**Title:** Impact of Selection Bias on Polygenic Risk Score Estimates in Healthcare Settings

**Authors:**

Younga Heather Lee, PhD[1,3,4]

Tanayott Thaweethai, PhD[2,4]

Yi-han Sheu, MD, MPH, ScD[1,3,4,7]

Yen-Chen Anne Feng, ScD[5]

Elizabeth W. Karlson, MD[4,6]

Tian Ge, PhD[1,3,4,7]

Peter Kraft, PhD[8,9]

Jordan W. Smoller, MD, ScD[1,3,4,7]

**Affiliations:**

[1] Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA

[2] Biostatistics Center, Massachusetts General Hospital, Boston, Massachusetts, USA

[3] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[4] Harvard Medical School, Boston, Massachusetts, USA

[5] Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University

[6] Division of Rheumatology, Immunity, and Inflammation, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA

[7] Center for Precision Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA

[8] Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

[9] Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

**Correspondence:**

Jordan W. Smoller (jsmoller@mgh.harvard.edu)

**Word count (main text):** 4,130/4,500 max

**ABSTRACT (shortened version; word count: 249/250 max)**

**Background:** Hospital-based biobanks have become an increasingly prominent resource for evaluating the clinical impact of disease-related polygenic risk scores (PRS). However, biobank cohorts typically rely on selection of volunteers who may differ systematically from non-participants.

**Methods:** PRS weights for schizophrenia, bipolar disorder, and depression were derived using summary statistics from the largest available genomic studies. These PRS were then calculated in a sample of 24,153 European ancestry participants in the Mass General Brigham (MGB) Biobank. To correct for selection bias, we fitted a model with inverse probability (IP) weights estimated using 1,839 sociodemographic and clinical features extracted from electronic health records (EHRs) of eligible MGB patients. Finally, we tested the utility of a modular specification of the IP weight model for selection.

**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%) in the unweighted analysis to 28.9% (25.8%-31.9%) after IP weighting. Modular correction for selection bias in intermediate selection steps did not substantially impact PRS effect estimates.

**Conclusions:** Non-random selection of participants into volunteer biobanks may induce clinically relevant selection bias that could impact implementation of PRS and risk communication in clinical practice. As efforts to integrate PRS in medical practice expand, recognition and mitigation of these biases should be considered.

## 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) (1–6) and, more recently, in psychiatric conditions (7). 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 (8,9). 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 (10–12). With respect to returning psychiatric PRS results, individuals living with bipolar disorder (BD) were highly accepting of polygenic risk information for BD—even more so when accompanied by comprehensive psychiatric genetic counseling (13).

With the prospect of using PRS to guide clinical decision making, optimizing the accuracy of the risk estimates they provide becomes especially important (14). 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. 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 (15). 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 (16). 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 (17–19), especially when working with high-dimensional electronic health records (EHRs) (20). 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, healthcare utilization, and clinical characteristics (captured in the longitudinal EHRs) and estimate IP weights for selection. 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 (see e**Figure 1**). Haneuse and Daniels proposed a general statistical framework that explicitly models the decisions made by patients and healthcare providers that collectively determine which data are available in the EHR relevant to a given research inquiry (20). They conceptualize selection bias as a missing data problem and encourage researchers to *modularize* the complex selection mechanism into a series of sub-mechanisms that are easier to characterize and model (21). Applying this modular IP weighting framework, we evaluate the discrepancy between PRS effect estimates for psychiatric conditions when using standard versus modular approaches to defining selection mechanisms.

## 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 HealthCare hospital system (formerly Partners HealthCare) including Brigham & Women's Hospital, Massachusetts General Hospital, and other affiliated hospitals in the greater Boston area. To assemble the cohort for this study, we queried the MGB RPDR for 1,546,440 patients who self-identified as non-Hispanic White (that is, 74% 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 (22,23) (see **Figure 1**). The race and ethnicity restriction was applied here because the subsequent PRS 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 (12). 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 more than 130,000 participants, collected 82,092 DNA samples, and generated genotyping microarray data for more than 56,923 participants (4,920 using the Illumina MEGA, 5,334 using the Illumina MEGA EX, 26,144 using the Illumina MEG, and 24,789 using the Illumina GSA) (23). This research was conducted as part of the PsycheMERGE Consortium (24), under approval from the MGB Institutional Review Board.

