Accounting for Selection Bias and Missing Data for Inferential Questions in Electronic Health Record-Linked Biobanks

by

Maxwell Mayer Salvatore

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Epidemiological Science)
in the University of Michigan
2024

Doctoral Committee:

Professor Bhramar Mukherjee, Co-Chair
Professor Celeste Leigh Pearce, Co-Chair
Professor Christopher R. Friese
Associate Professor Alison Mondul

Maxwell Salvatore

mmsalva@umich.edu

ORCiD ID: 0000-0002-3659-1514

Dedication

For my wife, Kristen, whose unwavering support, patience, and goofiness are my foundation; our perfect, precious dog, Sebastian; my brother, Geno; my parents and first teachers, Paula and Lou; and for all those who strive to make their corner of the world a better place.

Acknowledgments

This dissertation would not have been possible without the support of my family, colleagues, and professors.

First, I would like to thank my co-chair, Dr. Celeste Leigh Pearce. Leigh, you took a chance on me and took me under your wing. You have shared the highs and, perhaps more importantly, the lows of science. Our tennis chats kept me grounded when I struggled to see the light at the end of the tunnel. You are a credit to our field and have shown me the dedication it takes to be an impactful epidemiologist. I will carry the supportive spirit and understanding of others that I have learned from working with you into my professional and personal relationships.

I would also like to thank my co-chair, Dr. Bhramar Mukherjee. Bhramar, words will not do justice to your impact on my science and development. You supported my curiosity as a staff researcher and pushed me to pursue a doctoral degree. The world is better because of your thoughtfulness, integrity, vision, leadership, and patience. Your unwavering commitment to rigorous, impactful studies, especially while the world frantically sought answers to the raging COVID-19 pandemic, is admirable and inspiring. I aspire to conduct myself as a scientist and mentor with the same compassion and dedication you have shown me and all your students.

I would also like to thank my committee members, Drs. Christopher R. Friese and Alison M. Mondul, and Dr. David Hanauer, a frequent collaborator. I am grateful to have supportive mentors like yourselves who have provided unique perspectives and insight.

The supportive community and camaraderie of the Pearce and Mukherjee labs and my PhD cohort have kept my spirits up throughout this journey. Thank you to my brilliant peers for inspiring me, particularly Deesha Bhaumik, Lilah Khoja, Ritoban Kundu, and Jiacong Du. The most special thanks go to Drs. Lauren Beesley and Lars Fritsche, alumni of the Center for Precision Health Data Science, who treated me with kindness and respect and inspired me to pursue a doctoral degree in the first place.

My dissertation would not have been possible without the faculty at the University of Michigan. Your guidance and support have been invaluable to me. Thank you to Dr. Mark Wilson, my MPH cohort advisor; Dr. Rafael Meza, my MPH mentor and PhD cohort advisor; Dr. Sara Adar, my Epidemiology doctoral program chair; and Drs. Hal Morgenstern, Sung Kyun Park, Kelly Bakulski, and Lindsay Kobayashi, teachers of my favorite courses.

And a special thank you to my family, who have shown me unconditional love and support: my wife, Kristen, and our dog-son, Sebastian; my brother, Geno; my best friend, Aroh; and Carolyn and Steve, who might as well be my aunt and uncle. And, perhaps most importantly, my parents, Paula and Lou. And to my other family and friends who are too numerous to name here - I am blessed to have the best family.

I have had the privilege of being supported financially through various grants and fellowships. This support has been instrumental in my academic journey, allowing me to focus on my research and studies. I thank *Rogel Cancer Center Training, Education, and*

Career Development and the *Rackham Graduate School* for providing fellowship funding during my PhD. Additionally, I thank my co-chairs for providing me with graduate student research assistant funding through various sources, including the Michigan Cancer and Research on the Environment Study (MI-CARES), and Dr. Kelly Bakulski for providing me with a graduate student instructor position. It is due to your generosity that I have been able to grow, learn, and research, and can now make lifelong contributions to public health of my own.

Table of Contents

List of Tables

List of Figures

Abstract

This dissertation explored two major sources of systematic errors, namely, selection bias and missing data in electronic health record (EHR)-linked biobank research. EHR-linked biobanks are a tremendous resource for answering questions of public health and clinical significance. However, they are non-probability samples wrinkled with multiple sources of bias. Aim 1 explored the impact of selection weights on the potential for reducing selection bias for four common analyses (prevalence, dimensionality, and association estimation, and large-scale hypothesis testing) across three EHR-linked biobanks with different recruitment mechanisms: the NIH All of Us Research Program (AOU; n=244,071), the University of Michigan's Michigan Genomics Initiative (MGI; n=81,243), and the UK Biobank (UKB; n=401,167). In the US-based cohorts (AOU and MGI), inverse probability and poststratification selection weights were derived using National Health Interview Survey data to reflect the US adult population. Findings highlighted the importance of selection weights, especially when estimating prevalences and associations, and underscored the need for biobanks to disclose their recruitment and selection processes.

Aim 2 compared six approaches to constructing phenotype risk score (PheRS) in MGI for three digestive cancer diagnoses with no definitive screening tools currently available: esophageal, liver, and pancreatic. We assessed whether weighted approaches enhanced performance in the external evaluation cohort, AOU. No single PheRS approach uniformly performed better in terms of risk stratification, though elastic net and random forest tended to exhibit good properties. Additionally, in no setting did using weights meaningfully or consistently improve PheRS risk stratification performance. Notably, the results for liver cancer suggest that agnostic EHR-based approaches toward early detection have promise. Our findings suggest that EHR-linked biobank researchers should consider using health history summarized as PheRS, which contributed to risk stratification alongside other domains, including covariates, risk factors, and presenting symptoms, in risk prediction and stratification. The use of weights, however, did not conclusively alter the transferability of the risk prediction models.

Aim 3 explored the joint impacts of missing data and selection bias in EHR-linked biobanks using polygenic risk scores (PRS). Leveraging relatively complete genetic information available in EHR-linked biobanks, we compared three missing data methods: complete case analysis, multiple imputation without using exposure and outcome PRS (woPRS-imputed), and multiple imputation with exposure and outcome PRS (PRS-imputed). We evaluated association estimation performance with simulated data generated as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Our simulation analyses considered random and biased sampling of data with missingness in (a) exposure only and (b) both exposure and outcome. We found that weighting biased sample data was crucial to reducing bias in association estimation and recovering coverage rates. PRS-imputed analyses also had better association estimation properties than standard woPRS-imputed analyses, particularly for MAR data. These findings highlight the need to use sampling weights

representative of the intended target population to address selection bias and PRS to address missing data.

This dissertation has the potential to guide the application of principled methods that are appropriately applied to imperfect EHR-linked biobank data, thereby improving the quality of analyses, inference, and their impact on data-driven decision-making for relevant target populations. The role of selection bias differs across inferential goals and that studying multiple sources of bias, such as selection bias and missing data, together is critical for biobank analysis, a burgeoning field in epidemiology and clinical research.

Chapter 1 Introduction

1.1 Overview of the dissertation

This dissertation was focused on addressing the methodological concerns of selection bias and missing data in electronic health record (EHR)-linked biobanks.

- Aim 1 focused on the use of selection weights in common data tasks using EHR-linked biobank data. EHR-linked biobanks often do not represent a group of people about whom researchers are interested in drawing conclusions. This aim involved developing selection weights that can be used to correct for the lack of representativeness in EHR-linked biobanks to draw research conclusions that are generalizable to the target population.

- Aim 2 focused on comparing methods for risk prediction model development and whether weighted-based methods improved performance of prediction rules when they are transferred across EHR-linked biobanks. EHR-linked biobanks have multiple domains of data, including sociodemographic information, diagnosis codes, laboratory results, imaging data, and genotype data. We compared six risk prediction methods and explored whether weighting methods improved the performance of risk prediction models for three digestive cancers developed using diagnosis code data in the EHR. We compared the performance of the unweighted and weighted approaches alongside models using covariates, risk factors, and symptoms.

- Aim 3 jointly focused on missing data and selection bias in the EHR-linked biobanks. Exposure data, such as smoking and drinking status, body mass index, and glucose, are often incomplete or missing in EHR data. We explored (a) whether using largely complete genotype data available in EHR-linked biobanks reduced bias due to missing data using multiple imputation and (b) the joint impacts of missing data and selection bias on association analyses.

This dissertation includes six chapters. Chapter 1 introduces the dissertation and the three aims. Chapter 2 provides background and a critical evaluation of the literature on EHR-linked biobanks, methodological concerns of selection bias and missing data, and risk prediction models, three main themes that are explored in this dissertation. Chapters 3-5 present the scientific content relevant to the three aims. Chapter 6 is the conclusion, which summarizes the significance and public health relevance of the key findings for each of the aims and describes directions for future research.

1.2 Aim 1: Exploring the impact of selection weights on commonly conducted analyses in EHR-linked biobanks

EHR-linked biobanks link EHR data like diagnosis codes, procedures, laboratory results, and imaging data to genetic data and potentially other domains, including self-reported survey data, death registry data, residential history, and neighborhood-level characteristics.[1] EHR-linked biobanks, including All of Us (AOU)[2] (n > 760,000) and the UK Biobank (UKB)[3] (n > 500,000), are increasing in size and number and are publicly available to researchers. However, these cohorts are non-probability samples and are likely not representative of their target populations.[2,4]

Despite aiming to be population-based, the lack of representativeness in these EHR-linked biobanks induces selection bias and may lead to conclusions that are not generalizable.[5–7] Ideally, selection bias would be mitigated in the study design phase.[6] However, EHR-linked biobank administrators often do not (or cannot) take steps to prevent selection bias[2,4,8,9] in pursuit of large sample sizes. Further, many of the users of these data are not involved in data collection and study design.

There are analytic approaches to reduce the impact of selection bias; the most common are weighting-based methods.[5–7,10] Recently, Beesley and Mukherjee developed theory to address selection bias in EHR-linked biobanks.[7,11] Van Alten and colleagues estimated selection weights for the UK Biobank.[8] However, many EHR-linked biobanks do not provide selection weights for researchers, and it is unclear how important selection weights are in different analyses.

Aim 1: To explore the role of selection bias adjustment by weighting EHR-linked biobank data for commonly performed analyses.

- *Sub-aim 1.1:* To estimate inverse probability and poststratification weights in two US-based EHR-linked biobanks – AOU and MGI – to make them representative of the US adult population and to apply previously estimated weights in UKB to make it representative of the UKB-eligible population.
- *Sub-aim 1.2:* To compare demographics, EHR characteristics, and the diagnostic phenomes of three EHR-linked biobanks: AOU, MGI, and UKB.

- *Sub-aim 1.3:* To evaluate whether selection weights meaningfully change and improve descriptive (prevalence and latent dimension estimation) and analytic (association estimation and large-scale, agnostic hypothesis testing) tasks in the same three EHR-linked biobanks.

1.3 Aim 2: Exploring the impact of selection weights on the development of risk prediction models

Risk prediction and stratification is a hallmark of population and precision health.[12–15] A classic example is the Framingham Risk Score, which identifies individuals at high 10-year risk for cardiovascular disease and are candidates for drug interventions to reduce their risk.[14] While models, scores, and calculators for many diseases exist, computational and statistical advances in the analysis of multi-modal Big Data are transforming the promise of precision health into reality.[16]

Risk prediction methods include traditional regression, regularized regression, machine learning, ensemble methods, and pruning-and-thresholding in polygenic risk score (PRS) development. Some methods, like regularized regression and random forest, can accommodate high-dimensional data (like those found in EHR-linked biobanks) and involve tuning hyperparameters prior to model fitting. Recently, Iparragirre and colleagues developed theory and an R package to tune the λ hyperparameter in lasso models with complex survey weights, which can be modified to accommodate selection weights for several types of models.[17] Weighting-based methods can make analytic samples more representative of the target population (e.g., an external sample). Lack of transferability, a related problem, occurs when there are differences in the data distributions between internal and external samples, resulting in a reduction in model performance in the external sample.

Because EHR-linked biobanks contain multiple domains of data, they present an opportunity to compare risk prediction performance and combine information across domains. Recently, Salvatore and colleagues developed a phenotype risk score (PheRS) for pancreatic cancer using diagnosis data available in the EHR and found it contributed to risk prediction alongside covariate, risk factor, and genetic data.[18] Building on this work, different PheRS development approaches and the impact of weights on risk prediction model performance in an external validation cohort (i.e., transferability) are explored in Aim 2.

Aim 2: To determine optimal PheRS development approaches and whether weighting-based approaches improve transferability across EHR-linked biobanks.

- *Sub-aim 2.1:* To modify an existing R package to accommodate weighted hyperparameter tuning for regularized regression and random forest models.
- *Sub-aim 2.2:* To determine if there is an optimal PheRS development approach.
- *Sub-aim 2.3:* To determine whether weights improve the performance of diagnosis code-based risk prediction models in an external validation cohort.
- *Sub-aim 2.4:* To evaluate whether PheRS meaningfully improve risk stratification alongside models developed using covariate, risk factor, and symptom data.

1.4 Aim 3: Using genotype data to inform imputation of missing non-genetic exposure data

EHR-linked biobank data are subject to multiple concurrent sources of bias, like missing data and selection bias; however, these biases are often studied in isolation.

Missing data is a foundational problem in statistics[19] and a common practical problem in epidemiologic research.[20,21] Analyses restricted to individuals with non-missing data (i.e., complete case analyses) are commonplace[21] but can lead to biased conclusions if data missingness is informative (and dependent on unmeasured factors).[22] Observations in EHR data are further complicated by clinically informative visiting processes and EHR fragmentation (where different providers collect data on the same person).[23–28] Other analytic approaches to analyzing data with missing values are more robust, like inverse probability weighting and multiple imputation.[20,29–32]

Multiple imputation is a commonly used and easily implemented approach to reducing bias when analyzing data with missing values.[20,33–35] It fills in missing values by drawing from distributions informed by relationships between observed variables.[32] Imputation of medical and EHR data has received significant attention, though often without genetic data.[36–39]

Genetic data is often summarized as a PRS to predict health-related outcomes, most commonly disease risk.[40–42] Recently, Ma and colleagues developed exposure PRS (ExPRS) and found that the relationships between these ExPRS mirrored the relationships seen in the raw exposure data.[43] Concurrently, Li, Chen, and Moore found that using genetic information improved imputation in missing EHR data.[44] These findings present an opportunity to explore whether genotype-informed imputation of non-genetic data reduces biases in exposure-outcome association analyses.

At the same time, recruitment mechanisms (e.g., recruitment through specific clinics[45] or oversampling of underrepresented groups[2]) and participant-driven (e.g., healthy volunteers[4]) factors impact EHR-linked biobank participation leading to selection bias.

Aim 3: To evaluate (a) the degree to which genotype-informed multiple imputation of missing data and (b) selection weighting of the internal sample mitigate the joint impacts of missing data and selection bias and lead to bias reduction in association analyses using EHR-linked biobank data.

1.5 Objective

This dissertation explores methodological issues related to selection bias and missing data in common EHR-linked biobank analyses. As a broader range of researchers analyze EHR-linked biobank data, we aim to provide practical guidance on when and what methods researchers should implement to perform meaningful and impactful research. In total, this dissertation informs analytic approaches designed to achieve results that are readily applicable to intended target populations, reducing the time between study and the translation of its results among the people whom research is intended to benefit.

Chapter 2 Background

2.1 Promise and perils of electronic health record-linked biobanks

Electronic health record (EHR)-linked biobanks are repositories with biospecimen and/or related data linked to EHR and other forms of auxiliary data (e.g., medical and pharmacy claims, residential-level neighborhood characteristics; Section 2.2).[1,7,11,46] Notable examples include the National Institutes of Health All of Us Research Program[2] (AOU; beginning in 2018) and the UK Biobank[3] (UKB; starting in 2006), while efforts like the Global Biobank Meta-analysis Initiative (GMBI)[47] are fostering collaborations and meta-analyses across biobanks. Alongside improvements in biospecimen analysis and computational methods and capabilities, EHR-linked biobanks are rapidly growing in size and number (Beesley and colleagues (2020) highlighted 21 major EHR-linked biobanks[1] – only a fraction of those globally).

Because of their size, data linkage capabilities, immediacy, and accessibility, EHR-linked biobanks are attractive to clinical and health researchers. Central administration of these cohorts means researchers can avoid devoting resources like time (e.g., obtaining IRB approval for primary human subjects' data collection, writing grant applications for funding, managing study team personnel, data collection administration) and money towards primary data collection and instead focus on analyzing data and publishing findings. The wealth of and benefits of EHR-linked biobank data have not gone unnoticed;

PubMed citations dealing with EHR and EHR data have more than tripled in the last decade from 3,212 citations in 2013 to 9,824 in 2023.

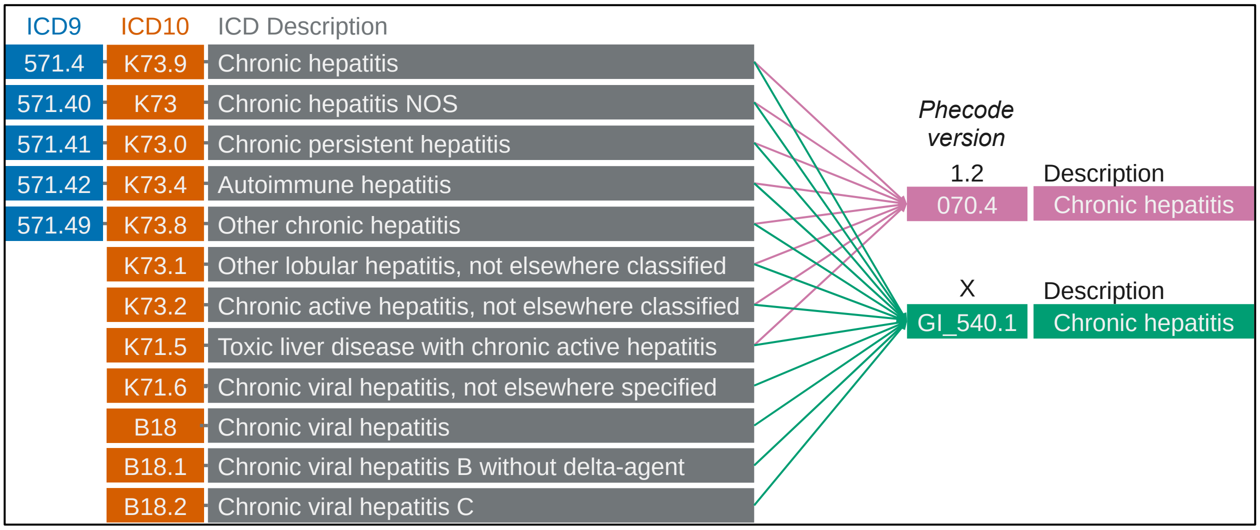
However, while careful study design before primary data collection can mitigate data quality issues, this is not always possible in EHR-linked biobanks. There are methodological issues researchers need to be aware of and critically consider before diving into analyses and drawing conclusions. Traditional concerns include selection bias (Section 2.3), missing data (Section 2.4), confounding,[48,49] and misclassification.[7,50,51] In contrast, other concerns arise like false (a) confidence in results simply due to large sample size and (b) discoveries (controlling false discovery rate due to multiple testing[52,53]), target validity,[54] and informed presence/absence bias[1,25,26,55,56] (other concerns in Section 2.5). Sections 0, 2.4, and 2.5 describe the current literature regarding these concerns. Section 2.6 discusses risk prediction model development, its application to EHR-linked biobank data, and challenges. Section 2.7 summarizes the dissertation motivation based on the background provided in the preceding sections.

2.2 Electronic health records: what are they?

EHRs, acting as massive longitudinal cohorts with passively collected data, hold immense potential for medical research. These data include structured metrics like diagnosis codes (International Classification of Disease [ICD]) and unstructured data like doctors' notes and imaging data. Structured data, with their standardized nomenclature, are readily accessible to researchers. Unstructured data, with the growth of processing and analysis methods like natural language processing, machine learning, and artificial intelligence, will only continue to augment and complement existing structured data. Significant efforts, like the Observational Medical Outcomes Partnership Common Data

Model[57] and SNOMED Clinical Terminology,[58] map information from several clinical sources and lexicons to a standard concept set for domains including drugs, conditions, procedures, and measurements, further increasing accessibility and interoperability.

Diagnoses captured as ICD codes during health care encounters serve as a foundation data domain in this dissertation. They are maintained by the World Health Organization, originally designed to facilitate the international comparability of morbidity and mortality data.[59,60] ICD codes are, as mentioned in Section 2.2, structured data, and the tens of thousands of alphanumeric codes are grouped into 22 chapters (in ICD 10th Revisions, or ICD-10; e.g., Chapter II: Neoplasms, Chapter XI: Diseases of the digestive system) and can be nested (e.g., C25: Malignant neoplasm of the pancreas; C25.0: Malignant neoplasm of head of pancreas). Moreover, as these codesets are periodically updated, codes and their usage are also updated (e.g., ICD-9 was used until ICD-10 came into effect in 2015). Because these codesets are standardized and used globally, they are attractive for use in research. However, the granularity encoded in ICD codes may not be helpful in research contexts, and changing codesets over time can hamper data harmonization. Realizing these limitations, a team at Vanderbilt University developed phecodes, which map both ICD-9 and ICD-10 codes into broader yet clinically meaningful phenotypes (the latest version, phecode X, defines 3,612 phenotypes; Figure 2.1), focused on common diagnoses to facilitate genome-wide association studies.[61–63]



ICD9	ICD10	ICD Description
571.4	K73.9	Chronic hepatitis
571.40	K73	Chronic hepatitis NOS
571.41	K73.0	Chronic persistent hepatitis
571.42	K73.4	Autoimmune hepatitis
571.49	K73.8	Other chronic hepatitis
	K73.1	Other lobular hepatitis, not elsewhere classified
	K73.2	Chronic active hepatitis, not elsewhere classified
	K71.5	Toxic liver disease with chronic active hepatitis
	K71.6	Chronic viral hepatitis, not elsewhere specified
	B18	Chronic viral hepatitis
	B18.1	Chronic viral hepatitis B without delta-agent
	B18.2	Chronic viral hepatitis C

Figure 2.1 Example mapping International Classification of Disease (ICD) 9th and 10th edition codes to phecode version 1.2 and phecode X codes for chronic hepatitis. Note the mapping for phecode X is *not* exhaustive (there are 23 unique ICD codes, only 17 are shown). Mapping tables can be found at https://phewascatalog.org. Phecodes are described in Denny and colleagues,[62] Bastarache,[63] and Shuey and colleagues.[61]

The first application of phecodes was to conduct disease-single nucleotide polymorphism (SNP) phenome-wide association studies (PheWAS) by Denny and colleagues in 2010, successfully reproducing 5 of 7 known disease-SNP associations and identifying 19 previously unidentified associations.[62] PheWAS have also been conducted on other types of genetic data besides SNPs, including gene expression levels and functional genetic variants. We have previously used phecodes to conduct a phenotype-phenotype PheWAS to identify associated diagnoses and construct a phenotype risk score (PheRS) for pancreatic cancer.[18] While phecodes are beneficial for large-scale, agnostic hypothesis testing and -omics-wide association studies, other high-throughput phenotyping algorithms exist (e.g., PheNorm[64] and PheCAP,[65] both of which also incorporate narrative notes) and curated (e.g., PheKB[66]) phenotyping algorithms which incorporate information across EHR data domains, including lab results and clinical notes, can be preferable in targeted analysis settings.

Moreover, EHR data are increasingly linked with biospecimens and associated genetic information, creating EHR-linked biobanks. These biobanks are also connected to other data sources, such as administrative claims data, vital status records, cancer registries, neighborhood-level characteristics, and self-report survey data (Figure 2.2).[67] The multi-modal nature of these data provides a unique opportunity to explore the relative importance and cumulative contribution of each domain in studies on a range of health-related questions, including association analyses and large-scale agnostic hypothesis testing (e.g., phenome-wide association studies, PheWAS),[68] risk prediction and stratification,[13,18,47,69] treatment response,[70–72] and time-to-event outcomes.[73–75]

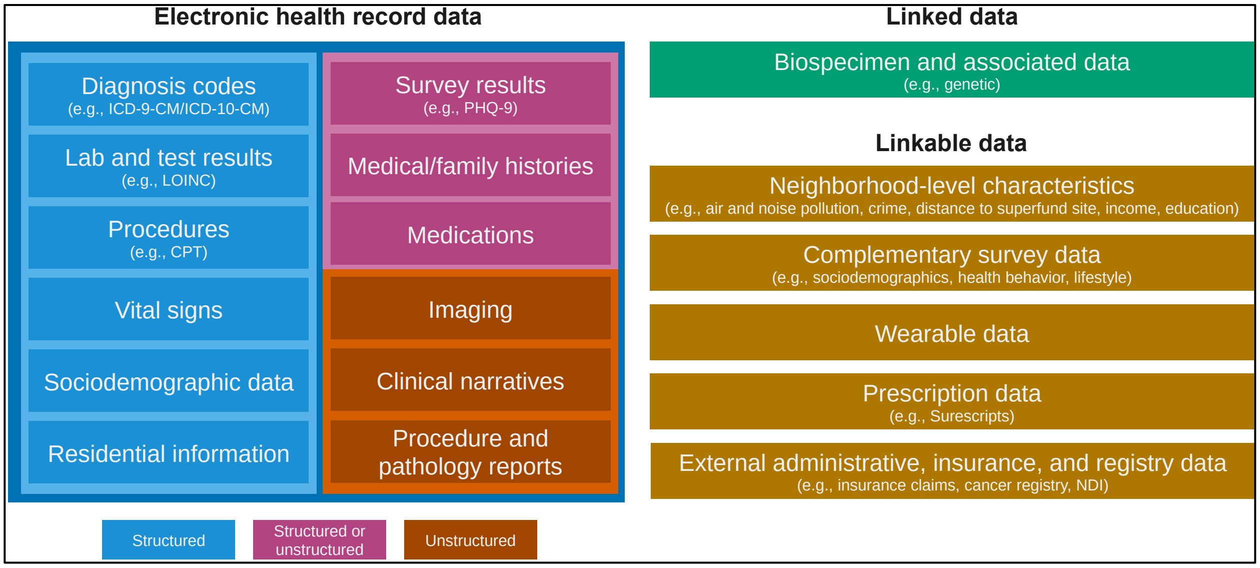

Electronic health record data
Diagnosis codes
(e.g., ICD-9-CM/ICD-10-CM)
Lab and test results
(e.g., LOINC)
Procedures
(e.g., CPT)
Vital signs
Sociodemographic data
Residential information
Survey results
(e.g., PHQ-9)
Medical/family histories
Medications
Imaging
Clinical narratives
Procedure and pathology reports
Linked data
Biospecimen and associated data
(e.g., genetic)
Linkable data
Neighborhood-level characteristics
(e.g., air and noise pollution, crime, distance to superfund site, income, education)
Complementary survey data
(e.g., sociodemographics, health behavior, lifestyle)
Wearable data
Prescription data
(e.g., Surescripts)
External administrative, insurance, and registry data
(e.g., insurance claims, cancer registry, NDI)
Structured
Structured or unstructured
Unstructured

Figure 2.2. Schematic representation of structured and unstructured electronic health record (EHR) data alongside linked and linkable data in EHR-linked biobanks.

2.3 Selection bias: who do your data represent?

Selection bias and cohort representativeness[7,11,24,76–79] are areas receiving significant attention – and for good reason. Large sample sizes and rich data, along with some EHR-linked biobank cohorts like the UKB being labeled as “population-based,” may distract researchers from considering the representativeness of the underlying data. For example, work has shown that UKB data are not representative of the general UK or UKB-eligible population,[8,80] and not accounting for this can lead to invalid estimates.

Selection bias is a distortion arising from a lack of representativeness in the study sample with respect to a population of interest (i.e., the target population, which sometimes is the source population).[7,81,82] The implication is that the result of a data task – whether it be estimating a prevalence or association, testing a hypothesis, or predicting an outcome – is not expected to align with the truth (e.g., a population parameter).[5,6,83] This theoretical concern is increasingly a practical problem given the rise of non-probabilistic samples of Big Data (e.g., web surveys, social media). Specifically, EHR-linked biobanks are appealing to researchers because they have large sample sizes, contain rich multimodal data, and are publicly available for secondary data analysis (e.g., researchers are not responsible for resources related to data collection).[84–86]

What is the potential impact of selection bias in these cohorts, and what analytic tools do researchers have at their disposal to address the bias? Selection bias is particularly troublesome because the magnitude and direction of its impact are hard to determine,[87] its effect cannot be mitigated by increasing sample size,[88,89] and it can be coupled with other data imperfections, including outcome and exposure misclassification,[7,77] missing data,[76,77] and immortality bias.[90,91] Until recently, based on the belief that genetics (e.g., single nucleotide polymorphisms) are not related to selection, it has been argued that genetic and downstream analyses may not be meaningfully affected by selection bias.[92,93] However, there is now evidence that even genetic analyses are not immune to selection bias.[94–96] For example, Schoeler and colleagues (2023) found both over- and under-estimation of genetic associations with behavioral, lifestyle, and social (and, to a lesser extent, physical and molecular) outcomes because of participation bias in the UKB.[97] The use of weights resulted in the identification of novel single nucleotide polymorphism (SNP) associations for 12 traits, small change in heritability estimates (maximum change in h^2, 5%), and substantial discrepancies for genetic correlations (maximum change in r_g, 0.31) and Mendelian randomization estimates (maximum change in β_{STD}, 0.15) for socio-behavioral traits.

Epidemiology textbooks stress the importance of control selection and well-defined source populations in the study design phase (i.e., before recruitment and data collection) to mitigate selection bias.[5,6,98] However, it is often unavoidable;[5] EHR-linked biobanks are subject to selection biases because of healthy volunteer bias[4] and recruitment strategies,[2,45] and researchers must grapple with this bias. There are three common analytic approaches to adjust for selection bias – stratification, bias analysis, and inverse probability weighting.[5,6,99,100] Stratification is easier to perform than weighting-based approaches because it simply involves adjusting for factors related to selection; however, it will only yield unbiased results in limited settings (when selection on the descendent of a collider[5]) where selection bias is present. Bias analysis is used to estimate the potential magnitude and direction of biases and to quantify uncertainty about these biases to combat overconfidence in results and guide future research.[99] The most common analytic approach to handling selection bias is inverse probability (IP) weighting (IPW).[5,7,8,77,97,101,102] IPW involves reweighting individuals in the sample by the inverse of the estimated probability of their inclusion (relative to the target population) conditional on factors impacting selection.[7,10,11] The estimation of IP weights relies on (a) access to representative individual-level data from the target population (which is often unavailable) and (b) correct specification of the selection model. Representative data can be in the form of probability samples like the National Health Interview Survey (NHIS).[103] Poststratification (PS) weighting is an alternative that relies on summary-level data from the target population, which is often more available.[104] Methods for combining information from probability samples (like NHIS) and non-probability samples (like AOU, MGI, and UKB) have been pioneered in survey literature.[105,106]

Selection bias in EHR-linked biobanks has been examined in many studies. Some have simply proposed adjusting for factors related to selection, like referral status or clinic type.[55,56] However, selection factors are often unknown and require reviewing study recruitment protocols and critical thinking about possible participant-driven factors. Further, there are often multiple selection mechanisms. Haneuse and Daniels proposed modeling each selection mechanism separately,[76] an approach that demonstrated reduced bias in an EHR-based study to address this.[77] Theory for weighting-based approaches to reducing bias due to selection exist,[7,79] and have been applied to the UK Biobank.[8,97] Notably, despite often being studied in isolation, Beesley and Mukherjee derived theory to address selection bias and misclassification jointly.[7,11] More research on the joint and relative impacts of selection and other sources of bias is needed because multiple biases co-occur in practice.

2.4 Missing data

Missing data is a foundational topic of statistics[19] and a common issue in health research,[21] receiving significant attention generally and within the context of EHR data. Rubin’s seminal paper introduced the concept of three missing data classes based on the reason why the data are missing: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR or NMAR).[22] These classes help summarize instances when bias is expected and what methods might help address them.

If the probability of being missing is the same for all participants, then the data are considered MCAR. That is, the probability of being missing is unrelated to the data.[32] While information is lost (i.e., larger variances), analyses based on MCAR data are expected to remain unbiased. However, assuming missing data are MCAR is often unrealistic. For example, in EHR-linked biobanks, patient-provider interactions are not random (sicker patients are observed more frequently[23,55,56]) and a clinician's judgment concerning a patient's risk can prompt a test order (meaning the absence of a test order is clinically informative).[107,108] EHR fragmentation, where providers document the patient's interactions with their system, as in the US, can result in incomplete pictures of a patient's health history across time and type of encounter, further complicating missing EHR data. Observations can be unobserved for other reasons, including failure to initiate or complete an encounter, financial costs associated with testing and diagnosis,[109] underdiagnosis,[110] and differential disease classification processes.[111] If, however, the probability of being missing is the same within groups defined by observed data, then the data are MAR (i.e., missingness is random conditional on observed values). Complete case analyses of MAR data are expected to be biased. However, methods exist to mitigate bias due to MAR data. Finally, suppose the probability of being missing depends on unobserved values. In that case, observed data cannot explain missingness, and the data are considered MNAR. Complete case analyses of MNAR data are expected to be biased and generally rely on sensitivity analyses to understand the robustness of the results.

Two common methods for analyzing MAR data are multiple imputation and IPW. Multiple imputation relies on drawing multiple plausible values from distributions and relationships between observed variables in the data.[112] A statistical analysis is carried out on each imputed dataset, and the results are pooled using Rubin's rules.[112] Common issues for multiple imputation arise when the outcome variable is omitted in the imputation model, when variables are not normally distributed, when the MAR assumption is not plausible, and when the data are MNAR.[113] While it can be computationally expensive,[114] multiple imputation and related analyses are easily and flexibly carried out through available software.[35,115] Importantly, multiple draws of missing values retain variation and relationships between observed variables while simply analyzing the single most accurate prediction of the missing values results in too-small standard errors and false positives.[32] These are several high-quality introductory reviews[32,116].

Alternatively, IPW, where complete cases are weighted by the inverse of their probability of being a complete case, can be used to reduce bias due to missing data.[117] As discussed in Section 2.3, IPW can also be used to simultaneously address lack of sample representativeness and biases due to missing data.[117] Instead of relying on the distribution of missing values conditional on observed data (as in multiple imputation), IPW relies on correctly specifying a model for being a complete case.[117] Some methodologists suggest IPW is preferable to multiple imputation for the analysis of missing data for several reasons: (a) IPW is arguably easier to explain, (b) the distribution of missingness predictors is very different between complete and incomplete observations, and (c) when individuals with missing values are missing data for several variables rather than one or two.[117] Other methods for analyzing missing data exist, like full information maximum likelihood.[118,119] Several papers compare the performance of different missing data methods,[34,117,120,121] and, while different methods may be preferred in different settings, multiple imputation is generally more efficient.[117]

Missing data in the context of EHR has been discussed. Petersen and colleagues[122] and Li and colleagues[44] have discussed adaptations of multiple imputation for EHR-linked biobank data. For example, Li and colleagues exploited non-missing genotype data available in EHR-linked biobank data to improve imputation of cardiovascular-related measurements.[44] Haneuse and colleagues introduced a modularization approach to thinking about missing EHR data to facilitate their analysis.[123] Importantly, missing EHR data can be thought of as inducing selection bias.[77] Peskoe and colleagues applied a modularization approach for handling selection bias due to missing EHR data by estimating a series of IPW.[77] Relatedly, Beesley and colleagues conceptualized an individual's true phenotype as "missing" data.[1] In this case, individuals lacking a diagnosis could be erroneously considered a non-case. Beesley and Mukherjee proposed several likelihood-based approaches for handling misclassification in EHR, potentially due to underdiagnosis or lack of provider observation (i.e., missing diagnoses).[7] They also jointly considered multiple biases in EHR-linked biobank data, a commonly encountered issue in practice.[7,11] Section 2.5 discusses a related concept of informed presence/absence.

2.5 Other electronic health record data concerns

Beyond introductory papers,[1,124–128] substantial work has focused specifically on other traditional methodological concerns, including confounding[48,49] and misclassification,[7,50,51] in EHR-based cohorts (Figure 2.3). For example, traits defined using the phecode framework have demonstrated reduced misclassification compared to ICD codes.[129] One method to further reduce the impact of misclassification, described by Hubbard and colleagues, relies on EHR-derived probabilistic phenotyping.[50] Others have described methods using manual chart review on a subset of data to improve EHR-derived phenotypes.[51,130,131] Teixeira and colleagues explored the incorporation of unstructured data like doctors' notes, which improved the identification of hypertensive individuals compared to using ICD codes and blood pressure reading cutoffs alone.[132]

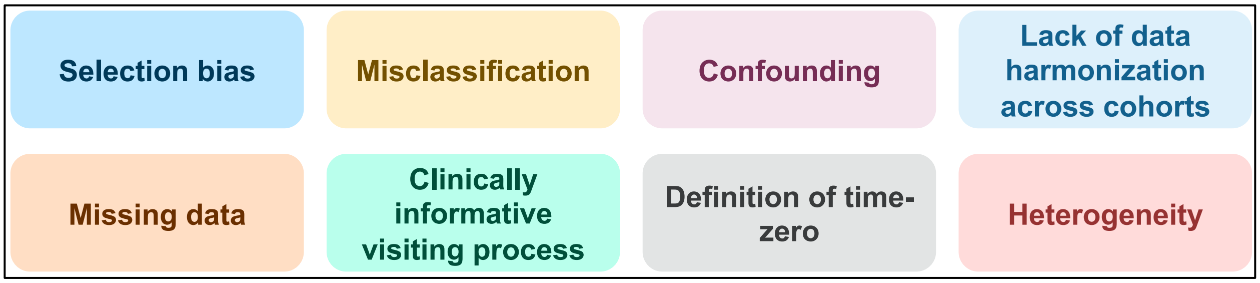


Figure 2.3 Common methodological concerns in the analysis of EHR-linked biobank data

Target validity is one consideration broadly applicable in health research but particularly acute in EHR-based analyses. Westreich and colleagues have defined this as a joint measure of internal and external validity of an effect estimate with respect to a specific target population.[54] Historically, internal validity, the notion that an estimate reflects the true underlying parameter in the study population, has taken precedence over external validity, that the parameter in the study population is representative of the true parameter in the target population. However, because of observation mechanisms and recruitment strategies into EHR-linked biobanks, the target population is almost certainly never (a) exactly the study sample or (b) the population of which the study sample is a simple random sample.[54] EHR researchers should think critically regarding whom the results are intended for or representative of before beginning an analysis and making their target populations explicit in their work. It is crucial for researchers to consider weighted approaches that account for both the observation and recruitment mechanisms in each cohort (including potential subcohorts) and differences in the distribution of key characteristics between the analytic cohort and the target population.

One unique concept in EHR is that of clinically informative observation processes, defined by Goldstein and colleagues as “the notion that inclusion in an EHR is not random but rather indicates that the subject is ill, making people in EHRs systematically different from those not in EHRs.”[55] This discrepancy harms generalizability to general populations who tend to be healthier than those in the EHR data sample and results in bias. This concept extends to individuals within the EHR – those who are sicker tend to have more encounters and records than those who are healthier[23,24] – and, in some cases, to records in the EHR (e.g., lab results). This phenomenon is illustrated by Agniel and colleagues, which shows that the presence and timing of laboratory results were more informative than the value of the laboratory results themselves.[133] Interested readers can learn more about informed presence elsewhere.[1,25,55,56,134] Including EHR metadata, like the length of follow-up, number of encounters, the density of laboratory measurements, and visit type (e.g., outpatient vs. inpatient vs. emergency), and careful selection or matching of controls in analyses are recommended to improve exchangeability and attempt to make EHR observation mechanisms comparable.

