepancy between PRS effect estimates for psychiatric conditions when using standard *v.* modular approaches to define selection mechanisms (see Figs 1b and c).

## Methods

### Study sample

#### *Mass General Brigham (MGB) Research Patient Data Registry (RPDR)*

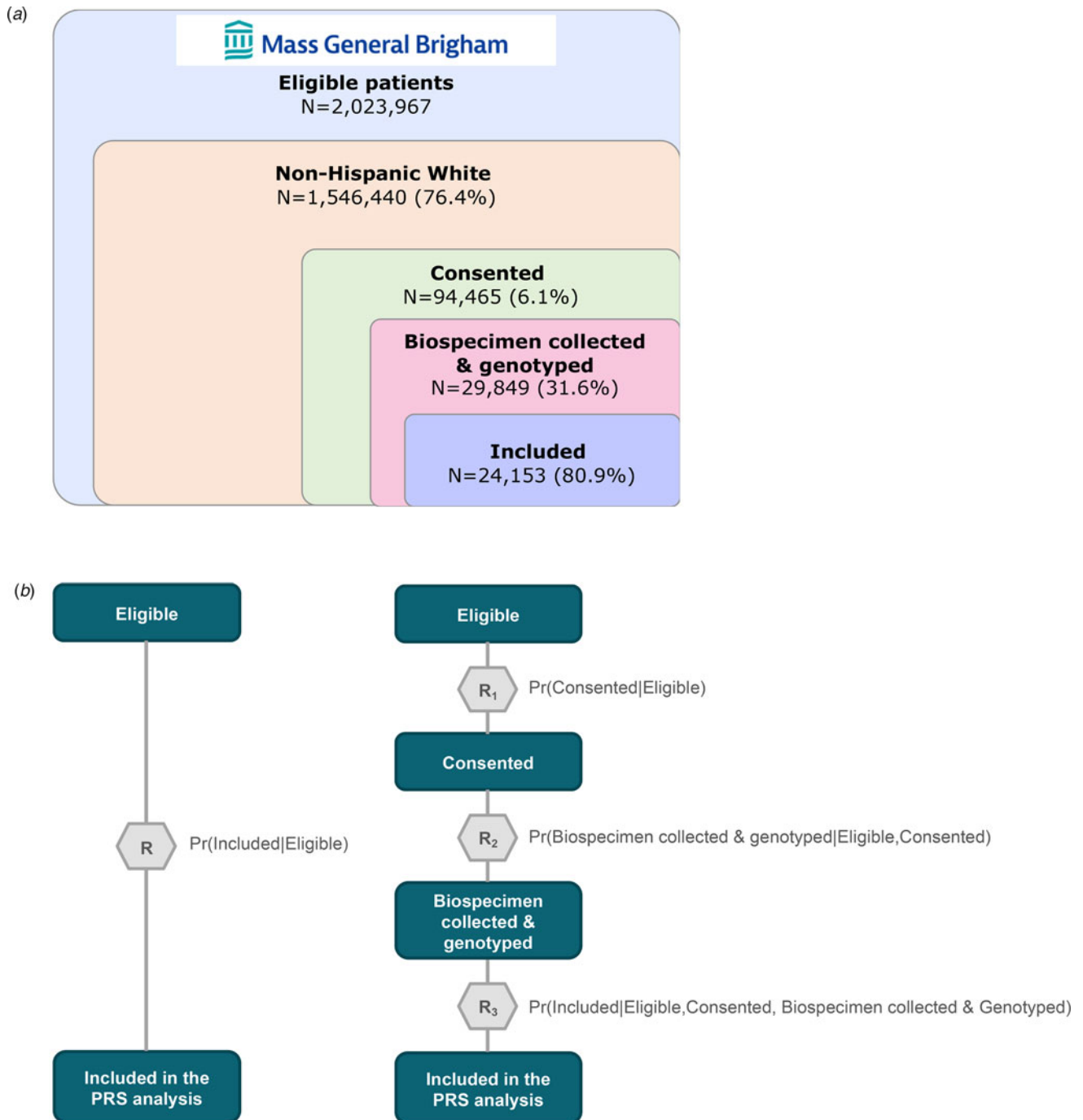
The primary data source was the MGB RPDR, an EHR data warehouse covering 4.6 million patients across the MGB integrated healthcare system (formerly Partners HealthCare) in the USA serving 1.5 million people annually across 11 inpatient hospitals, a rehabilitation network, 20 community health centers, a home-based service network, and hundreds of outpatient clinics. To assemble the cohort for this study, we queried the MGB RPDR for 1 546 440 patients who self-identified as non-Hispanic White (i.e. 76% of the overall MGB patient population) having at least three visits after 2005, more than 30 days apart between the first and last visits, and at least one visit greater than age 10 and less than age 90, as of February 2020 (Bayramli et al., 2021; Castro et al., 2021) (see Fig. 1). The race and ethnicity restriction was applied here because the subsequent PRS analyses were based on samples of European ancestry.

#### *MGB Biobank*

The MGB Biobank is a hospital-based research program launched in 2010 to empower genomic and translational research for human health (Boutin et al., 2022; Karlson et al., 2016). Participants are patients at MGB-affiliated hospital(s) above age 18 (at the time of the recruitment) who provided informed consent to join the Biobank study. Each consented participant was asked to provide blood samples (e.g. plasma, serum, DNA), which are then linked to their clinical data in the EHRs as well as survey data on lifestyle, behavioral and environmental factors, and family history. Leveraging in-person and electronic recruitment methods, the MGB Biobank has currently enrolled 141 451 participants (85% self-identify as White), collected 95 213 DNA samples, and generated genotyping microarray data for more than 65 081 participants (4919 using the Illumina MEGA, 5332 using the Illumina MEGA EX, 26 135 using the Illumina MEG, and 53 284 using the Illumina GSA) (Castro et al., 2021). Further details on the participant recruitment and consent process can be found in eMethods. This research was conducted as part of the PsycheMERGE Consortium (Smoller, 2018), under approval from the MGB Institutional Review Board (2018P002642).

#### *Exposure: polygenic risk scores for three psychiatric conditions*

We generated PRS for the 24 153 MGB Biobank participants of European ancestry using their genotype data and weights derived by applying PRS-CS-Auto (Ge, Chen, Ni, Feng, & Smoller, 2019), a Bayesian polygenic prediction method, to publicly available summary statistics from the largest genome-wide association studies (GWAS) of schizophrenia (The Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke, Walters, & O'Donovan, 2020), bipolar disorder (Mullins et al., 2021), and depression (Howard et al., 2019) on populations of



**Figure 1.** Schematic of sample curation for polygenic risk score analysis using Mass General Brigham (MGB) Biobank and visualization of standard and modular inverse probability (IP) weighting approaches. (a) Sample curation for polygenic risk score analysis using the MGB Biobank sample. (b) Visualization of the standard (left) and modular (right) inverse probability (IP) weighting approach.