### Data-driven approach to specify IP weight models for selection

We employed a machine learning modeling method—extreme gradient boosting (XGBoost) classification (25)—to empirically identify key determinants of non-random sampling of biobank participants and calculate the IP selection weights using a large set of demographic and clinical features extracted from high-dimensional EHRs including 15 sociodemographic, 10 healthcare utilization, and 1,814 diagnostic characteristics (see **eTable 4** for the full list of features used to train XGBoost models). XGBoost is an open-source library providing 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 (7).

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 vs. not included in the PRS analysis for the standard IP weighting approach). 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 (8).

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 **eFigure 1**). 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 weighting approach and performed weighted PRS analyses.

In addition, we applied a game theory-based algorithm called Shapley Additive Explanations (SHAP) method (https://github.com/slundberg/shap) to further elucidate the complex selection mechanism of the MGB Biobank. 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 direction (26). 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 rankings and directionality of contribution by the top 20 features in **eFigures 2 and 3**).

### PRS construction

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 (27), a Bayesian polygenic prediction method, to publicly available summary statistics from the largest genome-wide association studies (GWAS) of schizophrenia (28), bipolar disorder (29), and depression (30) on populations of European ancestry (see **eMethods** for details on genomic data processing and **eTable 2** for further information on discovery GWAS).

### Case definition

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 (31,32). 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 qualifying ICD codes for a given phecode (see the full list of qualifying diagnostic codes in **eTable 3**).

### Statistical analysis

We compared effect estimates of the associations between schizophrenia, bipolar disorder, and depression PRS and their respective target diagnoses using three approaches: unweighted, standard IP-weighted, and modular IP-weighted as described below (see **eFigure 1**). In the unweighted approach, PRS effect estimates are calculated without accounting for non-random sampling. 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. 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 weight model (see **Figure 2**). Prior to calculation of penetrance and discrimination of the PRS, 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 (33).

We first evaluated the impact of IP-weighting on the penetrance (i.e., case prevalence as a function of PRS) by comparing the weighted case prevalence against the unweighted case prevalence (34). Next, we evaluated the discrimination of the PRS using the area under the receiver operator characteristic curve (hereafter, the AUC) (35). 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. We then we calculated the AUC to compare performance of the unweighted and IP-weighted logistic regression models (36). Lastly, we estimated the discrimination of psychiatric PRS using the same method in subsamples defined by sex assigned at birth and current age to explore potential effect modification by these factors.

## 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 **eTable 1**). Consistent with their higher frequency of healthcare interactions, individuals included in the Biobank 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., 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 for schizophrenia, bipolar disorder, and depression in the analytic sample of 24,153 MGB Biobank participants were 1.0% ($n_{case}$=236), 4.5% ($n_{case}$=1,079), and 26.2% ($n_{case}$=6,329), respectively.

**Identification of key determinants of selection in the MGB Biobank**

In the XGBoost model under the standard IP weighting 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 informative features that differentiated those included and those not included in the PRS analysis, followed by treatment at Northshore Medical Center or Newton-Wellesley Hospital and median neighborhood income in 2010 (see **eFigure 2a**). The top features indicative of healthcare utilization from the standard IP weighting approach also appeared in the three XGBoost models under the modular IP weighting 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, treatment history at Faulkner Hospital, and rheumatoid arthritis and other inflammatory polyarthropathies (see **eFigures 2b-d**).

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 weighting approach. This was motivated in part by prior work showing that standard IP weighting can lead to biased estimates when a given feature plays a different role in each step of a sequential selection procedure (21,37,38). 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 selection probabilities across the three steps (see **eFigure 3b-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. This underscores the incomplete information captured by standard IP weighting when there are factors that impact selection probabilities differently across multiple phases of selection.

## 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 **Figure 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 (see **Figure 3b**). 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 IP weighting (see **Figure 3c**). Results using modular IP weighting based on intermediate selection steps were similar to those observed with standard IP weighting (see **eTables 5-7**).

### *Discrimination of psychiatric PRS*

We found the largest impact of IP-weighting on discrimination with respect to schizophrenia relative to bipolar disorder and depression (see **eTable 8**). When stratified by sex assigned at birth, AUC were generally higher among male participants than female participants regardless of the weighting scheme (see **eFigure 6a)**. 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 direction of the impact of IP weighting varied by age, especially for schizophrenia (see **eFigure 6b**). For example, among participants whose age was less than 40 years, the AUC of schizophrenia PRS from the unweighted IP-weighted 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

As interest grows in returning PRS-based risk estimates to participants in both research and clinical settings, the robustness of such estimates becomes increasingly important. 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, healthcare utilization, and clinical 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" (39–43), 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 the utility of an efficient machine learning algorithm to identify demographic and clinical variables that are significantly associated with selection in biobank PRS analyses and to adjust for selection bias.