Finally, a topic relevant to Big Data in general and EHR-linked biobank data specifically is the “Big Data Paradox.”[135] Xiao-Li Meng eloquently proposed a decomposition of mean bias into three parts: (a) data quality, (b) data quantity, and (c) problem difficulty.[135] Conceptually, the Big Data Paradox characterizes the idea that systematic error (e.g., selection bias, information bias, confounding bias; Figure 2.4) is not mitigated by increasing sample size as random error is. This paradox demands the thorough exploration and thoughtful application of methods to reduce biases associated with systematic error (described in Sections 2.3, 2.4, and 2.5).

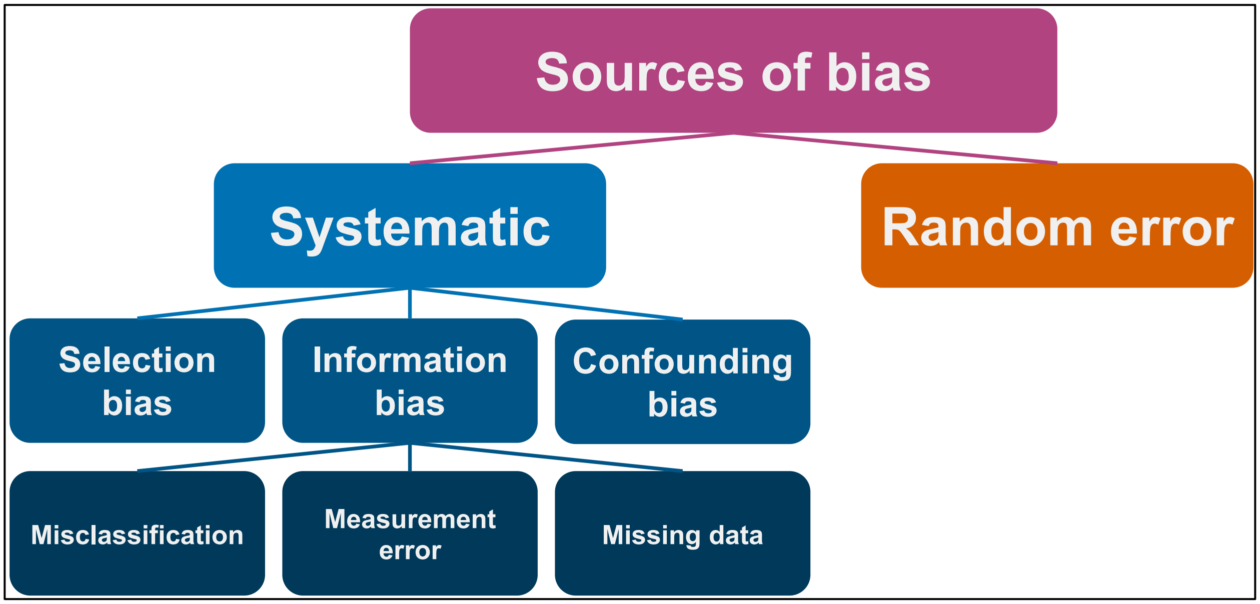


Figure 2.4 Flowchart depicting systematic and random errors.

2.6 Risk prediction: methods, challenges, and context

Risk prediction models, such as the widely recognized Framingham Risk Score,[14] are a fundamental tool in public health and precision medicine. These statistical models predict the likelihood of a health outcome, such as diagnosis, prognosis, or treatment response, to guide prevention, intervention, or treatment strategies.[136] Individuals with high scores, indicating high risk, are often recommended for preventive therapies like cholesterol-lowering statins.[137] Numerous models, including those for cardiovascular disease[14,138,139] and cancer,[15,140–147] exist to enhance both the quantity and quality of life.

Many methods for developing risk prediction models exist (Figure 2.5). Conventional linear, logistic, and Cox regression models are staples of an epidemiologist’s toolbox. Regularized regression methods, like ridge, lasso, and elastic net, which penalize the magnitude of coefficients, are increasingly used because they can perform variable selection (as in lasso) and handle multicollinearity (as in ridge).[148] Machine learning methods, like decision trees, support vector machines, and neural networks, trade interpretability for flexibility and performance.[149] Ensemble methods like bagging,[150] random forest,[151] boosting,[152] and SuperLearner,[153] combine results from multiple models are designed with the goal of achieving even better performance.[153,154]

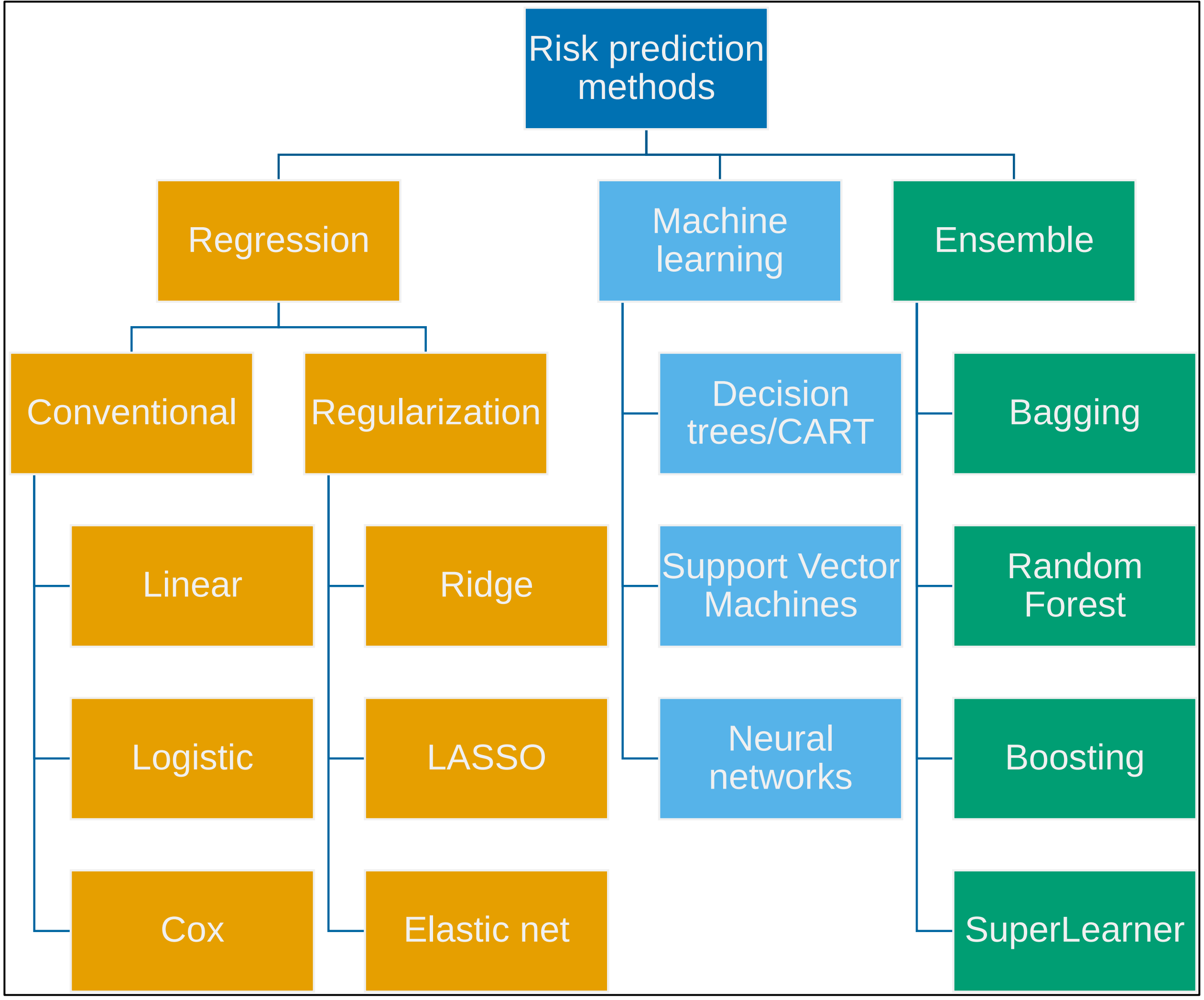


Figure 2.5 Schematic representation of some common risk prediction model methods.

EHR-linked biobanks present a promising avenue for expediting the development of risk prediction models. Their appeal lies in their large sample size, real-time, real-world clinical data access, and linkage to multiple data domains (e.g., genetics, administrative and insurance claims, complementary survey data; Section 2.2). Indeed, several risk prediction models, including those for cardiovascular disease,[155–157] cancer,[18,158,159] and COVID-19 diagnosis and outcomes,[160–162] have been developed using EHR-linked biobank data.

However, despite the numerous benefits of EHR-linked biobank-based risk prediction, several concerns remain. First, there is the issue of privacy and data sharing. EHR and genetic data are subject to institutional and government regulations. Institutional Review Boards limit the sharing of identifiable information in research involving human subjects and ensure compliance with federal rules.[163] Governments protect and restrict the sharing of protected health information (like EHR data; e.g., Health Insurance Portability and Accountability Act of 1996 (HIPAA) in the US[164]). These restrictions make it challenging to aggregate fragmented EHR into an individual's complete health history and, thus, to generate robust and generalizable predictions. Beyond attempts to deidentify or obfuscate data before sharing and analysis, federated learning has been developed to minimize data sharing.[165] Federated methods allow multiple institutions to collaboratively perform analyses without sharing their data, generally via an iterative process.[166] One example is where intermediate summary-level data are prepared and broadcast by a central site to other sites before synthesizing aggregated data (Figure 2.6). Many federated learning methods have been developed, like those for linear mixed,[167] logistic,[168–170] and Cox[171,172] regression. While iterative sharing of aggregated data makes their implementation challenging, these newer methods rely on few to no iterations, enhancing their use in practice.

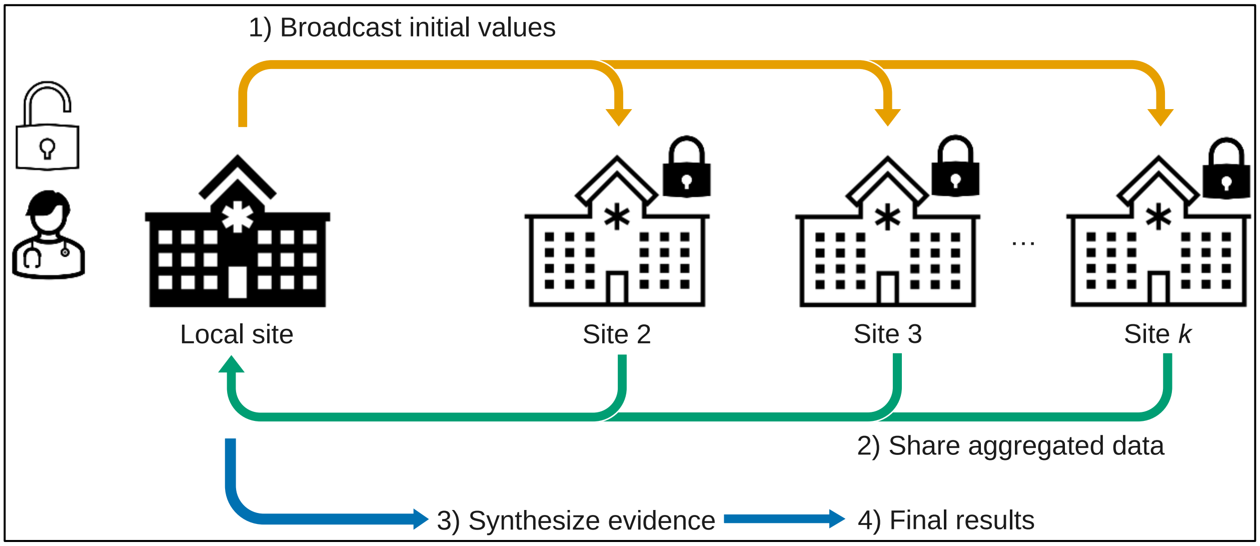


Figure 2.6 Schematic representation of a broadcasting federated learning approach. Figure adapted from Privacy-preserving Distributed Algorithms (https://pdamethods.org/).

Second, there are issues when there are differences between a sample used to develop a risk prediction model (e.g., an EHR-linked biobank like MGI) and its application to a target population of interest. When the sample and target population have different data distributions (i.e., different "case-mix"[136]), the model may not apply to the target population, a problem called lack of transferability. Transfer learning is a class of methods that adapt existing models for use in a new population.[173,174] Computer science literature has discussed issues of covariate shift and domain adaptation.[175–179] The related problem of lack of representativeness, called selection bias (Section 2.3), is most commonly dealt with using weighting-based methods. Steingrimsson and colleagues developed an inverse-odds weighting-based approach to tailoring risk prediction models for use in an external target population where outcome data are unavailable.[180] Weighting-based approaches allow for consideration of the target population during model development rather than incorporating information into an existing model, as in transfer learning. While methods addressing issues of privacy and transferability exist and are actively being developed, their application to and assessment in EHR-linked biobanks is needed.

Further, the growing number of researchers using EHR-linked biobank data need practical recommendations regarding their implementation.

It is crucial for researchers to deeply understand the clinical context of a model's outcome and its application in the development of risk prediction models. Cancers, such as esophageal, liver, and pancreatic, exhibit variations in their clinical presentation, particularly at advanced stages, and their diagnostic approach, with early detection and screening largely unavailable for these three cancers. Esophageal, liver, and pancreatic cancers, which currently lack screening mechanisms, are often diagnosed at a late stage. Thirty-nine percent of esophageal, 20% of liver, and 51% of pancreatic cancers are diagnosed after the cancer has metastasized when the 5-year relative survival is 5.3%, 3.31%, and 3.1%, respectively (SEER-22, 2014-2020, all races, both sexes[181]).

Current risk prediction models for these cancers tend to focus on high-risk populations, such as those with chronic hepatitis B virus infections[182–185] or chronic liver disease[186–188] for liver cancer and those with new-onset diabetes for pancreatic cancer.[189,190] Other models aim to identify individuals to screen for premalignant conditions, as in the case of Barrett's esophagus prior to the transition to esophageal cancer.[191,192] Importantly, models incorporating biomarkers and genetic factors to construct integrated and multi-factorial models generally exhibit better performance.[193–195]

These models play a crucial role in guiding surveillance and monitoring strategies, such as abdominal ultrasonography and α-fetoprotein (AFP) tests for high-risk individuals for liver cancer,[196] or endoscopic ultrasonography and MRI for high-risk individuals for pancreatic cancer.[197] In the absence of screening mechanisms, incorporating multi-modal data, including biomarker and genetic factors alongside demographics, risk factors, and diagnostic history, can facilitate the development of risk prediction models in the general population to identify high-risk individuals for targeted enhanced surveillance and prevention measures. Moreover, each cancer has multiple histological types with different risk factors, clinical features, genetic susceptibility, and pathogenesis, such as squamous cell carcinoma and adenocarcinoma for esophageal cancer.[198] Data-driven agnostic approaches to risk prediction model development need to acknowledge the heterogeneity across and within cancer types and consider the multitude of data domains available in EHR-linked biobanks. A focused and rigorous approach to model development and stratification can significantly enhance early detection, targeted surveillance, and patient outcomes for these and other cancers.

Finally, researchers must also be cognizant of obstacles to effective model deployment related to development, implementation, and adoption in health care settings. First, despite the growing number of publications on clinical risk prediction models, few ultimately are implemented because of issues in model development, including lack of reproducibility and replicability,[199,200] lack of model fairness,[201–203] heterogeneities in clinical data,[204,205] and improper model evaluation.[206,207] Chan and Wong recommend external validation (whenever possible), checking that the evaluation methodology is error-free (e.g., data leakage), and assessing model performance using metrics focused on achieving the prediction objective.[206] Second, barriers to adoption include skepticism around evaluating “black box” machine learning algorithms and lack of clear actionability, limiting clinical implementation.[208–210] Watson and colleagues recommended considering traditional versus machine learning methods for development, close collaboration with clinicians before and during model development, and the actionability of model results before model development to address these barriers.[208] Finally, predictive models can be a victim of their success.[211–214] The knowledge that a factor, like cholesterol levels, holds predictive value can change a clinician's consideration to measure that factor. Because the observation and measurement mechanism changes, so does the predictive value of the presence or absence of a measurement, which can degrade model performance.[211] In the context of prognostic prediction models, Lenert and colleagues found that along with model performance surveillance and updating, incorporating the intervention space and considering the model life cycle can mitigate performance degradation.[212]

2.7 Dissertation motivation

The dissertation aims to explore methods addressing two sources of bias – selection bias and missing data – in EHR-linked biobanks, separately and jointly. Through principled interrogation of these biases, this dissertation aims to provide researchers practical guidance for analyzing EHR-linked biobank data to produce relevant and impactful results.

Chapter 3 To Weight or Not to Weight? The Effect of Selection Bias in Three Large EHR-Linked Biobanks and Recommendations for Practice

This manuscript has been published in the Journal of the American Medical Informatics Association.[215]

3.1 Abstract

Objective: To develop recommendations regarding the use of weights to reduce selection bias for commonly performed analyses using electronic health record (EHR)-linked biobank data.

Materials and methods: We mapped diagnosis (ICD code) data to standardized phecodes from three EHR-linked biobanks with varying recruitment strategies: All of Us (AOU; n=244,071), Michigan Genomics Initiative (MGI; n=81,243), and UK Biobank (UKB; n=401,167). Using 2019 National Health Interview Survey data, we constructed selection weights for AOU and MGI to represent the US adult population more. We used weights previously developed for UKB to represent the UKB-eligible population. We conducted four common analyses comparing unweighted and weighted results.

Results: For AOU and MGI, estimated phecode prevalences decreased after weighting (weighted-unweighted median phecode prevalence ratio [MPR]: 0.82 and 0.61), while UKB estimates increased (MPR: 1.06). Weighting minimally impacted latent phenome dimensionality estimation. Comparing weighted versus unweighted PheWAS for colorectal cancer, the strongest associations remained unaltered, with considerable overlap in significant hits. Weighting affected the estimated log-odds ratio for sex and colorectal cancer to align more closely with national registry-based estimates.

Discussion: Weighting had a limited impact on dimensionality estimation and large-scale hypothesis testing but impacted prevalence and association estimation. When interested in estimating effect size, specific signals from untargeted association analyses should be followed up by weighted analysis.

Conclusion: EHR-linked biobanks should report recruitment and selection mechanisms and provide selection weights with defined target populations. Researchers should consider their intended estimands, specify source and target populations, and weight EHR-linked biobank analyses accordingly.

3.2 Background and significance

Electronic health record (EHR)-linked biobanks are repositories with biospecimen and/or related data linked to EHR and auxiliary data (e.g., medical and pharmacy claims, residential-level neighborhood characteristics).[1,7,11,46] Many EHR-linked biobanks are non-probability samples[1,2,9,45,80,216] drawn from a poorly defined source population (i.e., the population from which individuals are sampled). Because of their large sample size, linked multimodal data, immediacy, and accessibility,[84–86] researchers have used EHR data *en masse* for scientific research (from 3,212 PubMed citations in 2013 to 9,824 in 2023). EHR-linked biobanks are increasingly prevalent, and efforts like the Global Biobank Meta-analysis Initiative (GBMI)[47] facilitate global collaboration.[43,217–220]

As the research community gets excited about amassing data, two fundamental questions must be asked: (a) who is in the study, and (b) what is the target population of interest? If biobanks are not representative of the target population, they are vulnerable to selection bias,[7,79,81,82,221] a naïve analysis is not expected to align with the population truth.[5,6,83] Handling selection bias presents a challenge; it is difficult to pinpoint the magnitude and direction of its impact on estimates,[87] increasing the sample size does not mitigate its effect,[88,89] and it often occurs in concert with other data imperfections.[7,76,77,90,91] Moreover, contrary to previous arguments,[92,93] recent evidence suggests that even genetic association analyses with inherited germline susceptibility factors can also be prone to selection bias.[94–97]

There are three common analytic approaches for handling selection bias: stratification,[5,6] quantitative bias analysis,[6,99] and, by far the most common, inverse probability (IP)-weighting.[5,7,8,11,77,79,97,101] IP-weighting involves reweighting individuals in a given sample by the inverse of the estimated probability of their inclusion (relative to the target population) constructed as a function of variables that impact selection.[7,10,11] IP-weight estimation relies on (a) access to representative individual-level data from the target population and (b) correct specification of the selection probability model. Representative data can be probability samples drawn from the target population, like the National Health Interview Survey (NHIS; USA).[103] One can use poststratification (PS)-weights that rely on summary-level data when individual-level data on the target population is unavailable.[104]

In this paper, we consider three EHR-linked biobanks that have three different recruitment strategies/selection mechanisms: the National Institutes of Health All of Us Research Program (AOU),[2,222] our University of Michigan's Michigan Genomics Initiative (MGI),[45,223] and the UK Biobank (UKB).[3,224] We explore the impact of the use of a set of selection weights on common descriptive (prevalence estimation, principal components analysis) and inferential (agnostic large-scale association testing, estimation of targeted association parameters) tasks in EHR data (Supplementary Figure 3.1). First, we estimate selection weights in both US-based cohorts using NHIS data. Second, we characterize demographic and diagnosis (prevalences, latent dimensionality, partial correlation) data in AOU, MGI, and UKB, with and without selection weights. Third, we investigate how using weights impacts discovery in large-scale untargeted hypothesis testing by performing a phenome-wide association study (PheWAS). Fourth, we characterize the influence of weights on a targeted effect estimate in a fitted logistic regression model using colorectal cancer as a sample phenotype. Finally, we develop practical recommendations regarding using selection weights for researchers conducting analyses in and across biobanks.

Weighting-based methods are foundational to survey methodology. It stems from Horvitz and Thompson's 1952 work[225] and has been integral for over seven decades (see Pfefferman's review[226]). Extensions of weighting methods to EHR-linked biobanks are not new.[7,11,79] However, it often needs to be clarified how to create these weights in biobanks with incomplete knowledge of their recruitment strategies. Survey design weights are usually not available or applicable for biobanks. Investigators have tried to develop and apply weights in EHR-linked biobanks inconsistently.[8,97] To the best of our knowledge, there is no systematic evaluation of the effect of weights on downstream analyses across a range of tasks and multiple biobanks. Thus, our paper fills a critical gap in the literature by guiding the use of selection weights in biobank analysis based on empirical evidence.

3.3 Materials and methods

3.3.1 Cohorts

3.3.1.1 AOU: All of Us

AOU started in 2018 to enroll over one million adults via a combination of open invitations and a network of healthcare provider-based recruitment sites. Engagement efforts have focused on oversampling people from communities historically underrepresented in biomedical research based on 10 factors: age, sex, race/ethnicity, gender identity, sexual orientation, disability status, healthcare access, income, educational attainment, and geographic location.[2] We considered these selection factors (except gender identity (not collected in NHIS) and disability status (significant missingness (~61%) in AOU)) in the estimation of IP- and PS-based selection weights. As of January 1, 2024, there were over 760,000 participants, providing access to over 539,000 biosamples and 420,000 EHRs. The AOU subset used in these analyses comprises 244,071 participants with sociodemographic and ICD-9-CM/ICD-10-CM data as part of the curated data repository version 7 (Controlled Tier C2022Q4R9).

3.3.1.2 MGI: Michigan Genomics Initiative

The Michigan Medicine-based MGI (University of Michigan) began in 2012 recruiting adults primarily through appointments for procedures requiring anesthesia.[45] It evolved to include sub-cohorts through metabolism, endocrinology and diabetes (MEND) and mental health (MHB) clinics and a wearables cohort enriched with hypertensive individuals (MIPACT). Age, sex, and race/ethnicity were considered selection factors. Additionally, cancer, diabetes and body mass index (BMI), anxiety and depression, and hypertension were selection mechanisms into the original cohort and these sub-cohorts, respectively, and were also used in selection weight estimation. As of September 2023, there were ~100,000 consented participants in MGI. The MGI subset used in these analyses consists of 81,243 participants (August 22, 2022, data pull) with demographic and ICD-9-CM/ICD-10-CM data.

3.3.1.3 UKB: UK Biobank

The UKB recruited over 500,000 adults aged 40-69 by mailing over 9 million invitations to homes within ~40 kilometers of 22 assessment centers across the UK. Following evidence of healthy volunteer bias,[80] van Alten and colleagues developed a set of generic weights to reweight the UKB sample to the UKB-eligible population using UK Census Microdata.[8] Using an array of sociodemographic characteristics – age, sex, race/ethnicity, educational attainment, employment status, location of residence, tenure of dwelling, number of cars in household, self-reported health, and one-person household status – they estimated lasso regression-based IP-weights.[8] These weights were used in this paper. The UKB subset used in these analyses consists of 401,167 participants with sociodemographic and ICD-10 code data remaining after phenome curation (Supplementary Figure 3.2).

3.3.2 Phenome curation

For all cohorts, ICD-9-CM and ICD-10(-CM) codes were recoded into up to 3,612 phecodes across 18 phecode categories (i.e., phecodes, or "PheWAS codes"[62]), using the phecode X mapping tables (downloaded from GitHub[227] on 6 September 2023) and the PheWAS R package (version 0.99.6-1).[228] Cases were defined as individuals with a single occurrence of a corresponding phecode. 3,493, 3,354, and 2,660 phecodes were defined in AOU, MGI, and UKB, respectively; we restricted our analyses to the 2,042 phecodes with at least 20 cases in all three cohorts. Flowcharts depicting sample size changes following filtering and ICD-to-phecode mapping for all cohorts are shown in

Supplementary Figure 3.2. Phecode-derived trait mappings are shown in Supplementary Table 3.1.

3.3.3 Weight estimation

3.3.3.1 Inverse probability weighting

We constructed IP-weights, which require individual-level data in the target population, in the US-based cohorts. To do this, we used the 2019 NHIS, a probabilistic sample of US adults with self-reported health information. We estimated selection probabilities, ψ, using a simplex regression framework based on the Beta regression approach to weight estimation described in Kundu and colleagues[79]:

$$\psi = P(S = 1|\boldsymbol{X}) \approx P(S_{external} = 1|\boldsymbol{X}) \times \frac{P(S = 1|\boldsymbol{X}, S_{all} = 1)}{1 - P(S = 1|\boldsymbol{X}, S_{all} = 1)} \qquad \textit{Eq. (1)}$$

where, assuming there is no overlap between the internal and external data, S is an inclusion indicator in the internal cohort (i.e., AOU or MGI), $S_{external}$ is an indicator for inclusion in the external cohort (i.e., NHIS), S_{all} is an indicator for inclusion in either cohort, and $\boldsymbol{X}$ are selection factors as listed in the Cohorts section (page 31 and Figure 3.1). We estimated the first term, $P(S_{external} = 1|\boldsymbol{X})$, by fitting a simplex regression model for the known design probabilities using NHIS data. We estimated the numerator of the second term, $P(S = 1|\boldsymbol{X}, S_{all} = 1)$, using a logistic regression model using both internal and external data.

In AOU, we flexibly selected $\boldsymbol{X}$ by splitting the data in half and fitting a lasso-penalized logistic regression model on $\boldsymbol{X}$ and all possible pairwise interactions using the glmnet R package (version 4.1-8). Using 10-fold cross-validation, we selected λ such that the error is within 1 standard error of the minimum to result in a parsimonious model. The selected terms were then used as the final set of $\boldsymbol{X}$ to estimate IP weights in the other half of the data as described above.

3.3.3.2 Poststratification

Using weighted NHIS data, PS-weights were calculated using:

$$\omega = \frac{\Pr(\boldsymbol{X} = \boldsymbol{x})}{\Pr(\boldsymbol{X} = \boldsymbol{x} | S = 1)} \qquad \textit{Eq. (2)}$$

where $\boldsymbol{X}$ are the set of selection variables, and S is an indicator for membership in the internal sample (i.e., AOU or MGI). IP- and PS-weights were winsorized at the 2.5th and 97.5th percentile. Variable definitions are described in Supplementary Table 3.2 Definition of variables by cohort used throughout paper**,** and additional details of IP- and PS-weight estimation are described in Supplementary Methods.

Figure 3.1 summarizes the cohorts and their source populations, sampling strategies, presumed target populations, external data for weighting, and selection factors.

3.3.4 Statistical analyses

First, we obtained crude unweighted and IP-weighted estimates of prevalences. These are calculated as the number of cases over the number of individuals in the respective biobanks. For sex-specific phecodes, only individuals with the corresponding sex are considered.

Second, we estimated the latent dimensionality of the phenome by conducting unweighted and IP-weighted principal components analyses (PCA). We used the number of principal components explaining 95% and 99% of the cumulative variation in the data to represent its dimensionality. Additionally, we explored partial correlations, described in Section 3.10.3.3.

Third, we conducted a colorectal cancer (phecode CA_101.41) PheWAS to illustrate large-scale hypothesis testing. Here, the interest was in obtaining the test statistic and corresponding p-value. PheWAS were adjusted for age, sex, and length of EHR follow-up.

Fourth, we estimated the association between biological sex and colorectal cancer, where the interest was in estimating the log-odds ratio. The female-colorectal cancer association was selected because it is known to be negative (recent log-odds ratio estimate approximations range from -0.414 to -0.270) in the US[229] and the UK.[230] For hypothesis testing and targeted association analyses, after performing a weighted or unweighted analysis within each cohort, we conducted a meta-analysis across three cohorts by using inverse variance weights and a fixed effect model using the meta R package (version 6.5-0) (Supplementary Figure 3.3).[231] Additional data preparation detail is described in Supplementary Methods.

3.3.5 Software

All data cleaning, manipulation, and analysis were conducted using R version 4.2.2. Code and supplementary data are publicly available: https://github.com/maxsal/biobank_selection_weights.

3.4 Results

3.4.1 Descriptive characteristics

Of 244,071 AOU participants, 62.2% were female, with a mean (standard deviation (SD)) age of 54.0 (17.3) years old (Table 3.1). Additionally, 55.4% were non-Hispanic White, and 27.1% had a qualifying cancer phecode in their EHR. Of 81,243 MGI participants, 53.8% were female, with a mean age of 56.3 (17.0) years old. Most of MGI was non-Hispanic White (83.1%) and 49.2% had a cancer diagnosis on their EHR. MGI had substantially more EHR data points per person than AOU as measured by encounters per person (mean 103 in MGI vs. 32 in AOU), unique phecodes per person (77 vs. 72), and years of follow-up per person (9.9 vs. 9.3). Both IP- and PS-weighting brought AOU and MGI closer to NHIS-based estimates of the US population concerning age (47.7 years old), sex (51.7% female), and race/ethnicity (63.2% non-Hispanic White).

Of the 401,167 participants in UKB, 55.3% were female, and their mean age was 57.7 (8.0) years. Additionally, they were 94.2% White, and 25.9% had a qualifying cancer phecode on their EHR. The application of the IP-weights resulted in a cohort that was reflective of the UKB-eligible population concerning age (54.9 weighted vs. 54.8 UKB-eligible), sex (50.8% female weighted vs. 50.8% female UKB-eligible), and race/ethnicity (90.9% White weighted vs. 87.0% White UKB-eligible).

3.4.2 Phecode prevalences

3.4.2.1 Within cohort comparison

In AOU, unweighted phecode prevalences ranged from <0.01% to 52.07% with a median of 0.40%, while IP-weighted (hereafter "weighted" unless otherwise specified) prevalences ranged from 0% to 46.86% with a median of 0.20%. Weighted-to-unweighted phecode prevalence ratios (PR; Figure 3.2A) were down-weighted (i.e., below 1) phenome-wide with a median PR (MPR) of 0.82. In MGI, unweighted prevalences ranged from <0.01% to 50.69% with a median of 0.33%, while weighted prevalences ranged from 0% to 43.12% with a median of 0.21%. Weighting tended to down-weight prevalences with an MPR of 0.61 (Figure 3.2B). In UKB, unweighted prevalences spanned <0.01% to 33.68%, with a median of 0.06%, while weighted prevalences spread from 0% to 32.12%, with a median of 0.07%. Weighting tended to upweight prevalences with an MPR of 1.06 (Figure 3.2C).

3.4.2.2 Across cohorts comparison

Comparing unweighted phecode prevalences, MGI over AOU (Figure 3.3A), we calculated a median and mean PR of 1.15 and 1.70, respectively. On average, 13 of 17 phecode categories had higher prevalences in MGI than AOU except for infections, dermatological, pregnancy, and mental categories (MPRs 0.97, 0.92, 0.88, and 0.74, respectively). Neoplasms were substantially more common in MGI (MPR 2.69). After IP-weighting both cohorts (Figure 3.3D), median and mean PRs were 0.81 and 1.23, respectively. Only congenital and genetic (MPRs 1.70, 1.02, respectively) phecodes remained more common in MGI after weighting.

Using unweighted data (Figure 3.3B and C), phecodes in AOU and MGI were more common than in UKB (MPR: AOU/UKB 5.12; MGI/UKB: 6.37). After IP-weighting (Figure 3.3E and F), phecodes in AOU and MGI were still more common than in UKB (MPR: AOU/UKB: 3.87; MGI/UKB 3.39).

3.4.3 Phenome structure: PCA to estimate the effective number of phenotypes

The latent dimensionality of the diagnostic phenome (n = 2,042) was estimated using PCA in AOU, MGI, and UKB (Table 3.2; shown graphically in Supplementary Figure 3.4). Within cohorts, weighting nominally decreased the number of PCs explaining 95%

of the cumulative variation (CV) in AOU and MGI (from 732 to 711 in AOU; from 752 to 729 in MGI) and nominally increased in UKB (from 553 to 569). This trend was the same at the 99% CV threshold (from 1,262 to 1,236 in AOU; from 1,293 to 1,258 in MGI; from 1,065 to 1,080 in UKB). The dimensionality of the UKB data was noticeably smaller than the US-based cohorts with higher phecode prevalences (e.g., at the 95% CV threshold, 569 PCs weighted UKB phenome vs. 711 and 729 PCs in AOU and MGI, respectively).

We calculated unweighted and weighted partial correlations as a supplemental exploration (Unweighted and weighted partial correlations**,** pg. 81). Partial correlations were visualized as network graphs for AOU, MGI, and UKB in Supplementary Figure 3.5, Supplementary Figure 3.6, and Supplementary Figure 3.7, respectively, and did not show noticeable differences after weighting. Distributions of unweighted (Supplementary Figure 3.8) and weighted (Supplementary Figure 3.9) partial correlations showed that cohorts with higher phecode prevalences (e.g., MGI) had slightly stronger correlations than those with lower phecode prevalences (e.g., UKB).

3.4.4 Large-scale hypothesis testing: an "untargeted" PheWAS for colorectal cancer

In AOU, there were 25 phenome-wide significant hits in the unweighted PheWAS across 6 categories (Figure 3.4A). After IP-weighting, there were only 5 hits, all neoplasms (Figure 3.4D) – the same top 5 hits as in the unweighted PheWAS. In MGI, there were 9 phenome-wide significant hits in the unweighted PheWAS across 2 categories (Figure 3.4B). After IP-weighting, there were 26 hits across 4 categories (Figure 3.4E). The IP-weighted PheWAS identified 3 of the 9 unweighted hits. The IP-weighted PheWAS identified 23 hits not identified in the unweighted PheWAS. In UKB, there were 60 phenome-wide-significant hits in the unweighted PheWAS across 11 categories (Figure 3.4C). After IP-weighting, there were 34 hits across 8 categories (Figure 3.4F). Of the 60 unweighted hits, 30 were also identified in the weighted PheWAS. There were 4 new gastrointestinal hits in the weighted PheWAS. Venn diagrams show the overlaps in phenome-wide significant hits across weighting strategies within cohort in Supplementary Figure 3.10A-C.

Of the 96 unique hits identified in any unweighted or IP-weighted PheWAS, 21.9% (n = 21) appeared only in IP-weighted PheWAS. Most of these hits found only in weighted PheWAS were neoplasms (11), with others belonging to the gastrointestinal (4), neurological (3), mental (1), and musculoskeletal (2) categories. The only hit identified in all three IP-weighted PheWAS (CA_101: Malignant neoplasm of the digestive organs) was also identified in all three unweighted PheWAS. Of the 21 hits only identified in IP-weighted PheWAS, 71.4% (15) appeared only in MGI, and 14.3% (3) appeared only in UKB. Venn diagrams show the overlaps in phenome-wide significant hits across cohorts by weighting strategy in Supplementary Figure 3.10D-F.

The unweighted meta-PheWAS identified 37 hits across 9 categories, while the IP-weighted meta-PheWAS identified 22 hits across 5 categories. Of the 44 unique hits identified in both meta-PheWAS, 15.9% (7) appeared only in the IP-weighted meta-PheWAS. Notably, the IP-weighted meta-PheWAS identified a hit (NS_356.2: Aphasia and dysphasia) in a novel category (neurological). The overlaps in phenome-wide significant hits across weighting strategies are shown as Venn diagrams in Supplementary Figure 3.11.

Of the 101 unique hits identified in any unweighted or PS-weighted PheWAS, 25.7% (n = 26) appeared only in PS-weighted PheWAS (Supplementary Figure 3.12). PheWAS summary statistics are available in the supplementary data file in the GitHub repository.

3.4.5 “Targeted” estimation of the sex-colorectal cancer log-odds ratio

The unweighted age-adjusted log-odds ratio for female sex and colorectal cancer were -0.098 (-0.164, -0.033), -0.164 (-0.247, -0.082), and -0.389 (-0.431, 0.348) for AOU, MGI, and UKB, respectively. The unweighted UKB estimate overlapped with the benchmark range of -0.414 to -0.270 based on 2018-2020 US SEER[229] and UK[230] estimates. The unweighted meta-analytic estimate was -0.284 (-0.316, -0.252). IP- and PS-weighting did not improve estimation in AOU, resulting in null estimates of -0.047 (-0.198, 0.104) and -0.084 (-0.191, 0.024), respectively. However, in MGI, weighting improved estimation with the IP-weighted confidence interval overlapping with (-0.217 (-0.419, -0.014)) and the PS-weighted point estimate falling within (-0.342 (-0.629, -0.056)) the benchmark range. IP-weighting did not change the UKB estimate (-0.398 (-0.461, -0.334)). The IP- and PS-weighted meta-analytic estimates (-0.335 (-0.392, -0.279) and -0.318 (-0.371, -0.264), respectively) remained stable, driven by the UKB estimates. Along with unadjusted estimates, these results are shown in Figure 3.5 and Supplementary Table 3.3.

3.5 Discussion

EHR-linked biobanks – such as AOU, MGI, and UKB analyzed here – are transforming the fields of epidemiology and health research. They offer valuable resources comprising large longitudinal cohorts, with vast amounts of readily available structured and unstructured data and potential for data linkages at relatively low costs.[1,2,4,9,45,216] However, the varying sampling mechanisms across these cohorts require researchers to understand and address the impact of selection bias on various descriptive and inferential tasks (Supplementary Figure 3.1). We developed practical recommendations on constructing and applying weights to mitigate selection bias in standard EHR-linked biobank analyses. Furthermore, we advise biobank management bodies to define their recruitment strategies clearly and explain various forms of enrichment that can lead to departure from the source and target population, which essential for accurately developing selection weights.

To do this, we estimated IP- and PS-based selection weights for AOU and MGI and, along with previously described UKB IP-weights,[8] evaluated their impact on common analyses currently undertaken in the field (impact on prediction is the subject of a forthcoming manuscript). Estimates of latent phenome dimensionality were marginally lower in cohorts with relatively higher phecode prevalences (e.g., AOU and MGI). The practical implication in terms of reduction in the denominator of a Bonferroni-corrected p-value from the number of total tests to the PCA-estimated number of independent tests would not have a meaningful impact.[232] Further, p-value-identified results from untargeted hypothesis testing (as explored via a colorectal cancer PheWAS) for the strongest association signals remained largely unaltered following the introduction of selection weights. For example, the top 9 hits (and 12 total) from the unweighted meta-PheWAS were also identified in both weighted meta-PheWAS, and the top 5 hits were the same in all meta-PheWAS (Supplementary Figure 3.11). We also found that while weighting typically increases p-values, some p-values in MGI decreased, likely due to significant selection bias. These results indicate using selection weights for exploring phenome structure and large-scale hypothesis testing tasks is not crucial, particularly when such weights are not provided. If weights are readily available, using selection weights in this context is advisable. Significant hits from agnostic analyses should be followed by a targeted analysis where the importance of using weights is clearer.

