**Supplementary Online Content**

Lee YH, Thaweethai T, Sheu Y, Feng YCA, Karlson EW, Ge T, Kraft P, Smoller JW. Impact of Selection Bias on Polygenic Risk Score Estimates in Healthcare Settings.

**eMethod.** Supplemental methods.

**eFigure 1.** Feature importance of top 20 features from the modular IP weight models using the mean absolute Shapley values.

**eFigure 2.** SHAPbeeswarm plots showing directionality and magnitude of effects on the selection probabilities by the top 20 features from the standard and modular IP weight models.

**eFigure 3.** Case prevalence per deciles of standardized residuals of PRS, stratified by sex assigned at birth.

**eFigure 4.** Case prevalence per deciles of standardized residuals of PRS, stratified by current age.

**eTable 1-8 can be found in the accompanying Excel spreadsheets.**

This supplementary material has been provided by the authors to give readers additional information about their work.

**eMethods**

***Recruitment and consent process at the Mass General Brigham (MGB) Biobank.*** A small team established the foundational components for the Mass General Brigham (MGB) Biobank, including its protocols, governance, and procedures for conducting a baseline study, and set up the core systems required to manage a large-scale recruitment and biospecimen collection from 2008 through 2011. Initially, recruitment relied on in-person interactions with patients in clinical settings (e.g., clinic waiting rooms), but has since expanded through partnerships with other MGB studies and electronic consent via email invitations. From 2012 to 2020, some biospecimens were obtained as residual biospecimens from clinical blood draws (“clinical discards”), and the Biobank has been ordering research phlebotomies through the MGB electronic health record system (Epic, Epic Systems Corporation, Verona, WI) since 2018. Mass General Brigham (MGB) hospitals notify patients at registration that their data may be used for research with proper Institutional Review Board (IRB) approval, which was obtained here. Further details on the participant recruitment and consent process can be found in a recently published extensive discussion of the MGB Biobank (Boutin et al., 2022).

***Genomic data processing.*** The MGB Biobank samples were genotyped on Multi-Ethnic Global array (MEGA) from Illumina (Illumina Inc., San Diego, USA) and released in several batches. We performed batch-specific genotype data QC to remove single nucleotide polymorphisms (SNPs) with genotype missing rate >0.05, samples with genotype missing rate >0.02, and SNPs with differential missing rate >0.01 between any two batches, after which different batches were merged for subsequent QC steps. As MGB Biobank included individuals from diverse populations, we inferred genetic ancestry of biobank participants using 1000 Genomes samples (1KG) as the population reference panel (1000 Genomes Project Consortium et al., 2015). Specifically, we computed principal components (PCs) for biobank samples and 1KG samples combined and trained a random forest classifier to assign a “super population” label for biobank samples with a prediction probability ≥0.9 using the first 6 PCs of the 1KG samples as the training data. This resulted in 26,677 individuals whose ancestry was classified as European (EUR), 1,607 as African (AFR), 1,840 as Admixed American (AMR), 504 as East Asian (EAS) and 297 as South Asian (SAS) ancestry. Within each ancestry, we removed samples with a mismatched reported and genetic sex, outliers of the absolute value of heterozygosity (>5 standard deviations from the mean), and one from each pair of related individuals (identity-by-descent (IBD) >0.2); SNPs that showed significant batch associations at P < 1 × 10−4, had a missing rate > 0.02 or Hardy–Weinberg equilibrium (HWE) test P < 1 × 10−10 were also discarded. Next, we used the Michigan Imputation Server (Minimac4) to impute genotype dosages for biobank samples, with the Haplotype Reference Consortium (HRC) as the reference panel for EUR ancestry. Lastly, we removed markers with imputation quality INFO score <0.8, minor allele frequency (MAF) <0.01, a significant deviation from HWE with P < 1 × 10−10, and missing rate >0.02. The dataset uses genome build 37 (hg19). Further information about genotyping, QC, imputation, and population assignment procedures for the MGB Biobank is available on the GitHub repository (<https://github.com/Annefeng/PBK-QC-pipeline>).