Using IP weighting procedures, we found that selection bias can produce meaningful effects on estimates of penetrance and discrimination 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. Using the example of a bipolar disorder PRS, **Figure 2** depicts a causal diagram (directed acyclic graph) to illustrate how selection bias might inflate PRS effect estimates in hospital-based biobanks, such as the MGB Biobank. 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 (44). 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 direction 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 (43,45). 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 diseases, 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 selection mechanisms. For instance, Haneuse and Daniels (20,21) conceptualized selection bias as a missing data problem and encouraged researchers to modularize complex selection mechanisms into a series of sub-mechanisms that are easier to characterize and model. In the current study, we adapted this statistical framework to accommodate the selection procedures unique to PRS analyses conducted in hospital-based biobanks, though results of modular IP weighting did not differ substantially from standard IP weighting in our sample. As an alternative, Goldstein and colleagues (46) proposed controlling for the number of healthcare encounters. 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 (47). 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 population-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 EHR (Epic) system to 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 results should be interpreted in light of several limitations. First, our approach does not address another threat to the validity of PRS risk estimates implemented in healthcare settings—the distributional mismatch between the sample in which the PRS is trained and the samples in which the PRS is being validated or implemented. Such estimates are typically derived from validation samples to which allelic weights found in an independent discovery GWAS are applied. To the extent that the discovery and validation samples differ from the implementation sample (here, the MGB biobank) on a range of factors (e.g., age, sex,

socioeconomic status) that can affect PRS penetrance, risk estimates may be miscalibrated (48). Second, we examined only three psychiatric traits and our results suggest that the impact of selection bias will vary across different clinical conditions. Lastly, the current investigation was limited to subjects with self-reported white race. Thus, further investigation and validation in other ancestry populations as well as in non-psychiatric conditions are necessary to evaluate the generalizability of our results.

In conclusion, our analyses demonstrate a novel approach for detecting and accounting for unrecognized selection bias in polygenic risk estimation in hospital-based biobank samples. With the growing interest in the return of genomic risk information in clinical practice, it will be important to address such biases to avoid adverse impacts on clinical practice and patient outcomes.

## 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.

YAF is supported by the National Taiwan University Higher Education Sprout Project (NTU-110L8810) within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. EWK was supported by 5U01HG008685. TG was supported in part by NIA R00AG054573, NHGRI U01HG008685 and NHGRI U01HG011723. JWS was supported in part by NIMH R01MH118233, NHGRI U01HG008685, and a gift from the Demarest Lloyd, Jr. Foundation.

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.

## DISCLOSURES

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

1. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, et al. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. N Engl J Med. 2016 Dec 15;375(24):2349–58.

2. Sharp SA, Rich SS, Wood AR, Jones SE, Beaumont RN, Harrison JW, et al. Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. Diabetes Care. 2019 Feb 1;42(2):200–7.

3. Läll K, Mägi R, Morris A, Metspalu A, Fischer K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. Genet Med. 2017 Mar;19(3):322–9.

4. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet. 2019 Jan 3;104(1):21–34.

5. Pashayan N, Pharoah PDP, Schleutker J, Talala K, Tammela TLJ, Määttänen L, et al. Reducing overdiagnosis by polygenic risk-stratified screening: findings from the Finnish section of the ERSPC. Br J Cancer. 2015 Aug 20;113(7):1086–93.

6. Khera AV, Chaffin M, Wade KH, Zahid S, Brancale J, Xia R, et al. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. Cell. 2019 Apr 18;177(3):587-596.e9.

7. Murray GK, Lin T, Austin J, McGrath JJ, Hickie IB, Wray NR. Could Polygenic Risk Scores Be Useful in Psychiatry?: A Review. JAMA Psychiatry. 2021 Feb 1;78(2):210–9.

8. Electronic Medical Records and Genomics (eMERGE) Network [Internet]. [cited 2021 Apr 29]. Available from: https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE

9. Wiesner GL, Kulchak Rahm A, Appelbaum P, Aufox S, Bland ST, Blout CL, et al. Returning Results in the Genomic Era: Initial Experiences of the eMERGE Network. J Pers Med [Internet]. 2020 Apr 27;10(2). Available from: http://dx.doi.org/10.3390/jpm10020030

10. Pet DB, Holm IA, Williams JL, Myers MF, Novak LL, Brothers KB, et al. Physicians' perspectives on receiving unsolicited genomic results. Genet Med. 2019 Feb;21(2):311–8.