For estimation tasks, like prevalence and effect size estimation, using selection weights to reduce potential selection bias is recommended. Regarding phecode prevalence estimation, we saw considerable changes in prevalence estimates after weighting (e.g., prevalence of MB_286.2: Major depressive disorder dropped 24 percentage points after IP-weighting in MGI), and these changes were phenome-wide (e.g., IP-weighted over unweighted MPR in AOU: 0.82). Sampling strategies that are health system-based (e.g., MGI) or that target groups with elevated disease burdens due to healthcare disparities and negative social determinants (e.g., AOU) can lead to overrepresentation of individuals with more diseases and comorbidities. Such enrichment can explain the marked deflation in within-cohort (particularly in MGI and to a lesser extent in AOU; Figure 3.2) and cross-cohort (AOU/UKB, MGI/UKB; Figure 3.3) prevalence ratios after applying weights to align with the target population. Regarding association estimation, we saw that using generic selection weights moved sex log-odds ratio estimates for colorectal to within the benchmark interval in MGI. However, AOU estimates remained outside the benchmark interval even after weighting, likely because of substantial racial/ethnic heterogeneity (Supplementary Figure 3.13). Stratified analyses are preferable when there is expected or known heterogeneity, especially when the data are powered to do so (e.g., race/ethnicity-specific analyses in AOU). In the case of targeted association estimation, we also recommend that weights be curated based on the outcome of interest, a conclusion supported by recent literature.[7,11,79] Finally, in all settings, selection weights are more critical in samples whose characteristics differ more from the target population than in smaller and non-population-based cohorts, like the MGI.

3.5.1 Achievable goal is to reduce, not remove, bias

Weighted analyses are, historically, attempts to remove the impact of selection bias (e.g., on an association estimate) with respect to a defined target population.[5,101] We developed selection weights based on explicit selection factors that were either publicly reported to have influenced recruitment strategies (as in AOU) or known to impact eligibility (as in MGI). However, these selection mechanisms are complex, and the true mechanisms are not fully known. Thus, using selection weights aims to reduce rather than remove bias. This is particularly important in the case of Big Data where, while confidence intervals are narrow, effects of selection bias are not mitigated by increasingly large sample sizes.[233] Additionally, some associations may be more or less prone to selection biases, but which associations are affected and how are unknown. See section 3.10.5 for comments on methodological considerations in EHR-based data analysis.

3.5.2 Strengths and limitations

This study has multiple strengths. First, we utilized AOU and UKB data, which are large-scale, public, and frequently used EHR-linked biobanks. Second, we utilized various methods to visualize and characterize EHR-linked biobanks. Third, we estimated IP- and PS-weights in AOU and provided code for recreating them. Fourth, the weights are based on NHIS data, a public resource with individual-level data representing a probabilistic sample of the US adult population. Fifth, we used the new phecode X mapping table, which is more granular than its predecessor (version 1.2), is built on ICD-10 data, and appears to have more accurate phecode definitions (an earlier version of this manuscript used phecode 1.2 mappings and found unexpected consequences of its phecode definitions; see section 3.10.6).

However, our study also has several limitations. First, we cannot fully account for selection bias because the selection mechanisms are not fully known. Thus, our selection weights attempt to reduce selection bias. Second, the cohorts used vary in terms of geographical location, recruitment mechanisms, and access to EHR data (e.g., single medical system vs. primary care EHR). Future studies could examine more comparable cohorts to derive nuanced insights. Third, we performed meta-analyses of the US- and UK-based cohorts following the assumption of a fixed/common effect meta-analysis.[234] However, the phenome has salient socio-behavioral, economic, infrastructural, and environmental contributors that are divergent across the cohorts. As such, for interpreting pooled estimates and p-values from meta-analyses, investigators should consider the heterogeneity of estimates within and across cohorts. Fourth, we focus on generalizing results to an overall target population. However, having statistical power to conduct analyses in historically underrepresented groups in biomedical research (e.g., AOU) is a strength, not a liability. Aggregating data by reporting overall rather than stratified results can mask health inequities. Researchers can consider constructing strata-specific weights to obtain results generalizable to national level subpopulation (e.g., all adult non-Hispanic Blacks in the US; Supplementary Figure 3.14). Fifth, analyses focused on or combined with genetic data are common and a strength of EHR-linked biobanks, which we do not address. Schoeler and colleagues explore the impact of selection weights on genetic analyses like GWAS, heritability estimation, and Mendelian randomization in UKB, which may interest readers.[97] Further research should explore the effect of selection bias on the genome-by-phenome landscape. Finally, while our focus was on approaches to reduce the impact of selection bias, multiple sources of bias[7,235,25,48,236,237] need to be considered when conducting EHR analysis. Future studies should investigate these biases jointly, how they affect analytic tasks, and their relative importance on the final inferential conclusion.

3.6 Conclusion

We have introduced methods for assessing and comparing the effect of selection bias in EHR-linked biobanks and computed IP- and PS-weights for two US-based biobanks. These weights can potentially reduce – not remove – selection bias as the selection mechanisms are not fully known. Our findings suggest that using generic selection weights for exploring phenome structure (i.e., latent dimensionality, partial correlation across phecodes) and large-scale hypothesis testing is not crucial. EHR-linked biobanks should provide detailed guidance on sampling and recruitment processes and, where possible, make selection weights publicly available. Researchers should also clearly state their intended target population and estimand and describe recruitment and selection mechanisms from the source population. Systematic and rigorous exploration and comparisons of cohorts should be standard in analyses using multi-center EHR-linked biobank data.

3.7 Acknowledgments

Funding Statement:

This work was funded by National Cancer Institute grant P30CA046592 and the Training, Education, and Career Development Graduate Student Scholarship of the University of Michigan Rogel Cancer Center**.**

Competing Interests Statement:

LGF is a Without Compensation (WOC) employee at the VA Ann Arbor, a United States government facility. All other authors declare that they have no competing financial or non-financial interests related to this research.

Contributorship statement:

MS and BM contributed to conceptualization and investigation. MS, RK, and BM contributed to methodology. MS and LGF were involved in data curation. MS conducted the formal analysis and produced visualizations. BM supervised this study. MS and BM drafted the initial manuscript. All authors contributed to the critical revision of the manuscript for content, interpretation, and presentation.

Data availability statement:

Patient confidentiality prevents the sharing of data publicly. However, the data underlying the study's results are available from the All of Us Research Program at https://www.researchallofus.org/register/, the Michigan Genomics Initiative at https://precisionhealth.umich.edu/ourresearch/michigangenomics/, and the UK Biobank at http://www.ukbiobank.ac.uk/register-apply/ for researchers who meet the criteria for confidential data access. Code and supplementary data are publicly available: https://github.com/maxsal/biobank_selection_weights.

Ethics Statement:

This study is covered by IRB application HUM00155849 and received approval from the University of Michigan Institutional Review Board of the Medical School (IRBMED). Standard approvals for accessing and analyzing All of Us and UK Biobank (application number 88536) were received. Deidentified data was used to conduct these analyses.

Acknowledgments:

Michigan Genomics Initiative: The authors acknowledge the Michigan Genomics Initiative participants, Precision Health at the University of Michigan, the University of Michigan Medical School Central Biorepository, the University of Michigan Medical School Data Office for Clinical and Translational Research, and the University of Michigan Advanced Genomics Core for providing data and specimen storage, management, processing, and distribution services, and the Center for Statistical Genetics in the Department of Biostatistics at the School of Public Health for genotype data curation, imputation, and management in support of the research reported in this publication/grant application/presentation.

UK Biobank: This research has been conducted using the UK Biobank Resource under application number 88536.

All of Us: The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research

Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants.

We want to thank Sjoerd van Alten and colleagues for generously sharing their code and for assisting in replicating their UK Biobank selection weights for use in this manuscript.

3.8 Tables

Table 3.1 Descriptive characteristics of the Michigan Genomics Initiative, the UK Biobank, and All of Us. For unweighted metrics, mean (standard deviation) and percent (n) are provided for continuous and categorical/binary variables, respectively. For weighted metrics, mean (standard error) and percent (standard error) are provided for continuous and categorical/binary variables, respectively.

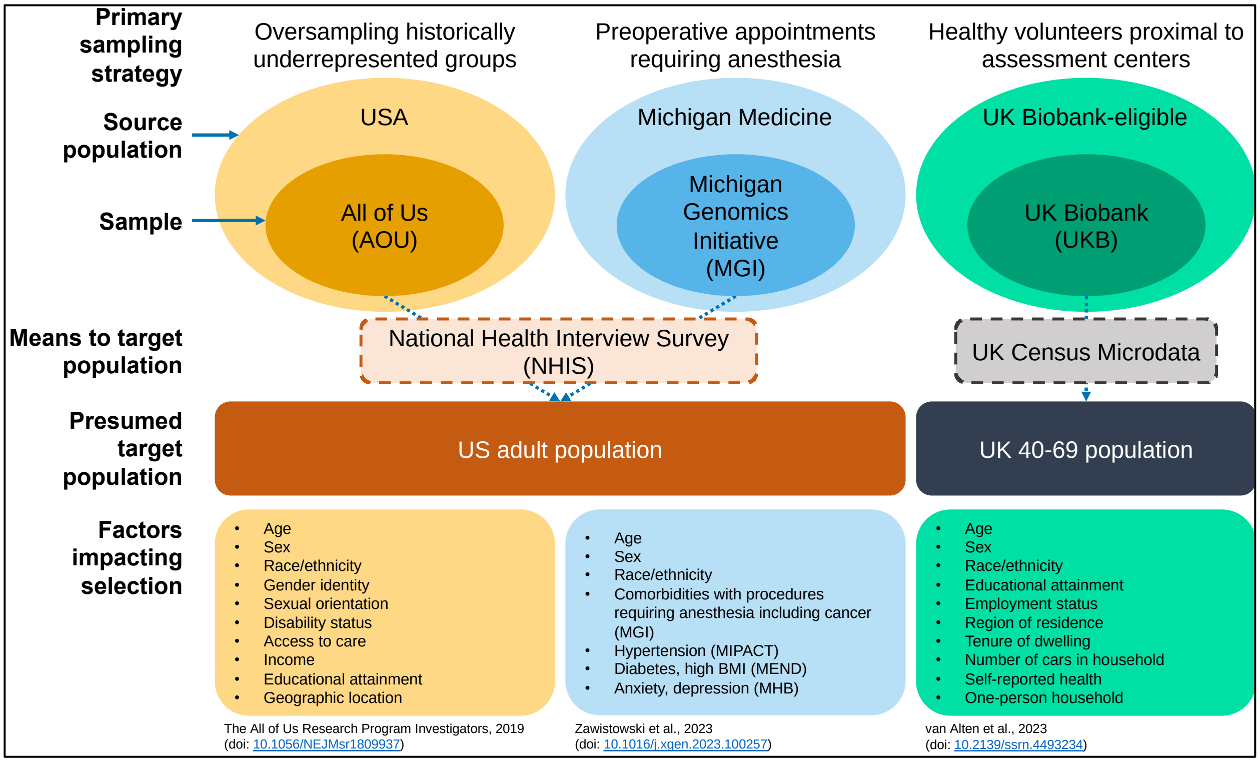
This table can be viewed via this dissertation's corresponding repository at https://www.doi.org/10.17605/OSF.IO/SBMN2

Table 3.2 Number of principal components by proportion of cumulative variation (CV) in diagnostic phenome (n = 2,042) explained by cohort.

	95% CV explained		**99% CV explained**	
	Unweighted	**Weighted**	**Unweighted**	**Weighted**
All of Us	732	711	1,262	1,236
Michigan Genomics Initiative	752	729	1,293	1,258
UK Biobank	553	569	1,065	1,080

Weighted results were conducted using inverse probability (IP)-weights. Out of 2,042 phecodes with at least 20 cases in all three cohorts.

3.9 Figures



The All of Us Research Program Investigators, 2019 (doi: 10.1056/NEJMsr1809937)

Zawistowski et al., 2023 (doi: 10.1016/j.xgen.2023.100257)

van Alten et al., 2023 (doi: 10.2139/ssrn.4493234)

Figure 3.1 Schematic representation of the All of Us, the Michigan Genomics Initiative, and the UK Biobank cohorts, their sampling strategies, potential target populations, and selection factors. All three cohorts are non-probability samples of their source populations for different reasons: oversampling, procedures requiring anesthesia, and healthy volunteers, respectively. External data like NHIS or UK Census Microdata can be used in selection weight construction to make inferences regarding presumed target populations. Factors known to influence recruitment strategy or eligibility criteria are listed.

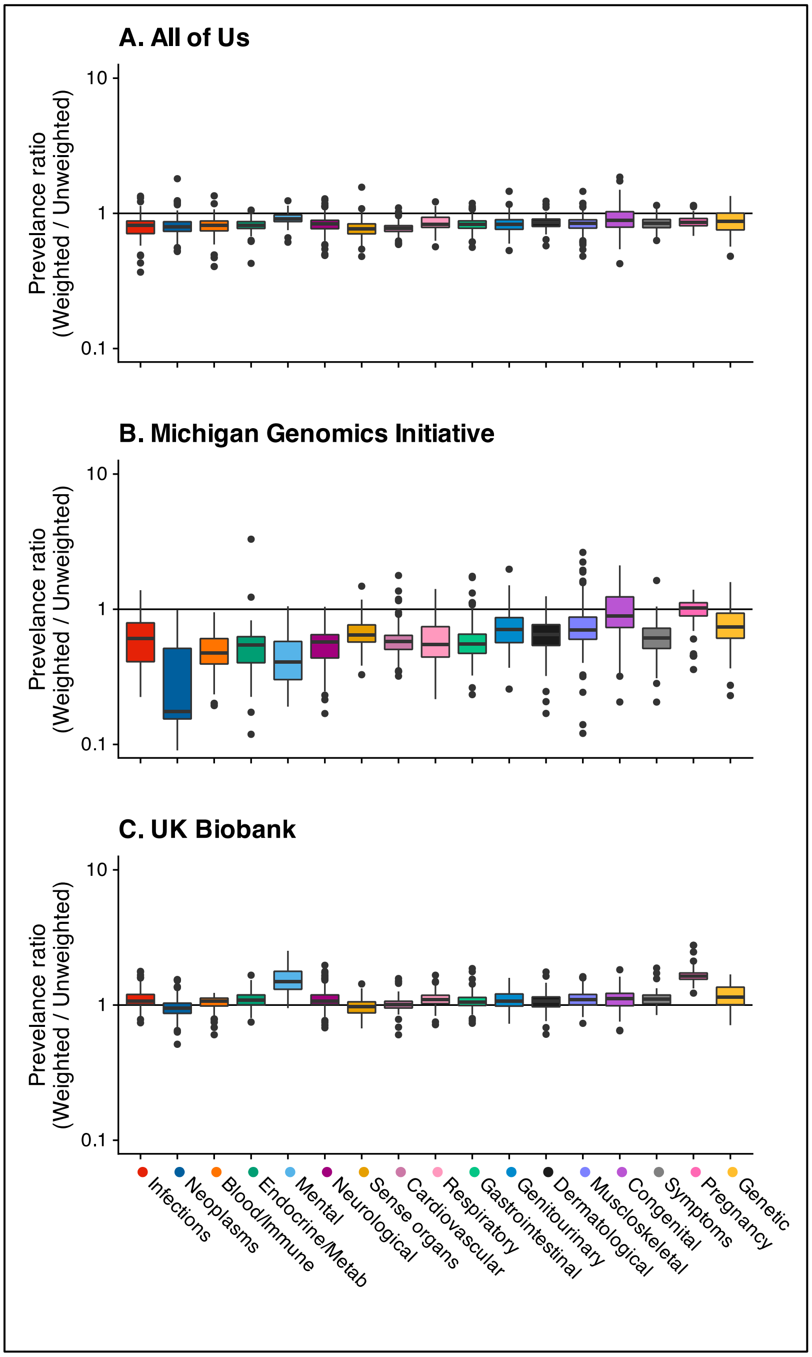


Figure 3.2 Side-by-side boxplots of the inverse probability (IP)-weighted over unweighted phecode prevalence ratios within cohorts by 17 defined phecode categories. Panel A shows the ratio of IP-weighted/unweighted prevalences in AOU, panel B shows the ratio of IP-weight/unweighted prevalences in MGI, and panel C shows the ratio of IP-weighted/unweighted prevalances in UKB. IP-weights were used in AOU and MGI and IP-weights described in van Alten et al.[8] were used in UKB.

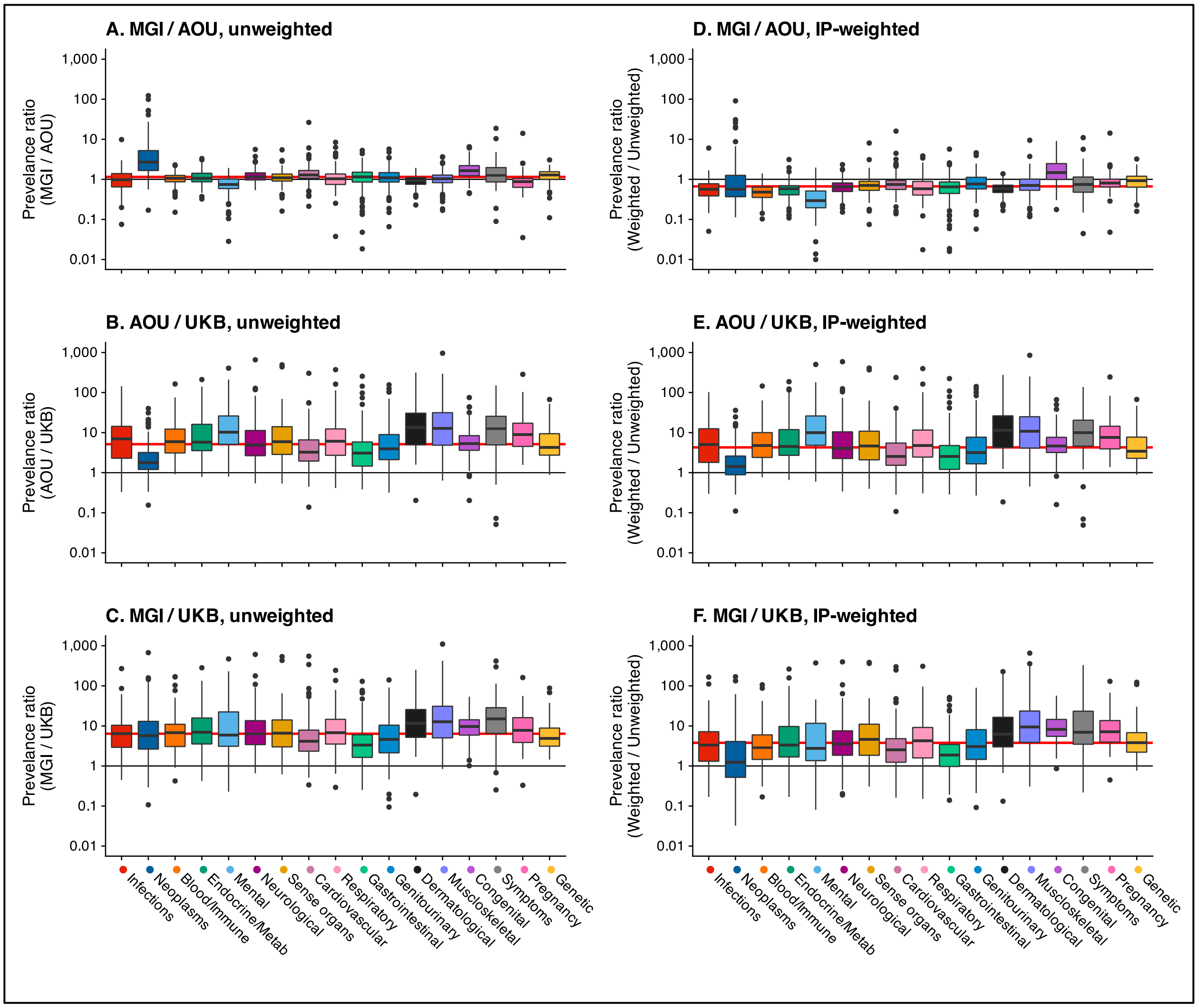


Figure 3.3 Side-by-side boxplots of the unweighted and inverse probability (IP)-weighted phecode prevalence ratios across cohorts by 17 defined phecode categories. Panel A shows the ratio of unweighted prevalences in MGI over AOU, panel B shows the ratio of unweighted prevalences in AOU / UKB, and panel C shows the ratio of unweighted prevalances in MGI / UKB. Panel D shows the ratio of IP-weighted prevalences in MGI over AOU, panel E shows the ratio of IP-weighted prevalences in AOU / UKB, and panel F shows the ratio of IP-weighted prevalances in MGI / UKB. The horizontal red line indicates the median phenome-wide prevalence ratio value. IP-weights were used in AOU and MGI and IP-weights described in van Alten et al.[8] were used in UKB.

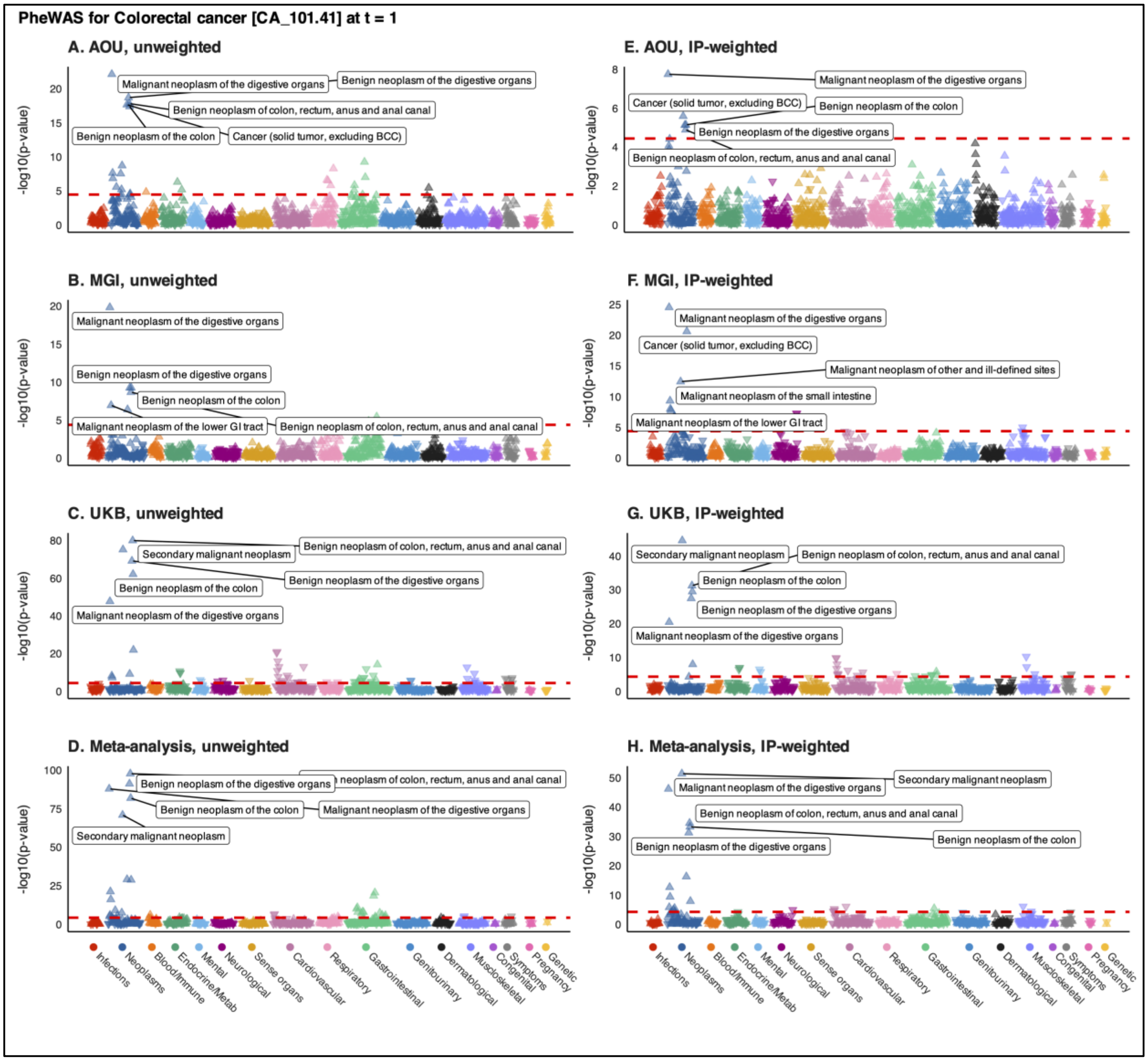
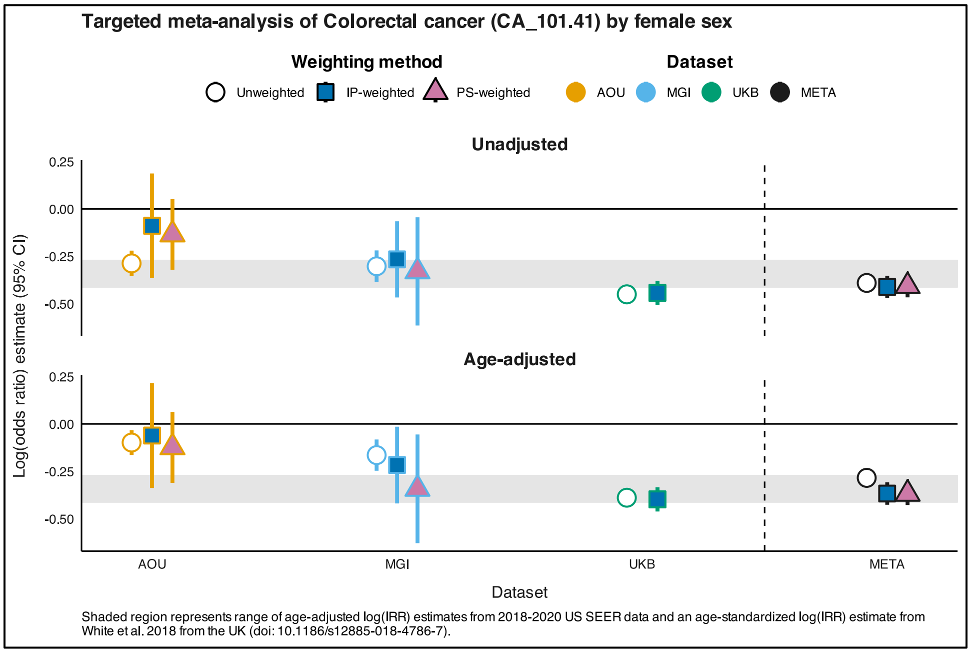


Figure 3.4 Manhattan plots summarizing unweighted (panels A-C) and inverse probability (IP)-weighted (panels E-G) phenome-wide association studies (PheWAS) for colorectal cancer in All of Us, the Michigan Genomics Initiative, and UK Biobank using 1:2 case:non-case matched data restricted to one year prior to initial diagnosis. Panels D and H show the unweighted and IP-weighted meta-analysis PheWAS, respectively. The dashed red line represents the Bonferroni-corrected p-value threshold (-log10(0.05/number of traits)). The five traits with the smallest p-values are labeled. The upward (downward) orientation of the triangle indicates a positive (negative) association. Plots corresponding to poststratification-weighted PheWAS are presented in Supplementary Figure 3.12.



Shaded region represents range of age-adjusted log(IRR) estimates from 2018-2020 US SEER data and an age-standardized log(IRR) estimate from White et al. 2018 from the UK (doi: 10.1186/s12885-018-4786-7).

Figure 3.5 Within cohort and meta-analysis unadjusted and age-adjusted female log-odds ratio estimates (95% confidence interval) for colorectal cancer (phecode CA_101.41). Point estimate shapes and fill colors correspond to the weighting method (white circle, unweighted; dark blue square, inverse probability (IP)-weighted; pink triangle, poststratification (PS)-weighted). Line colors correspond to the cohort (orange, AOU; blue, MGI; green, UKB; black, meta-analysis). Shaded region represents range of age-adjusted log(incidence rate ratio [IRR]) estimates from 2018-2020 US SEER data[229] and an age-standardized log(IRR) estimate from White et al. 2018 from the UK.[230]

3.10 Supplementary materials

3.10.1 Supplementary tables

Supplementary Table 3.1 Phenotypes defined in paper and their qualifying phecode definitions

Variable	Phecode	Description
Anxiety	MB_288	Anxiety and anxiety disorders
Cancer	CA_100	Malignant neoplasm of the head and neck
	CA_100.1	Malignant neoplasm of the oral cavity
	CA_100.12	Malignant neoplasm of the tongue
	CA_100.13	Malignant neoplasm of the gums
	CA_100.14	Malignant neoplasm of the floor of mouth
	CA_100.15	Malignant neoplasm of the palate
	CA_100.2	Malignant neoplasm of the oropharynx
	CA_100.3	Malignant neoplasm of the nasopharynx
	CA_100.4	Malignant neoplasm of the hypopharynx
	CA_100.5	Malignant neoplasm of nasal cavities, middle ear, and accessory sinuses
	CA_100.6	Malignant neoplasm of the larynx
	CA_100.7	Malignant neoplasm of the pharynx
	CA_100.8	Malignant neoplasm of the lip
	CA_100.9	Malignant neoplasm of the salivary glands
	CA_101	Malignant neoplasm of the digestive organs
	CA_101.1	Malignant neoplasm of the esophagus
	CA_101.2	Malignant neoplasm of stomach
	CA_101.21	Malignant neoplasm of cardia
	CA_101.3	Malignant neoplasm of the small intestine
	CA_101.4	Malignant neoplasm of the lower GI tract
	CA_101.41	Colorectal cancer
	CA_101.411	Malignant neoplasm of colon
	CA_101.412	Malignant neoplasm of appendix
	CA_101.42	Malignant neoplasm of anus
	CA_101.6	Malignant neoplasm of the liver and intrahepatic bile ducts
	CA_101.61	Malignant neoplasm of the liver
	CA_101.62	Malignant neoplasm of the intrahepatic bile ducts
	CA_101.7	Malignant neoplasm of the gallbladder and extrahepatic bile ducts
	CA_101.71	Malignant neoplasm of the gallbladder
	CA_101.8	Malignant neoplasm of the pancreas
	CA_102	Malignant neoplasm of the thoracic and respiratory organs
	CA_102.1	Malignant neoplasm of the of bronchus and lung
	CA_102.3	Malignant neoplasm of the trachea
	CA_102.5	Malignant neoplasm of the heart, mediastinum, thymus, and pleura
	CA_102.51	Malignant neoplasm of the heart
	CA_102.52	Malignant neoplasm of the mediastinum
	CA_102.53	Malignant neoplasm of the of pleura
	CA_102.54	Malignant neoplasm of the thymus
	CA_103	Malignant neoplasm of the skin
	CA_103.1	Melanomas of skin
	CA_103.2	Keratinocyte carcinoma
	CA_103.21	Basal cell carcinoma
	CA_103.22	Squamous cell carcinoma of the skin
	CA_103.3	Carcinoma in situ of skin
	CA_104	Malignant sarcoma-related cancers
	CA_104.1	Malignant neoplasm of the bone and/or cartilage

Variable	Phecode	Description
	CA_104.2	Malignant neoplasm of retroperitoneum and peritoneum
	CA_104.3	Malignant neoplasm of connective and soft tissue
	CA_104.4	Malignant neoplasm of peripheral nerves*
	CA_104.5	Gastrointestinal stromal tumor*
	CA_104.6	Kaposi's sarcoma
	CA_105	Malignant neoplasm of the breast
	CA_105.1	Malignant neoplasm of the breast, female
	CA_105.2	Malignant neoplasm of the breast, male
	CA_106	Gynecological malignant neoplasms
	CA_106.1	Malignant neoplasm of external female genital organs and cervix
	CA_106.11	Malignant neoplasm of the vulva
	CA_106.12	Malignant neoplasm of the vagina
	CA_106.13	Malignant neoplasm of the cervix
	CA_106.2	Malignant neoplasm of the uterus
	CA_106.21	Malignant neoplasm of endometrium
	CA_106.3	Malignant neoplasm of the ovary
	CA_106.4	Malignant neoplasm of the fallopian tube and uterine adnexa
	CA_106.6	Malignant neoplasm of the placenta
	CA_107	Malignant neoplasm of male genitalia
	CA_107.1	Malignant neoplasm of the penis
	CA_107.2	Malignant neoplasm of the prostate
	CA_107.3	Malignant neoplasm of the testis
	CA_107.4	Malignant neoplasm of epididymis
	CA_107.5	Malignant neoplasm of spermatic cord
	CA_107.6	Malignant neoplasm of the scrotum
	CA_108	Malignant neoplasm of the urinary tract
	CA_108.4	Malignant neoplasm of the kidney
	CA_108.41	Malignant neoplasm of kidney, except pelvis
	CA_108.42	Malignant neoplasm of renal pelvis
	CA_108.5	Malignant neoplasm of the bladder
	CA_108.6	Malignant neoplasm of urethra
	CA_108.7	Malignant neoplasm of ureter
	CA_109	Malignant neoplasm of the eye, brain and other parts of central nervous system
	CA_109.1	Malignant neoplasm of eye
	CA_109.11	Malignant neoplasm of orbit
	CA_109.12	Malignant neoplasm of lacrimal gland and duct
	CA_109.13	Malignant neoplasm of conjunctiva
	CA_109.14	Malignant neoplasm of cornea
	CA_109.15	Malignant neoplasm of retina
	CA_109.16	Malignant neoplasm of choroid
	CA_109.2	Malignant neoplasm of meninges
	CA_109.3	Malignant neoplasm of brain
	CA_109.4	Malignant neoplasm of spinal cord
	CA_109.5	Malignant neoplasm of cranial nerve
	CA_110	Malignant neoplasm of the endocrine glands
	CA_110.1	Malignant neoplasm of the thyroid
	CA_110.3	Malignant neoplasm of the parathyroid gland
	CA_110.4	Malignant neoplasm of the pituitary gland and craniopharyngeal duct
	CA_110.5	Malignant neoplasm of the pineal gland
	CA_112	Malignant neoplasm of other and ill-defined sites
	CA_112.1	Mesothelioma*
	CA_114	Neuroendocrine tumors
	CA_114.1	Malignant neuroendocrine tumors
	CA_114.11	Exocrine pancreatic cancer

Variable	Phecode	Description
	CA_114.12	Merkel cell carcinoma
	CA_114.2	Benign neuroendocrine tumors
	CA_114.4	Carcinoid tumors
	CA_114.41	Intestinal carcinoid
	CA_114.42	Carcinoid tumor of the bronchus and lung
	CA_114.43	Carcinoid tumor of the thymus
	CA_114.44	Carcinoid tumor of the stomach
	CA_114.45	Carcinoid tumor of the kidney
	CA_114.5	Paraganglioma
	CA_114.6	Pheochromocytoma
	CA_116	Secondary malignant neoplasm
	CA_120	Hemo onc - by cell of origin
	CA_120.1	Myeloid
	CA_120.11	Plasma cell
	CA_120.12	Monocyte
	CA_120.13	Erythroid
	CA_120.14	Megakaryoblast
	CA_120.15	Mast cell
	CA_120.2	Lymphoid
	CA_120.21	Mature B-cell
	CA_120.22	Mature T-Cell
	CA_120.3	Histocytes
	CA_121	Leukemia
	CA_121.1	Acute leukemia
	CA_121.11	Acute lymphoid leukemia
	CA_121.12	Acute myeloid leukemia
	CA_121.2	Chronic leukemia
	CA_121.21	Chronic lymphoid leukemia
	CA_121.22	Chronic myeloid leukemia
	CA_121.23	Chronic myelomonocytic (monocytic) leukemia
	CA_122	Lymphoma
	CA_122.1	Hodgkin lymphoma
	CA_122.11	Nodular sclerosis Hodgkin lymphoma
	CA_122.2	Non-Hodgkin lymphoma
	CA_122.21	Follicular lymphoma
	CA_122.22	Diffuse large B-cell lymphoma*
	CA_122.23	Burkitt lymphoma
	CA_122.24	T-cell lymphoma
	CA_122.25	Anaplastic large cell lymphoma
	CA_122.26	Extranodal NK/T-cell lymphoma, nasal type*
	CA_123	Multiple myeloma and malignant plasma cell neoplasms
	CA_123.1	Multiple myeloma
	CA_124	Myeloproliferative disorder
	CA_124.3	Polycythemia vera
	CA_124.5	Essential thrombocythemia
	CA_124.6	Myelodysplastic syndrome
	CA_124.7	Chronic myeloproliferative disease*
	CA_124.8	Myelofibrosis
	CA_125	Other malignant neoplasms of lymphoid, hematopoietic and related tissue
	CA_128	Estrogen receptor status
	CA_128.1	Estrogen receptor positive status [ER+]
	CA_128.2	Estrogen receptor negative status [ER-]
	CA_130	Cancer (solid tumor, excluding BCC)
	CA_132	Sequelae of cancer

Variable	Phecode	Description
Coronary artery disease	CV_404.2	Coronary atherosclerosis [Atherosclerotic heart disease]
Depression	MB_286.2	Major depressive disorder
Diabetes	EM_202	Diabetes mellitus

Visit https://phewascatalog.org (phecodeX) and https://github.com/PheWAS/PhecodeX

Supplementary Table 3.2 Definition of variables by cohort used throughout paper

	AOU	MGI	UKB	NHIS (2019)*
Age	Age at last diagnosis	Age at last diagnosis	Age at consent: date of consent (field ID 200) minus date of birth (field IDs 34, 52)	Age at screening (AGEP_A)
Sex	Self-reported sex at birth (field name: sex_at_birth_concept_id)	Self-report EHR	Acquired by central registry at recruitment, may be updated by individual (field ID 31)	SEX_A
Race/ethnicity	Self-reported race ethnicity (field names: race_source_concept_id, ethnicity_source_concept_id)	Self-report EHR	Self-report survey (field ID 21000)	HISPALLP_A
BMI	Median of EHR values	Median of EHR values	Median of assessed values (field ID 21001)	BMICAT_A (HEIGHTTC_A, WEIGHTLBTC_A)
Smoking status	Self-report (concept IDs: 1585857, 1585860)	Self-report EHR	Survey (field ID 20116)	SMKCIGST_A
Anxiety	Phecode MB_288: Anxiety and anxiety disorders			GADCAT_A
Cancer	See Supplementary Table 3.1			CANEV_A
Coronary artery disease	Phecode CV_404.2: Coronary atherosclerosis [Atherosclerotic heart disease]			CHDEV_A
Depression	Phecode MB_286.2: Major depressive disorder			PHQCAT_A
Diabetes	Phecode EM_202: Diabetes mellitus			DIBEV_A

* visit https://www.cdc.gov/nchs/nhis/2019nhis.htm for more information

Supplementary Table 3.3 Female log odds ratio estimate (95% confidence interval) for colorectal cancer (phecode CA_101.41).