European ancestry (see online Supplementary eMethods for details on genomic data processing and online Supplementary eTable 2 for further information on discovery GWAS). In preparation for the main analysis, we fitted linear regression models with age, sex assigned at birth, top 20 genetic principal components, and genotyping microarray as predictors of each respective psychiatric PRS. We then extracted and standardized the residuals from each regression model and generated a categorical version of the PRS using deciles. In the current study, we primarily

focus on disease risk for the top decile of the standardized residuals of PRS, a threshold commonly used to define high genetic risk in the context of clinical translation (Lewis & Vassos, 2017).

**Outcome: clinical diagnosis of three psychiatric conditions**

We identified cases of the three psychiatric traits by mapping the entire longitudinal health records available on all patients at MGB-affiliated hospital(s) to the phecode system using the

*PheWAS* R package (Carroll, Bastarache, & Denny, 2014; Wei *et al.*, 2017). We identified qualifying ICD-9CM and ICD-10CM codes for schizophrenia (phecode 295.1), bipolar disorder (phecode 296.1), and depression (phecode 296.2), and defined cases as those having at least two phecode for a given outcome occurring on different dates (see the full list of qualifying diagnostic codes in online Supplementary eTable 3).

### Statistical approach

We compared effect estimates of the associations between schizophrenia, bipolar disorder, and depression PRS and their respective target diagnoses using three weighting schemes defined by how the IP weight models were specified: (1) unweighted, (2) standard IP-weighted, and (3) modular IP-weighted. In the unweighted approach, PRS effect estimates are calculated without accounting for non-random sampling (i.e. standard PRS analysis). In contrast, the latter two approaches involve a systematic evaluation and adjustment for differential probabilities of being selected into the analytic sample for the PRS analyses, with the modular approach involving additional specification of the intermediate steps of selection (see Figs 1b and c). The application of IP weights allows us to construct a hypothetical population in which we can estimate the effects of PRS in the absence of spurious associations induced by participation-related factors specified in the IP weight model (see the causal diagram in Fig. 2b).

We first evaluated the impact of IP-weighting on PRS penetrance, which represents the case prevalence as a function of PRS and provides an estimate of the absolute disease risk (Bigdeli *et al.*, 2022; Zheutlin *et al.*, 2019). To assess this impact, we compared the IP-weighted penetrance against the unweighted penetrance. Next, we evaluated the discrimination ability of the PRS using the area under the receiver operator characteristic curve (hereafter, the AUC) (Robin *et al.*, 2011). Under the unweighted approach, we fitted standard logistic regression models adjusting for covariates. Under the IP-weighted approaches, we inputted the standard and modular IP-weights, respectively, and fitted weighted logistic regression models (Lumley, 2021). We then calculated the AUC to compare the discrimination ability of the unweighted and IP-weighted logistic regression models (Mangiafico, 2022). Lastly, we explored potential effect modification of the discrimination ability of psychiatric PRS by sex assigned at birth and current age.

### Data-driven specification of IP weight models for selection

We utilized a large set of demographic and clinical features extracted from high-dimensional EHRs, including 15 sociodemographic, 1814 diagnostic, and 10 healthcare utilization characteristics to identify the key determinants of non-random sampling of biobank participants and calculate the IP selection weights (refer to online Supplementary eTable 4 for the full list of features and online Supplementary eMethods for how they were curated). To achieve this, we employed a machine learning approach, Extreme Gradient Boosting (XGBoost) classification (Chen & Guestrin, 2016), which is an open-source library that provides a computationally efficient and high-performance implementation of gradient-boosted decision trees (<https://github.com/dmlc/xgboost>).

In the first set of IP-weighted analyses (i.e. standard IP-weighted approach), we fitted an XGBoost model classifying the inclusion into the PRS analysis ( $N = 24\,153$ ) from a pool of

1 546 440 adult patients at MGB-affiliated hospital(s) self-identifying as non-Hispanic White. Considering that a very small proportion of the patient population participated in the Biobank study, we ensured that the training and test sample (with a split ratio of 80:20) had the same proportion of the target outcome in a given selection step (e.g. included *v.* not included in the PRS analysis). After fitting the model, we derived weights by taking the inverse of the predicted probabilities of being selected into the final PRS analysis. We further stabilized the IP weights by dividing the predicted probabilities by the marginal probability of selection and truncated the top and bottom 1% of the distribution to account for extreme weights.