11. Allen NL, Karlson EW, Malspeis S, Lu B, Seidman CE, Lehmann LS. Biobank participants' preferences for disclosure of genetic research results: perspectives from the OurGenes, OurHealth, OurCommunity project. Mayo Clin Proc. 2014 Jun;89(6):738–46.

12. Karlson EW, Boutin NT, Hoffnagle AG, Allen NL. Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. J Pers Med [Internet]. 2016 Jan 14;6(1). Available from: http://dx.doi.org/10.3390/jpm6010002

13. Putt S, Yanes T, Meiser B, Kaur R, Fullerton JM, Barlow-Stewart K, et al. Exploration of experiences with and understanding of polygenic risk scores for bipolar disorder. J Affect Disord. 2020 Mar 15;265:342–50.

14. Polygenic Risk Score Task Force of the International Common Disease Alliance. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. Nat Med [Internet]. 2021 Nov 15; Available from: http://dx.doi.org/10.1038/s41591-021-01549-6

15. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013 Jun;22(3):278–95.

16. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol. 2008 Sep 15;168(6):656–64.

17. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997 Feb 28;16(4):385–95.

18. Zou H. The Adaptive Lasso and Its Oracle Properties. J Am Stat Assoc. 2006 Dec 1;101(476):1418–29.

19. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. J R Stat Soc Series B Stat Methodol. 2008 Jan 4;70(1):53–71.

20. Haneuse S, Daniels M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why? EGEMS (Wash DC). 2016 Aug 31;4(1):1203.

21. Haneuse S, Arterburn D, Daniels MJ. Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task. JAMA Netw Open. 2021 Feb 1;4(2):e210184.

22. Bayramli I, Castro V, Barak-Corren Y, Madsen EM, Nock MK, Smoller JW, et al. Temporally informed random forests for suicide risk prediction. J Am Med Inform Assoc. 2021 Dec 28;29(1):62–71.

23. Castro VM, Gainer V, Wattanasin N, Benoit B, Cagan A, Ghosh B, et al. The Mass General Brigham Biobank Portal: an i2b2-based data repository linking disparate and high-dimensional patient data to support multimodal analytics. J Am Med Inform Assoc [Internet]. 2021 Nov 28; Available from: http://dx.doi.org/10.1093/jamia/ocab264

24. Smoller JW. The use of electronic health records for psychiatric phenotyping and genomics. Am J Med Genet B Neuropsychiatr Genet. 2018 Oct;177(7):601–12.

25. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 785–94. (KDD '16).

26. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions [Internet]. arXiv [cs.AI]. 2017. Available from: http://arxiv.org/abs/1705.07874

27. Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019 Apr 16;10(1):1776.

28. The Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke S, Walters JTR, O'Donovan MC. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia [Internet]. bioRxiv. medRxiv; 2020. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.09.12.20192922

29. Mullins N, Forstner AJ, O'Connell KS, Coombes B, Coleman JRI, Qiao Z, et al. Genome-wide association study of over 40,000 bipolar disorder cases provides new insights into the underlying biology [Internet]. bioRxiv. medRxiv; 2020. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.09.17.20187054

30. Howard DM, Folkersen L, Coleman JRI, Adams MJ, Glanville K, Werge T, et al. Genetic stratification of depression in UK Biobank. Transl Psychiatry. 2020 May 24;10(1):163.

31. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS One. 2017 Jul 7;12(7):e0175508.

32. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics. 2014 Aug 15;30(16):2375–6.

33. Lewis CM, Vassos E. Prospects for using risk scores in polygenic medicine. Genome Med. 2017 Nov 13;9(1):96.

34. survey: Analysis of Complex Survey Samples [Internet]. Comprehensive R Archive Network (CRAN). [cited 2022 Feb 14]. Available from: https://cran.r-project.org/web/packages/survey/index.html

35. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011 Mar 17;12:77.

36. Mangiafico S. Functions to Support Extension Education Program Evaluation [R package rcompanion version 2.4.13]. 2022 Jan 3 [cited 2022 Feb 14]; Available from: https://CRAN.R-project.org/package=rcompanion

37. Thaweethai T, Arterburn DE, Coleman KJ, Haneuse S. 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. aoas. 2021 Mar;15(1):126–47.

38. Peskoe SB, Arterburn D, Coleman KJ, Herrinton LJ, Daniels MJ, Haneuse S. Adjusting for selection bias due to missing data in electronic health records-based research. Stat Methods Med Res. 2021 Oct;30(10):2221–38.