***Construction of Bayesian polygenic risk scores (PRS).*** We generated PRS for the European ancestry participants of the MGB Biobank using PRS-CS-Auto (Ge, Chen, Ni, Feng, & Smoller, 2019), a Bayesian polygenic prediction method, based on their genotype data and publicly available summary statistics from the largest, European ancestry genome-wide association studies (GWAS) of schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), bipolar disorder (Mullins et al., 2021), and depression (Howard et al., 2019). PRS-CS-Auto places a continuous shrinkage (CS) prior on SNP effect sizes and infers posterior SNP weights using GWAS summary statistics and an external linkage disequilibrium (LD) reference panel (e.g., 1000 Genomes Project European Samples). Allowing multivariate modeling of local LD patterns, PRS-CS-Auto is robust to diverse underlying genetic architectures and can increase the accuracy of PRS over conventional approaches (Ge et al., 2019). We generated PRS for each individual by summing all risk-associated variants weighted by their posterior effect size estimates inferred by PRS-CS-Auto using PLINK, version 2.0 (Chang et al., 2015). In addition to the continuous PRS, we calculated deciles of the standardized residuals of PRS after adjusting for top 10 ancestry PCs, sex assigned at birth, age, and genotyping microarray.

***Curation of input features for inverse probability (IP) weight models.*** The inverse probability (IP) weight models included three main types of features: sociodemographic, clinical, and healthcare utilization characteristics. The proportion of missing observations for sociodemographic variables varied from 0.0004% to 3.3%, with the exception of veteran status which was 18.1% among the 1,546,440 non-Hispanic White patients eligible for the MGB Biobank study at the time of their first visit to the MGB-affiliated hospitals. To address missingness in sociodemographic features, we imputed the median household income using the population median and kept missing observations as a separate category for those who selected "unknown" categories for sociodemographic characteristics such as gender, veteran status, and marital status. For clinical features, we utilized the phecode system (Wu et al., 2019) to group 57,665 unique ICD codes into 1,814 hierarchical phenotype codes, which allowed us to minimize the number of low-frequency codes without applying an arbitrary threshold based on case prevalence. Cases were defined as those having at least two phecodes for a given outcome occurring on different dates, and we made the strong assumption that having no evidence of a condition was considered as not having the condition. While acknowledging that this assumption may lead to undercounting due to false negatives, we do not expect it to bias the overall findings of the manuscript unless this form of missingness is systematically different for the biobank versus the underlying patient population. Finally, it is worth noting that healthcare utilization features (e.g., code count, visit count) cannot be entirely missing, as participants with entirely missing healthcare utilization features would not be eligible for the study population.

**eReferences**

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. doi:10.1038/nature15393

Boutin, N. T., Schecter, 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

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7. doi:10.1186/s13742-015-0047-8

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

Howard, D. M., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., Shirali, 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

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 more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nature Genetics*, *53*(6), 817–829. doi:10.1038/s41588-021-00857-4

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421–427. doi:10.1038/nature13595

Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., … Wei, W.-Q. (2019). Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Medical Informatics*, *7*(4), e14325. doi:10.2196/14325

**eFigure 1.** Evaluation of the marginal contribution by top 20 features from the standard and modular IP weight models based on Shapley values. Features having higher mean absolute Shapley values would have a more significant impact on the model’s decision than those having lower values. On the vertical axis, features are rank-sorted based on the magnitude of the mean absolute Shapley values from high (top) to low (bottom).

1. **Modular IPW** –Step 1: Pr(Consented|Eligible)**A picture containing chart

   Description automatically generated**
2. **Modular IPW** – Step 2: Pr(Biospecimen genotyped|Eligible,Consented)**Chart

   Description automatically generated**
3. **Modular IPW** – Step 3: Pr(Included|Eligible,Consented,Biospecimen genotyped)**Chart

   Description automatically generated**

**eFigure 2**. SHAP beeswarm plot showing directionality of effects for the top 20 features in the corresponding IP weight models. The features are rank-sorted in the same order as in the eFigure 2 with the `visit count` feature having the highest mean absolute Shapley values (i.e., most important feature in the standard IP-weighted model). The position of dots on the horizontal axis is determined by the Shapley value of that feature (averaged across all participants), and dots pile up along each feature row to show density. The color of dots is used to display the original value of a feature (i.e., higher feature values would be redder).