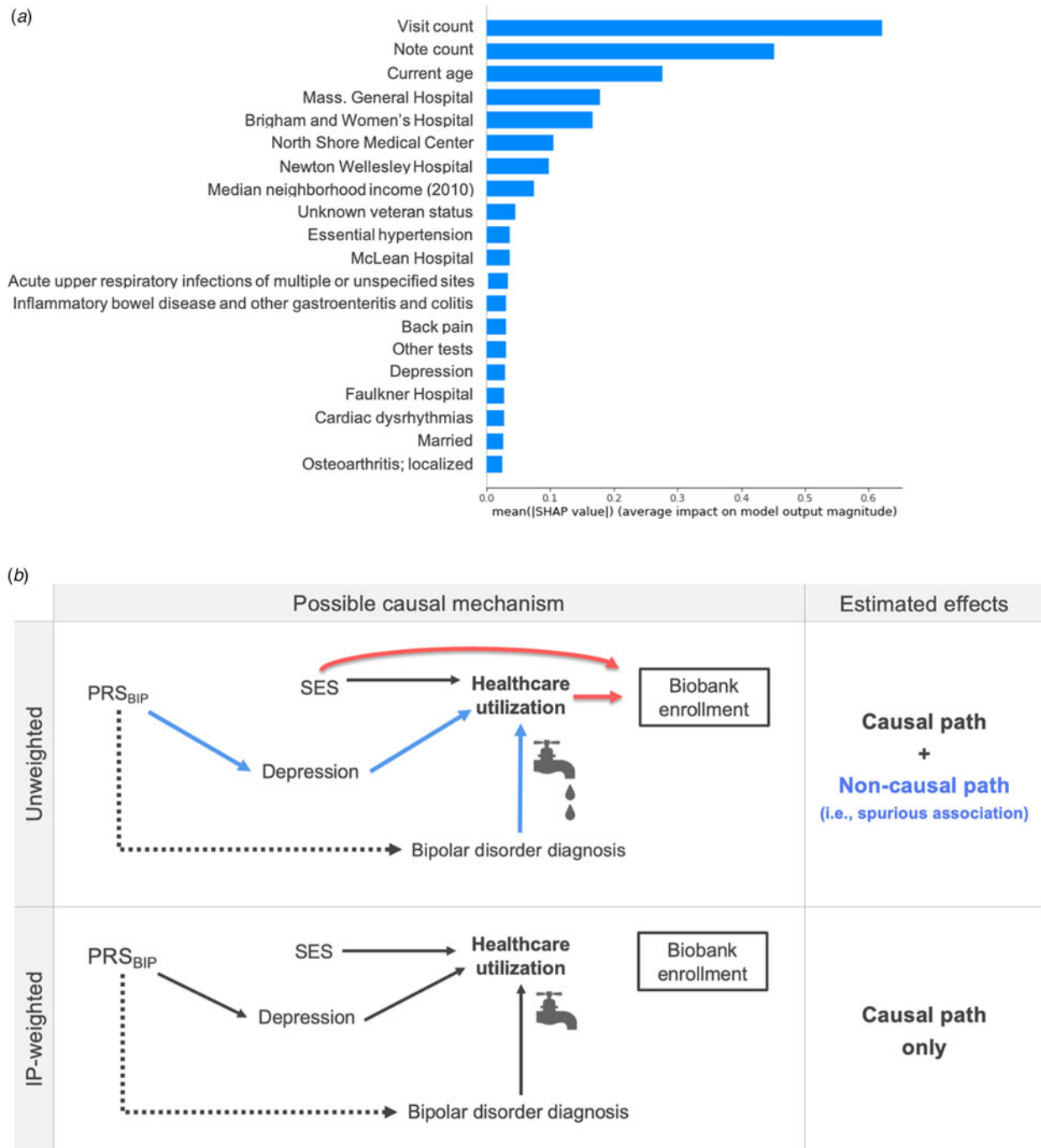
In the second set of IP-weighted analyses (i.e. modular IP-weighted approach), we fit *three* separate sets of XGBoost models classifying each of the three selection steps (see Fig. 1c). The three targets for classification were: 1) consent status among eligible participants, 2) biospecimen collection and genotyping status among consented participants, and 3) inclusion in the PRS dataset among participants who are eligible, consented, and had biospecimens collected and genotyped. We extracted predicted probabilities from each of the three models and took the product of these conditional probabilities to calculate the joint probabilities of being included in the final analytic sample given the three sequential steps of selection. We then stabilized and truncated the inverse of the joint probabilities in the same way as we did for the standard IP-weighted approach, and performed weighted PRS analyses.

In addition, we applied a game theory-based algorithm called Shapley Additive Explanations (SHAP) method to further elucidate the complex selection mechanism of the MGB Biobank (Lundberg & Lee, 2017). We calculated Shapley values, which are the weighted average of the marginal contribution of each feature value toward the model's decision, to explain how changes in a feature value would shift the models' decision both in terms of absolute magnitude and directionality. This way, we characterized the importance of each feature to the predicted probability of being retained in the study sample at each step of selection (see magnitude and directionality of contribution by the top 20 features in online Supplementary eFigs 1 and 2, respectively).

## Results

### Descriptive statistics

As shown in Table 1, we first compared participants in the analytic (Biobank) sample ( $N = 24\,153$ ) for the PRS analyses against those who were not included (from the broader pool of eligible patients in the healthcare system). In general, the included individuals were significantly more likely to be male, veterans, and married, have publicly funded insurance, and have markedly greater healthcare utilization compared to those excluded and those in the overall source population. Additionally, we compared the prevalence estimates for common health conditions of those included in the final analytic sample against those of excluded participants (see online Supplementary eTable 1). Consistent with their higher frequency of healthcare interactions, individuals included in the PRS analysis were more likely to have clinical diagnoses of all disease conditions examined, including up to three times higher rates of endocrine, nutritional, and metabolic diseases (e.g. type 1 and 2 diabetes mellitus, obesity), neuropsychiatric conditions (e.g. neurological disorders, major depressive disorder, suicidal behavior), and circulatory conditions (e.g.



**Figure 2.** Identification of key features predicting inclusion (e.g. healthcare utilization, income) and a causal diagram illustrating how stratification on healthcare utilization may introduce bias in standard PRS estimation in hospital-based biobanks. (a) A visual demonstration of top 20 features from the standard IP weight model based on mean absolute Shapley values. Features with higher mean absolute Shapley values have a greater impact on the model's decision than those with lower values. The vertical axis shows the features rank-sorted according to the magnitude of the mean absolute Shapley values, from highest (top) to lowest (bottom). (b) A causal diagram (directed acyclic graph or DAG) illustrating how non-random sampling into hospital-based biobanks may introduce bias in a standard PRS estimation. Using the example of a bipolar disorder PRS, the figure depicts two DAGs to illustrate how selection bias could inflate PRS effect estimates in an unweighted PRS analysis. The relationship of interest is denoted by the dotted line connecting PRS<sub>BIP</sub> (bipolar disorder polygenic risk score) with bipolar disorder diagnosis. Restriction of PRS analysis to biobank participants is represented as a box around biobank enrollment in the causal diagram. Healthcare utilization is a common effect of PRS<sub>BIP</sub> (through the effect of PRS<sub>BIP</sub> on depression) and clinical diagnosis of bipolar disorder. In this example, stratification on biobank enrollment, a descendant of healthcare utilization, can induce a spurious association between the PRS and the target trait (represented as a dripping faucet in the figure). Thus, the estimated effect could include not only true causal effects but also the spurious association, resulting in larger estimates in standard PRS analysis when non-random sampling is not addressed. In contrast, when selection bias is accounted for using IP weighting, socioeconomic status (SES) and healthcare utilization are no longer associated with biobank enrollment, and so biobank enrollment is no longer a descendant of a collider. Therefore, stratifying on biobank enrollment would not open the non-causal path blocked by healthcare utilization (represented as a tight faucet in the figure). Thus, IP-weighted PRS estimates would likely represent effects through the causal path only.