Weighting	Covariates	AOU	MGI	UKB	META
Unweighted	None	**-0.287** (-0.354, -0.220)	**-0.303** (-0.387, -0.219)	**-0.450** (-0.492, -0.409)	**-0.390** (-0.423, -0.358)
	Age	**-0.098** (-0.164, -0.033)	**-0.164** (-0.247, -0.082)	**-0.389** (-0.431, -0.348)	**-0.284** (-0.316, -0.252)
IP-weighted	None	-0.037 (-0.188, 0.113)	**-0.266** (-0.467, -0.065)	**-0.443** (-0.506, -0.380)	**-0.373** (-0.429, -0.317)
	Age	-0.047 (-0.198, 0.104)	**-0.217** (-0.419, -0.014)	**-0.398** (-0.461, -0.334)	**-0.335** (-0.392, -0.279)
PS-weighted	None	-0.135 (-0.321, 0.052)	**-0.329** (-0.615, -0.044)	**-0.443** (-0.506, -0.380)	**-0.408** (-0.466, -0.349)
	Age	-0.123 (-0.311, 0.064)	**-0.342** (-0.629, -0.056)	**-0.398** (-0.461, -0.334)	**-0.368** (-0.427, -0.310)

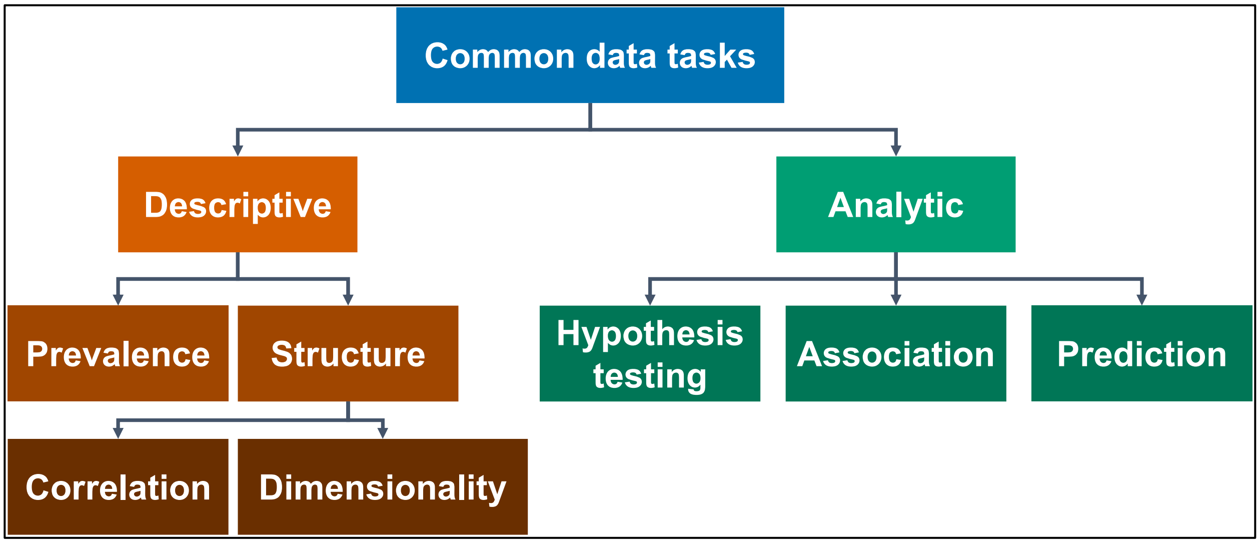
* Meta-analysis results include IP-weighted estimate from UKB

Abbrevs: AOU, All of Us; IP, inverse probability; META, meta-analysis; MGI, Michigan Genomics Initiative; PS, poststratification; UKB, UK Biobank

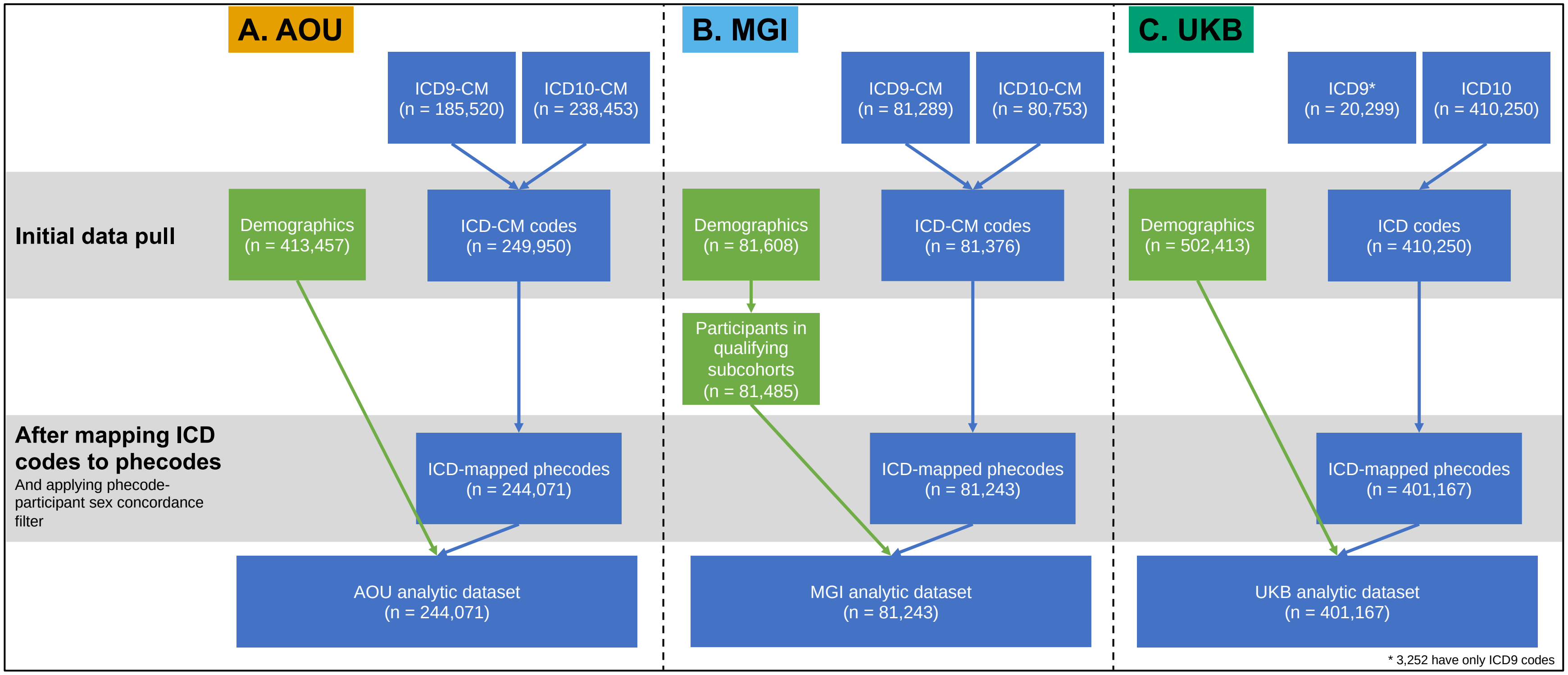
Bolded point estimates are statistically significant at the 95% confidence level

Supplementary Table 3.4 Comparison between ICD codes by colorectal cancer phecode mapping table, count with ICD code, and overlap with individuals who have HIV phecode (sorted by proportion of overlap). *This table can be viewed via this dissertation's corresponding repository at https://www.doi.org/10.17605/OSF.IO/SBMN2*

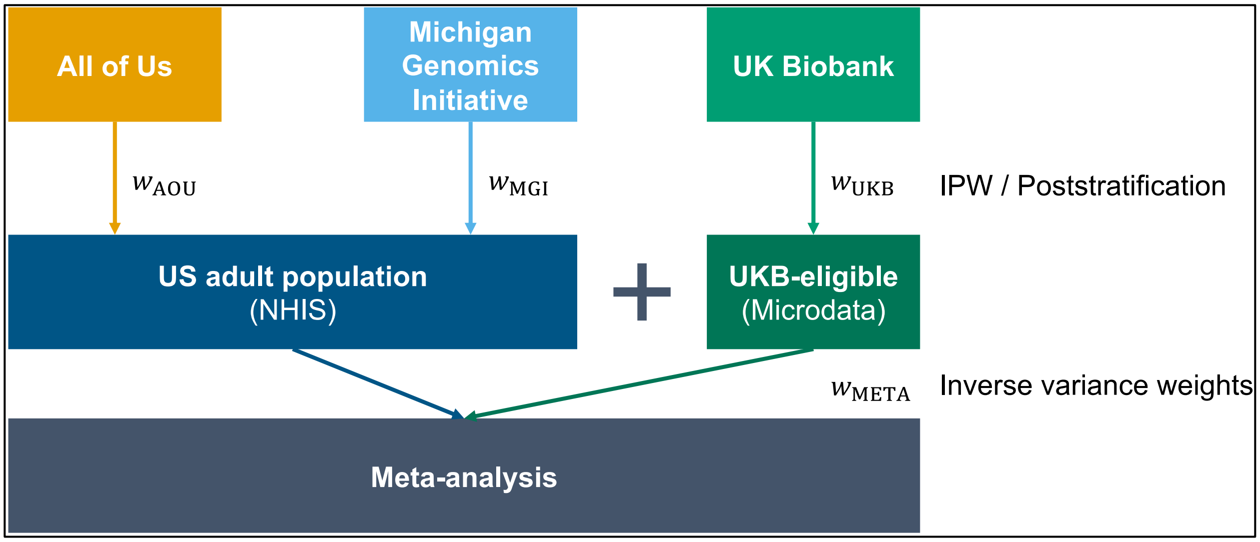
3.10.2 Supplementary figures



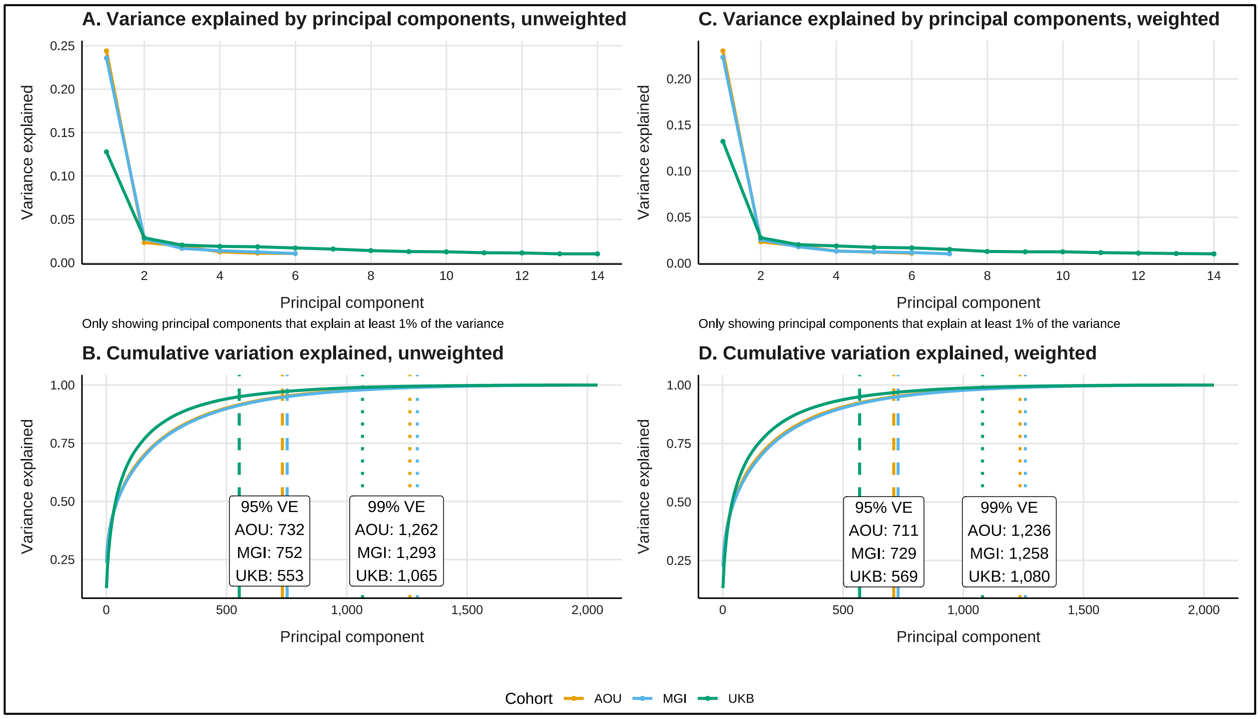
Supplementary Figure 3.1 Flowchart depicting several common data tasks. This flowchart is subjective and not exhaustive.


A. AOU
B. MGI
C. UKB
Initial data pull
After mapping ICD codes to phecodes
And applying phecode-participant sex concordance filter
ICD9-CM (n = 185,520)
ICD10-CM (n = 238,453)
Demographics (n = 413,457)
ICD-CM codes (n = 249,950)
ICD-mapped phecodes (n = 244,071)
AOU analytic dataset (n = 244,071)
ICD9-CM (n = 81,289)
ICD10-CM (n = 80,753)
Demographics (n = 81,608)
ICD-CM codes (n = 81,376)
Participants in qualifying subcohorts (n = 81,485)
ICD-mapped phecodes (n = 81,243)
MGI analytic dataset (n = 81,243)
ICD9* (n = 20,299)
ICD10 (n = 410,250)
Demographics (n = 502,413)
ICD codes (n = 410,250)
ICD-mapped phecodes (n = 401,167)
UKB analytic dataset (n = 401,167)
* 3,252 have only ICD9 codes

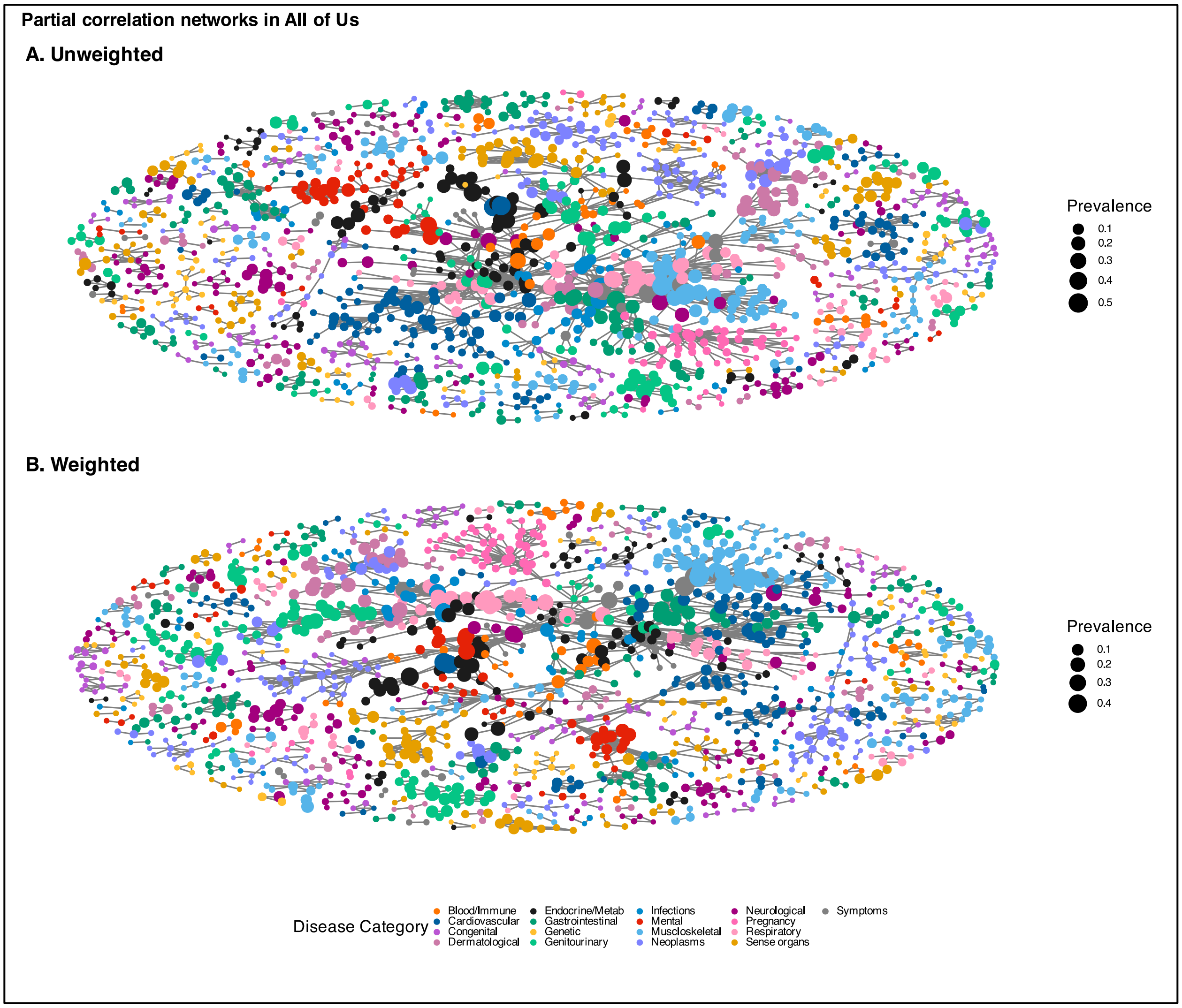
Supplementary Figure 3.2 Flowcharts depicting samples sizes before and after filter and ICD-to-phecode mapping in AOU (panel A), MGI (panel B), and UKB (panel C).



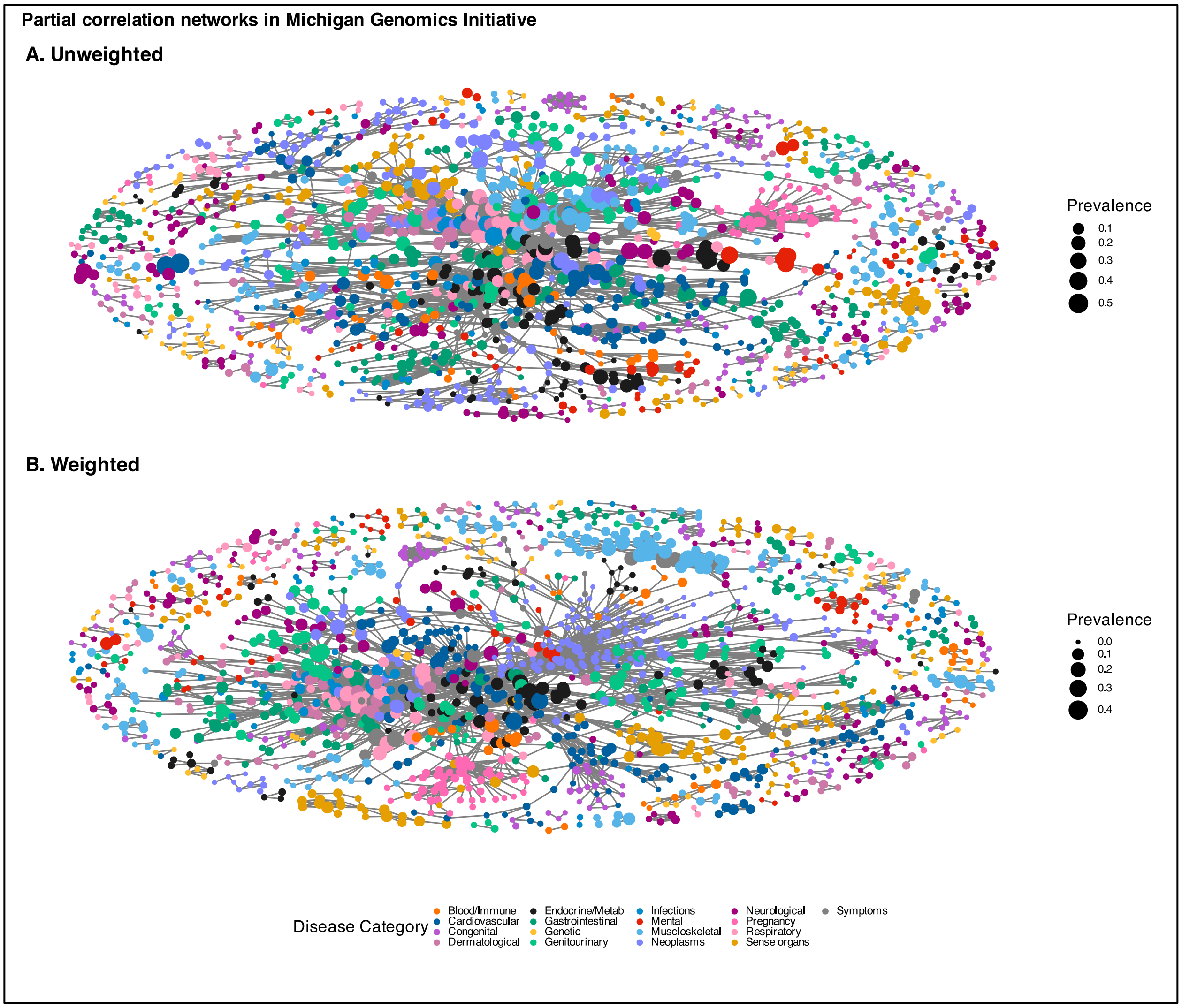
Supplementary Figure 3.3 A schematic representation of the targeted and untargeted association analyses pipelines carried out in the manuscript.



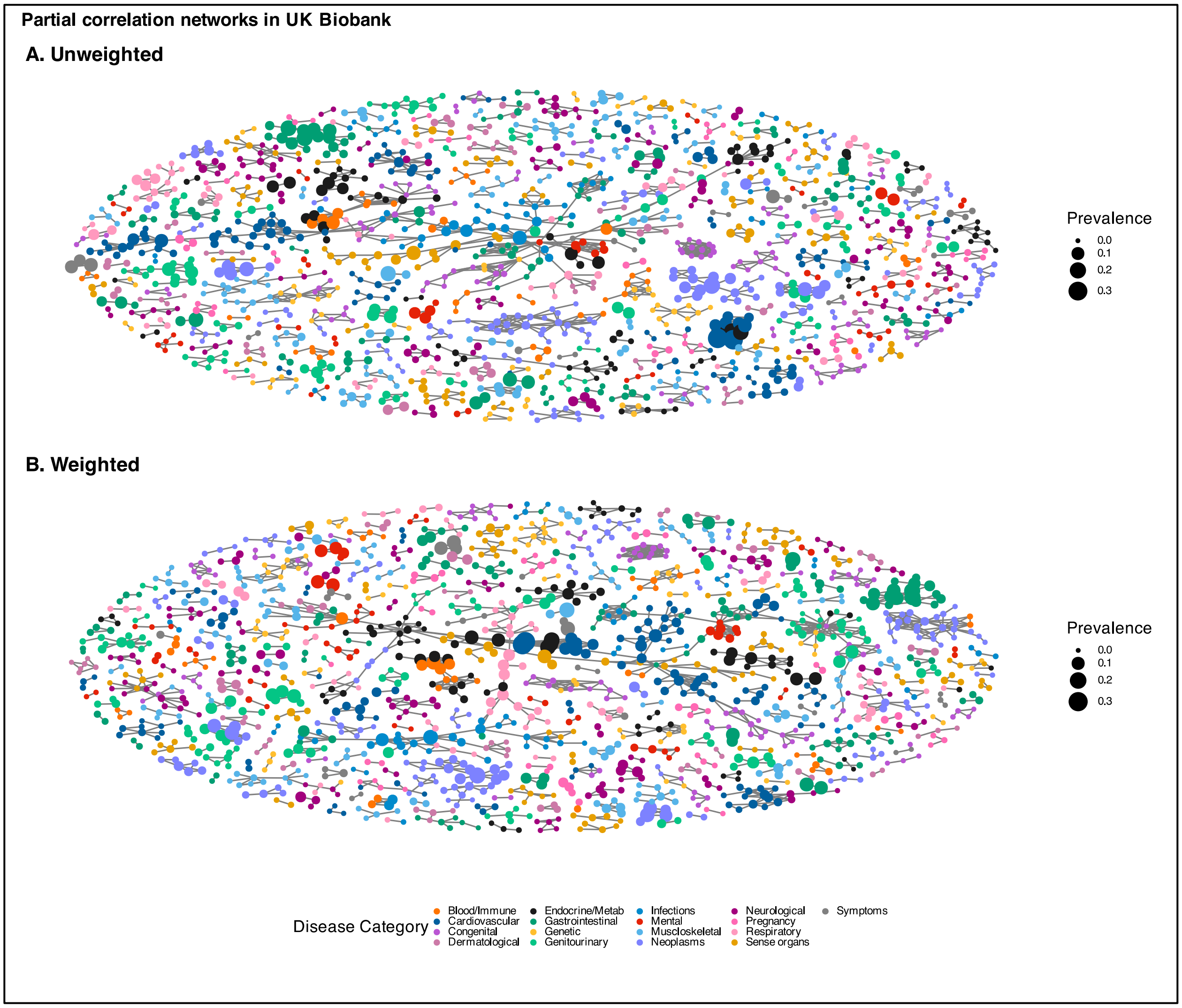
Supplementary Figure 3.4 Principal components (PC) analysis in All of Us (AOU), the Michigan Genomics Initiative (MGI), and the UK Biobank (UKB). Panel A shows all principal components explain at least 1% of variation. Panel B shows the cumulative proportion of variance explained (VE) and reports variance explanation thresholds. The vertical dashed lines represent the number of PCs that explain at least 95% of total variance. The vertical dotted lines represent the number of PCs that explain at least 99% of the total variance.



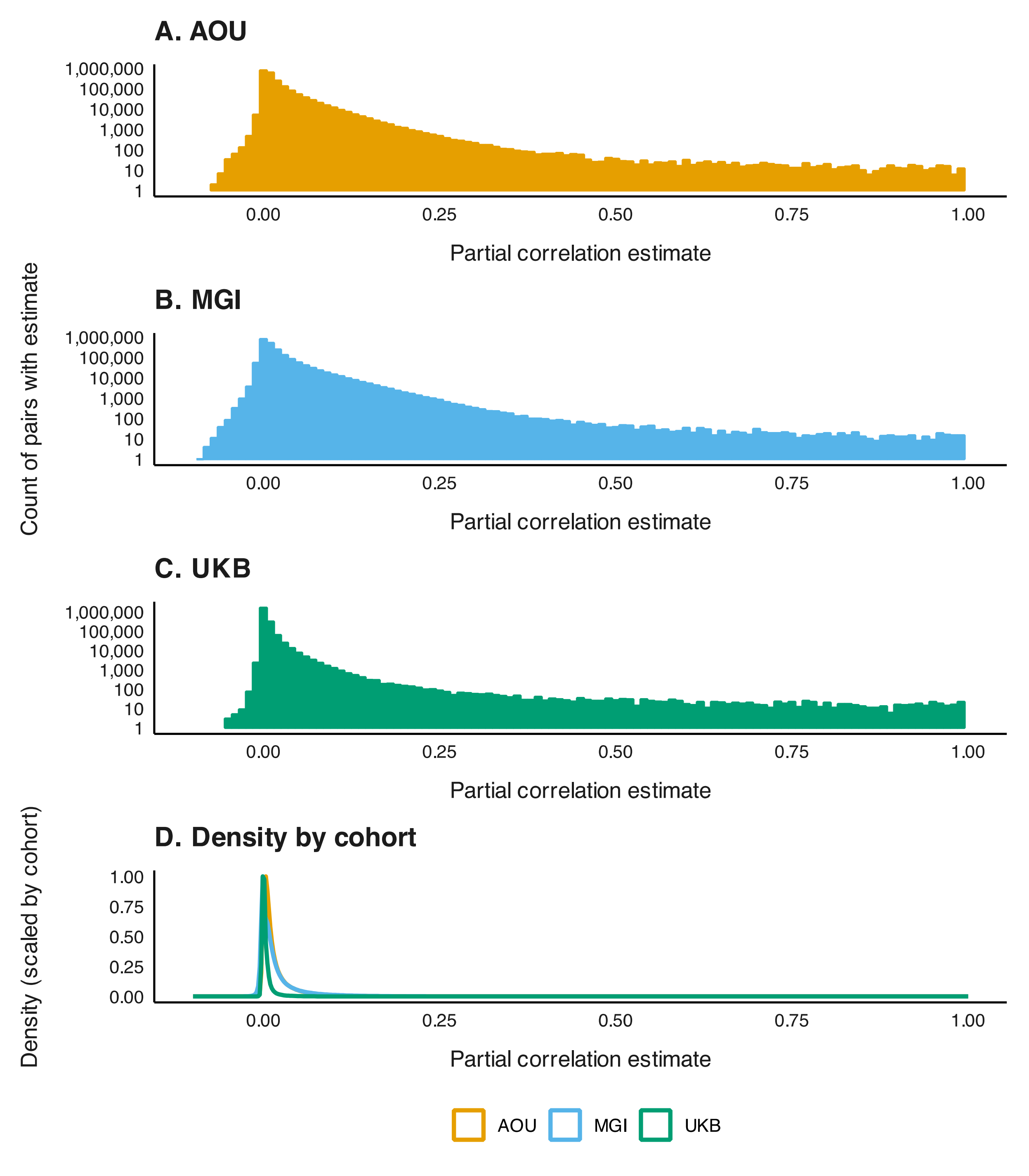
Supplementary Figure 3.5 Unweighted (panel A) and inverse probability-weighted (panel B) network plots of the partial correlation structure of medical phenomes in All of Us. Correlation coefficients are adjusted for age and sex. Only correlations with an absolute value greater than or equal to 0.3 are shown. The size of the nodes corresponds to the prevalence of the trait in its cohort and the color corresponds to the phecode category. Corresponding figures for MGI and UKB are in Supplementary Figure 3.6 and Supplementary Figure 3.7, respectively.



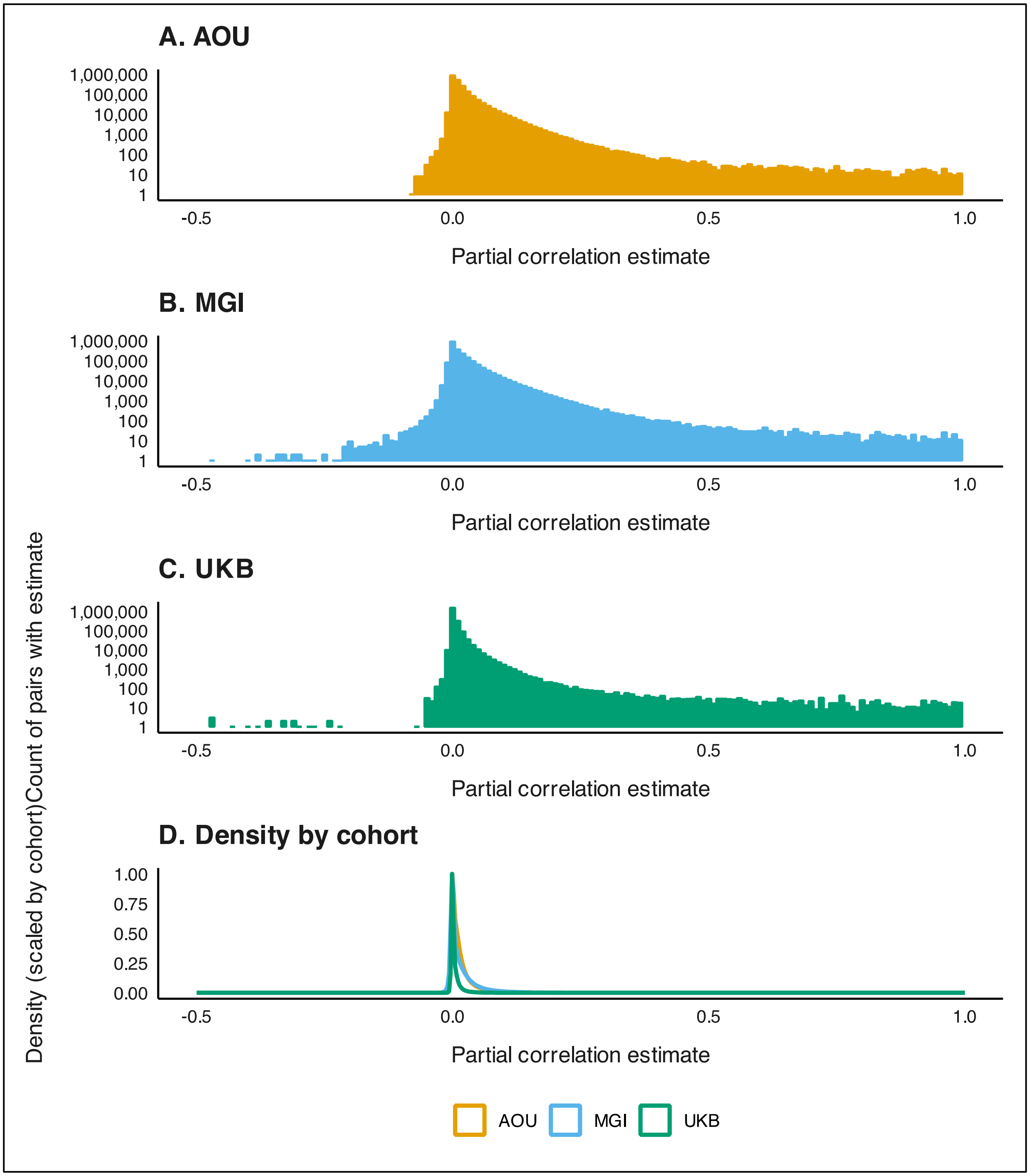
Supplementary Figure 3.6 Unweighted (panel A) and inverse probability-weighted (panel B) network plots of the partial correlation structure of medical phenomes in MGI. Correlation coefficients are adjusted for age and sex. Only correlations with an absolute value greater than or equal to 0.3 are shown. The size of the nodes corresponds to the prevalence of the trait in its cohort and the color corresponds to the phecode category. Corresponding figures for AOU and UKB are in Supplementary Figure 3.5 and Supplementary Figure 3.7, respectively.



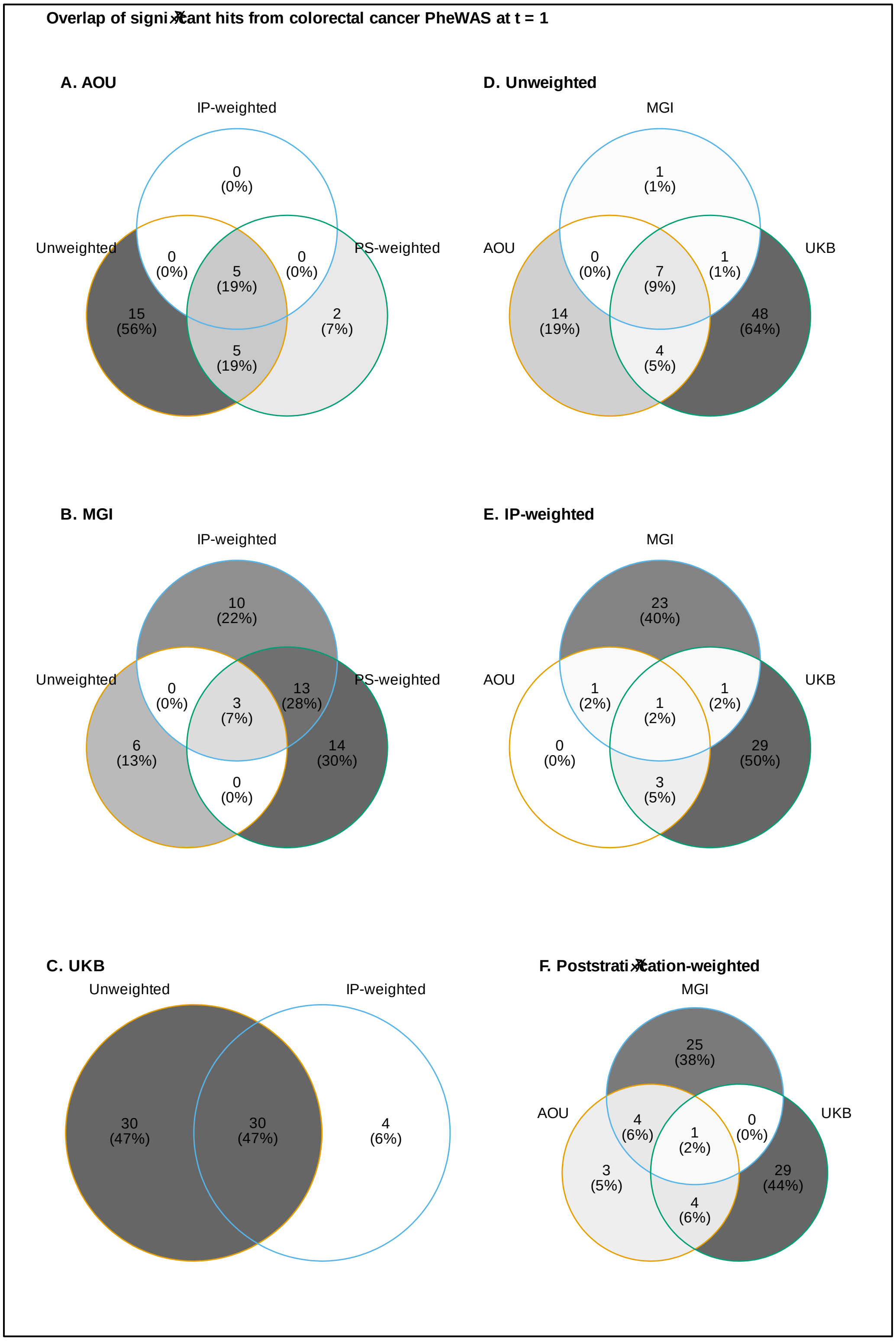
Supplementary Figure 3.7 Unweighted (panel A) and inverse probability-weighted (panel B) network plots of the partial correlation structure of medical phenomes in UKB. Correlation coefficients are adjusted for age and sex. Only correlations with an absolute value greater than or equal to 0.3 are shown. The size of the nodes corresponds to the prevalence of the trait in its cohort and the color corresponds to the phecode category. Corresponding figures for AOU and MGI are in Supplementary Figure 3.5 and Supplementary Figure 3.6, respectively.



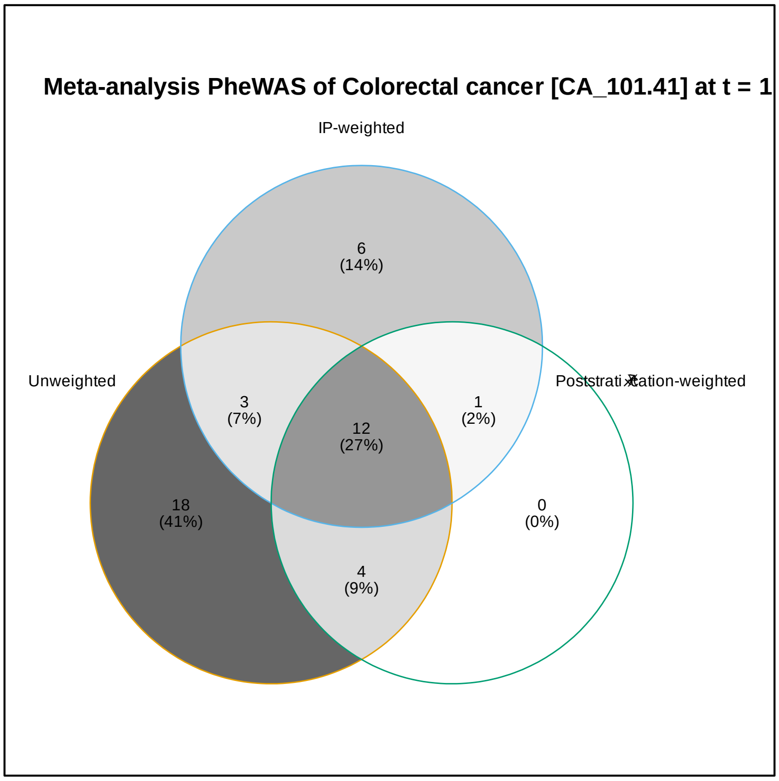
Supplementary Figure 3.8 Distribution of unweighted partial correlations across medical phenomes. Partial correlations were adjusted for age and, if both codes in the pair applied to both sexes, sex.