39. Andreeva VA, Salanave B, Castetbon K, Deschamps V, Vernay M, Kesse-Guyot E, et al. Comparison of the sociodemographic characteristics of the large NutriNet-Santé e-cohort with French Census data: the issue of volunteer bias revisited. J Epidemiol Community Health. 2015 Sep;69(9):893–8.

40. Struijk EA, May AM, Beulens JWJ, van Gils CH, Monninkhof EM, van der Schouw YT, et al. Mortality and cancer incidence in the EPIC-NL cohort: impact of the healthy volunteer effect. Eur J Public Health. 2014 Apr 15;25(1):144–9.

41. Brown WJ, Bryson L, Byles JE, Dobson AJ, Lee C, Mishra G, et al. Women's Health Australia: recruitment for a national longitudinal cohort study. Women Health. 1998;28(1):23–40.

42. Mishra GD, Hockey R, Powers J, Loxton D, Tooth L, Rowlands I, et al. Recruitment via the Internet and social networking sites: the 1989-1995 cohort of the Australian Longitudinal Study on Women's Health. J Med Internet Res. 2014 Dec 15;16(12):e279.

43. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol. 2017 Nov 1;186(9):1026–34.

44. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615–25.

45. van Alten S, Domingue BW, Galama T, Marees AT. Reweighting the UK Biobank to reflect its underlying sampling population substantially reduces pervasive selection bias due to volunteering. medRxiv. 2022 May 16;2022.05.16.22275048.

46. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record. Am J Epidemiol. 2016 Dec 1;184(11):847–55.

47. Beesley LJ, Mukherjee B. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. Biometrics [Internet]. 2020 Nov 12; Available from: http://dx.doi.org/10.1111/biom.13400

48. Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M. Variable prediction accuracy of polygenic scores within an ancestry group. Elife [Internet]. 2020 Jan 30;9. Available from: http://dx.doi.org/10.7554/eLife.48376

## TABLES

**Table 1.** Comparison of demographic and healthcare utilization characteristics of patients self-identifying as non-Hispanic White in the overall MGB patient population[†] against those included in the PRS analysis (shown in number of participants and prevalence of a given condition).

| | | Overall patient population[†] (N=1,546,440) | Not included (N=1,522,287) | Included (N=24,153) | p-value |
|---|---|---|---|---|---|
| **Sociodemographic characteristics, N (%)** | | | | | |
| Mean age (SD) | | 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) | 96322 (6.3) | 2,401 (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) | 1,707 (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) | 1,773 (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 | 7,395 (0.5) | 7,206 (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) | 6,022 (24.9) | |
| | Widowed | 83,703 (5.4) | 82,469 (5.4) | 1234 (5.1) | |
| ***Healthcare utilization, mean (SD)*** | | | | | |
| Visit count | | 73.71 (114.9) | 71.34 (110.5) | 223.27 (229.3) | <0.001 |
| ICD code count | | 185.1 (332.7) | 178.25 (317.2) | 615.95 (743.7) | <0.001 |
| CPT code count | | 140.63 (255.2) | 135.38 (244.5) | 471.15 (537.9) | <0.001 |
| Note count | | 360.70 (566.7) | 349.81 (545.8) | 1,047.00 (1,145.2) | <0.001 |

Abbreviation(s): ICD, International Statistical Classification of Diseases and Related Health Problems; CPT, Current Procedural Terminology (CPT).

[†] 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 **Figure 1**).