**Table 1.** Comparison of demographic and healthcare utilization characteristics of European ancestry patients in the overall MGB patient population<sup>1</sup> against those included in the PRS analysis (shown in number of participants and prevalence of a given condition)

Category		Overall patient population <sup>1</sup> (N = 1 546 440)	Not included (N = 1 522 287)	Included (N = 24 153)	p value
Sociodemographic characteristics [N (%)]					
Mean age (s.d.)		58.2 (19.3)	58.1 (19.3)	63.0 (16.3)	<0.001
Gender, N (%)	Female	887 810 (57.4)	874 943 (57.5)	12 867 (53.3)	<0.001
	Male	658 565 (42.6)	647 279 (42.5)	11 286 (46.7)	
	Unknown	65 (0.0)	65 (0.0)	0 (0.0)	
Veteran status, N (%)	Yes	98 723 (6.4)	96 322 (6.3)	2401 (9.9)	<0.001
	No	1 191 436 (77.0)	1 171 391 (76.9)	20 045 (83.0)	
	Unknown	256 281 (16.6)	254 574 (16.7)	1707 (7.1)	
Health insurance, N (%)	Private	888 548 (57.5)	878 586 (57.7)	99,62 (41.2)	<0.001
	Public	657 892 (42.5)	643 701 (42.3)	14 191 (58.8)	
Marital status, N (%)	Divorced	93 297 (6.0)	91 524 (6.0)	1773 (7.3)	<0.001
	Married	823 131 (53.2)	808 753 (53.1)	14 378 (59.5)	
	Other/unknown	48 181 (3.1)	47 843 (3.1)	338 (1.4)	
	Partner	7395 (0.5)	7206 (0.5)	189 (0.8)	
	Separated	12 331 (0.8)	12 112 (0.8)	219 (0.9)	
	Single	478 402 (30.9)	472 380 (31.0)	6022 (24.9)	
	Widowed	83 703 (5.4)	82 469 (5.4)	1234 (5.1)	
Healthcare utilization [mean (s.d.)]					
Visit count	73.71 (114.9)	71.34 (110.5)	223.27 (229.3)	<0.001	
ICD code count <sup>2</sup>	185.1 (332.7)	178.25 (317.2)	615.95 (743.7)	<0.001	
CPT code count <sup>2</sup>	140.63 (255.2)	135.38 (244.5)	471.15 (537.9)	<0.001	
Note count	360.70 (566.7)	349.81 (545.8)	1047.00 (1145.2)	<0.001	

<sup>1</sup>The denominator ('overall MGB patient population') is defined as adult patients (18 years and older by 2010) of European ancestry having at least three visits after 2005 and more than 40 days apart with at least one clinical note (N = 1 546 440; see Fig. 1).

<sup>2</sup>ICD, International Classification of Diseases; CPT, Current Procedural Terminology.

essential hypertension, myocardial infarction). Of note, the prevalence of rheumatoid arthritis was up to five times greater among those included than those not included in the PRS analyses, likely reflecting recruitment into the MGB Biobank from rheumatology clinics. Lastly, the prevalence estimates of the three target traits in the analytic sample of 24 153 MGB Biobank participants were 1.0% ( $n_{\text{case}} = 236$ ), 4.5% ( $n_{\text{case}} = 1079$ ), and 26.2% ( $n_{\text{case}} = 6329$ ) for schizophrenia, bipolar disorder, and depression, respectively.

### Identification of key determinants of selection

In the XGBoost model under the standard IP-weighted approach, visit count, note count, current age, and clinical encounters at Massachusetts General Hospital (MGH) or Brigham and Women's Hospital (BWH) were the five most important features that differentiated those included and those not included in the PRS analysis, followed by clinical encounters at Northshore Medical Center or Newton-Wellesley Hospital and median neighborhood income in 2010 (see Fig. 2a). The top features indicative of healthcare utilization from the standard IP-weighted approach also appeared in the three XGBoost models under the modular IP-weighted approach. The modular approach identified additional features that contributed to the probability of being retained in each step of selection, such as anxiety, phobic, and

dissociative disorders, ischemic heart disease, clinical encounters at Faulkner Hospital, and rheumatoid arthritis and other inflammatory polyarthropathies (see online Supplementary eFigs 1a–c).

In addition to overall feature importance, we further examined the *directionality* of feature contributions to being retained in each step of selection in the modular IP-weighted approach. This was motivated in part by prior work showing that standard IP weighting may lead to biased estimates when a given feature plays a different role in each step of a sequential selection procedure (Haneuse et al., 2021; Peskoe et al., 2021; Thaweethai, Arterburn, Coleman, & Haneuse, 2021). To address this, we calculated Shapley values at every observed value of each feature across all possible combinations with other features and evaluated whether key features had dynamic contributions across the three selection steps. Interestingly, visit count, which was the most important feature in every step of selection, exhibited different directions of associations with retention probabilities across the three selection steps (see online Supplementary eFigs 2b–d). For example, an increasing number of visits was associated with a *higher* likelihood of providing consent to participate in the Biobank but a *lower* likelihood of being retained in the subsequent steps of selection. Although the modularization of the IP weight model did not substantially improve the adjustment of selection bias in the PRS analysis relative to standard IP weighting, our results underscore the

importance of considering the possibility that some factors may affect retention probability differently across multiple phases of selection in biobank studies.

### Polygenic risk estimation

#### Case prevalence per deciles of standardized residuals of psychiatric PRS

After standardizing PRS by principal components, sex, age, and genotyping microarray, case prevalence for schizophrenia in the top decile of standardized residuals of schizophrenia PRS was 2.7% (2.1–3.3) in the unweighted analysis, and 2.0% (1.2–2.7) in the standard IP-weighted analysis (see Fig. 3a). The unweighted and IP-weighted estimates differed more substantially in the case of bipolar disorder; case prevalence of bipolar disorder in the top PRS decile was 10.0% (8.8–11.2) in the unweighted analysis, but only 6.2% (5.0–7.5) when selection bias was accounted for using IP weights. Finally, case prevalence of depression in the top decile of standardized residuals of depression PRS was 33.5% (31.7–35.4) in the unweighted analysis but was reduced to 28.9% (25.8–31.9) after standard IP weighting. Results using modular IP weighting based on intermediate selection steps were similar to those observed with standard IP weighting (see online Supplementary eTables 5–7).

#### Discrimination ability at the top decile of psychiatric PRS

We found the largest impact of IP weighting on tail discrimination with respect to schizophrenia relative to bipolar disorder and depression (see online Supplementary eTable 8). When stratified by sex assigned at birth, the AUC estimates were generally higher among male participants than female participants regardless of the weighting scheme (see Fig. 3b). The impact of IP weighting was also greater among males (AUC = 0.792 and 0.711 from unweighted and modular IP-weighted models, respectively) than females (AUC = 0.711 and 0.675 from unweighted and modular IP-weighted models, respectively).