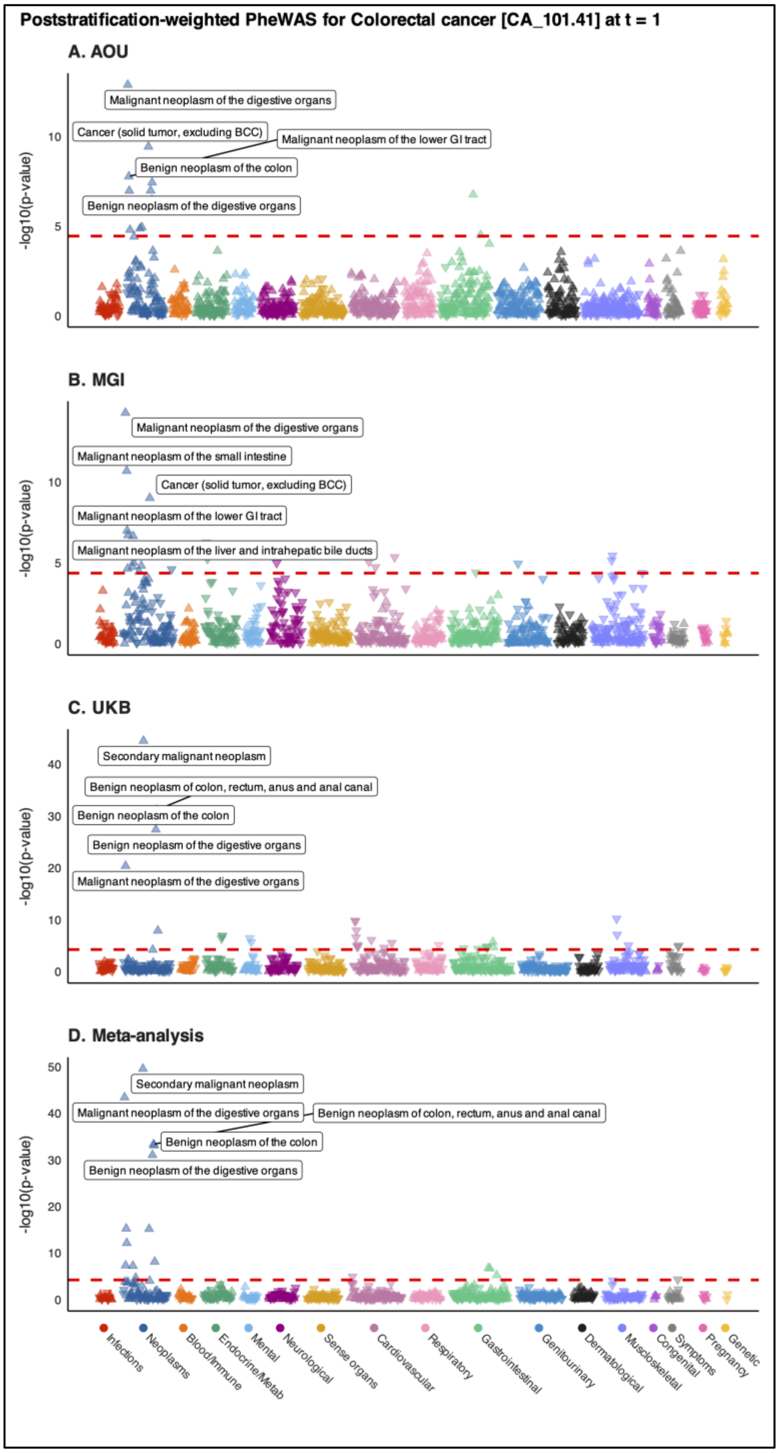
Supplementary Figure 3.9 Distribution of weighted partial correlations across medical phenomes. Partial correlations were adjusted for age and, if both codes in the pair applied to both sexes, sex. IP-based weights were used for AOU and MGI and IP-based weighted developed by van Alten and colleagues^{8} were used for UKB.



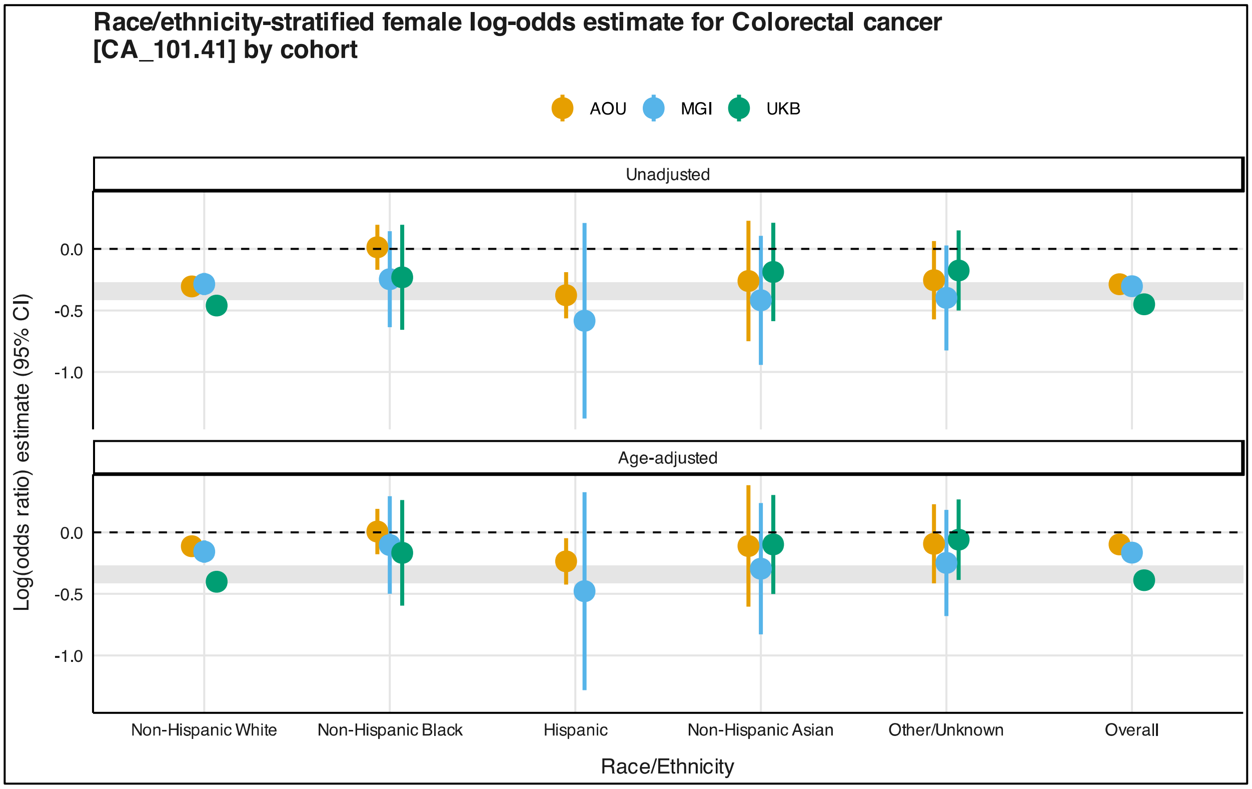
Supplementary Figure 3.10 Venn diagrams comparing the overlap in phenome-wide significant hits from unweighted and weighted colorectal cancer PheWAS in AOU, MGI, and UKB.



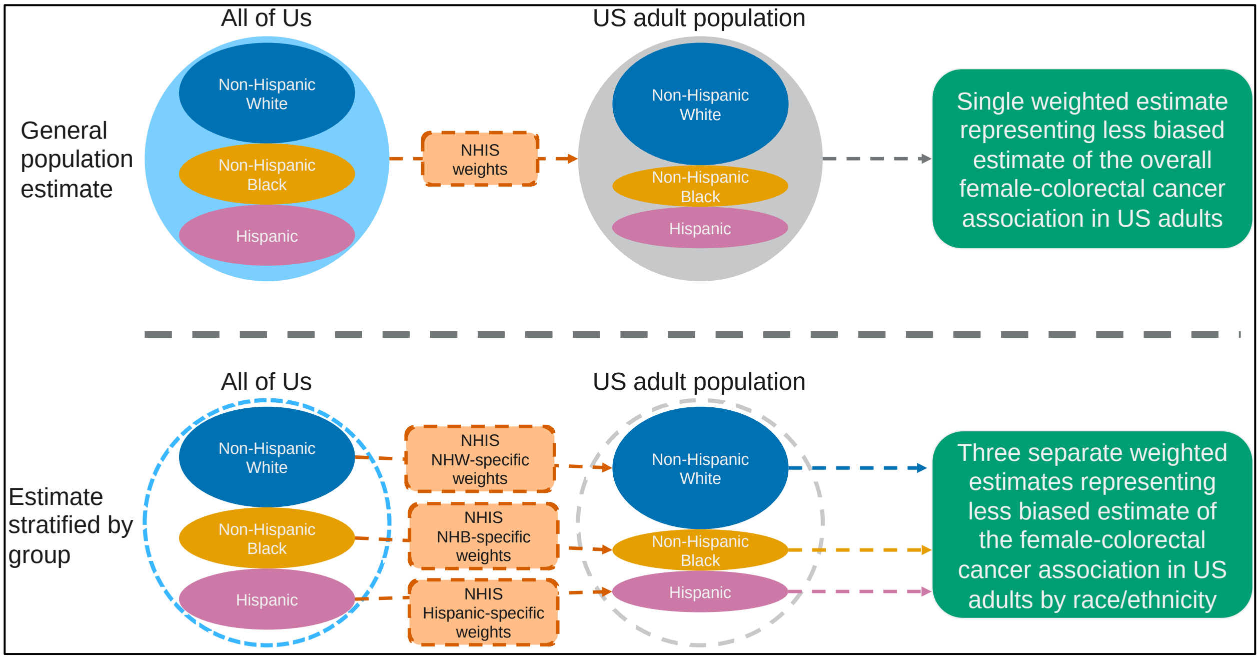
Supplementary Figure 3.11 Venn diagrams comparing the overlap in phenome-wide significant hits from meta-analysis PheWAS.



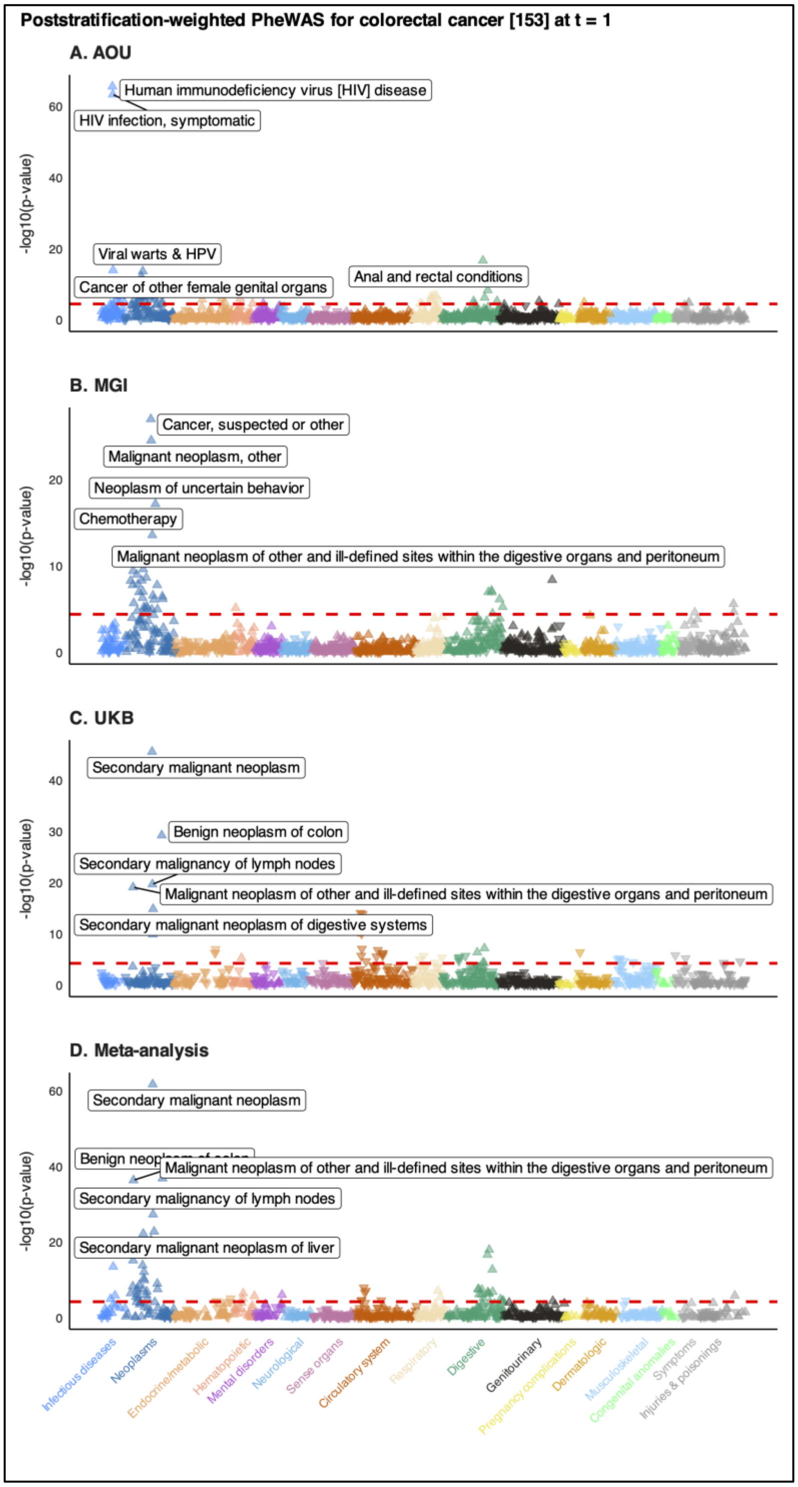
Supplementary Figure 3.12 Manhattan plots summarizing poststratification-weighted (panels A-D) phenomewide association studies for colorectal cancer in All of Us and the Michigan Genomics Initiative and the inverse probability weighted UK Biobank using 1:2 case:non-case matched data restricted to one year prior to initial diagnosis along with the corresponding meta-analysis. The dashed red line represents the Bonferroni-corrected p-value threshold (-log10(0.05/number of traits)). The five traits with the smallest p-values are labeled. The upward (downward) orientation of the triangle indicates a positive (negative) association. Plots corresponding to unweighted and IP-weighted PheWAS are presented in Figure 3.4.



Supplementary Figure 3.13 Unweighted unadjusted and age-adjusted female log-odds ratio estimate (95% confidence interval) for colorectal cancer (phecode CA_101.41) by race/ethnicity and cohort. Line colors correspond to the cohort (orange, AOU; blue, MGI; green, UKB). Shaded region represents range of age-adjusted log(incidence rate ratio [IRR]) estimates from 2018-2020 US SEER data[229] and an age-standardized log(IRR) estimate from White et al. 2018 from the UK.[230]


All of Us
US adult population
General population estimate
Non-Hispanic White
Non-Hispanic Black
Hispanic
NHIS weights
Non-Hispanic White
Non-Hispanic Black
Hispanic
Single weighted estimate representing less biased estimate of the overall female-colorectal cancer association in US adults
All of Us
US adult population
Estimate stratified by group
Non-Hispanic White
Non-Hispanic Black
Hispanic
NHIS NHW-specific weights
NHIS NHB-specific weights
NHIS Hispanic-specific weights
Non-Hispanic White
Non-Hispanic Black
Hispanic
Three separate weighted estimates representing less biased estimate of the female-colorectal cancer association in US adults by race/ethnicity

Supplementary Figure 3.14 Schematic representation of analysis pipelines where All of Us data is used to generate an association estimate representative of the general US adult population (upper half) and association estimates representative of the US adult population by race/ethnicity category (lower half). Abbreviations: NHB, non-Hispanic Black; NHIS, National Health Interview Survey; NHW, non-Hispanic White.



Supplementary Figure 3.15 Manhattan plots summarizing unweighted (panels A-C) phenomewide association studies for colorectal cancer in All of Us, the Michigan Genomics Initiative, and UK Biobank using 1:2 case:non-case matched data restricted to one year prior to initial diagnosis. Panel D shows the unweighted meta-analysis PheWAS, respectively. The dashed red line represents the Bonferroni-corrected p-value threshold (-log10(0.05/number of traits)). The five traits with the smallest p-values are labeled. The upward (downward) orientation of the triangle indicates a positive (negative) association.

3.10.3 Supplementary Methods

3.10.3.1 Inverse probability weighting

In MGI, we estimated the first term, $P(S_{external} = 1|\boldsymbol{X})$, by fitting a simplex regression model for the known design probabilities using NHIS data. We estimated the numerator of the second term, $P(S = 1|\boldsymbol{X}, S_{all} = 1)$, using a logistic regression model. We considered the set of selection factors, $\boldsymbol{X}$: age (≥ 50 indicator), female sex, BMI (categorical), non-Hispanic White race/ethnicity, and EHR-derived binary indicators for anxiety, depression, diabetes, cancer, and hypertension (variable definitions in Supplementary Table 3.2). Cancer was not included directly in the estimation procedure above because the small prevalence of cancer in NHIS led to unstable model fitting.[11] Instead, a cancer factor, γ_{cancer}, defined as $\frac{P(\text{Cancer}|\boldsymbol{X},S=1)}{P(\text{Cancer}|\boldsymbol{X})}$, was estimated by fitting logistic regression models with the same $\boldsymbol{X}$. The probabilities, ψ, were multiplied by this factor (i.e., $\psi\gamma_{cancer}$).

In AOU, we flexibly selected $\boldsymbol{X}$ by splitting the data in half and fitting a lasso-penalized logistic regression model on $\boldsymbol{X}$ and all possible pairwise interactions using the glmnet R package (version 4.1-8). We considered a set of selection factors, $\boldsymbol{X}$: age ((≥ 50 indicator), female sex, non-Hispanic White race/ethnicity, non-heterosexual orientation (yes/no), health insurance coverage status (yes/no), annual household/family income (≥ \$75,000), educational attainment (at least high school graduate or equivalent), and region of residence (indicators for West, South, and Northeast) (variable definitions in Supplementary Table 3.2). Using 10-fold cross-validation, we selected the largest λ such that the error is within 1 standard error of the minimum to result in a parsimonious model. Of the 55 possible main effect and interaction terms, 39 were selected by this model

(Supplementary Table 3.3) and (along with the main effect for West region) were then used as the final set of $\boldsymbol{X}$ to estimate IP weights in the other half of the data as described for MGI above. The indicator variables for income, health insurance status, and non-Hispanic White race/ethnicity were the three most important variables (Supplementary Figure 3.2). In both cohorts, the resulting probabilities were winsorized at the 2.5th and 97.5th percentiles.

We note that augmented inverse probability weighting (AIPW) is a doubly robust weighting method that may be of interest to the reader; see [238–240].

3.10.3.2 Poststratification weighting

In AOU, we considered the set of $\boldsymbol{X}$: age (≥ 50 indicator), female sex, non-Hispanic White race/ethnicity, sexual orientation (non-heterosexual indicator), health insurance coverage status (yes/no), annual household/family income (≥ \$75,000 indicator), and region of residence (categorical). In MGI we considered the set of selection factors, $\boldsymbol{X}$: age (≥ 50 indicator), female sex, non-Hispanic White race/ethnicity, BMI (categorical), smoking status (ever/never), and EHR-derived history of anxiety, cancer, depression, diabetes, and hypertension.

We note that other there are other weighting methods relying only on summary statistics like calibration, raking, and pseudolikelihood that may be of interest to the reader; see [7,79,104].

3.10.3.3 Correlations

We also explored the correlation structure of unweighted and weighted phenomes through partial correlations. Unweighted partial correlations were calculated between pairs of traits, X and Y, adjusted for age and sex, using the ppcor R package (version

1.1). [241] Weighted partial correlations were approximated as the coefficient β_X from the weighted multiple linear regression model $Y = \beta_0 + \beta_X X + \boldsymbol{\beta_Z Z}$, where X, Y, and $\boldsymbol{Z}$ were mean standardized and $\boldsymbol{Z}$ were age and female sex. For X, Y pairs where one trait was sex-specific, the other trait was limited to individuals of that sex, and sex was not included as a covariate. Network graphs of correlations with absolute values greater than 0.3 were constructed to visually inspect the structure. All traits were treated as binary based on the presence of a single phecode in the EHR. (see Section 3.10.4 for results).

3.10.3.4 PheWAS

The data were prepared as described in Salvatore and colleagues [18] at the one-year prior to colorectal cancer diagnosis threshold. For sex-specific phecodes, those with discordant sex were treated as missing. (Of note, some ICD codes do not map to phecodes). Logistic regression models were fit as follows:

$$logit\big(P(\mathrm{CA}_{101.41} = 1 | k, \boldsymbol{covariates})\big) = \beta_0 + \beta_k k + \boldsymbol{\beta}_{\mathbf{covariates}}\mathbf{covariates} \qquad \textit{Eq. (S1)}$$

where CA_101.41 (the phecode for colorectal cancer) is an indicator for the outcome, k represents the exposure phecode k (indicator), and covariates are age at one-year prior to colorectal cancer diagnosis (continuous), female sex (indicator), and length of EHR follow-up (continuous).

Phenomewide significant hits were identified using a conservative multiple testing corrected threshold of 0.05 divided by the number of *total* tests. Weighted logistic regression models were fit using svyglm from the survey R package.[242] In cases where a given exposure phecode did not have both (1) at least 20 occurrences and (2) at least 10 individuals with the exposure and colorectal cancer, weighted Firth bias-corrected logistic regression (logistf R package version 1.26.0) was used to address concerns about separation.

3.10.4 Unweighted and weighted partial correlations

Network diagrams depicting unweighted and weighted partial correlation coefficients with absolute values greater than 0.3 (an arbitrary threshold) in AOU is shown in Supplementary Figure 3.5 (MGI and UKB shown in Supplementary Figure 3.6 and Supplementary Figure 3.7). We can see clusters of correlated traits within endocrine/metabolic and musculoskeletal categories, as well as a cluster including both digestive and neurological traits. A small reduction in correlations with absolute values greater than 0.3 were observed after weighting (2,533 vs. 2,474). Interestingly, we see strong correlations with neoplasm traits in MGI (Supplementary Figure 3.6), which largely disappear after weighting. There are distinct clusters within musculoskeletal traits and across circulatory system and endocrine/metabolic traits in UKB, which remain after weighting. The number of strong (absolute value > 0.3) correlations in UKB slightly increases after weighting (1,674 vs 1,757). Supplementary Figure 3.8 and Supplementary Figure 3.9 depict the distribution of the unweighted and weighted partial correlation coefficients in each cohort, respectively. Generally, correlations tend to be highest in MGI followed by AOU and then UKB. Comparing the two US-based cohorts, AOU (Supplementary Figure 3.5) and MGI (Supplementary Figure 3.6), we see that, while the prevalences of traits involved in these networks are comparable, the network in MGI is denser compared to AOU.

3.10.5 Comments on methodological considerations in EHR-based data analysis

Weighting-based analytic approaches present a relatively simple way for researchers to improve the generalizability of their results and help *reduce* (not *remove*) selection bias. IP weights are preferred to PS weights though they rely on the assumption that the weighting model is correctly specified. Regression-based weights can be made more flexible through the use of indicator variables (as in our AOU IP-weights and in van Alten and colleagues[8]), though non-parametric methods like random forest can be used. When individual-level data from the target population is not available, PS weights can be estimated using summary-level strata probabilities (provided these probabilities are conditionally independent). When selection weights are unavailable, methods like covariate or propensity score adjustment, which are simple to implement, can be considered to address in some situations where selection bias is a concern.

Beyond introductory papers,[1,124–128] substantial work has focused specifically on traditional methodological concerns including confounding,[48,49] misclassification,[7,50,51] missing data,[20,23,117,123,235,243,244] and selection bias and cohort representativeness[7,11,24,76–79] related to EHR-based cohorts. For example, traits defined using the phecode framework have demonstrated reduced misclassification compared to ICD codes.[129] One method to further reduce the impact of misclassification, described by Hubbard and colleagues, relies on EHR-derived probabilistic phenotyping.[50] Others have described methods using manual chart review on a subset of data to improve EHR-derived phenotypes.[51,130,131] Beesley and Mukherjee developed three novel likelihood-based bias correction strategies to address outcome misclassification of EHR-derived disease status.[7] Teixeira and colleagues explored incorporation of unstructured data like doctors notes, which improved the identification of hypertensive individuals compared to using ICD codes and blood pressure reading cutoffs alone.[132] Missing data is another issue that has received attention to avoid loss of power and inducing selection bias (via complete case analyses) and aid in meeting assumptions necessary for multiple imputation.[243] One avenue is using non-missing genotype data available in EHR-linked biobanks to inform imputation, which demonstrated improvements in imputation of cardiovascular related measurements.[44] This idea could be extended using exposure polygenic risk scores[43] to inform imputation of missing exposure data.

Target validity is one consideration broadly applicable in health research but particularly acute in EHR-based analyses. Westreich and colleagues have defined this as a joint measure of internal and external validity of an effect estimate with respect to a specific target population.[54] Historically, internal validity, the notion that an estimate reflects the true underlying parameter in the study population, has taken precedence over external validity, that the parameter in the study population is representative of the true parameter in the target population. However, because of observation mechanisms and recruitment strategies into EHR-linked biobanks, the target population is almost certainly never (1) exactly the study sample or (2) the population of which the study sample is a simple random sample.[54] EHR researchers should think critically regarding who the results are intended for or representative of before beginning an analysis and make their target populations explicit in their work. We believe it is critical for researchers to consider weighted approaches that account for both the observation and recruitment mechanisms in each cohort (including potential subcohorts) and differences in the distribution of key characteristics between the analytic cohort and the target population.

We want to highlight some considerations that are hallmarks of EHR analysis. One such consideration is *informed presence*, defined by Goldstein and colleagues as "the notion that inclusion in an EHR is not random but rather indicates that the subject is ill, making people in EHRs systematically different from those not in EHRs."[55] This resulting discrepancy harms generalizability to general populations who tend to be healthier than those in the EHR data sample and results in bias. This concept extends to individuals within the EHR – those that are sicker tend to have more encounters and records than those who are healthier – and, in some cases, to records in the EHR (e.g., lab results). This phenomenon is illustrated by Agniel and colleagues, which shows that the presence and timing of laboratory results was more informative than the value of the laboratory results themselves.[133] Interested readers can learn more about informed presence elsewhere.[1,25,55,56,134] Including EHR metadata, like length of follow-up, number of encounters, density of laboratory measurements, and visit type (e.g., outpatient vs inpatient vs emergency), and careful selection or matching of controls in analyses are recommended to improve exchangeability and attempt to make EHR observation mechanisms comparable.

3.10.6 Investigation into infectious diseases peak in AOU PheWAS using phecode 1.2 mapping tables

An earlier version of the manuscript was performed using the phecode 1.2 mapping tables instead of phecode X. The Manhattan plot representing the colorectal cancer PheWAS in AOU in Supplementary Figure 3.15 shows a peak in the infectious disease category. The top hit is Human immunodeficiency virus [HIV] disease, or phecode 071 in the phecode 1.2 mapping tables. It is well established that there is no association between HIV status and colorectal cancer.[245,246] We investigated the underlying ICD codes that qualify as a colorectal cancer case. Our analyses in the manuscript use the phecode mapping table present in the PheWAS R package (version 1.2).[228,247] We also show qualifying ICD codes for a different phecode mapping table (version X),[61,227] which defines over 3,600 traits. The results of the differences in qualifying ICD codes, number of individuals with the ICD code, and the number (and percent) overlap with individuals who have HIV according to their version 1.2 defined phecode are summarized in Supplementary Table 3.4. We see that there is significant overlap between individuals with ICD codes for anal Pap smears, inconclusive results and *carcinoma in situ* and HIV status. These codes are present in the version 1.2 mapping table, but not in the version X mapping table. Codes present in the version 1.2 definition also include malignant neoplasms of the anus, but not in the version X definition. And there is evidence that people living with HIV experience higher incidence of anal cancer.[248] Because version X has more traits, there is greater separation between colorectal cancer and anal cancer.

Chapter 4 The Impact of Sample-Weighting on Risk Prediction and Risk Stratification Properties of Prediction Models Trained in One EHR-Linked Biobank When Applied to Another Biobank with a Different Recruitment Strategy: A Case Study in the United States

4.1 Abstract

Should weights be considered for developing risk prediction/stratification models using electronic health record (EHR)-linked biobank data when the external test cohort has a different sampling strategy than the internal training sample? To answer this question, we calculated two sets of poststratification (PS) weights to make a hospital-based biobank, the Michigan Genomics Initiative (MGI; n=76,757) in the United States, resemble a nationally recruited biobank in the US that oversamples groups historically underrepresented in biomedical research, All of Us (AOU; n=226,764) and assessed the impact of using these weights on the performance of risk scores constructed based on EHR data in MGI. Basic PS weights (PS_{BASIC}) included age, sex, and race/ethnicity; full PS weights (PS_{FULL}) additionally included smoking, alcohol consumption, BMI, depression, hypertension, and the Charlson Comorbidity Index. We compared weighted and unweighted versions of six commonly used methods, including lasso, ridge, elastic net, and random forest, to diagnosis code-derived phecode X data to develop phenotype risk scores (PheRS). We developed risk prediction models using MGI EHR data from 0, 1, 2, and 5 years prior to the index diagnosis date for three cancer types: esophageal, liver, and pancreatic, where there is a pressing need for early detection and screening

tools. PheRS were considered in concert with three other groups of predictors: basic covariates (age, sex, race/ethnicity), known risk factors (e.g., alcohol consumption for liver cancer; curated for each outcome), and a presenting symptom (e.g., weight loss for no known reason for liver cancer; curated for each outcome). The primary risk stratification metric of interest was the odds ratio (OR), comparing the top decile to the middle 40th-60th percentile of the risk score distribution. We also calculated other measures for evaluating prediction models, such as the area under the receiver operating curve (AUC), Hosmer-Lemeshow goodness-of-fit statistic, and Brier Score. While no single PheRS construction approach uniformly performed better in terms of risk stratification or discrimination, elastic net and random forest tended to exhibit good properties in general. In no setting did the use of PS weights consistently or meaningfully improve risk stratification performance (e.g., unweighted random forest PheRS alone OR (95% CI) for liver cancer at t=1: unweighted: 13.73 (8.97, 21.01), PS_{BASIC}-weighted: 14.55 (9.45, 22.42), PS_{FULL}-weighted: 13.62 (8.90, 20.85)). The indeterminate impact of PS weights applied to other prediction diagnostics and to the predictive performance of other data domains. PheRS was the most important in risk stratification compared to the other three domains of predictors (e.g., unweighted OR (95% CI) for liver cancer at t=1: covariates and risk factors: 1.75 (1.16, 2.63), plus random forest PheRS: 7.02 (4.75, 10.38), plus presenting symptom: 6.26 (4.18, 9.39)). The results for liver cancer are indeed encouraging for an agnostic EHR-based approach towards early detection. Researchers should consider EHR-embedded health history (i.e., PheRS) alongside other data domains like genetics, laboratory results, and medication data to improve risk stratification and predictive properties of clinical prediction models. The use of weights does not conclusively alter the performance of the risk scores we considered when the transferability of prediction models from one biobank to the other is considered.

4.2 Introduction

Risk prediction models are classic tools in clinical medicine and precision health. Prediction models and resultant risk scores have been developed to identify or stratify individuals at elevated risk for many health-related outcomes to prioritize prevention, screening, diagnostic, and treatment approaches.[14,249] A classic example is the Framingham Risk Score, which predicts 10-year risk for cardiovascular disease;[14] individuals at higher risk are often recommended to start preventive treatment based on this score.[137] Many risk score models exist for cancer, including breast,[140–142] ovarian,[141] colorectal,[15,143,144,250–252] esophageal,[13,144,253,254] liver,[146,147,193,255,256] and pancreatic cancers,[18,144,250,257,258] that identify high-risk individuals who might benefit differential preventive or treatment approaches than those offered to individuals at baseline risk. Risk prediction and stratification are critical for cancers where early detection is poor, and screening is generally unavailable.

Electronic health record (EHR)-linked biobanks, which are cohorts combining EHR, survey, and genetic data with other linkable data (e.g., cancer and vital status registries, prescription and insurance claims data, neighborhood-level environmental exposures), are rapidly increasing in size and number.[1] Examples include the UK Biobank (UKB)[3] and the US-based NIH All of Us Research Program (All of Us; AOU),[2] each containing over 500,000 participants. Researchers have used diagnosis codes in EHR-linked biobanks to summarize an individual's health history for risk prediction.[18,259–262] For example, we used diagnosis code data to develop a pancreatic cancer phenotype risk score (PheRS) in the University of Michigan's Michigan Genomics Initiative (US-based; MGI) using a parametric pruning-and-thresholding approach and assessed the performance of these scores in the UKB.[18]

However, EHR-linked biobanks often adopt sampling mechanisms such as recruiting patients who are awaiting surgery (MGI),[45] oversampling groups historically underrepresented in research (as in AOU),[2] or enacting strategies that result in healthy volunteer self-selection (as in UKB),[4] making them not representative of their respective source populations (or comparable to one another). One problem that can arise with predictions is lack of transferability. This happens when differences exist in the underlying distributions between the data used to create a risk prediction model and the sample to which the model is applied. As a result, the model may have sub-optimal predictive performance in the target sample.[199] One approach to addressing this problem is called transfer learning, which can adapt models developed in one sample for use in a second sample by using a relatively small amount of information from the second sample.[173,174] Alternatively, a sample weighting-based approach to transferable model building could be considered.[180]

Weighting-based methods are commonly employed to address the lack of representativeness between the analytic sample and its source population (i.e., selection bias).[5,7,77,81,82] They can also address differences between the sample and an external target population of interest (i.e., transportability).[54] Steingrimsson and colleagues developed an inverse-odds weight-based framework for transporting risk prediction models for use in an external target population where outcome data is unavailable.[180] The lack of readily available software to implement variable selection and machine learning methods to complex multi-stage samples has been a bottleneck in developing risk prediction models for weighted data. Recently, Iparragirre and colleagues developed a method for tuning hyperparameters for lasso models for risk prediction in weighted settings.[17] Their general framework can be extended to other regularized regression and random forest risk prediction models, which could enhance the transferability of risk prediction models through the use of weights. Such transferability would accelerate the integration of risk prediction models developed in one biobank into another healthcare system, making them available to more clinicians at the point of care.[263]

Building off our work[18] and that of Iparragirre and colleagues,[17] we assessed the impact of poststratification (PS) weights on the performance of PheRS for esophageal, liver, and pancreatic cancers at four time thresholds (t=0,1,2,5 years) before diagnosis of the index cancer. We considered six different methods commonly used for risk score construction. One-step PheRS methods included regularized regression (lasso, ridge, elastic net) and random forest models, and two-step PheRS methods included univariable and multivariable pruning-and-thresholding-like approaches. We developed PheRS using EHR data in MGI, a cohort enriched with cancer diagnoses because of its perioperative recruitment strategy. We assessed their performance in the US-based cohort AOU. MGI was split 50/50 such that hyperparameter tuning or feature selection was performed in 50% of the data, and model fitting to obtain the beta-coefficients or estimates of fitted parameters was conducted in the other 50%. Two sets of PS weights were calculated: $PS_{BASIC,}$ which accounted for age, sex, and race/ethnicity, and PS_{FULL}, which additionally included smoking, alcohol consumption, BMI, depression, hypertension, and Charlson Comorbidity Index (CCI, which captures local and metastatic cancers). We had weighted and unweighted versions of the fitted models, with weights designed to make MGI resemble the target cohort AOU. Evaluating in AOU, we had three aims: (a) to compare different PheRS construction approaches, (b) to determine whether using PS weights improved PheRS performance, and (c) to contrast PheRS performance with that of three other predictor domains: basic demographic covariates, risk factors, and presenting symptoms. For each aim, we chose several metrics associated with prediction (namely discrimination, calibration, and accuracy measures) and measures of risk stratification (Figure 4.1). Finally, we made recommendations regarding (a) the choice of methods for constructing PheRS, (b) the use of PS weights to make PheRS more transferrable to another cohort, and (c) the role of PheRS in comparison to other sets of predictors in the development of EHR-based risk scores.

4.3 Results

4.3.1 Characteristics of the training and assessment cohorts

Unweighted, AOU (n=226,764) had an average age of 54, was 62% female, and was 55% non-Hispanic White (Table 4.1). Additionally, 12% were high on the CCI, 88% had reported ever consuming alcohol, and 26% had a record of depression (phecode MB_286.2). MGI (n=76,757) had a similar average age of 57 but was less female (54%), more White (84%), had more individuals with high CCI (33%), self-reported less alcohol (69%), and had a higher rate of depression (32%).

Using PS weights that accounted for age, sex, and race/ethnicity (i.e., PS_{BASIC}), MGI looked more similar with respect to age (mean 55 years old), sex (60% female), and race/ethnicity (60% non-Hispanic White), but still had more individuals with high CCI (30%), less reported alcohol consumption (68%), and higher rates of depression (33%).

PS weights that additionally accounted for smoking, alcohol consumption, body mass index, depression, hypertension, and CCI (i.e., PS_{FULL}), MGI generally mimicked the marginal frequencies more like AOU, including age (mean 55 years old), sex (61% female), race/ethnicity (65% non-Hispanic White), high CCI (13%), reported alcohol consumption (89%), and rate of depression (27%).

Regarding the three digestive cancer outcomes we consider, there were 193 esophageal, 599 liver, and 385 pancreatic cancer diagnoses in AOU. In MGI, there were 389 esophageal, 337 liver, and 311 pancreatic cancer diagnoses.

4.3.2 Methods for constructing phenotype risk score

Overall finding: No single PheRS approach is best for risk stratification or discrimination

Risk stratification: PheRS risk stratification capacity was assessed by measuring the top decile of risk score compared to the middle 40th-60th percentile in terms of the relevant cancer outcome odds ratio (hereafter, simply OR). Different unweighted PheRS approaches performed best for esophageal cancer depending on the time threshold (Table 4.2). For example, the two-step multivariable PheRS (2.40 (1.30, 4.43)) performed best, while the ridge PheRS performed worst (1.36 (0.62, 2.99)) at the t=1 threshold. At the t=2 threshold, the univariable PheRS performed best (2.26 (1.16, 4.42)), and the random forest PheRS performed worst (0.81 (0.35, 1.88)). This pattern was seen for unweighted PheRS for the other cancer outcomes. For liver cancer, the random forest PheRS performed best (t=1: 13.73 (8.97, 21.01); t=2: 16.42 (10.19, 26.46)) and the elastic net PheRS performed worst (t=1: 2.66 (1.95, 3.63); t=2: 1.37 (0.95, 1.99)) at both t=1 and t=2. For pancreatic cancer, the ridge PheRS performed best at both t=1 and t=2; random forest (1.15 (0.64, 2.08)) and univariable PheRS (0.98 (0.53, 1.78)) performed worst at t=1 and t=2, respectively.

The following section reports results regarding the impact of weights on the risk stratification and discriminatory ability of PheRS.

Discrimination: The area under the receiver-operator characteristics curve (AUC) was considered a summary measure of each PheRS' discriminatory ability. Different unweighted esophageal cancer PheRS approaches performed better at different time thresholds. For example, the elastic net PheRS performed best at the t=1 threshold (AUC (95% CI): 0.594 (0.552, 0.636)) while the multivariable PheRS performed best at the t=2 threshold (0.610 (0.545, 0.674); Table 4.3). The best PheRS approach in terms of AUC at a given time threshold for one outcome was not necessarily the best PheRS approach for a different outcome. For example, while the elastic net PheRS for esophageal and pancreatic cancers had the highest AUC at t=1, the lasso (0.771 (0.742, 0.800)) and random forest (0.771 (0.741, 0.801)) PheRS were highest for liver cancer. Though no clear winner exists, elastic net and random forest generally exhibited better discrimination than the other approaches.

While all PheRS approaches exhibited fair calibration (by Hosmer-Lemeshow goodness-of-fit test) and comparable accuracy (by Brier score; highest accuracy was observed for liver cancer (e.g., mean Brier score across all PheRS at t=1: 0.170; followed by esophageal (0.214) and pancreatic (0.220)), no discernible patterns of consistent, substantial differences in these metrics by PheRS methods were observed (t=0,1,2,5 in Supplementary Table 4.3, Supplementary Table 4.4, Supplementary Table 4.5, and Supplementary Table 4.6, respectively).