## FIGURES

**Figure 1.** Schematic of sample curation for polygenic risk score analysis using the MGB Biobank sample.

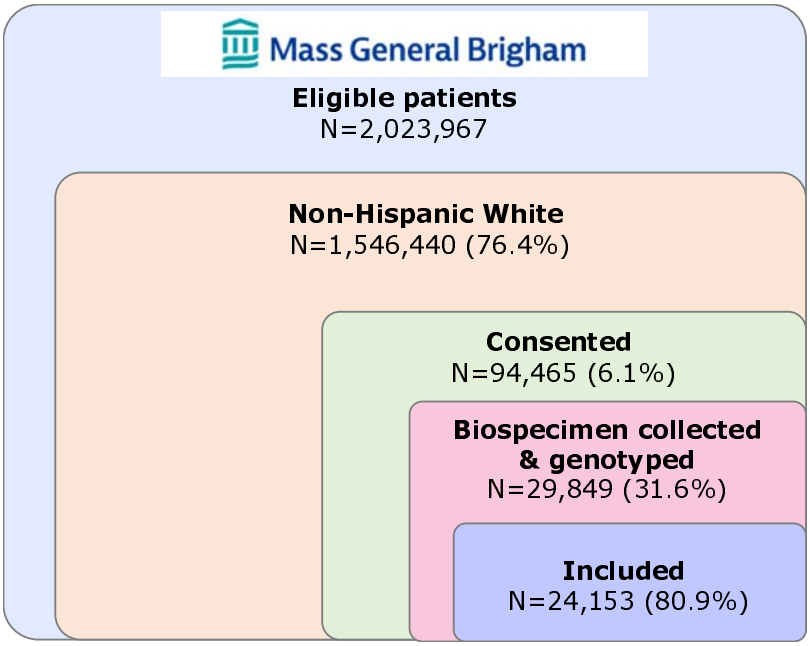

**Figure 2.** Causal diagram (directed acyclic graph or DAG) illustrating how non-random sampling into hospital-based biobanks may introduce bias in PRS estimation.

Using the example of a bipolar disorder PRS, this figure depicts two DAGs to illustrate how selection bias might inflate PRS effect estimates in a hospital-based biobank in unweighted PRS analysis. The relationship of interest is denoted by the dotted line connecting $PRS_{BIP}$ (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 $PRS_{BIP}$ (through the effect of $PRS_{BIP}$ 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 below). Thus, 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. In contrast, when selection bias is accounted for using inverse probability (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 below). Thus, IP-weighted PRS estimates would likely represent effects through the causal path only.

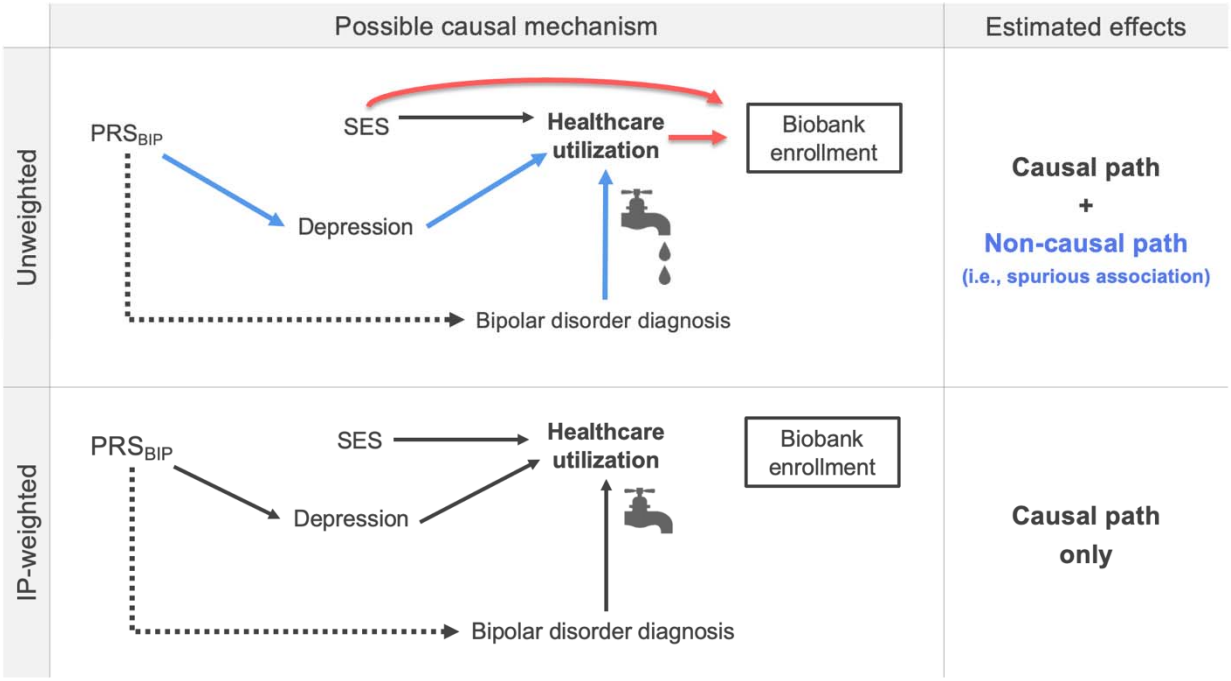

**Figure 3.** Case prevalence by polygenic risk score (PRS) decile for three psychiatric traits using two different weighting schemes—unweighted and modular IP-weighted.

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 **eTables 5** through **7**.