In addition, we found that both the magnitude and directionality of the impact of IP weighting varied by age, especially for schizophrenia (see Fig. 3c). For example, among participants whose age was less than 40 years, the AUC of schizophrenia PRS from the unweighted model was lower than the AUC from the modular IP-weighted model. Conversely, the AUC from the unweighted model was higher than the AUC from the modular IP-weighted model among participants whose age was greater than or equal to 40.

## Discussion

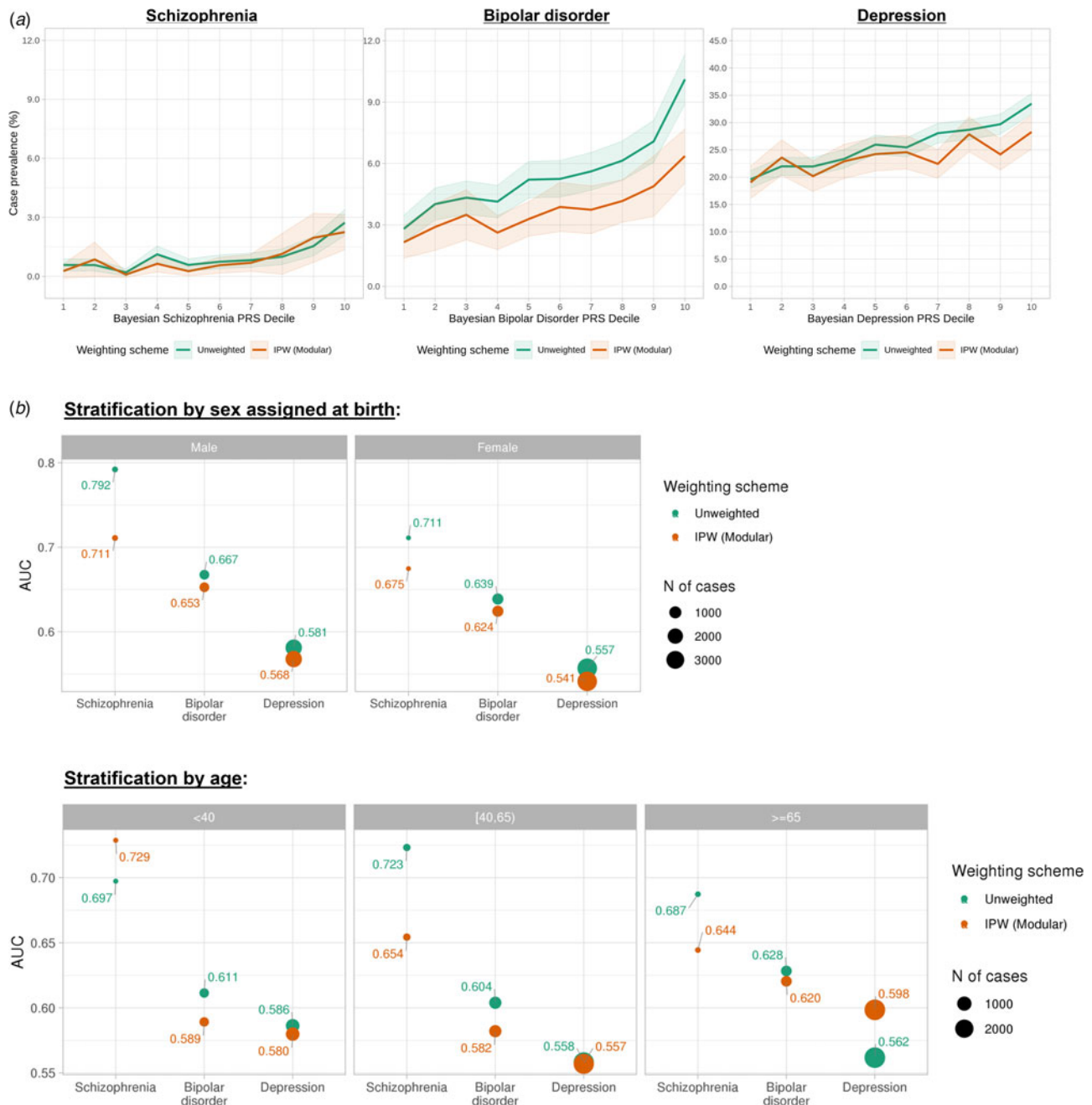
In the present study, we demonstrated that effect estimates of psychiatric PRS can be sensitive to selection bias, using the MGB Biobank as a case example. First, we showed that volunteer-based biobank participants may substantially differ from patients in the underlying healthcare system with respect to a wide range of patient profiles including sociodemographic, clinical, and healthcare utilization characteristics. Notably, prevalence of disease conditions and rates of healthcare utilization were substantially higher in the analytic sample than in the overall MGB patient population. This suggests that, in contrast to the well-known phenomenon of ‘healthy volunteer bias’ in population-based biobanks (Fry et al., 2017; Swanson, 2012; Tyrrell et al., 2021), patients enrolled in hospital-based biobanks may have a greater burden of illness than those in the underlying healthcare system

from which they were selected. In addition, we demonstrated that an efficient machine learning algorithm can help discover key sociodemographic, clinical, and healthcare utilization characteristics associated with the probability of retention in each selection step of PRS analyses, allowing for a more comprehensive adjustment of selection bias.

Using IP weighting procedures, we found that selection bias can produce meaningful impact on estimates of penetrance and discrimination ability of psychiatric PRS in biobank samples derived from healthcare system populations. Overall, unweighted effect estimates of psychiatric PRS were larger than the IP-weighted estimates for the three psychiatric traits examined in the current study. In the example of a bipolar disorder PRS, Fig. 2b shows a causal diagram that illustrates how selection bias could inflate PRS effect estimates in hospital-based biobanks that tend to be enriched with patients with higher frequency of healthcare interactions compared to the underlying patient population. Restriction of PRS analysis to biobank participants is represented as a box around biobank enrollment in the causal diagram. In this example, stratification on the descendent of healthcare utilization, a common effect (i.e. collider) of bipolar disorder PRS and clinical diagnosis of bipolar disorder, can induce a spurious association between the PRS and the target trait – a phenomenon commonly referred to as ‘collider stratification bias’ and known to pose a potential threat to the internal validity (Hernán, Hernández-Díaz, & Robins, 2004). As such, the estimated effect could include not only true causal effects but also the spurious association, thereby resulting in larger estimates in standard PRS analysis when non-random sampling is not addressed.