4.3.3 Influence of weights on the performance of optimal phenotype risk score

Overall Finding: Weights do not substantially alter risk stratification or discriminatory properties of PheRS

Risk stratification: Considering weights in the development of PheRS did not consistently or meaningfully change risk stratification performance across outcomes, PheRS approaches, or time thresholds. For example, consider the random forest liver cancer PheRS approach across time thresholds (Figure 4.3). The unweighted approach had the highest OR point estimate (unweighted random forest PheRS: 63.80 (36.78, 110.68)) compared to both weighted versions for each PheRS approach (PS_{BASIC}-weighted random forest PheRS: 49.94 (30.52, 81.73); PS_{FULL}-weighted random forest PheRS: 31.14 (20.43, 47.47)) at t=0, though the confidence intervals overlapped. At the t=1 threshold, the PS_{BASIC}-weighted random forest PheRS (14.55 (9.45, 22.42)) yielded a higher OR than the unweighted (13.73 (8.97, 21.01)) and PS_{FULL}-weighted (13.62 (8.90, 20.85)) versions. In this instance, comparing the two weighted approaches, the unweighted version performed better at t=0, while the PS_{BASIC}-weighted version performed better at t=1.

A lack of a uniformly superior PheRS approach was seen when looking across liver cancer PheRS at the same time threshold. For example, at the t=1 threshold, among random forest PheRS, the PS_{FULL}-weighted version performed best (unweighted: 1.25 (0.99, 1.58); PS_{BASIC}-weighted: 1.19 (0.95, 1.50); PS_{FULL}-weighted: 1.50 (1.20, 1.88)). However, among lasso PheRS, the unweighted version performed best (unweighted: 12.24 (8.74, 17.14); PS_{BASIC}-weighted: 1.00 (0.99, 1.01); PS_{FULL}-weighted:7.86 (5.97, 10.34)).

In general, unweighted and weighted PheRS OR confidence intervals tended to overlap within (and, to a lesser extent, across) PheRS approaches, with similar conclusions regarding magnitude and statistical significance. These observations were seen across all three cancer outcomes (Supplementary Figure 4.1 and Supplementary Figure 4.2).

4.3.4 Assessing the relative contribution of PheRS with other domains of data for risk stratification and discrimination

Overall Finding: PheRS contributes significantly to risk stratification and discrimination alongside demographic covariates, risk factors, and presenting symptoms

Risk stratification: A series of models that we refer to as the model cascade were fit to determine the individual, cumulative, and combined risk stratification capacity of multiple domains of predictive data: covariates (age, sex, race/ethnicity), risk factors (obesity, alcohol, and smoking status), PheRS, and a presenting symptom (curated for each outcome based on literature; see Materials and methods). For example, for liver cancer at t=1, unweighted OR (95% CI) for covariates, risk factors, random forest PheRS, and presenting symptom alone were 1.27 (0.82, 1.95), 1.49 (1.07, 2.09), 13.73 (8.97, 21.01), and 1.53 (1.21, 1.95), respectively (Figure 4.2). The OR increased from covariates alone to covariates and risk factors combined (1.75 (1.16, 2.63)) and again, substantially, after adding the random forest PheRS (7.02 (4.75, 10.38)). The increase in OR after the inclusion of PheRS indicates that it adds to risk stratification capacity alongside covariates and risk factors. Additionally, including the presenting symptom slightly attenuated the OR estimate to 6.26 (4.18, 9.39), possibly due to correlation with PheRS (ρ = 0.22). Finally, considering all factors jointly (i.e., covariates, risk factors, phecodes, and symptom simultaneously in a random forest model) performed almost as well as PheRS alone (13.57 (8.90, 20.70)). Among sequential and joint models, we see that PheRS substantially contributes to risk stratification for those in the top decile compared to those in different parts of the risk score distribution (Figure 4.4, left panel). The contribution of PheRS to risk stratification is most pronounced among those in the highest decile rather than those in the second-, third-, and fourth-highest deciles.

To explore the consistency of feature selection across time thresholds, we gathered the top 10 features by variable importance from unweighted random forest PheRS for liver cancer (Table 4.4). As expected, the top features primarily came from the gastrointestinal phecode category. Established risk factors, chronic liver disease, fibrosis and cirrhosis of liver, and cirrhosis of liver, were selected at all time thresholds. Hepatitis and hepatovirus were top 10 features at all non-0 time thresholds. Several peri-liver cancer diagnosis features were identified at t=0 that ranked lower at other time thresholds, including hepatomegaly, diseases of the pancreas, and obstruction of bile duct. Important peri-diagnosis features present an opportunity for targeted follow-up to determine specific presenting symptoms. Top features at the t=5 threshold, including back pain, hyperlipidemia, diseases of spleen, and nonspecific abnormal results of function study of liver, were identified and remained relatively important as time threshold decreased.

Results from the model cascade for other outcomes at t=1 were attenuated for pancreatic and, to a lesser extent, esophageal cancer. For example, for pancreatic cancer, PS_{BASIC}-weighted OR for covariates, risk factors, random forest PheRS, and presenting symptom alone were 0.94 (0.51, 1.74), 1.15 (0.80, 1.66), 1.20 (0.66, 2.18), and 1.00 (0.99, 1.01), respectively. Adding risk factors to covariates (1.78 (0.98, 3.23)) improved the OR compared to each alone. Adding the random forest PheRS to covariates decreased the OR to 0.88 (0.47, 1.64) while adding the presenting symptom increased the OR to 0.96 (0.52, 1.78). One potential explanation for the decrease in OR point estimate after adding the PheRS is collinearity between the risk factors and PheRS. Considering covariates, risk factors, diagnosis history, and a presenting symptom jointly did not change the OR (0.93 (0.53, 1.65)). For comparison, the joint PS_{BASIC}-weighted random forest OR for esophageal cancer at t=1 was 3.31 (1.58, 6.94).

We also compared the top 10 features according to unweighted random forest variable importance across outcomes at t=1 (Table 4.5). While they are all digestive cancers, there is substantial heterogeneity in their risk factors and presentation. None of the top 10 features for one cancer appeared in the top 10 for another and were often outside the top 100 most important features. For example, the most important feature for liver cancer at t=1, chronic nonalcoholic liver disease, was the 56th and 184th most important feature for pancreatic and esophageal cancer, respectively. We also saw differences in phecode group representation. While 8 of the top 10 features were gastrointestinal for liver cancer, only 4 were for esophageal cancer and 3 for pancreatic. Interestingly, there were 4 cardiovascular features in the top 10 for esophageal cancer, including atrial fibrillation and flutter, abnormal results of cardiovascular function studies, essential hypertension, and ischemic heart disease.

Discrimination: Similar results were observed concerning discriminatory ability as measured by AUC. For example, for liver cancer at t=1, unweighted AUC (95% CI) for covariates, risk factors, random forest PheRS, and presenting symptoms alone were 0.573 (0.541, 0.605), 0.572 (0.540, 0.603), 0.771 (0.741, 0.801), and 0.548 (0.521, 0.575), respectively. Combining covariates and risk factors improved AUC compared to each domain individually (0.601 (0.570, 0.633)). However, the subsequent addition of the random forest resulted in a significant improvement (0.701 (0.669, 0.732); Figure 4.4, right panel). Adding the presenting symptom at this time threshold did not improve AUC (0.673 (0.640, 0.705)). The jointly constructed model exhibited the highest discriminatory ability (0.776 (0.747, 0.806). The models where PheRS was included exhibited a noticeable shift in the AUC curve towards the upper left corner of the plot (Figure 4.4, right panel).

Results for other outcomes at t=1 were attenuated for esophageal and pancreatic cancers. For example, for pancreatic cancer, unweighted AUC (95% CI) for covariates, risk factors, random forest PheRS, and presenting symptom alone were 0.516 (0.470, 0.562), 0.519 (0.474, 0.564), 0.532 (0.486, 0.578), and 0.497 (0.493, 0.500), respectively. Adding risk factors to covariates did not change AUC (0.487 (0.441, 0.534)), while the subsequent addition of PheRS saw a nominal increase (0.511 (0.465, 0.557). Adding the presenting symptom lowered the AUC (0.492 (0.446, 0.538). The joint model (along with PheRS alone) exhibited statistically significant discriminatory ability (0.549 (0.504, 0.595)). For comparison, the joint unweighted random forest AUC for esophageal cancer at t=1 was 0.575 (0.513, 0.637).

Supplementary Table 4.3, Supplementary Table 4.4, Supplementary Table 4.5, and Supplementary Table 4.6 contain diagnostics for all models in the model cascade for all outcomes and PheRS and weighting approaches at t=0,1,2, and 5, respectively.

4.4 Discussion

Summary contributions: We explored the impact of sample weights on the performance of diagnosis-code-based risk predictions for esophageal, liver, and pancreatic cancers developed in one EHR-linked biobank (the University of Michigan's Michigan Genomics Initiative or MGI) for use in an external EHR-linked biobank with a different recruitment mechanism (the NIH All of Us Research Program). Two sets of poststratification (PS) weights were estimated in MGI to make it representative of the AOU population. We modified an existing R package developed to perform hyperparameter tuning in complex survey design settings for lasso models[17] to accommodate lasso, ridge, elastic net, and random forest models. Using time-restricted ICD code-derived phecode data, we constructed unweighted and weighted one-step (lasso, ridge, elastic net, and random forest) and two-step (univariable and multivariable; pruning-and-thresholding-analogous) risk score models (called PheRS). We compared and combined these models with covariates, risk factors, and symptoms to (a) compare PheRS construction approaches, (b) assess the impact of PS weights on optimal PheRS performance, and (c) discern the relative performance of PheRS alongside demographic covariates, risk factors, and presenting symptoms when evaluating in an external sample using EHR-linked biobank data. Thus, the paper presents a comprehensive empirical assessment of transferring predictions from one biobank to another using sample weighting.

Methodological novelty: This work contributes to the growing literature regarding weights, selection bias, and the challenge of transferable risk prediction[264] in EHR-linked biobank and clinical settings. We previously (a) developed a framework for estimating time-based two-step risk scores using diagnostic data[18] and (b) explored the use of selection weights in making common EHR-linked biobank analyses with non-probabilistic recruitment mechanisms more generalizable.[215] Beesley and Mukherjee[7,11] and Kundu and colleagues[79] investigated weighting-based methods to account for selection bias in EHR-linked biobank analyses. None of these studies explore the impact of weights on prediction. Steingrimsson and colleagues proposed a weighting-based approach to developing and assessing a risk score model in an external sample in which one does not have outcome information.[180] Of relevance to developers of clinical prediction models is the fact that we expanded the framework developed by Iparragirre and colleagues to tune the λ hyperparameter for lasso models in weighted settings to ridge, lasso, elastic net, and random forest models. We made the R code publicly available via a GitHub repository (https://github.com/maxsal/weighted_prediction).

Other non-weighting-based methods have demonstrated promise, including semi-supervised models,[265] cross-site feature selection,[266] and multi-site model building.[267] These approaches can be explored to create more generalizable risk prediction models. Alternatively, when models have been developed using large datasets, transfer learning can be applied to improve prediction performance by using relatively small amounts of information from the target cohort.[173,174,268]

PheRS exhibits risk stratification and discriminatory ability, but no single approach uniformly performed best in the AOU test cohort

No PheRS approach was best regarding risk stratification or discriminatory capacity, and their performance varied by outcome and worsened as time threshold increased. For example, ORs (95% CI) for unweighted random forest PheRS at the t=1 threshold ranged from 1.15 (0.64, 2.08) for pancreatic cancer to 13.73 (8.97, 21.01) for liver cancer. These estimates fell to 1.11 (0.53, 2.34) and 6.84 (4.22, 11.11), respectively, when the time threshold increased to t=5 (Supplementary Table 4.6). An alternative approach, elastic net, performed slightly better for pancreatic cancer (1.19 (0.66, 2.14) and much worse for liver cancer (2.64 (1.79, 3.90) at t=1. However, regardless of approach, health history captured by diagnosis codes and summarized as PheRS can generally perform risk stratification. At least one unweighted PheRS approach resulted in a statistically significant OR for all outcomes through t=2 and all by pancreatic cancer through t=5 (Table 4.2). Our risk stratification results align with epidemiology: some cancers have strong, long-term signals (e.g., liver cancer[269]) while others (e.g., pancreatic cancer[270,271]) remain hard to predict. Our results demonstrate that no single PheRS approach performs best by outcome or time threshold.

Using weights did not consistently improve PheRS risk stratification or discrimination in the AOU test cohort

Risk stratification capacity did not consistently or meaningfully change when the model development process considered PS_{BASIC}- or PS_{FULL}-weights (Figure 4.3). For example, for the random forest pancreatic cancer PheRS at t=1, the OR for the unweighted approach was 1.15 (0.64, 2.08), compared to 1.20 (0.66, 2.18) and 1.41 (0.80, 2.48) for the PS_{BASIC}- and PS_{FULL}-weighted approaches, respectively (Supplementary Figure 4.3). While the PS_{FULL}-weighted random forest at t=1 had the highest OR for pancreatic cancer, the PS_{BASIC}-weighted version was highest for esophageal cancer (Table SD2). At t=2, the PS_{FULL}-weighted and unweighted versions were highest for esophageal and liver cancer, respectively (Table SD3). The indeterminate impact of weighting applied to risk stratification and discrimination for all data domains.

PheRS contributes independently and significantly to risk stratification and discriminatory ability alongside other data domains in a model cascade

By building sequential and joint models via a "model cascade," we assessed the role of different domains of data to determine whether PheRS were additive to covariates (like age, sex, and race/ethnicity, which are deemed non-modifiable) and risk factors (which vary by outcome and may be preventable/modifiable) (Figure 4.2). We found that when PheRS alone can perform risk stratification, as for the random forest PheRS for liver cancer at t=1 (Figure 4.2), they generally improve the risk stratification ability after covariates and risk factors are considered. For example, the unweighted covariates and risk factors model for liver cancer at t=1 had an OR (95% CI) of 1.75 (1.16, 2.63). After adding the random forest PheRS, the OR (95% CI) increased to 7.02 (4.75, 10.38). PheRS can also contribute to risk stratification even when covariates and risk factors do not. For example, for pancreatic cancer at t=2, unweighted covariates and risk factors had an OR (95% CI) of 1.85 (0.97, 3.51). After adding the unweighted ridge PheRS, the OR became statistically significant (1.97 (1.09, 3.53)). A single acute symptom demonstrated risk stratification capacity (e.g., unweighted chest pain OR (95% CI) for esophageal cancer at t=0: 1.68 (1.13, 2.50)) but can quickly become null as the time-threshold increases (e.g., unweighted chest pain OR (95% CI) for esophageal cancer at t=1: 1.00 (0.97, 1.03)). These conclusions align with previous literature that found PheRS additive alongside covariates and risk factors for pancreatic cancer.[18]

Comparison with existing risk prediction models for these three cancers

For each cancer outcome, we identified previously published risk prediction models developed using (at least in part) a general US or UK adult population estimating 5- or 10-year risk that reported AUCs (Supplementary Figure 4.3). Several features of these studies differ such as the incorporation of genetic information,[18,257,272,273] restriction to older adults (40+ years old),[143,147,257,272,274–276] or creation of sex-stratified models.[143,274,275,277] Several used internal set-aside[147,272,277] or cross-validation[143,273,276] data to evaluate model performance. Generally, models were developed using cohort[143,147,272,278] or case-control[18,272–274,276] study designs and Cox[143,274,277] or logistic regression[18,272–276] models and used demographic, lifestyle, and personal and health history information. Details on comparison studies are presented in Supplementary Table 4.7 and Supplementary Table 4.8.

For esophageal cancer, we saw that only the model at t=0 (0.820 (0.781, 0.860); 0.605 (0.543, 0.667) at t=1, 0.540 (0.476, 0.604) at t=2, 0.538 (0.459, 0.616) at t=5) was able to achieve comparable AUC compared to Dong and colleagues (0.745 (0.721, 0.769)).[273] Dong and colleagues developed a logistic regression model that focused on esophageal adenocarcinoma (circumventing challenges with potentially conflicting risk factors with squamous cell carcinoma[279] as in a data-driven approach), had a much larger sample size (n=2,511 cases vs. n=389 in MGI), and included established risk factors (e.g., use of non-steroidal anti-inflammatory drugs (NSAIDs)). Moreover, they used self-reported data collected at or near the time of the cancer diagnosis, which can be more complete and accurate than comparable data available in EHR,[280,281] capturing important presenting symptoms, including heartburn and regurgitation symptoms. These factors highlight the importance of specificity in the outcome definition, the timing of predictors relative to the outcome, and using established risk factors (e.g., NSAID use) and symptoms (e.g., heartburn and regurgitation symptoms) in enhancing the predictive performance of models for esophageal cancer.

For liver cancer, we observed comparable AUCs in our models (0.909 (0.893, 0926) at t=0, 0.776 (0.747, 0.806) at t=1, 0.762 (0.731, 0.793) at t=2, 0.713 (0.675, 0.752)) with those found by Liu and colleagues (0.771 (0.702, 0.840)).[147] Liu and colleagues developed a Fine-Gray regression model for 5-year incident liver cancer using self-reported survey data supplemented by EHR in the UKB (n=113 cases in development dataset vs. n=337 in MGI). The selection of factors was driven by clinical knowledge and a literature review, and it focused on socioeconomic status, anthropomorphic measurements, lifestyle factors, and personal and family health history. Notably, there are strong, long-term predictors of liver cancer,[282] including history of viral hepatitis and liver disease, which were included in Liu and colleagues' model and selected in ours. These common factors can explain the common AUCs between the two models despite very different model development approaches.

For pancreatic cancer, we observed lower AUCs in our models (0.842 (0.814, 0.867) at t=0, 0.574 (0.529, 0.620) at t=1, 0.558 (0.511, 0.605) at t=2, 0.530 (0.474, 0.585) at t=5) compared to those by Salvatore and colleagues[18] (0.732 (0.710, 0.754) at t=5) and by Hippisley-Cox and Coupland[277] (0.857 (0.846, 0.867) in males, 0.865 (0.855, 0.875) in females). Salvatore and colleagues developed time-restricted models, using MGI data a pruning-and-thresholding and multivariable regression framework (like those in this paper), but had several limitations, including being assessed in time-unrestricted data in UKB and the development and validation cohorts being very different geographically and in terms of age. Hippisley-Cox and Coupland used a flexible, data-driven approach and a large dataset comprising over 6 million adult patients from the United Kingdom (n=7,117 cases vs. n=311 in MGI). Their Cox model used fractional polynomials for non-linear relationships, considered interactions between risk factors selected based on literature, and had richer covariate information (e.g., categorical smoking and alcohol variables instead of binary). Unlike data in this paper, the UK data comes from a country with universal health care, possibly increasing EHR completeness, and their hold-out validation cohort is likely to contain data collected similar to the development cohort. Notably, Salvatore and colleagues and Hippisley-Cox and Coupland considered individuals with a history of other cancers. However, we restricted to individuals without a history of cancer and sought to predict first primary pancreatic cancer diagnoses. These factors can explain why we found relatively lower AUCs at non-0 time thresholds than others reported in the literature.

Understanding clinical context is paramount in risk prediction model development

It is crucial to consider the clinical context of the outcome and the use of the risk prediction model. The outcomes considered here vary greatly in their clinical presentation, particularly at advanced stages, and their diagnostic approach. These differences can explain why different approaches are better suited for different outcomes. All three outcomes currently do not have screening mechanisms and are often diagnosed late when prognosis is poor. Thirty-nine percent of esophageal, 20% of liver, and 51% of pancreatic cancers are diagnosed after the cancer has metastasized when the 5-year relative survival is 5.3%, 3.3%, and 3.1%, respectively (SEER-22, 2014-2020, all races, both sexes[181]).

The current risk prediction models for these cancers tend to focus on high-risk populations, such as those with chronic hepatitis B virus infections[182–185] or chronic liver disease[186–188] for liver cancer and those with new-onset diabetes for pancreatic cancer.[189,190] Other models aim to identify individuals to screen for premalignant conditions, as in the case of Barrett's esophagus prior to the transition to esophageal cancer.[191,192] Importantly, models incorporating biomarkers and genetic factors to construct integrated and multi-factorial models generally exhibit better performance.[193–195]

These models can inform surveillance and monitoring strategies, such as abdominal ultrasonography and α-fetoprotein (AFP) tests for high-risk individuals for liver cancer,[196] or endoscopic ultrasonography or MRI for high-risk individuals for pancreatic cancer.[197] In the absence of screening mechanisms, incorporating biomarker and genetic factors alongside demographics, risk factors, and diagnostic history can aid in developing risk prediction models in the general population to identify high-risk individuals for targeted enhanced surveillance and prevention measures. Furthermore, each cancer has multiple histological types with different risk factors, clinical features, genetic susceptibility, and pathogenesis, such as squamous cell carcinoma and adenocarcinoma for esophageal cancer.[198] Designing risk prediction models by subtype theoretically could improve model performance by allowing models to consider heterogeneous tumor behavior separately for each histological type. However, the rarity of subtypes for these cancers makes such models challenging to develop and diminishes their utility. A focused, comprehensive approach to risk prediction model development and stratification can improve early detection, targeted surveillance, and, ultimately, patient outcomes for these challenging cancers.

4.4.1 Strengths and Limitations

This paper has several strengths. First, we used the relatively new phecode X mapping table, designed with ICD-10-CM in mind, and almost doubled the number of defined phecodes compared to its predecessor.[61] Second, we explored many commonly used risk prediction modeling approaches. Third, we developed and shared code for tuning hyperparameters in weighted settings for regularized regression and random forest models in R.

There are also several limitations of our work. First, despite the size of cohorts, sample sizes for some outcomes were small. This contributed to convergence issues for lasso and elastic net models, which we handled by considering multiple hyperparameter (i.e., λ) values, screening out highly correlated predictors, and fitting models with weights as a predictor. It also resulted in some models with little variation in the assessment cohort. Future work should consider outcomes with larger sample sizes. Second, we used joint strata proportions from MGI and AOU to estimate PS weights. Inverse probability weights could not be estimated because the aggregation of individual-level data across cohorts was restricted due to privacy concerns. For PS weights with many strata, like those in our PS_{FULL}-weights, proportions are typically only available with access to individual-level data. However, in practice, the number of factors used for poststratification is often limited because accurately estimating proportions within each stratum becomes more difficult as the complexity of stratification increases with additional factors. Third, there are many risk prediction methods that we did not consider, like neural networks, support vector machines, and SuperLearner, that can be used and have shown promise in many clinical settings.[283–287] The performance of these models can be compared with other approaches, including transfer, semi-supervised, and federated learning.[173,174,265,267] Future work should apply transfer learning to risk prediction models developed with and without weights. Fifth, Firth bias-corrected logistic regression was used to fit sequential models, which could have hampered performance in the presence of collinearity between terms. Alternative approaches like ridge regression that handle collinearity could be considered. Sixth, we ignored genetic data, a core feature of EHR-linked biobanks, and other potentially predictive data domains, including laboratory results and medications. Future work should include polygenic risk scores (PRS) and combine PRS and PheRS with approaches to enhance transferability, as in Zhao and colleagues' transfer learning PRS (TL-PRS) approach.[268] Seventh, the two-step PheRS approach identifies an initial set of candidate phecodes with the 50 smallest p-values. Future work, preferably with larger data, could employ a p-value threshold cutoff (e.g., a phenome-wide significance threshold corrected for multiple testing).

4.5 Conclusion

Using two EHR-linked biobanks, we have explored the role of poststratification sampling weights on the risk stratification performance of diagnosis code-derived PheRS for esophageal, liver, and pancreatic cancer. We constructed two sets of poststratification weights to make MGI look more like AOU. We modified an existing R package to accommodate hyperparameter tuning for regularized regression and random forest models in weighted settings. No PheRS approach consistently or meaningfully improved risk stratification or discriminatory ability; using weights in PheRS construction did not change this observation. Because PheRS contributes to risk stratification and discriminatory ability alongside demographic covariates, risk factors, and a presenting symptom, PheRS should be considered when developing risk prediction models. Researchers should instead carefully consider the population represented by their data and refrain from overgeneralizing the risk stratification capacity of models developed using EHR-linked biobank data. Future work should consider a wider array of outcomes, different cohorts, non-weighting-based machine learning approaches for transferring predictions, and integrating genetic and other linkable data available in EHR-linked biobanks.

4.6 Materials and methods

4.6.1 Cohorts

4.6.1.1 All of Us (AOU)

AOU is a US-based EHR-linked biobank organized by the National Institutes of Health. It began recruitment in 2018, attempting to enroll over 1,000,000 adults from 340 recruitment centers nationwide. As of January 31, 2024, there are over 760,000 participants, including over 539,000 and 420,000 who have biosamples and EHR data, respectively. Its EHR data are shared by participants and based on the ICD-9/ICD-10-CM codesets. Our analysis focuses on 226,764 people with non-missing sociodemographic and mappable EHR data. After mapping ICD codes to phecodes (see Section 4.6.2.1), 3,489 traits were defined.

4.6.1.2 Michigan Genomics Initiative (MGI)

MGI is an academic medical center (Michigan Medicine)-based EHR-linked biobank at the University of Michigan. It began recruitment in 2012, primarily recruiting adult patients through pre-/peri-operative appointments requiring anesthesia. Over time, additional Precision Health cohorts have been launched, recruiting adults through mental health, endocrinology (diabetes), and outpatient clinics). As of September 2023, there are ~100,000 enrolled, with ~10,000 enrolled yearly. The EHR data comprise patients' Michigan Medicine EHR based on the ICD-9/ICD-10-CM codesets. Our analysis focuses on 76,757 people with non-missing sociodemographic and mappable EHR data. After mapping ICD codes to phecodes, 3,347 traits were defined.

4.6.2 Data

4.6.2.1 Construction of the phenome

ICD-9-CM and ICD-10-CM codes were aggregated to broader yet clinically meaningful phenotypes called PheWAS codes, or phecodes,[61,288] using the phecode X[61,227] mapping tables. Observations of sex-specific phecodes that were discordant with the individual's EHR-recorded sex were removed. After curating the time-based phecode data (see Phenome time-restriction section below), phecode indicator matrices were constructed containing case status indicator variables across all defined phecodes where a single occurrence of a phecode was sufficient to determine case status.

4.6.2.2 Phenome time-restriction

Time-restricted phenomes were created for a given outcome and time threshold combination based on Salvatore and colleagues.[18] For a given outcome, say, pancreatic cancer (CA_101.8), cases were identified as individuals whose pancreatic cancer diagnosis was the first specific malignant neoplasm diagnosis in their EHR (Supplementary Table 4.1; could co-occur with other malignant neoplasm diagnoses). All cases were matched with 2 non-cases based on age at first diagnosis (nearest neighbor), sex (exact), and length of EHR follow-up (nearest neighbor). Eligible non-cases were individuals who never had a malignant neoplasm diagnosis. For each matched group, the days since birth corresponding to the initial diagnosis in the case were used as the index threshold. Each group's phenomes were restricted to observations that occurred t years (time threshold) before the index threshold. Phecode indicator matrices are reconstructed using the time-restricted data. We considered time thresholds of 0, 1, 2, and 5 years before the diagnosis of interest. The MGI sample was split 50-50 at each time threshold into training (for hyperparameter tuning and phecode selection) and testing (for phecode weight estimation and model fitting) sets.

4.6.2.3 Outcomes, exposures, covariates, risk factors, symptoms, and sampling weight variables

Outcomes: Risk prediction models are constructed for the following digestive organ cancers: esophageal (CA_101.1), liver and intrahepatic bile duct (CA_101.6), and pancreatic (CA_101.8). Gastrointestinal cancers account for 26% of incident cancer cases and 35% of cancer deaths globally.[289] Because screening mechanisms exist, colorectal cancer was not considered. Other digestive cancers (anal, small intestine, stomach, and gallbladder and extrahepatic bile duct) were not considered because they did not have a sufficient sample size (at least 300 cases in MGI occurring as their initial specific, malignant cancer diagnosis). A single occurrence of a phecode was considered adequate to identify a case. The index threshold was identified at a case's first occurrence of the phecode.

Exposures, covariates, risk factors, and symptoms: Exposures were considered the set of non-outcome phecodes with at least 20 cases in both cohorts (n = 2,728). We compared PheRS performance with models that included combinations of covariates, risk factors, and a presenting symptom. Covariates were age (at time threshold), sex, and whether an individual was non-Hispanic White; factors considered non-modifiable. Risk factors included history of overweight/obesity, smoking, and alcohol consumption (all binary), factors considered modifiable. We selected a single non-specific symptom that is commonly associated with each cancer: chest pain (SS_800), jaundice (SS_814), and abdominal pain (phecode GI_527) for esophageal, liver, and pancreatic cancer, respectively.

Sampling weight variables: Several additional variables were curated for poststratification weighting (following section). Age at last EHR diagnosis was categorized into bins: [0-18), [18,35), [35,65), [65-80), [80+). Race/ethnicity was categorized into non-Hispanic Asian, non-Hispanic Black, Hispanic, non-Hispanic White, and Other/Unknown. Smoking and alcohol consumption status were ever/never indicator variables. Body mass index (BMI) was categorized into underweight (<18.5), healthy weight ([18,25)), overweight ([25,30)), and obese (30+). History of depression and hypertension were indicators of the presence of phecodes MB_286.2 and CV_401, respectively. Finally, using raw ICD-9-CM/ICD-10-CM data, Quan-weighted Charlson Comorbidity Index (CCI) scores[290] were calculated using the comorbidity R package (version 1.0.7)[291] and categorized into: very low (0), low (1-2), moderate (3-4), and high (5+). Importantly, because MGI is enriched for cancer patients,[45] information regarding the presence of local or metastasized tumors is captured in the CCI. Additional information on variable definitions in AOU and MGI is described in Supplementary Table 4.2.

4.6.2.4 Poststratification weighting

To account for differences between MGI and AOU, AOU PS-weights in MGI were estimated. PS-weights rely on summary-level data on the target population (i.e., AOU). PS-weights were estimated using the following formula:

$$\omega = \frac{\Pr(\boldsymbol{X} = \boldsymbol{x} | S = 0)}{\Pr(\boldsymbol{X} = \boldsymbol{x} | S = 1)} \qquad \textit{Eq. (1)}$$

where $\boldsymbol{X}$ are the set of factors over which to make MGI representative of AOU and S is an indicator variable for membership in AOU ($S = 0$) or in MGI ($S = 1$). PS-weights corresponding to two sets of $\boldsymbol{X}$ were estimated. First, basic weights ($\mathrm{PS}_{\mathrm{BASIC}}$) were calculated using categorical age at last EHR diagnosis, sex, and race/ethnicity. Basic weights mimic the setting where limited summary statistics are available in the target population. Full weights ($\mathrm{PS}_{\mathrm{FULL}}$) included basic factors and added smoking and alcohol consumption status, BMI, history of depression and hypertension, and categorized Charlson Comorbidity Index. Strata with a proportion equal to 0 were set equal to the minimum non-zero stratum proportion in each cohort. PS-weights were winsorized at the 2.5th and 97.5th percentiles.

4.6.2.5 Modeling approaches

Two-step approaches

Our two-step approaches are analogous to pruning-and-thresholding in constructing polygenic risk scores. First, we performed unweighted and weighted phenome-wide association studies (PheWAS) in the discovery cohort training sample at each time threshold. That is, for each of k phecodes (up to 2,728 phecodes with at least 20 cases in both cohorts exclusive of Y), we fit the following standard logistic regression model:

$$logit\big(P(Y_i = 1|Phecode_{ik}, \boldsymbol{Z_i}; \omega_i)\big) = \beta_0 + \beta_k Phecode_{ik} + \boldsymbol{\beta_Z Z_i} \qquad \textit{Eq. (2)}$$

where Y_i is an indicator variable for cancer outcome case status, $\boldsymbol{Z_i}$ are the set of covariates age, sex, and length of follow-up included in the model for individual i. For each PheWAS, we selected the top 50 hits based on smallest p-value. To remove highly correlated phecodes, we ranked pairwise Pearson correlations (calculated using full discovery cohort phenome) in descending order. We removed phecodes with the larger p-value in each pair until correlation coefficients were no greater than 0.5.

After phecodes were selected and screened, we estimated phecode weights in the discovery cohort testing sample. For univariable phecode weights, models shown in Eq. 2 were fit for each screened phecode. For multivariable phecode weights, a multivariable logistic regression including all screened phecodes following Eq. 2 was fit. Using the log-odds estimates from these models as phecode weights; we calculated the two-step phenotype risk score (PheRS) in the assessment cohort as:

$$\mathrm{PheRS}_i = \sum_j \hat{\beta}_j D_{ij} \qquad \textit{Eq. (3)}$$

where $\hat{\beta}_j$ is the log-odds estimate corresponding to phecode j and D_{ij} is an indicator variable that equals 1 if individual i has ever been diagnosed with phecode j. Two-step PheRS are “unweighted”/”weighted” when both the PheWAS and the phecode weight model are unweighted/weighted.

One-step approaches

One-step approaches included the full-time-restricted phenome in regularized regression (ridge, lasso, and elastic net) and random forest models. First, hyperparameters for these models were tuned in the discovery cohort training sample, including λ in regularized regression models, α in elastic net, and (a) the number of randomly sampled features considered at each split (mtry) and (b) the minimum number of observations in a leaf node in random forest (min.node.size). We modified the weighted lasso approach described by Iparragirre and colleagues,[17] which selects λ based on performance in replicate weighted subsamples, to accommodate regularized regression (via the glmnet package) and random forest (via the ranger package). The following pseudocode describes the process for obtaining weighted hyperparameter values:

```
1. Split data into 10 folds and generate 10 sets of replicate weights
2. Initialize a set of hyperparameter values to search over
    a. For regularized regression models
        i. lambda: default lambda grid provided by glmnet
    b. For random forest models
        i. mtry: number of variables to possibly split at in each node
            1. $\lfloor \boldsymbol{x}\sqrt{p} \rfloor$, where $p$ is the dataset dimension and $\boldsymbol{x}$ are values
               0.25, 0.5, 1, 2, and 4
        ii. min.node.size: minimal terminal node size to split at
            1. Considered 1, 3, 5, 10, 20
3. For each set of replicate weights
    a. Split data 90/10 into training/testing sets from folds defined in (1)
    b. For each possible value of hyperparameter in grid search
        i. Fit model in 90% training set
            1. For regularized regression
                a. A regularized regression model in glmnet with replicate
                   weights as observation weights (weights)
            2. For random forest
                a. A random forest model in ranger with replicate weights
                   as case weights (case.weights)
    c. Estimate errors in each 10% testing set
    d. Average over the errors across the testing sets
4. Select hyperparameter value(s) with minimum average error in test data
```

Default wlasso, glmnet, and ranger settings were used unless otherwise specified. Code for loading the modified package is available at https://github.com/maxsal/wglmnet.

Next, models for the selected hyperparameters were fit in the discovery cohort testing sample using the time-restricted phenome. For regularized regression models, they were first fit using λ that minimized the loss function, then using the λ within one standard error of the minimum, and then again, in order, after screening the predictors until the mean absolute value of correlation was less than 0.25 (via findCorrelation from the caret R package) until the model converged and had non-zero degrees of freedom and a positive deviance ratio. In weighted models, the weights were included in the weight argument of the glmnet function. If none of these models were selected, models with screened predictors were fit with weights as a predictor variable (additional information in Supplementary Section 4.1). For weighted random forest, the weights were included in the case.weights argument of the ranger function. Standardized predicted values from these fitted models were obtained using time-restricted phenome data in the assessment cohort and served as the PheRS.

4.6.2.6 Fitting the model cascade

The model cascade consists of 8 models: 4 alone, 3 sequential, and 1 joint. The first 4 models were fit for covariates, risk factors, PheRS, and symptoms alone in the discovery test cohort using Firth bias-corrected logistic regression. The 3 sequential models added covariates and risk factors, PheRS, and finally, a presenting symptom in the discovery test cohort using Firth bias-corrected logistic regression. Finally, the “joint search” model considered all covariates, risk factors, phecodes, and presenting symptoms jointly following the process for the corresponding PheRS approach. The mean-standardized predicted values from these models in the assessment cohort were used in evaluation.

4.6.2.7 Evaluation

All prediction models were summarized and evaluated using the same set of criteria. Our primary indicator of performance was each model’s ability to perform risk stratification assessed by estimating an odds ratio for those in the top decile compared to those in the middle 20% (i.e., 40th to 60th percentile; distribution estimated in the controls). The area under the receiver-operator characteristics curve (AUC) (pROC R package), Brier score (DescTools R package), and Hosmer-Lemeshow Chi-square goodness-of-fit test (ResourceSelection R package) are reported to describe discriminatory ability, accuracy, and calibration, respectively. Additionally, we reported model fit using Nagelkerke’s pseudo-R^2 and the continuous OR for the risk scores.

4.6.2.8 Software

Analyses were conducted using R version 4.3.1. Codes corresponding to studies carried out in this paper are available at https://github.com/maxsal/weighted_prediction.

4.7 Tables

Table 4.1 Comparison of unweighted All of Us and Michigan Genomics Initiative cohorts and All of Us-weighted Michigan Genomics Initiative.

		Michigan Genomics Initiative		
			Weighted	
Variable	**All of Us**	**Unweighted**	**PS$_{BASIC}$**	**PS$_{FULL}$**
N	226,764	76,757	-	-
Age	54.2 (17.2)	56.7 (16.8)	54.7 (17.2)	54.5 (17.0)
Female	62.1 (140,785)	53.9 (41,369)	60.3	60.9
Race/ethnicity				
Asian, non-Hispanic	2.6 (5,887)	3.6 (2,752)	2.8	2.9
Black, non-Hispanic	18.8 (42,607)	6.1 (4,657)	20.4	17.1
Hispanic	18.8 (42,736)	2.6 (2,008)	11.7	10.4
White, non-Hispanic	54.9 (124,488)	83.5 (64,089)	59.7	64.5
Other/Unknown	4.9 (11,046)	4.2 (3,251)	5.3	5.1
BMI category				
Underweight	1.3 (2,878)	1.0 (803)	1.0	0.8
Healthy weight	25.6 (57,968)	24.5 (18,818)	24.1	26.2
Overweight	30.6 (69,414)	32.0 (24,597)	30.8	31.0
Obese	42.6 (96,504)	42.4 (32,539)	44.1	42.0
Charlson Comorbidity Index				
Very low [0]	46.3 (104,962)	30.1 (23,117)	31.7	44.5
Low [1-2]	29.1 (66,089)	26.6 (20,452)	27.6	30.2
Moderate [3-4]	12.5 (28,299)	10.0 (7,679)	10.3	12.2
High [5+]	12.1 (27,414)	33.2 (25,509)	30.4	13.1
Alcohol (ever)	88.3 (200,139)	69.4 (53,285)	68.1	89.1
Smoking (ever)	41.0 (92,992)	47.0 (36,052)	44.5	41.6
Depression (phecode)	25.5 (57,821)	32.2 (24,748)	33.4	26.7
Hypertension (phecode)	47.5 (107,716)	51.7 (39,719)	51.1	47.7
Length of EHR follow-up (years)	9.4 (8.3)	10.2 (7.6)	-	-
Cancer				
Esophageal cancer	0.1 (193)	0.5 (389)	-	-
Liver cancer	0.3 (599)	0.4 (337)	-	-
Pancreatic cancer	0.2 (385)	0.4 (311)	-	-

Notes:

- Charlson Comorbidity Index calculated with Quan weights via the comorbidity R package and ICD-9-CM/ICD-10-CM codes.
- Basic weights include age at last diagnosis (categorical), sex, and race/ethnicity.
- Full weights include age at last diagnosis (categorical), sex, race/ethnicity, smoking (ever), alcohol (ever), body mass index (categorical), depression (binary phecode), hypertension (binary phecode), and Charlson Comorbidity Index (categorical).
- Additional information on variable definitions can be found in Supplementary Table 4.1 and Supplementary Table 4.2.

Table 4.2 Unweighted PheRS top decile to middle 20% (40th-60th percentiles) odds ratio (95% confidence interval) by outcome, PheRS approach, and time threshold in All of Us.