1. **A picture containing chart

   Description automatically generatedStandard IP weighting approach** – Pr(Included|Eligible)

|  |
| --- |
| Visit count |
| Notes count |
| Current age |
| Massachusetts General Hospital |
| Brigham & Women’s Hospital |
| North Shore Medical Center |
| Newton-Wellesley Hospital |
| Median neighborhood income (2010) |
| Unknown veteran status |
| Essential hypertension |
| McLean Hospital |
| Acute upper respiratory infection |
| Faulkner Hospital |
| Back pain |
| Depression |
| Other tests |
| Other features combined |

1. **A picture containing chart

   Description automatically generatedModular IP weighting approach** – Step 1: Pr(Consented|Eligible)

|  |
| --- |
| Visit count |
| Current age |
| Brigham & Women’s Hospital |
| Massachusetts General Hospital |
| North Shore Medical Center |
| Unknown veteran status |
| Notes count |
| Other tests |
| Public insurance |
| Median neighborhood income (2010) |
| Depression |
| Newton-Wellesley Hospital |
| Female gender |
| Acute upper respiratory infection |
| Other features combined |

1. **Chart

   Description automatically generated with medium confidenceModular IP weighting approach** – Step 2: Pr(Biospecimen genotyped| Eligible,Consented)

|  |
| --- |
| Current age |
| Median neighborhood income (2010) |
| Visit count |
| Newton-Wellesley Hospital |
| Massachusetts General Hospital |
| Other tests |
| Brigham & Women’s Hospital |
| Hyperplasia of prostate |
| Essential hypertension |
| Skin cancer |
| Mass Eye and Ear |
| Notes count |
| Cardiac dysrhythmias |
| Degenerative skin conditions and other dermatoses |
| Other features combined |

1. **Chart

   Description automatically generated with medium confidenceModular IP weighting approach** – Step 3: Pr(Included|Eligible,Consented, Biospecimen genotyped)

|  |
| --- |
| Notes count |
| Visit count |
| Massachusetts General Hospital |
| Essential hypertension |
| Newton-Wellesley Hospital |
| Ischemic heart disease |
| North Shore Medical Center |
| Current age |
| Faulkner Hospital |
| Disorders of lipoid metabolism |
| Brigham & Women’s Hospital |
| Secondary malignant neoplasm |
| Rheumatoid arthritis and other inflammatory polyarthropathies |
| Hyperlipidemia |
| Other features combined |

**eFigure 3.** Case prevalence by polygenic risk scores (PRS) decile for three psychiatric traits using two different weighting schemes, stratified by sex assigned at birth. PRS were adjusted for potential confounding by population stratification, sex, age, and genotyping microarray (see **eTable 7a-c** for numeric estimates used to generate the respective figure). The solid lines indicate point estimates, and the bands indicate 95% confidence intervals for corresponding point estimates.

1. **Schizophrenia**

**Chart, line chart

Description automatically generated**

1. **Bipolar disorder**

**Chart, line chart

Description automatically generated**

1. **Depression**

**Chart, line chart

Description automatically generated**

**eFigure 4.** Case prevalence by polygenic risk scores (PRS) decile for three psychiatric traits using two different weighting schemes, stratified by current age. PRS were adjusted for potential confounding by population stratification, sex, age, and genotyping microarray (see **eTable 7a-c** for numeric estimates used to generate the respective figure). The solid lines indicate point estimates, and the bands indicate 95% confidence intervals for corresponding point estimates.

1. **Schizophrenia**

**Chart, line chart

Description automatically generated**

1. **Bipolar disorder**

**Chart, line chart

Description automatically generated**

1. **Depression**

**Chart, line chart

Description automatically generated**