These findings underscore the complex nature of selection bias and the difficulty of predicting the magnitude or directionality of the effects by this type of bias on PRS estimates in real-world settings. For example, individuals who are more health-conscious or better informed about the clinical utility of genomic findings may be more willing to participate in a biobank, as has been shown in the UK Biobank (Fry et al., 2017; Swanson, 2012; Tyrrell et al., 2021; van Alten, Domingue, Galama, & Marees, 2022). Conversely, patients whose illness leads to more frequent encounters with the healthcare system may have more opportunities to be selected for biobank participation, leading to an overrepresentation of less healthy individuals. In addition, some individuals may enroll in genetic studies because they have a family history of heritable conditions, such as cancer, and are thus motivated to learn about their risk of illness; enrichment for family history of specific diseases may contribute to differences between biobank cohorts and their underlying source populations.

Recently, several analytic approaches to model and mitigate selection bias in EHR data have been proposed, with varying conceptual definitions of selection bias and statistical approaches to modeling underlying selection mechanisms. For instance, Haneuse and Daniels encourage researchers to modularize complex selection mechanisms into a series of sub-mechanisms that are easier to characterize and model (Haneuse & Daniels, 2016; Haneuse et al., 2021). In the current study, we adapted this statistical framework to accommodate the selection procedures unique to PRS analyses conducted in hospital-based biobanks, though modular IP weighting did not differ substantially from standard IP weighting in its impact on polygenic risk estimation. Nevertheless, the modular approach revealed that certain features (e.g. visit count) may have differing impacts on retention probability at different stages of selection and provided useful insights



**Figure 3.** Evaluation of the impact of the modular inverse probability (IP) weighting approach on the polygenic risk estimation of schizophrenia, bipolar disorder, and depression. (a) Case prevalence by polygenic risk score (PRS) decile for three psychiatric traits using two different weighting schemes - unweighted and modular IP-weighted. PRS were adjusted for potential confounding by top genetic principal components, sex, age, and genotyping microarray. The solid lines indicate point estimates, and the bands indicate 95% confidence intervals for corresponding point estimates. Note that the standard IP-weighted model is not shown in this figure, since the estimates were nearly identical to the modular IP-weighted model. Numeric estimates from all three models can be found in online Supplementary eTables 5–7. (b) Comparison of discrimination by psychiatric PRS (area under the receiver operating characteristic curve or AUC) across groups defined by sex assigned at birth and age.

into the variable contributions of features that would not have been identified otherwise. As an alternative to Haneuse and Daniels' approach, Goldstein and colleagues proposed controlling for the number of healthcare encounters (Goldstein, Bhavsar, Phelan, & Pencina, 2016). However, as they note, stratification on healthcare utilization may actually induce spurious association between two disease phenotypes in cases where healthcare

encounters may be the common outcome of the exposure and outcome (i.e. collider stratification bias).

More recently, Beesley and Mukherjee proposed calibration weighting and IP weighting methods to account for selection bias in EHR-linked biobank studies (Beesley & Mukherjee, 2022). They focus on the form of selection bias that arises from the lack of representativeness and propose constructing weights



from external data that better represent the demographic and clinical characteristics of the source population, such as national disease registries for target traits of interest. However, different healthcare systems serve different patient populations, each characterized by unique profiles of sociodemographic, clinical, and healthcare utilization characteristics. As such, it may not be feasible to directly transport selection weight models trained in one healthcare system to another. Instead, adjustment may require a context-specific examination of underlying distributions of the key determinants leading to retention in the analytic sample for PRS analyses. To that end, we leveraged the longitudinal EHRs linked to genomic data collected to derive a set of weights that are specific to the underlying selection mechanism for the MGB Biobank.

Relatedly, different health system biobanks may rely on varying strategies for recruitment and biospecimen collection. For example, at the MGB Biobank, participant enrollment is conducted using a range of procedures including recruitment via (a) outpatient primary care or specialty clinics; (b) inpatient settings; (c) at centralized phlebotomy services; (d) online enrollment; or (e) collaborating studies. For a subset of patients, biospecimen collection was obtained by placing an order into the Epic EHR system (Epic Systems Corporation, n.d.) collect a sample concurrently with a clinically ordered blood draw. Although an overrepresentation of less healthy individuals could be a general characteristic of hospital-based biobanks given that they originate from patient populations, the degree of overrepresentation may further vary depending on the distinct method of recruitment and sample collection used in each biobank study.

Our study has several limitations that should be considered when interpreting the results. First, our approach does not address the distributional mismatch between samples used to train and validate the PRS and the sample in which the PRS is implemented. This can lead to miscalibration of PRS estimates if the samples differ with respect to sample characteristics, such as age, sex, and socioeconomic status (Mostafavi et al., 2020). Although this is an important issue in the implementation of PRS in clinical practice, it was beyond the scope of our current study. Second, caution is advised when generalizing our findings to other traits as our study focused on three psychiatric traits. It is possible that the impact of selection bias may vary across different clinical conditions. Third, our study was restricted to participants of European ancestry, primarily due to the limited availability of non-European ancestry participants in the MGB Biobank, particularly in the subsample with genotype data available for analysis. As a result, we were unable to investigate the impact of selection bias in non-European populations or potential variations in participation by genetic ancestry. Notably, patients from diverse populations are more likely to be exposed to socio-demographic disadvantages, such as low income, low health literacy, lack of access to healthcare, mistrust in biomedical research, and cultural beliefs, which can contribute to low participation rates among diverse populations in biobank studies (Prictor, Teare, & Kaye, 2018). Therefore, further investigation and validation in more diverse samples and contexts are necessary to ensure equitable translation of PRS into clinical practice (Landry, Ali, Williams, Rehm, & Bonham, 2018).

In conclusion, our analyses demonstrate a novel, interdisciplinary approach for detecting and accounting for unrecognized selection bias in hospital-based biobanks, particularly in the context of PRS analyses. As efforts to integrate PRS into research and

clinical settings continue to expand, recognizing and mitigating these biases is increasingly important, since these biases may have implications for patient care and outcomes. Moreover, further research and validation in more diverse populations will be essential to ensure the generalizability and applicability of our approach in different contexts.

**Supplementary material.** The supplementary material for this article can be found at <https://doi.org/10.1017/S0033291723001186>.