		Time threshold			
Outcome	**PheRS approach**	**0**	**1**	**2**	**5**
Esophageal cancer [CA_101.1]	Univariable	6.87 (4.36, 10.84)	2.40 (1.17, 4.93)	2.26 (1.16, 4.42)	0.90 (0.34, 2.39)
	Multivariable	-	2.40 (1.30, 4.43)	1.45 (0.80, 2.62)	0.69 (0.26, 1.84)
	Lasso	-	-	NA	2.66 (1.02, 6.91)
	Ridge	5.79 (3.16, 10.61)	1.28 (0.60, 2.76)	1.13 (0.48, 2.64)	0.51 (0.15, 1.67)
	Elastic net	10.25 (6.41, 16.41)	-	1.27 (0.70, 2.30)	1.29 (0.50, 3.35)
	Random forest	10.14 (5.47, 18.81)	1.36 (0.62, 2.99)	0.81 (0.35, 1.88)	2.08 (0.70, 6.19)
Liver cancer [CA_101.6]	Univariable	22.69 (16.71, 30.80)	9.60 (7.11, 12.97)	11.01 (7.55, 16.06)	5.43 (3.61, 8.17)
	Multivariable	21.26 (15.52, 29.13)	8.28 (6.00, 11.44)	9.27 (6.39, 13.44)	3.18 (2.10, 4.80)
	Lasso	36.87 (26.25, 51.77)	12.24 (8.74, 17.14)	8.54 (6.14, 11.87)	3.68 (1.32, 10.22)
	Ridge	26.81 (17.65, 40.72)	12.08 (7.93, 18.39)	12.22 (7.81, 19.12)	6.61 (4.04, 10.81)
	Elastic net	37.25 (26.53, 52.30)	2.66 (1.95, 3.63)	1.37 (0.95, 1.99)	3.68 (1.32, 10.22)
	Random forest	63.80 (36.78, 110.68)	13.73 (8.97, 21.01)	16.42 (10.19, 26.46)	6.84 (4.22, 11.11)
Pancreatic cancer [CA_101.8]	Univariable	19.01 (12.97, 27.87)	1.75 (0.97, 3.14)	0.98 (0.53, 1.78)	1.05 (0.50, 2.21)
	Multivariable	16.23 (11.22, 23.49)	1.58 (0.88, 2.82)	1.22 (0.71, 2.11)	1.23 (0.64, 2.37)
	Lasso	18.22 (12.71, 26.11)	-	1.62 (0.94, 2.80)	1.19 (0.74, 1.94)
	Ridge	12.45 (7.79, 19.91)	1.88 (1.00, 3.53)	2.03 (1.11, 3.71)	1.20 (0.58, 2.48)
	Elastic net	18.70 (12.85, 27.21)	1.19 (0.66, 2.14)	1.77 (0.98, 3.19)	0.56 (0.25, 1.24)
	Random forest	16.88 (10.82, 26.35)	1.15 (0.64, 2.08)	1.36 (0.76, 2.44)	1.11 (0.53, 2.34)

Abbreviations: PheRS, phenotype risk score

Notes: A dash ('-') indicates the PheRS distribution among controls in All of Us was unable to distinguish between the middle 20% (40th-60th percentiles) and top decile. NA indicates that the model was unable to converge.

Table 4.3 Unweighted PheRS area under the receiver-operator characteristics curve (AUC, 95% confidence interval) by outcome, PheRS approach, and time threshold in All of Us.

		Time threshold			
Outcome	**PheRS approach**	**0**	**1**	**2**	**5**
Esophageal cancer [CA_101.1]	Univariable	0.767 (0.725, 0.809)	0.561 (0.498, 0.625)	0.540 (0.474, 0.607)	0.426 (0.349, 0.502)
	Multivariable	0.515 (0.475, 0.555)	0.528 (0.463, 0.592)	0.610 (0.545, 0.674)	0.572 (0.495, 0.648)
	Lasso	0.777 (0.737, 0.818)	0.572 (0.540, 0.605)	NA	0.569 (0.493, 0.646)
	Ridge	0.653 (0.603, 0.703)	0.505 (0.443, 0.567)	0.507 (0.443, 0.571)	0.529 (0.453, 0.604)
	Elastic net	0.775 (0.733, 0.818)	0.594 (0.552, 0.636)	0.561 (0.498, 0.625)	0.477 (0.400, 0.555)
	Random forest	0.781 (0.740, 0.823)	0.528 (0.467, 0.590)	0.521 (0.458, 0.585)	0.520 (0.440, 0.600)
Liver cancer [CA_101.6]	Univariable	0.845 (0.825, 0.865)	0.734 (0.704, 0.764)	0.764 (0.734, 0.795)	0.700 (0.661, 0.738)
	Multivariable	0.818 (0.795, 0.842)	0.687 (0.655, 0.720)	0.732 (0.700, 0.765)	0.642 (0.603, 0.680)
	Lasso	0.860 (0.839, 0.882)	0.771 (0.742, 0.800)	0.751 (0.719, 0.782)	0.712 (0.674, 0.749)
	Ridge	0.880 (0.862, 0.899)	0.770 (0.740, 0.799)	0.765 (0.735, 0.796)	0.689 (0.649, 0.729)
	Elastic net	0.862 (0.841, 0.884)	0.562 (0.538, 0.586)	0.540 (0.516, 0.564)	0.711 (0.674, 0.749)
	Random forest	0.909 (0.892, 0.925)	0.771 (0.741, 0.801)	0.753 (0.721, 0.785)	0.692 (0.652, 0.732)
Pancreatic cancer [CA_101.8]	Univariable	0.833 (0.805, 0.861)	0.530 (0.482, 0.577)	0.537 (0.490, 0.584)	0.477 (0.420, 0.535)
	Multivariable	0.760 (0.726, 0.794)	0.554 (0.507, 0.600)	0.504 (0.457, 0.551)	0.492 (0.435, 0.548)
	Lasso	0.810 (0.780, 0.840)	0.531 (0.492, 0.570)	0.546 (0.498, 0.593)	0.511 (0.467, 0.556)
	Ridge	0.753 (0.720, 0.785)	0.537 (0.490, 0.584)	0.536 (0.488, 0.585)	0.519 (0.462, 0.576)
	Elastic net	0.808 (0.777, 0.838)	0.560 (0.514, 0.606)	0.560 (0.513, 0.608)	0.511 (0.460, 0.561)
	Random forest	0.854 (0.827, 0.880)	0.532 (0.486, 0.578)	0.535 (0.488, 0.583)	0.517 (0.460, 0.574)

Abbreviations: PheRS, phenotype risk score

Notes: NA indicates that the model was unable to converge.

Table 4.4 Comparison of top 10 features by permutation-based variable importance for unweighted random forest liver cancer PheRS by time threshold.

		Variable importance rank by time threshold			
Phenotype [Phecode]	**Phecode group**	**0**	**1**	**2**	**5**
Other disorders of liver [GI_546]	Gastrointestinal	**1**	**8**	**6**	24
Other diseases of biliary tract [GI_552]	Gastrointestinal	**2**	**9**	14	45
Fibrosis and cirrhosis of liver [GI_542.1]	Gastrointestinal	**3**	**4**	**3**	**9**
Chronic liver disease [GI_542]	Gastrointestinal	**4**	**2**	**1**	**1**
Cirrhosis of liver [GI_542.11]	Gastrointestinal	**5**	**3**	**4**	**3**
Chronic nonalcoholic liver disease [GI_542.7]	Gastrointestinal	**6**	**1**	**2**	11
Abnormal findings on examination of blood [SS_829]	Symptoms	**7**	68	75	1293
Hepatomegaly [GI_546.3]	Gastrointestinal	**8**	376	265	192
Diseases of the pancreas [GI_554]	Gastrointestinal	**9**	56	56	135
Obstruction of bile duct [GI_552.2]	Gastrointestinal	**10**	1599	291	213
Hepatitis [GI_540]	Gastrointestinal	12	**5**	**5**	**2**
Chronic hepatitis [GI_540.1]	Gastrointestinal	13	11	**7**	**7**
Viral hepatitis [GI_540.3]	Gastrointestinal	14	**7**	**8**	13
Hepatovirus [ID_054]	Infections	18	**6**	**10**	**4**
Chronic hepatitis C [ID_054.31]	Infections	23	17	**9**	15
Back pain [MS_718]	Musculoskeletal	28	38	113	**6**
Hyperlipidemia [EM_239]	Endocrine/Metab	35	62	39	**10**
Diseases of spleen [BI_174]	Blood/Immune	41	18	17	**8**
Nonspecific abnormal results of function study of liver [GI_545]	Gastrointestinal	56	**10**	15	**5**

Notes: Ranks are **bolded** indicating they were identified as a top 10 feature for the given time threshold. Variable importance was estimated using the permutation method in the ranger package and extracted using the vip package. Phenotypes, phecodes, and phecode groups are defined using Phecode X mapping tables from https://phewascatalog.org.

Table 4.5 Comparison of top 10 features by permutation-based variable importance for unweighted random forest PheRS at t=1 by outcome.

		Variable importance rank by outcome		
Phenotype [Phecode]	**Phecode group**	**Esophageal**	**Liver**	**Pancreatic**
Hyperlipidemia [EM_239]	Endocrine/Metab	**1**	62	71
Disorders of prostate [GU_602]	Genitourinary	**2**	1615	184
Diseases of esophagus [GI_510]	Gastrointestinal	**3**	221	161
Esophageal obstruction (Stricture and stenosis of esophagus) [GI_514.1]	Gastrointestinal	**4**	808	673
Atrial fibrillation and flutter [CV_416.2]	Cardiovascular	**5**	270	105
Abnormal results of cardiovascular function studies [CV_418]	Cardiovascular	**6**	64	28
Barrett's esophagus [GI_510.8]	Gastrointestinal	**7**	804	668
Essential hypertension [CV_401.1]	Cardiovascular	**8**	1697	93
Ischemic heart disease [CV_404]	Cardiovascular	**9**	177	70
Diverticula of colon [GI_523.2]	Gastrointestinal	**10**	1666	84
Pain in joint [MS_713.3]	Musculoskeletal	35	70	**5**
Asthma [RE_475]	Respiratory	44	357	**8**
Cardiac arrhythmia and conduction disorders [CV_416]	Cardiovascular	51	26	**2**
Hepatitis [GI_540]	Gastrointestinal	81	**5**	1299
Obesity [EM_236.1]	Endocrine/Metab	98	60	**7**
Sinusitis [RE_462]	Respiratory	124	1685	**6**
Chronic nonalcoholic liver disease [GI_542.7]	Gastrointestinal	184	**1**	56
Chronic liver disease [GI_542]	Gastrointestinal	246	**2**	169
Other disorders of liver [GI_546]	Gastrointestinal	253	**8**	314
Other diseases of biliary tract [GI_552]	Gastrointestinal	256	**9**	1406
Nonspecific abnormal results of function study of liver [GI_545]	Gastrointestinal	348	**10**	126
Fibrosis and cirrhosis of liver [GI_542.1]	Gastrointestinal	687	**4**	699
Cirrhosis of liver [GI_542.11]	Gastrointestinal	688	**3**	700
Diseases of the pancreas [GI_554]	Gastrointestinal	698	56	**1**
Pancreatitis [GI_554.1]	Gastrointestinal	699	171	**3**
Acute pancreatitis [GI_554.11]	Gastrointestinal	700	98	**9**
Overweight and obesity [EM_236]	Endocrine/Metab	1191	66	**4**
Hepatovirus [ID_054]	Infections	1206	**6**	1292
Viral hepatitis [GI_540.3]	Gastrointestinal	1229	**7**	1287
Other disorders of bone [MS_727]	Musculoskeletal	1315	76	**10**

Notes: Ranks are **bolded** indicating they were identified as a top 10 feature for the given outcome. Variable importance was estimated using the permutation method in the ranger package and extracted using the vip package. Phenotypes, phecodes, and phecode groups are defined using Phecode X mapping tables from https://phewascatalog.org.

4.8 Figures

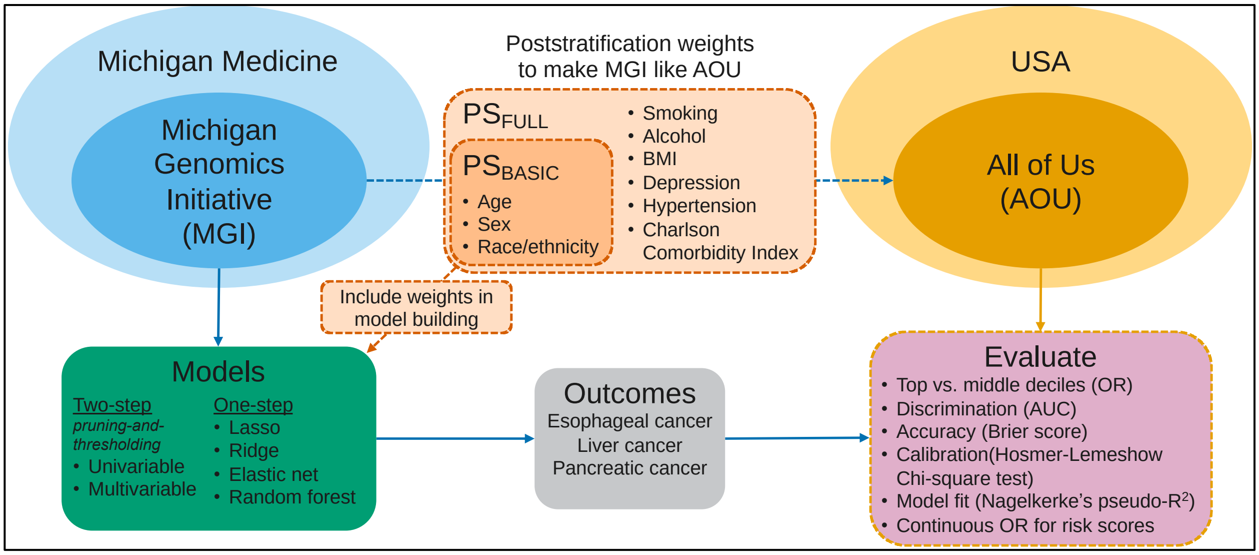


Figure 4.1 Schematic representation of analytic framework. Michigan Genomics Initiative (MGI) and All of Us (AOU) are samples of presumptive source populations – Michigan Medicine and the US adult population, respectively. In theory, Michigan Medicine is a subset of the US adult population, but we assume minimal overlap. First, poststratification weights are calculated to make the MGI sample representative of AOU. These weights are then used across several one- and two-step modeling approaches. Outcomes for these risk prediction models are esophageal, liver, and pancreatic cancers. Finally, the models developed using MGI data are applied to the AOU sample and evaluated for risk stratification capacity and other performance measures. Abbreviations: AOU, All of Us; AUC, area under the receiver-operator characteristics curve; MGI, Michigan Genomics Initiative; OR, odds ratio; PS, poststratification.

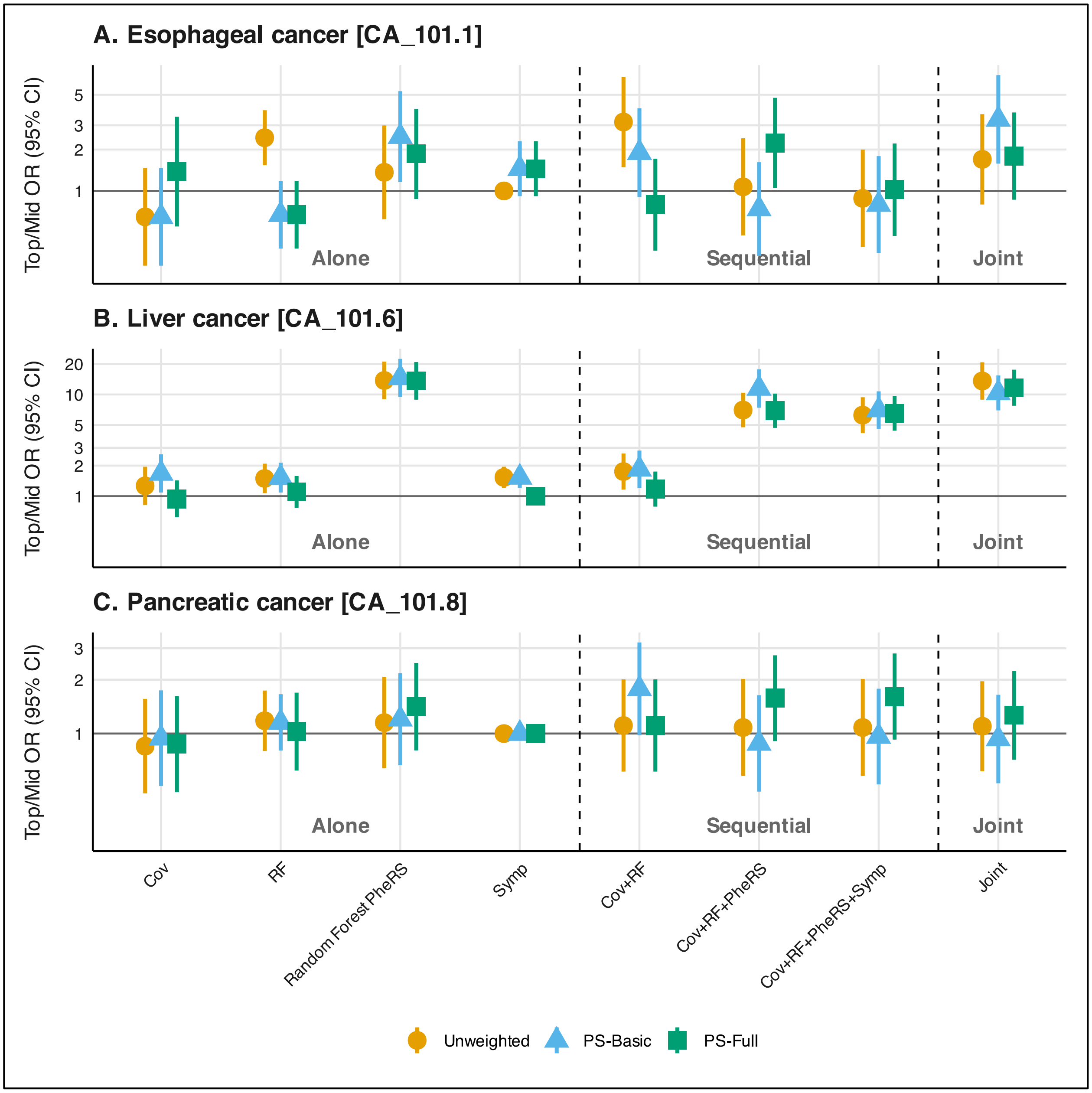


Figure 4.2 Comparison of model cascade for esophageal (panel A), liver (panel B), and pancreatic (panel C) cancer using the random forest PheRS approach by weighting approach at t = 1. The top decile compared to the middle 20% (40th-60th percentiles) odds ratio (95% confidence interval) is shown. Abbreviations: Cov, covariates; PheRS, phenotype risk score; RF, risk factors; Symp, presenting symptom

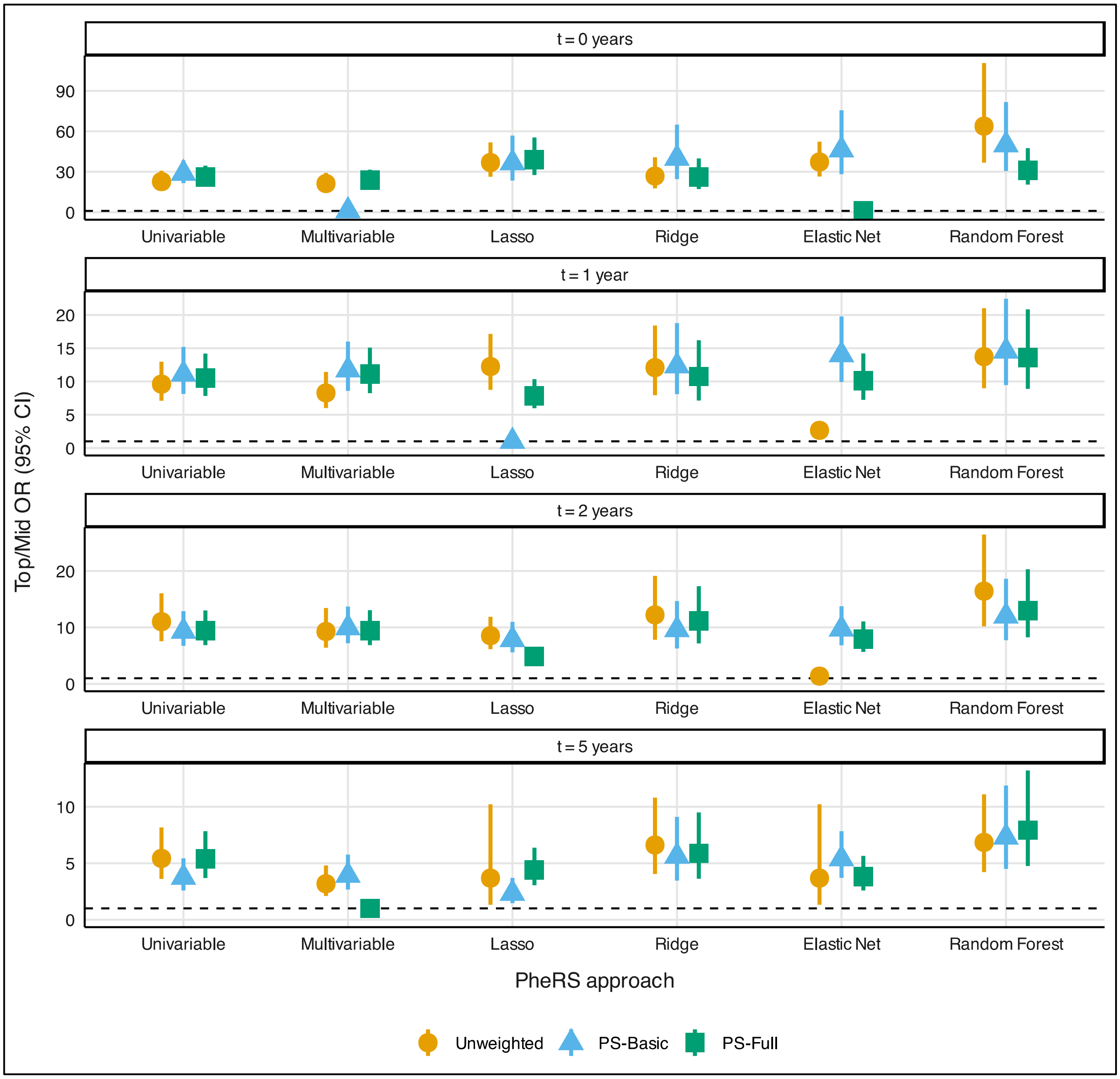


Figure 4.3 Top decile-to-middle 20% odds ratio (95% confidence interval) for by phenotype risk score (PheRS) and weighting approach by time threshold for liver cancer [CA_101.6]. Corresponding plots for esophageal [CA_101.1] and pancreatic [CA_101.8] cancers are shown in Supplementary Figure 4.1 and Supplementary Figure 4.2, respectively.

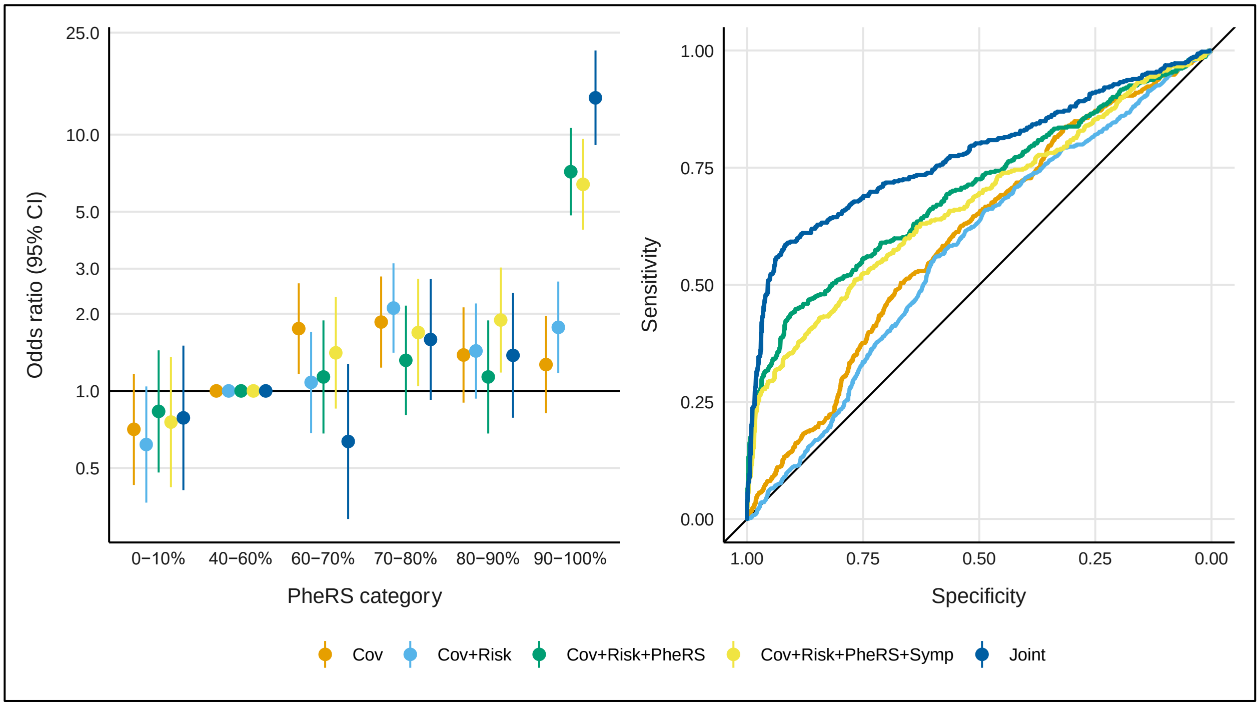


Figure 4.4 Comparison of unweighted liver cancer risk score stratified odds ratios (95% CI) (left panel) and area under the receiver-operator characteristics curves (AUC) (right panel) for the random forest PheRS by different predictor sets at the t=1 time threshold. The 40th-60th risk score percentile is used as the reference group in the left panel. Covariates (Cov) included age, sex, and a non-Hispanic White indicator. Risk factors (Risk) included alcohol consumption, obesity, and smoking. The presenting symptom (Symp) was abdominal pain. The first four models were constructed sequentially in a logistic regression model. The “Joint” model was built considering all predictors during PheRS construction (see Materials and Methods for details).

4.9 Supplementary materials

Supplementary Table 4.1 Phecode X neoplasm phecodes and whether they are considered malignant or specific.

Phecode	Description	Malignant	Specific
CA_100	Malignant neoplasm of the head and neck	1	1
CA_100.1	Malignant neoplasm of the oral cavity	1	1
CA_100.12	Malignant neoplasm of the tongue	1	1
CA_100.13	Malignant neoplasm of the gums	1	1
CA_100.14	Malignant neoplasm of the floor of mouth	1	1
CA_100.15	Malignant neoplasm of the palate	1	1
CA_100.2	Malignant neoplasm of the oropharynx	1	1
CA_100.3	Malignant neoplasm of the nasopharynx	1	1
CA_100.4	Malignant neoplasm of the hypopharynx	1	1
CA_100.5	Malignant neoplasm of nasal cavities, middle ear, and accessory sinuses	1	1
CA_100.6	Malignant neoplasm of the larynx	1	1
CA_100.7	Malignant neoplasm of the pharynx	1	1
CA_100.8	Malignant neoplasm of the lip	1	1
CA_100.9	Malignant neoplasm of the salivary glands	1	1
CA_101	Malignant neoplasm of the digestive organs	1	1
CA_101.1	Malignant neoplasm of the esophagus	1	1
CA_101.2	Malignant neoplasm of stomach	1	1
CA_101.21	Malignant neoplasm of cardia	1	1
CA_101.3	Malignant neoplasm of the small intestine	1	1
CA_101.4	Malignant neoplasm of the lower GI tract	1	1
CA_101.41	Colorectal cancer	1	1
CA_101.411	Malignant neoplasm of colon	1	1
CA_101.412	Malignant neoplasm of appendix	1	1
CA_101.42	Malignant neoplasm of anus	1	1
CA_101.6	Malignant neoplasm of the liver and intrahepatic bile ducts	1	1
CA_101.61	Malignant neoplasm of the liver	1	1
CA_101.62	Malignant neoplasm of the intrahepatic bile ducts	1	1
CA_101.7	Malignant neoplasm of the gallbladder and extrahepatic bile ducts	1	1
CA_101.71	Malignant neoplasm of the gallbladder	1	1
CA_101.8	Malignant neoplasm of the pancreas	1	1
CA_102	Malignant neoplasm of the thoracic and respiratory organs	1	1
CA_102.1	Malignant neoplasm of the of bronchus and lung	1	1
CA_102.3	Malignant neoplasm of the trachea	1	1
CA_102.5	Malignant neoplasm of the heart, mediastinum, thymus, and pleura	1	1
CA_102.51	Malignant neoplasm of the heart	1	1
CA_102.52	Malignant neoplasm of the mediastinum	1	1
CA_102.53	Malignant neoplasm of the of pleura	1	1
CA_102.54	Malignant neoplasm of the thymus	1	1
CA_103	Malignant neoplasm of the skin	1	0
CA_103.1	Melanomas of skin	1	1
CA_103.2	Keratinocyte carcinoma	1	0
CA_103.21	Basal cell carcinoma	1	0

Phecode	Description	Malignant	Specific
CA_103.22	Squamous cell carcinoma of the skin	1	0
CA_103.3	Carcinoma in situ of skin	0	0
CA_104	Malignant sarcoma-related cancers	1	1
CA_104.1	Malignant neoplasm of the bone and/or cartilage	1	1
CA_104.2	Malignant neoplasm of retroperitoneum and peritoneum	1	1
CA_104.3	Malignant neoplasm of connective and soft tissue	1	1
CA_104.4	Malignant neoplasm of peripheral nerves*	1	1
CA_104.5	Gastrointestinal stromal tumor*	1	1
CA_104.6	Kaposi's sarcoma	1	1
CA_105	Malignant neoplasm of the breast	1	1
CA_105.1	Malignant neoplasm of the breast, female	1	1
CA_105.2	Malignant neoplasm of the breast, male	1	1
CA_106	Gynecological malignant neoplasms	1	1
CA_106.1	Malignant neoplasm of external female genital organs and cervix	1	1
CA_106.11	Malignant neoplasm of the vulva	1	1
CA_106.12	Malignant neoplasm of the vagina	1	1
CA_106.13	Malignant neoplasm of the cervix	1	1
CA_106.2	Malignant neoplasm of the uterus	1	1
CA_106.21	Malignant neoplasm of endometrium	1	1
CA_106.3	Malignant neoplasm of the ovary	1	1
CA_106.4	Malignant neoplasm of the fallopian tube and uterine adnexa	1	1
CA_106.6	Malignant neoplasm of the placenta	1	1
CA_107	Malignant neoplasm of male genitalia	1	1
CA_107.1	Malignant neoplasm of the penis	1	1
CA_107.2	Malignant neoplasm of the prostate	1	1
CA_107.3	Malignant neoplasm of the testis	1	1
CA_107.4	Malignant neoplasm of epididymis	1	1
CA_107.5	Malignant neoplasm of spermatic cord	1	1
CA_107.6	Malignant neoplasm of the scrotum	1	1
CA_108	Malignant neoplasm of the urinary tract	1	1
CA_108.4	Malignant neoplasm of the kidney	1	1
CA_108.41	Malignant neoplasm of kidney, except pelvis	1	1
CA_108.42	Malignant neoplasm of renal pelvis	1	1
CA_108.5	Malignant neoplasm of the bladder	1	1
CA_108.6	Malignant neoplasm of urethra	1	1
CA_108.7	Malignant neoplasm of ureter	1	1
CA_109	Malignant neoplasm of the eye, brain and other parts of central nervous system	1	1
CA_109.1	Malignant neoplasm of eye	1	1
CA_109.11	Malignant neoplasm of orbit	1	1
CA_109.12	Malignant neoplasm of lacrimal gland and duct	1	1
CA_109.13	Malignant neoplasm of conjunctiva	1	1
CA_109.14	Malignant neoplasm of cornea	1	1
CA_109.15	Malignant neoplasm of retina	1	1
CA_109.16	Malignant neoplasm of choroid	1	1
CA_109.2	Malignant neoplasm of meninges	1	1
CA_109.3	Malignant neoplasm of brain	1	1
CA_109.4	Malignant neoplasm of spinal cord	1	1
CA_109.5	Malignant neoplasm of cranial nerve	1	1
CA_110	Malignant neoplasm of the endocrine glands	1	1
CA_110.1	Malignant neoplasm of the thyroid	1	1
CA_110.3	Malignant neoplasm of the parathyroid gland	1	1
CA_110.4	Malignant neoplasm of the pituitary gland and craniopharyngeal duct	1	1

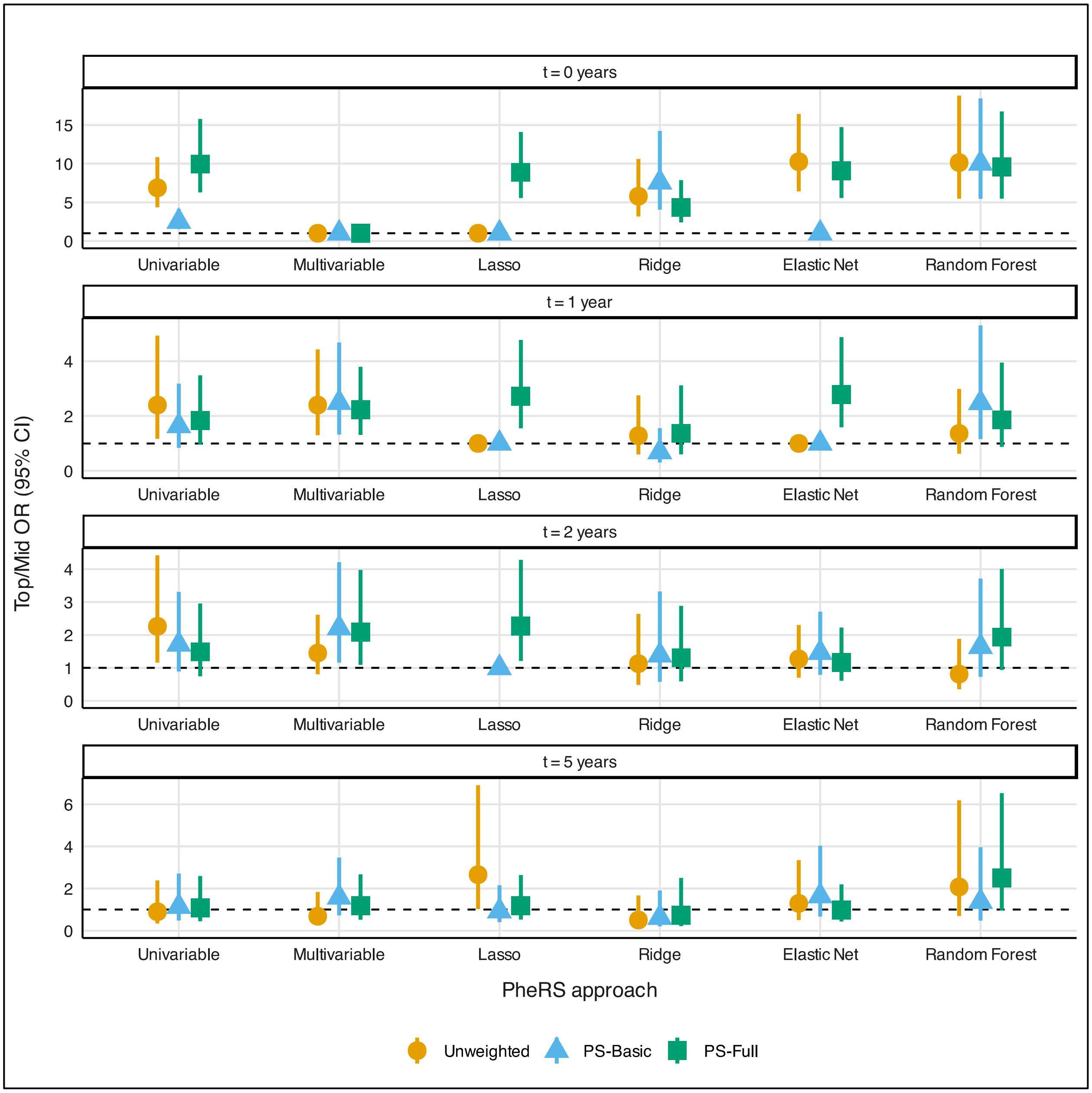
Phecode	Description	Malignant	Specific
CA_110.5	Malignant neoplasm of the pineal gland	1	1
CA_112	Malignant neoplasm of other and ill-defined sites	1	1
CA_112.1	Mesothelioma*	1	1
CA_114	Neuroendocrine tumors	0	0
CA_114.1	Malignant neuroendocrine tumors	1	1
CA_114.11	Exocrine pancreatic cancer	1	1
CA_114.12	Merkel cell carcinoma	1	1
CA_114.2	Benign neuroendocrine tumors	0	0
CA_114.4	Carcinoid tumors	1	1
CA_114.41	Intestinal carcinoid	1	1
CA_114.42	Carcinoid tumor of the bronchus and lung	1	1
CA_114.43	Carcinoid tumor of the thymus	1	1
CA_114.44	Carcinoid tumor of the stomach	1	1
CA_114.45	Carcinoid tumor of the kidney	1	1
CA_114.5	Paraganglioma	1	1
CA_114.6	Pheochromocytoma	1	1
CA_116	Secondary malignant neoplasm	1	0
CA_120	Hemo onc - by cell of origin	1	1
CA_120.1	Myeloid	1	1
CA_120.11	Plasma cell	1	1
CA_120.12	Monocyte	1	1
CA_120.13	Erythroid	1	1
CA_120.14	Megakaryoblast	1	1
CA_120.15	Mast cell	1	1
CA_120.2	Lymphoid	1	1
CA_120.21	Mature B-cell	1	1
CA_120.22	Mature T-Cell	1	1
CA_120.3	Histocytes	1	1
CA_121	Leukemia	1	1
CA_121.1	Acute leukemia	1	1
CA_121.11	Acute lymphoid leukemia	1	1
CA_121.12	Acute myeloid leukemia	1	1
CA_121.2	Chronic leukemia	1	1
CA_121.21	Chronic lymphoid leukemia	1	1
CA_121.22	Chronic myeloid leukemia	1	1
CA_121.23	Chronic myelomonocytic (monocytic) leukemia	1	1
CA_122	Lymphoma	1	1
CA_122.1	Hodgkin lymphoma	1	1
CA_122.11	Nodular sclerosis Hodgkin lymphoma	1	1
CA_122.2	Non-Hodgkin lymphoma	1	1
CA_122.21	Follicular lymphoma	1	1
CA_122.22	Diffuse large B-cell lymphoma*	1	1
CA_122.23	Burkitt lymphoma	1	1
CA_122.24	T-cell lymphoma	1	1
CA_122.25	Anaplastic large cell lymphoma	1	1
CA_122.26	Extranodal NK/T-cell lymphoma, nasal type*	1	1
CA_123	Multiple myeloma and malignant plasma cell neoplasms	1	1
CA_123.1	Multiple myeloma	1	1
CA_124	Myeloproliferative disorder	1	1
CA_124.3	Polycythemia vera	1	1
CA_124.5	Essential thrombocythemia	1	1
CA_124.6	Myelodysplastic syndrome	1	1
CA_124.7	Chronic myeloproliferative disease*	1	1
CA_124.8	Myelofibrosis	1	1