**Acknowledgements.** This work was conducted with support from Harvard Catalyst | The Harvard Clinical and Translational Science Center (National Center for Advancing Translational Sciences, National Institutes of Health Award UL1 TR002541) and financial contributions from Harvard University and its affiliated academic healthcare centers. The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University and its affiliated academic healthcare centers, or the National Institutes of Health. J. W. S. was supported in part by NIMH R01MH118233, NHGRI U01HG008685, and a gift from the Demarest Lloyd, Jr. Foundation. T. G. was supported in part by NIA R00AG054573, NHGRI U01HG008685, and NHGRI U01HG011723. E. W. K. was supported by 5U01HG008685. This study would not be possible without the contributions of Mass General Brigham (MGB) patients and Biobank participants. We would also like to thank the research coordinators and the Biobank study for their tremendous effort in participant recruitment and sample collection. Lastly, we would like to acknowledge the RPDR team for their work maintaining the enterprise research patient data warehouse.

**Financial support.** Dr Smoller is a member of the Leon Levy Foundation Neuroscience Advisory Board, the Scientific Advisory Board of Sensorium Therapeutics, and has received honoraria for internal seminars at Biogen, Inc and Tempus Labs. He is 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.

## References

- Allen, N. L., Karlson, E. W., Malspeis, S., Lu, B., Seidman, C. E., & Lehmann, L. S. (2014). Biobank participants' preferences for disclosure of genetic research results: Perspectives from the OurGenes, OurHealth, OurCommunity project. *Mayo Clinic Proceedings*, *89*(6), 738–746. doi:10.1016/j.mayocp.2014.03.015
- Bayramli, I., Castro, V., Barak-Corren, Y., Madsen, E. M., Nock, M. K., Smoller, J. W., & Reis, B. Y. (2021). Temporally informed random forests for suicide risk prediction. *Journal of the American Medical Informatics Association: JAMIA*, *29*(1), 62–71. doi:10.1093/jamia/ocab225
- Beesley, L. J., & Mukherjee, B. (2022). Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics*, *78*(1), 214–226. doi:10.1111/biom.13400
- Bigdeli, T. B., Voloudakis, G., Barr, P. B., Gorman, B. R., Genovese, G., & Peterson, R. E., ... Cooperative Studies Program (CSP) #572 and Million Veteran Program (MVP). (2022). Penetrance and pleiotropy of polygenic risk scores for schizophrenia, bipolar disorder, and depression among adults in the US veterans affairs health care system. *JAMA Psychiatry*, *79*(11), 1092–1101. doi:10.1001/jamapsychiatry.2022.2742
- Boutin, N. T., Schechter, S. B., Perez, E. F., Tchamitchian, N. S., Cerretani, X. R., Gainer, V. S., ... Smoller, J. W. (2022). The evolution of a large biobank at Mass General Brigham. *Journal of Personalized Medicine*, *12*(8), 1323. doi:10.3390/jpm12081323
- Carroll, R. J., Bastarache, L., & Denny, J. C. (2014). R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*, *30*(16), 2375–2376. doi:10.1093/bioinformatics/btu197
- Castro, V. M., Gainer, V., Wattanasin, N., Benoit, B., Cagan, A., Ghosh, B., ... Murphy, S. N. (2021). The Mass General Brigham Biobank Portal: An i2b2-based data repository linking disparate and high-dimensional patient data to support multimodal analytics. *Journal of the American Medical Informatics Association: JAMIA*, *29*(4), 643–651. doi:10.1093/jamia/ocab264

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Retrieved from <http://arxiv.org/abs/1603.02754>.
- Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, *168*(6), 656–664. doi:10.1093/aje/kwn164
- Electronic Medical Records and Genomics (eMERGE) Network. (n.d.). Retrieved from 29 April 2021 <https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE>.
- Epic Systems Corporation. (n.d.). Epic electronic health record. Verona, WI.
- Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., ... Allen, N. E. (2017). Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *American Journal of Epidemiology*, *186*(9), 1026–1034. doi:10.1093/aje/kwx246
- Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, *10*(1), 1776. doi:10.1038/s41467-019-09718-5
- Goldstein, B. A., Bhavsar, N. A., Phelan, M., & Pencina, M. J. (2016). Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American Journal of Epidemiology*, *184*(11), 847–855. doi:10.1093/aje/kww112
- Haneuse, S., Arterburn, D., & Daniels, M. J. (2021). Assessing missing data assumptions in EHR-based studies: A complex and underappreciated task. *JAMA Network Open*, *4*(2), e210184. doi:10.1001/jamanetworkopen.2021.0184
- Haneuse, S., & Daniels, M. (2016). A general framework for considering selection bias in EHR-based studies: What data are observed and why? *EGEMS*, *4*(1), 1203. doi:10.13063/2327-9214.1203
- Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, *15*(5), 615–625. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/15308962>.
- Howard, D. M., Adams, M. J., Clarke, T. K., Hafferty, J. D., Gibson, J., Shiralil, M., ... McIntosh, A. M. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature Neuroscience*, *22*(3), 343–352. doi:10.1038/s41593-018-0326-7
- Karlson, E. W., Boutin, N. T., Hoffnagle, A. G., & Allen, N. L. (2016). Building the partners HealthCare biobank at partners personalized medicine: Informed consent, return of research results, recruitment lessons and operational considerations. *Journal of Personalized Medicine*, *6*(1). doi:10.3390/jpm6010002
- Khera, A. V., Chaffin, M., Wade, K. H., Zahid, S., Brancale, J., Xia, R., ... Kathiresan, S. (2019). Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell*, *177*(3), 587–596.e9. doi:10.1016/j.cell.2019.03.028
- Khera, A. V., Emdin, C. A., Drake, I., Natarajan, P., Bick, A. G., Cook, N. R., ... Kathiresan, S. (2016). Genetic risk, adherence to a healthy lifestyle, and coronary disease. *The New England Journal of Medicine*, *375*(24), 2349–2358. doi:10.1056/NEJMoa1605086
- Läll, K., Mägi, R., Morris, A., Metspalu, A., & Fischer, K. (2017). Personalized risk prediction for type 2 diabetes: The potential of genetic risk scores. *Genetics in Medicine*, *19*(3), 322–329. doi:10.1038/gim.2016.103
- Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., & Bonham, V. L. (2018). Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Affairs*, *37*(5), 780–785. doi:10.1377/hlthaff.2017.1595
- Leppig, K. A., Kulchak Rahm, A., Appelbaum, P., Aufox, S., Bland, S. T., Buchanan, A., ... Wiesner, G. L. (2022). The reckoning: The return of genomic results to 1444 participants across the eMERGE3 network. *Genetics in Medicine*, *24*(5), 1130–1138. doi:10.1016/j.gim.2022.01.015
- Lewis, C. M., & Vassos, E. (2017). Prospects for using risk scores in polygenic medicine. *Genome Medicine*, *9*(1), 96. doi:10.1186/s13073-017-0489-y
- Lumley, T. (2021). survey: Analysis of complex survey samples (Version 4.1-1). Retrieved from University of Auckland website. Retrieved from <http://r-survey.r-forge.r-project.org/survey/>.
- Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Retrieved from <http://arxiv.org/abs/1705.07874>.
- Madden, J. A., Brothers, K. K., Williams, J. L., Myers, M. F., Leppig, K. A., Clayton, E. W., ... Holm, I. A. (2022). Impact of returning unsolicited genomic results to nongenetic health care providers in the eMERGE III network. *Genetics in Medicine*, *24*(6), 1297–1305. doi:10.1016/j.gim.2022.02.018
- Mangiafico, S. (2022). Functions to support extension education program evaluation [R package rcompanion version 2.4.13]. Retrieved from <https://CRAN.R-project.org/package=rcompanion>.
- Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., ... Easton, D. F. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *American Journal of Human Genetics*, *104*(1), 21–34. doi:10.1016/j.ajhg.2018.11.002
- Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *70*(1), 53–71. doi:10.1111/j.1467-9868.2007.00627.x
- Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J. K., & Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. *ELife*, *9*. doi:10.7554/eLife.48376
- Mullins, N., Forstner, A. J., O'Connell, K. S., Coombes, B., Coleman, J. R. I., Qiao, Z., ... Andreassen, O. A. (2021). Genome-wide association study of over 40000 bipolar disorder cases provides new insights into the underlying biology. *Nature Genetics*, *53*(6), 817–829. doi:10.1101/2020.09.17.20187054
- Murray, G. K., Lin, T., Austin, J., McGrath, J. J., Hickie, I. B., & Wray, N. R. (2021). Could polygenic risk scores be useful in psychiatry?: A review. *JAMA Psychiatry*, *78*(2), 210–219. doi:10.1001/jamapsychiatry.2020.3042
- Pashayan, N., Pharoah, P. D. P., Schleutker, J., Talala, K., Tammela, T. L. J., Mänttinen, L., ... Auvinen, A. (2015). Reducing overdiagnosis by polygenic risk-stratified screening: Findings from the Finnish section of the ERSPC. *British Journal of Cancer*, *113*(7), 1086–1093. doi:10.1038/bjc.2015.289
- Peskoe, S. B., Arterburn, D., Coleman, K. J., Herrinton, L. J., Daniels, M. J., & Haneuse, S. (2021). Adjusting for selection bias due to missing data in electronic health records-based research. *Statistical Methods in Medical Research*, *30*(10), 2221–2238. doi:10.1177/09622802211027601
- Pet, D. B., Holm, I. A., Williams, J. L., Myers, M. F., Novak, L. L., Brothers, K. B., ... Clayton, E. W. (2019). Physicians' perspectives on receiving unsolicited genomic results. *Genetics in Medicine*, *21*(2), 311–318. doi:10.1038/s41436-018-0047-z
- Polygenic Risk Score Task Force of the International Common Disease Alliance (2021). Responsible use of polygenic risk scores in the clinic: Potential benefits, risks and gaps. *Nature Medicine*, *27*(11), 1876–1884. doi:10.1038/s41591-021-01549-6
- Pictor, M., Teare, H. J. A., & Kaye, J. (2018). Equitable participation in biobanks: The risks and benefits of a 'dynamic consent' approach. *Frontiers in Public Health*, *6*, 253. doi:10.3389/fpubh.2018.00253
- The Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke, S., Walters, J. T. R., & O'Donovan, M. C. (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *Nature*, *604*(7906), 502–508. doi:10.1101/2020.09.12.20192922
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. doi:10.1186/1471-2105-12-77
- Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, *22*(3), 278–295. doi:10.1177/0962280210395740
- Sharp, S. A., Rich, S. S., Wood, A. R., Jones, S. E., Beaumont, R. N., Harrison, J. W., ... Oram, R. A. (2019). Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. *Diabetes Care*, *42*(2), 200–207. doi:10.2337/dc18-1785
- Smoller, J. W. (2018). The use of electronic health records for psychiatric phenotyping and genomics. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics*, *177*(7), 601–612. doi:10.1002/ajmg.b.32548
- Swanson, J. M. (2012). [Review of *The UK Biobank and selection bias*]. *The Lancet*, *380*(9837), 110. doi:10.1016/S0140-6736(12)61179-9
- Thawethai, T., Arterburn, D. E., Coleman, K. J., & Haneuse, S. (2021). Robust inference when combining inverse-probability weighting and multiple imputation to address missing data with application to an electronic health records-based study of bariatric surgery. *The Annals of Applied Statistics*, *15*(1), 126–147. doi:10.1214/20-AOAS1386
- Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, *16*(4), 385–395. doi:10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

- Tyrrell, J., Zheng, J., Beaumont, R., Hinton, K., Richardson, T. G., Wood, A. R., ... Tilling, K. (2021). Genetic predictors of participation in optional components of UK Biobank. *Nature Communications*, *12*(1), 886. doi:10.1038/s41467-021-21073-y
- van Alten, S., Domingue, B. W., Galama, T., & Marees, A. T. (2022). Reweighting the UK Biobank to reflect its underlying sampling population substantially reduces pervasive selection bias due to volunteering (p. 2022.05.16.22275048). doi:10.1101/2022.05.16.22275048
- Wei, W.-Q., Bastarache, L. A., Carroll, R. J., Marlo, J. E., Osterman, T. J., Gamazon, E. R., ... Denny, J. C. (2017). Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS ONE*, *12*(7), e0175508. doi:10.1371/journal.pone.0175508
- Wiesner, G. L., Kulchak Rahm, A., Appelbaum, P., Aufox, S., Bland, S. T., Blout, C. L., ... Leppig, K. A. (2020). Returning results in the genomic era: Initial experiences of the eMERGE network. *Journal of Personalized Medicine*, *10*(2). doi:10.3390/jpm10020030
- Zheutlin, A. B., Dennis, J., Karlsson Linnér, R., Moscatti, A., Restrepo, N., Straub, P., ... Smoller, J. W. (2019). Penetrance and pleiotropy of polygenic risk scores for schizophrenia in 106 160 patients across four health care systems. *The American Journal of Psychiatry*, *176*(10), 846–855. doi:10.1176/appi.ajp.2019.18091085
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. doi:10.1198/016214506000000735