Phecode	Description	Malignant	Specific
CA_125	Other malignant neoplasms of lymphoid, hematopoietic and related tissue	1	1
CA_125.1	Cutaneous mastocytosis*	0	1
CA_128	Estrogen receptor status	1	1
CA_128.1	Estrogen receptor positive status [ER+]	1	1
CA_128.2	Estrogen receptor negative status [ER-]	1	1
CA_130	Cancer (solid tumor, excluding BCC)	1	0
CA_132	Sequelae of cancer	1	0
CA_135	Benign neoplasm of the head and neck	0	1
CA_135.1	Benign neoplasm of the oral cavity	0	1
CA_135.11	Benign neoplasm of the lip	0	1
CA_135.12	Benign neoplasm of the tongue	0	1
CA_135.14	Benign neoplasm of the floor of mouth	0	1
CA_135.16	Benign neoplasm of the salivary glands	0	1
CA_135.2	Benign neoplasm of the oropharynx	0	1
CA_135.3	Benign neoplasm of the nasopharynx	0	1
CA_135.4	Benign neoplasm of the hypopharynx	0	1
CA_135.5	Benign neoplasm of the paranasal sinus and nasal cavity	0	1
CA_135.6	Benign neoplasm of vocal cord or larynx	0	1
CA_136	Benign neoplasm of the digestive organs	0	1
CA_136.1	Benign neoplasm of the esophagus	0	1
CA_136.2	Benign neoplasm of stomach	0	1
CA_136.3	Benign neoplasm of the small intestine	0	1
CA_136.4	Benign neoplasm of colon, rectum, anus and anal canal	0	1
CA_136.41	Benign neoplasm of the colon	0	1
CA_136.42	Benign neoplasm of rectum and anus	0	1
CA_136.6	Benign neoplasm of the liver and intrahepatic bile ducts	0	1
CA_136.61	Benign neoplasm of the liver*	0	1
CA_136.8	Benign neoplasm of the pancreas	0	1
CA_137	Benign neoplasm of the thoracic and respiratory organs	0	1
CA_137.1	Benign neoplasm of the of bronchus and lung	0	1
CA_137.3	Benign neoplasm of the trachea	0	1
CA_137.5	Benign neoplasm of the heart, mediastinum, thymus, and pleura	0	1
CA_137.51	Benign neoplasm of the heart	0	1
CA_137.52	Benign neoplasm of the mediastinum	0	1
CA_137.53	Benign neoplasm of the of pleura	0	1
CA_137.54	Benign neoplasm of the thymus	0	1
CA_138	Benign neoplasm of the skin	0	1
CA_138.1	Nevus, non-neoplastic	0	1
CA_138.2	Melanocytic nevi*	0	1
CA_139	Benign sarcoma-related cancers	0	1
CA_139.1	Benign neoplasm of the bone and/or cartilage	0	1
CA_139.2	Benign neoplasm of retroperitoneum and peritoneum	0	1
CA_139.3	Benign neoplasm of other connective and soft tissue	0	1
CA_139.4	Benign neoplasm of peripheral nerves*	0	1
CA_139.5	Lipoma	0	1
CA_139.51	Lipomatosis*	0	1
CA_139.52	Lipoma of intrathoracic organs	0	1
CA_139.53	Lipoma of skin subcutaneous tissue	0	1
CA_139.54	Testicular lipoma	0	1
CA_139.6	Hemangioma and lymphangioma	0	1
CA_139.61	Hemangioma	0	1
CA_139.62	Lymphangioma	0	1
CA_140	Benign neoplasm of the breast	0	1

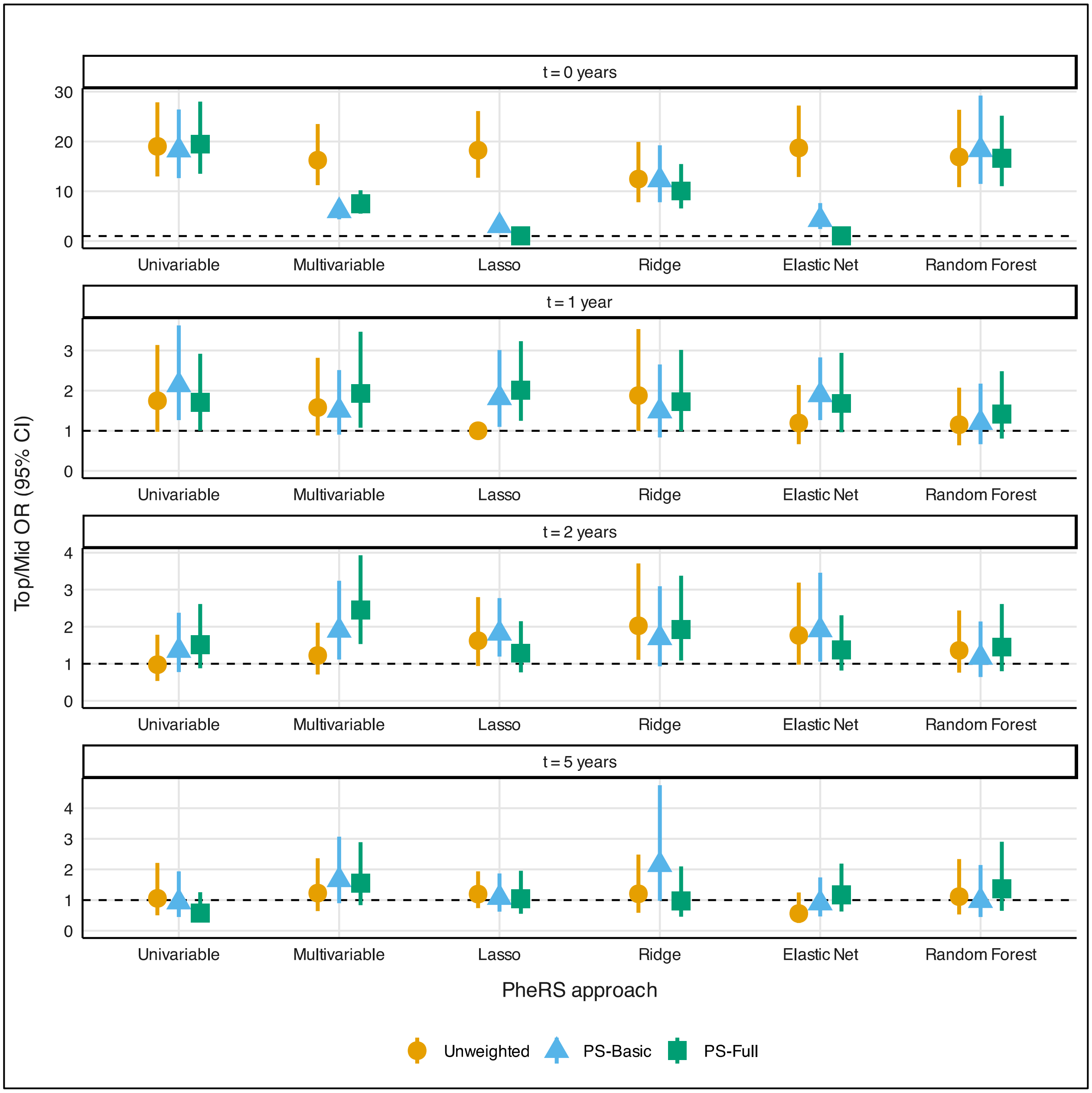
Phecode	Description	Malignant	Specific
CA_142	Lump or mass in breast or nonspecific abnormal breast exam	0	0
CA_142.1	Lump or mass in breast	0	0
CA_142.2	Abnormal mammogram	0	0
CA_142.21	Mammographic microcalcification	0	0
CA_144	Gynecological benign neoplasms	0	1
CA_144.1	Benign neoplasms of external female genital organs and cervix	0	1
CA_144.11	Benign neoplasms of the vulva	0	1
CA_144.12	Benign neoplasms of the vagina	0	1
CA_144.13	Benign neoplasms of the cervix	0	1
CA_144.2	Benign neoplasms of the uterus	0	1
CA_144.21	Leiomyoma of uterus	0	1
CA_144.3	Benign neoplasms of the ovary	0	1
CA_144.4	Benign neoplasm of the fallopian tube and uterine adnexa	0	1
CA_146	Benign neoplasm of male genital organs	0	1
CA_146.1	Benign neoplasm of the penis	0	1
CA_146.2	Benign neoplasm of the prostate	0	1
CA_146.3	Benign neoplasm of the testis	0	1
CA_146.31	Benign neoplasm of epididymis and spermatic cord	0	1
CA_146.32	Benign neoplasm of the scrotum	0	1
CA_147	Benign neoplasm of kidney and urinary organs	0	1
CA_147.1	Benign neoplasm of the kidney	0	1
CA_147.2	Benign neoplasm of ureter	0	1
CA_147.3	Benign neoplasm of bladder	0	1
CA_147.4	Benign neoplasm of urethra	0	1
CA_148	Benign neoplasm of the eye, brain and other parts of central nervous system	0	1
CA_148.1	Benign neoplasm of eye	0	1
CA_148.11	Benign neoplasm of orbit	0	1
CA_148.12	Benign neoplasm of lacrimal gland and duct	0	1
CA_148.13	Benign neoplasm of conjunctiva	0	1
CA_148.14	Benign neoplasm of cornea	0	1
CA_148.15	Benign neoplasm of retina	0	1
CA_148.16	Benign neoplasm of choroid	0	1
CA_148.2	Benign neoplasm of meninges (Meningioma)	0	1
CA_148.3	Benign neoplasm of brain	0	1
CA_148.4	Benign neoplasm of spinal cord	0	1
CA_148.5	Benign neoplasm of cranial nerve	0	1
CA_149	Benign neoplasm of the endocrine glands	0	1
CA_149.1	Benign neoplasm of the thyroid	0	1
CA_149.3	Benign neoplasm of the parathyroid gland	0	1
CA_149.4	Benign neoplasm of the pituitary gland and craniopharyngeal duct	0	1
CA_149.5	Benign neoplasm of pineal gland	0	1
CA_152	Benign neoplasm of lymph nodes	0	1

Supplementary Table 4.2 Definition of variables by cohort used throughout paper

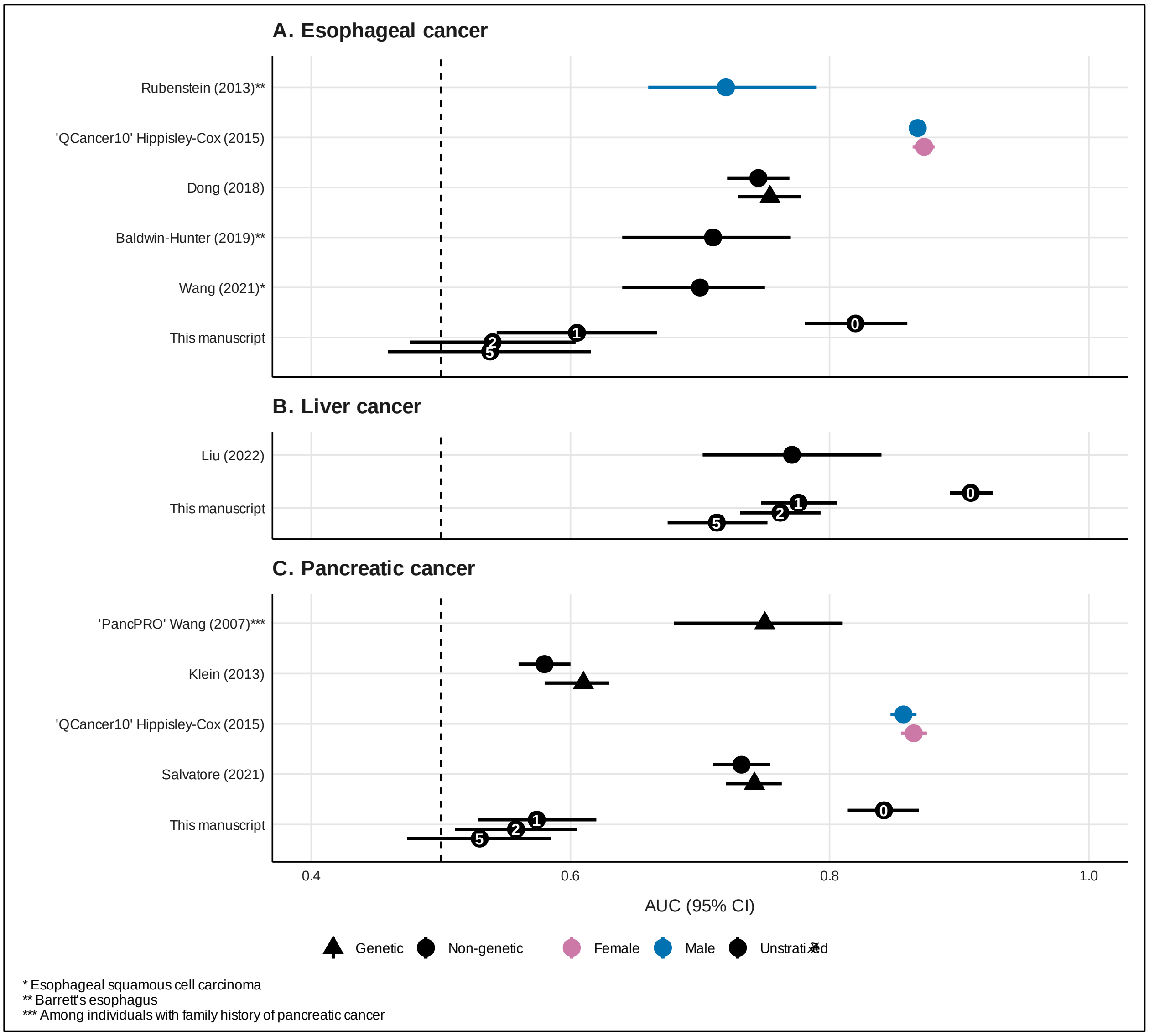
Variable	AOU	MGI	Categories
Age	Age at last diagnosis		[0-18), [18,35), [35,65), [65-80), [80+)
Sex	Self-reported sex at birth (field name: sex_at_birth_concept_id)	Self-report EHR	non-Hispanic Asian, non-Hispanic Black, Hispanic, non-Hispanic White, and Other/Unknown
Race/ethnicity	Self-reported race ethnicity (field names: race_source_concept_id, ethnicity_source_concept_id)	Self-report EHR	underweight (<18.5), healthy weight ([18,25)), overweight ([25,30)), and obese (30+)
Body mass index	Median of EHR values		
Drinking status	Self-report (concept ID: 1586198)	Self-report EHR	
Smoking status	Self-report (concept IDs: 1585857, 1585860)	Self-report EHR	
Depression	Phecode MB_286.2: Major depressive disorder		
Hypertension	Phecode CV_401: Hypertension		
Charlson Comorbidity Index	comorbidity R package using Quan weights and ICD-9-CM/ICD-10-CM data		very low (0), low (1-2), moderate (3-4), and high (5+)



Supplementary Figure 4.1 Top decile-to-middle 20% odds ratio (95% confidence interval) for by phenotype risk score (PheRS) and weighting approach by time threshold for esophageal cancer [CA_101.1]. Corresponding plots for liver[CA_101.6] and pancreatic [CA_101.8] cancers are shown in Figure 4.3 and Supplementary Figure 4.2, respectively.



Supplementary Figure 4.2 Top decile-to-middle 20% odds ratio (95% confidence interval) for by phenotype risk score (PheRS) and weighting approach by time threshold for pancreatic cancer [CA_101.8]. Corresponding plots for liver [CA_101.6] and esophageal [CA_101.1] cancers are shown in Figure 4.3 and Supplementary Figure 4.1, respectively.



Supplementary Figure 4.3 Comparison of AUCs (95% CI) in this manuscript with those published in the literature. We identified published risk prediction models developed for use in a general US or UK adult population that reported AUCs for esophageal (panel A),[273,275–278] liver (panel B),[147] and pancreatic cancer (panel C).[18,257,272,277] Point shapes indicate whether genetic information was (triangle) or was not (circle) used in model development. Point estimate color indicates whether the model was stratified by sex (blue for male; purple for female) or not (unstratified; black). Estimates from this manuscript (shaded background) are taken from the joint random forest model using the weighting approach that yielded the highest AUC. The numbers in the point estimates indicate the time threshold. Study details are provided in Supplementary Table 4.7 and Supplementary Table 4.8.

Supplementary Table 4.3 Risk stratification and diagnostic performance for cascade models by PheRS approach, weighted approach and outcome at t=0 threshold.

This table can be viewed via this dissertation's corresponding repository at https://www.doi.org/10.17605/OSF.IO/SBMN2

Supplementary Table 4.4 Risk stratification and diagnostic performance for cascade models by PheRS approach, weighted approach and outcome at t=1 threshold.

This table can be viewed via this dissertation's corresponding repository at https://www.doi.org/10.17605/OSF.IO/SBMN2

Supplementary Table 4.5 Risk stratification and diagnostic performance for cascade models by PheRS approach, weighted approach and outcome at t=2 threshold.

This table can be viewed via this dissertation's corresponding repository at https://www.doi.org/10.17605/OSF.IO/SBMN2

Supplementary Table 4.6 Risk stratification and diagnostic performance for cascade models by PheRS approach, weighted approach and outcome at t=5 threshold.

This table can be viewed via this dissertation's corresponding repository at https://www.doi.org/10.17605/OSF.IO/SBMN2

Supplementary Table 4.7 Comparison of AUCs (95% CI) in this manuscript with those published in the literature. We identified published 5- or 10-year risk prediction models developed for use in a general US or UK adult population that reported AUCs for esophageal, liver, and pancreatic cancers. Estimates from this manuscript are taken from the joint random forest model using the weighting approach that yielded the highest AUC.

Esophageal cancer

Study	**Country**	**AUC (95% CI)**
Rubenstein et al. (2013) doi: 10.1038/ajg.2012.446[275] *predicts Barrett's esophagus*	US	0.72 (0.66, 0.79)
QCancer10 Hippisley-Cox & Coupland (2015) doi: 10.1136/bmjopen-2015-007825[277]	US	0.868 (0.862, 0.874) (male) 0.873 (0.864, 0.881) (female)
Dong et al. (2018) doi: 10.1053/j.gastro.2017.12.003[273]	Multiple	0.745 (0.721, 0.769) (without genetic information) 0.754 (0.729, 0.778) (with genetic information)
Baldwin-Hunter et al. (2019) doi: 10.1007/s10620-019-05707-2[276] *predicts Barrett's esophagus*	US	0.71 (0.64, 0.77)
Wang et al. (2021) doi:10.14309/ajg.0000000000001094[278] *predicts esophageal squamous cell carcinoma*	Norway/UK	0.70 (0.64, 0.75)
This manuscript	US	0.820 (0.781, 0.860) (t=0) 0.605 (0.543, 0.667) (t=1) 0.540 (0.476, 0.604) (t=2) 0.538 (0.459, 0.616) (t=5)

Liver cancer

Study	**Country**	**AUC (95% CI)**
Liu et al. (2022) doi: 10.3389/fpubh.2022.955287[147]	UK	0.771 (0.702, 0.840)
This manuscript	US	0.909 (0.893, 0.926) (t=0) 0.776 (0.747, 0.806) (t=1) 0.762 (0.731, 0.793) (t=2) 0.713 (0.675, 0.752) (t=5)

Pancreatic cancer

Study	**Country**	**AUC (95% CI)**
Klein et al. (2013) doi: 10.1371/journal.pone.0072311	Multiple	0.58 (0.56, 0.60) (without genetic information) 0.61 (0.058, 0.63) (with genetic information)
QCancer10 Hippisley-Cox & Coupland (2015) doi: 10.1136/bmjopen-2015-007825[277]	UK	0.857 (0.846, 0.867) (male) 0.865 (0.855, 0.875) (female)
Salvatore et al. (2021) doi: 10.1016/j.jbi.2020.103652[18]	US	0.732 (0.710, 0.754) (t=5, without genetic information) 0.742 (0.720, 0.763) (t=5, with genetic information)
This manuscript	US	0.842 (0.814, 0.869) (t=0) 0.574 (0.529, 0.620) (t=1) 0.558 (0.511, 0.605) (t=2) 0.530 (0.474, 0.585) (t=5)

Supplementary Table 4.8 Summary of digestive cancer (5-/10-year) risk prediction models developed/validated using data from the US/UK for use in a general adult population and reported area-under-the-curve (AUC) metrics.

This table can be viewed via this dissertation's corresponding repository at https://www.doi.org/10.17605/OSF.IO/SBMN2

Supplementary Section 4.1 Handling of glmnet convergence issues

There are several settings that might cause glmnet models in R to not converge including large penalty factors, sparse data, collinearity in predictors, inadequate iterations, ill-scaled features, extreme values or outliers, choice of convergence criteria, and algorithmic instability. Convergence issues exclusively took place when the weights argument was specified (i.e., there were no convergence issues fitting unweighted glmnet models). glmnet models were fit for hyperparameter tuning and for prediction model fitting. We here describe the sequence in which models were fit if there were convergence issues.

For hyperparameter tuning

1. Attempt to perform tuning using wglmnet (based on Iparragirre et al.'s wlasso, described in Materials and methods one-step approaches section in main text)
2. If there is a warning or error, revert to glmnet::cv.glmnet with the weights as an unpenalized predictor

For prediction model fitting

Model fitting was carried out in the following order. If the model (a) failed to converge, (b) had a negative deviance ratio, (c) had 0 degrees of freedom, or (d) retained only the weights as a predictor (see 5 and 6), the subsequent model was fit. If none of the models met these criteria, no model was fit (happened in one instance: unweighted lasso for esophageal cancer [CA_101.1] at t=2).

1. Attempt to fit glmnet model with lambda.min from hyperparameter tuning and weights in weight argument
2. Attempt to fit glmnet model with lambda.1se from hyperparameter tuning and weights in weight argument

Screen out highly correlated predictors using caret::findCorrelation such that absolute values of pair-wise correlations are less than 0.25 (retaining variable with smaller mean absolute correlation).

3. Attempt to fit glmnet model with screened predictors with lambda.min from hyperparameter tuning and weights in weight argument
4. Attempt to fit glmnet model with screened predictors with lambda.1se from hyperparameter tuning and weights in weight argument
5. Attempt to fit glmnet model with screened predictors with lambda.min from hyperparameter tuning and with weights as unpenalized predictor
6. Attempt to fit glmnet model with screened predictors with lambda.1se from hyperparameter tuning and with weights as unpenalized predictor

Diagnostics corresponding to attempted and final model fits are described for esophageal,

liver, and pancreatic cancers in

Supplementary Table 4.9, Supplementary Table 4.10, and

Supplementary Table 4.11, respectively.

Supplementary Table 4.9 Diagnostics corresponding to glmnet predictive model fitting for esophageal cancer [CA_101.1].

This table can be viewed via this dissertation's corresponding repository at https://www.doi.org/10.17605/OSF.IO/SBMN2

Supplementary Table 4.10 Diagnostics corresponding to glmnet predictive model fitting for liver cancer [CA_101.6].

This table can be viewed via this dissertation's corresponding repository at https://www.doi.org/10.17605/OSF.IO/SBMN2

Supplementary Table 4.11 Diagnostics corresponding to glmnet predictive model fitting for pancreatic cancer [CA_101.8].

This table can be viewed via this dissertation's corresponding repository at https://www.doi.org/10.17605/OSF.IO/SBMN2

Chapter 5 Impact of Polygenic Risk Score-Informed Multiple Imputation and Sample Weighting for Handling Missing Data and Selection Bias on Association Estimation in EHR-Linked Biobanks

5.1 Abstract

Electronic health records (EHR) are valuable public health and clinical research resources. However, analyses based on EHR data are subject to multiple sources of bias, including those due to missing data and non-probability selection. Missing data in EHRs is particularly problematic because it is challenging to distinguish between data that were not recorded and data that are absent/present due to clinically informative missingness/observation process. The patient's visit process may be driven by patient-level covariates such as age, sex, self-reported race/ethnicity, partnered status, access to healthcare, or underlying health status. In the US, this is exacerbated by EHR fragmentation across healthcare providers, where single centers maintain incomplete snapshots of an individual's health history. EHR-linked biobanks like the Michigan Genomics Initiative (MGI) link EHR with genetic information and can lead to exposure and outcome proxies by constructing polygenic risk scores (PRS). Data from MGI are subject to selection bias because 90% of MGI participants are recruited while awaiting surgery. It is an interesting question whether PRS that are relatively complete on the recruited participants can help handling exposure or outcome missing values.

While multiple sources of bias in EHR-linked biobanks are often studied in isolation, in reality, they co-occur. This paper investigates: (a) whether PRS-informed

multiple imputation reduces bias in estimating association parameters due to missing data in a biobank with germline genetics available for nearly all participants, and (b) the joint impact of PRS-informed multiple imputation and sample weighting on exposure-outcome association estimation. To study these questions, we conducted simulations inducing (a) exposure only and (b) exposure and outcome missingness under different missingness (e.g., missing at random (MAR)) and sampling (random and covariate-driven) mechanisms, and analyzed the bias, coverage, width of confidence interval, and root mean square error properties of covariate-adjusted exposure-outcome association estimates across several sample sizes (n=1,000, 2,500, 5,000, and 10,000). We then presented a case study using MGI data to estimate the association of BMI with blood glucose in individuals 40 years or older without diabetes. We fit the association model separately in non-Hispanic Whites (n=42,999) and non-Hispanic Blacks (n=2,297). We compared association estimates using complete case analysis and multiple imputation with (PRS-imputed) and without PRS (woPRS-imputed).

In our simulation study, PRS-imputed analyses exhibited better properties than woPRS-imputed with MAR data under random sampling (e.g., woPRS-imputed and PRS-imputed percent bias for n=2,500 with exposure and outcome missingness: 2.1% and 1.4%, respectively; n=5,000: 1.8%, and 1.4%, respectively). PRS-imputed analyses, bolstered by sample weighting, exhibited superior properties when analyzing MAR data where covariates drove sample selection. However, it did not fully recover nominal coverage rates (e.g., coverage rate for n=2,500 and n=5,000: 0.860 and 0.895, respectively).

In our MGI case study, we estimated the regression coefficient of BMI on glucose in non-Hispanic White and non-Hispanic Black strata. Among non-Hispanic Whites, unweighted, adjusted coefficient estimates changed somewhat between complete case, woPRS-imputed, and PRS-imputed analyses (e.g., unweighted complete case: 0.288 (0.264, 0.312); woPRS-imputed: 0.300 (0.277, 0.323); PRS-imputed: 0.302 (0.280, 0.324)). However, weighted analyses using stratum-specific selection weights changed the estimates considerably (e.g., PRS-imputed increased to 0.338 (0.302, 0.374)), more closely aligning with a benchmark range (0.375, 0.423) estimated using All of Us data. A smaller extent of change after weighting was observed in non-Hispanic Blacks (e.g., unweighted PRS-imputed: 0.203 (0.116, 0.290); weighted PRS-imputed: 0.214, (0.096, 0.332); benchmark range: (0.196, 0.297)).

Our study suggests that EHR-linked biobanks can effectively use genetic data, such as PRS, as proxies in multiple imputation strategies to address missing data. Researchers should employ multiple analytic techniques to address multiple biases, like PRS-informed multiple imputation for missing data and sample weighted analysis for selection bias, which exhibited the best association estimation properties in our simulation study.

5.2 Introduction

Electronic health records (EHR) represent a rich, longitudinal resource that researchers increasingly use to address questions of public health and clinical significance. EHR-linked biobanks, which often contain genetic information linked to other data sources, including administrative and insurance claims, neighborhood-level characteristics, and complementary survey data, are growing in both the number of participants (n) and the breadth of measured variables (p). However, EHR data has not been designed for research purposes, so researchers must carefully consider potential biases/systematic errors. Potential sources of bias include selection bias,[7,76] misclassification,[7,51,292] confounding,[48,293] missing data,[20,44,107] clinically informative visiting process,[107,211,294] immortal time bias,[295,296] and heterogeneity across EHRs.[297,298] Although the advent of large-scale secondary data (colloquially, “Big Data”[299]) effectively neutralizes the threat of random error, systematic sources of bias are ever-present adversaries, unphased by ever-increasing sample sizes. In fact, large sample sizes amplify these biases relative to the very small variance, frequently making inference erroneous. While we typically study one source of bias at a time, these biases exist simultaneously in practice; our study jointly considered missing data and selection bias. Most importantly, we explored whether the observed genetic data, measured on almost all participants in the internal study sample, when related to the variables with missing data or to the variables driving sample selection, can help us reduce the double jeopardy of biases.

Missing Data: Missing data is a foundational statistical problem,[19] ubiquitous in epidemiology,[300–304] and almost universally encountered in health research.[21,305] Complete case analyses, which remove observations with missing values for the variables of interest, are the most commonly reported missing data method in randomized clinical trials[305] and epidemiology studies.[21] However, depending on why the data are missing, ignoring missing data (as in complete case analyses) can lead to biased parameter estimation, leading to invalid conclusions.[20,300,306]

Why data are missing (i.e., the missing data mechanism(s)) are classically defined into three classes: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).[22] Complete case analyses are expected to produce unbiased results when data are MCAR. However, assuming MCAR in EHR-linked biobanks is unreasonable for several reasons. First, patient-provider interactions are not random – for example, sicker patients and patients with easy access to healthcare are observed more frequently[23,55,56]). Second, clinicians judge a patient's risk, which can prompt a test order; thus, the presence/absence of a test order itself is clinically informative.[107,108] Third, EHR fragmentation, an issue in the US, means providers document only the patient interactions within their own health care system. This can result in incomplete capture of a patient's health history across time and type of encounter. Data in an EHR can be missing for other reasons, including failure to initiate or complete an encounter,[76] financial costs associated with testing and diagnosis,[109] underdiagnosis,[110] and differential disease classification processes.[111] Methods like inverse probability weighting[34,117,307] and full-information maximum likelihood[118,119,308,309] can improve precision, reduce bias, and result in valid conclusions in MAR data, a more plausible assumption in EHR data.[310] However, MNAR EHR data cannot be ruled out, particularly for laboratory tests, such as cholesterol tests, which may be ordered when a clinician suspects the patient is at elevated risk or when the result is more likely to be abnormal based on presenting symptoms.

Perhaps the most common method for handling missing data (after complete case analysis), multiple imputation uses information from observed data to create a distribution of possible values for missing data, is flexible, and is readily implementable in statistical software.[20,112,114,311] Multiple imputation aims to produce unbiased and efficient estimates for population parameters of interest by retaining uncertainty in the missing values rather than simply attempting to achieve the best estimate for a missing value.[300] Multiple imputation has been recommended as a method for handling the common problem of missing data in EHR-linked biobanks.[1,38,44,310,312] However, it is empirically impossible to distinguish between MAR and MNAR settings, both of which are plausible in EHR-linked biobanks. While multiple imputation is generally effective in MAR settings, it may not work well in MNAR settings.

Selection Bias: EHR-linked biobanks often do not represent their source (or target) population, introducing potential selection bias. Recruitment mechanisms like recruiting patients awaiting surgery (as in the Michigan Genomics Initiative (MGI)[45]) and oversampling groups historically underrepresented in biomedical research (as in the NIH All of Us Research Program[2]) as well as participant-driven factors like healthy volunteer bias (as in the UK Biobank[4,8]) explain differences between the cohorts and their underlying source populations.[4,313] (We note that selection bias due to missing data or non-response differs because it induces differences between the analytic and study samples). Weighting-based methods like inverse probability weighting and poststratification weighting are often employed to reduce selection bias in parameter estimation. Recent papers have shown that weighted analyses reduce (but not remove) bias due to selection in EHR-linked biobanks.[8,97,215]

The role of genetic information: Analyses of genetic data, which are often completely observed in EHR-linked biobanks, were shown to be vulnerable to selection bias.[97] Using the non-missing genetic data to predict missing phenotypes/outcomes in

EHR-linked biobanks has been explored. Li, Chen, and Moore demonstrated that using genetic data available in EHR-linked biobanks improved imputation for missing cardiovascular phenotypes.[44] Concurrently, Ma and colleagues developed exposure polygenic risk scores (PRS) (ExPRS), which were associated with 27 common exposures, like body mass index (BMI) and glucose, in MGI and the UK biobank.[43] These works motivate, to the best of our knowledge, an unexplored question: can PRS-informed multiple imputation reduce bias due to missing exposure data in association estimation?

Building off these studies, we investigated (a) whether PRS-informed multiple imputation meaningfully reduces bias due to missing data and (b) the joint impact of PRS-informed multiple imputation and sample weighting on exposure-outcome association estimation (Figure 5.1). We calculated unweighted and weighted complete case- and multiple imputation-based estimates of the BMI coefficient for glucose in simulations and a stratified case study using data among non-Hispanic White (n=42,999) and non-Hispanic Black (n=2,297) participants in MGI. First, our simulation studies explored the joint impacts of complete case analysis and multiple imputation with and without exposure and outcome PRS and weighted analysis for selection bias on association estimation. We evaluated these methods in terms of percent bias, coverage rate, confidence interval width, and root mean square error (RMSE). Our case study applied these methods to MGI data to estimate the BMI coefficient for glucose using the same missing data methods and stratum-specific selection weights (as described previously[215]) to demonstrate differences in association estimates in real-world data. We concluded by highlighting the utility of genetic data to reduce bias due to missing data and calling for applying multiple approaches to address the issue of multiple biases.

5.3 Materials and Methods

5.3.1 Simulations

5.3.1.1 Generating outcome, exposure, covariates, and polygenic risk scores jointly

We simulated 1,000 replicates of a pseudo-population with size 100,000 (Figure 5.2). To achieve this, we first generated an 8-dimensional multivariate normal distribution, $\mathbf{X} \sim \mathcal{N}_8(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, mimicking the joint distribution of age, sex, non-Hispanic White (NHW), smoking status (ever/never), BMI, glucose, BMI PRS, and glucose PRS, assuming mean standardized variables ($\boldsymbol{\mu} = 0$) and $\boldsymbol{\Sigma}$ as observed in MGI (see Eq.1 below).

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.998 & 0.209 & 0.254 & -0.177 & 0.060 & 0.092 & 0.084 & 0.092 \\ 0.209 & 0.997 & 0.051 & 0.020 & -0.006 & 0.022 & 0.301 & -0.005 \\ 0.254 & 0.051 & 0.923 & -0.126 & 0.160 & 0.165 & -0.052 & -0.004 \\ -0.177 & 0.020 & -0.126 & 1.001 & -0.038 & -0.115 & 0.022 & -0.008 \\ 0.060 & -0.006 & 0.160 & -0.038 & 0.907 & 0.078 & -0.033 & 0.023 \\ 0.092 & 0.0217 & 0.165 & -0.115 & 0.078 & 1.005 & 0.076 & 0.001 \\ 0.084 & 0.301 & -0.052 & 0.022 & -0.033 & 0.076 & 0.990 & 0.061 \\ 0.092 & -0.005 & -0.004 & -0.008 & 0.023 & 0.001 & 0.061 & 1.005 \end{pmatrix} \qquad \textit{(Eq.1)}$$

Binary variables were recoded (sex, NHW, smoking status) from the generated continuous variables to preserve the observed variance-covariance structure.

5.3.1.2 Sample selection

For each pseudo-population, we performed sampling under two scenarios: random and biased/covariate-informed. Covariate-informed sampling probabilities depended on observed age, glucose, and BMI (e.g., $logit\left(P(S = 1|age, glucose, BMI)\right) = \gamma_0 + \gamma_{age}age + \gamma_{glucose}glucose + \gamma_{BMI}BMI$, where S is an indicator variable for selection into the sample; $\gamma_{age}, \gamma_{glucose}, \gamma_{BMI} = 1$). For each scenario, the intercept, γ_0, was selected to draw a probability sample of approximately 1,000, 2,500, 5,000, and 10,000 (γ_0 = -6.92, -5.83, -4.94, -3.93, respectively) from the pseudo-population where individual i had selection probability $P\big(S_i = 1|age_i, glucose_i, BMI_i)\big)$. In practice, selection could be dependent on unmeasured variables, but this was not considered in our simulation.

5.3.1.3 Missingness generation

We simulated (a) exposure only (i.e., BMI) and (b) exposure and outcome (i.e., BMI and glucose) missingness under MCAR, MAR, and MNAR mechanisms for each selected sample size n. Approximately 25% missingness was generated for each variable. Under MCAR, the probability of missingness (e.g., $P(R_{BMI} = 1)$ where R_{BMI} is an indicator for whether BMI is missing) was 25% for all observations. Under MAR, exposure (i.e., BMI) missingness depended on the outcome and covariates (age, sex, race/ethnicity, smoking status) and outcome missingness dependent only on covariates (e.g., $P(R_{BMI} = 1|glucose, \boldsymbol{covariates}) = \alpha_0 + \alpha_{glucose} glucose + \boldsymbol{\alpha_{covariates} covariates}$). Under MNAR, exposure and outcome missingness was dependent on the exposure, outcome, and covariates (e.g., $P\big(R_{glucose} = 1\big|glucose, BMI, \boldsymbol{cov}\big) = \zeta_0 + \zeta_{glucose} glucose + \zeta_{BMI} BMI + \boldsymbol{\zeta_{cov} cov}$). In all settings, all non-intercept coefficients were set equal to 1 and only the intercept varied. Supplementary Table 5.1 shows the intercept coefficient values by missingness mechanism to (approximately) achieve the desired sample sizes.

5.3.1.4 Methods

We performed unweighted and weighted analyses. Weighted analyses were employed to address selection bias. We used the covariate-informed sampling probabilities meaning the true sample weights were known, not estimated. The weights were proportional to the inverse of sampling probabilities for each individual i ($\omega_i \propto \frac{1}{P(S_i=1|age_i,glucose_i,BMI_i)}$), and weighted analyses were carried out using the survey R package (version 4.4-2).[242]

We also performed multiple imputation to address missing data. For each sample, multiple imputation using age, sex, NHW, smoking status, BMI, glucose (woPRS-imputed) and additionally exposure (BMI) and outcome (glucose) PRS (PRS-imputed) was carried out using the R package mice (version 3.16.0)[35,115] with default settings (5 imputations using predictive mean matching). Beta coefficients across imputations were pooled using Rubin's rule, with confidence intervals calculated from pooled standard errors based on within and between imputation variances.[112,314] For multiply imputed analyses of biased/covariate-informed samples, weighted analyses were conducted on each imputed dataset before pooling.

Our target quantity was the true coefficient of BMI in a linear regression model for glucose (β_{BMI}) adjusted for age, sex, non-Hispanic White race/ethnicity, and smoking status (ever/never) (Eq. 2).

$$Glucose_i = \beta_0 + \beta_{BMI}BMI_i + \beta_{age}age_i + \beta_{sex}sex_i + \beta_{NHW}NHW_i + \beta_{smoke}smoke_i + \epsilon_i \qquad \textit{(Eq.2)}$$

For each replicate, the true β_{BMI} was obtained from the pseudo-population of size 100,000 and the sample estimates were obtained in the selected samples of sizes 1,000, 2,500, 5,000, and 10,000. In each sample, we conducted complete case, woPRS-imputed, and PRS-imputed analyses, extracting the coefficient estimate of BMI for glucose ($\hat{\beta}_{BMI}$). We evaluated association estimation properties using percent bias, coverage rate, average 95% confidence interval width, and RMSE, averaged over the 1,000 replicates.

5.3.2 Case study: Michigan Genomics Initiative

5.3.2.1 Description of the study cohort

MGI is an EHR-linked biobank that began in 2012, initially recruiting adult patients through pre-/peri-operative appointments requiring anesthesia from the University of Michigan Health System. As of September 2023, ~100,000 consented participants have provided access to their EHR and a biospecimen for genotyping, with a recent follow-up effort collecting complementary survey data.[67] This paper included 42,999 (25,520 complete cases) non-Hispanic White and 2,297 (1,240 complete cases) non-Hispanic Black participants aged 40 or older without a diabetes diagnosis (phecode EM_202) and with demographic, health measurement, laboratory, and polygenic risk score data. MGI protocols were reviewed and approved by the University of Michigan Medical School Institutional Review Board (IRB ID HUM00099605 and HUM00155849).

5.3.2.2 Outcome, exposure, covariates, and polygenic risk score

The outcome and exposure of interest were glucose (mg/dL; logical observation identifiers names and codes (LOINC) code: 2345-7) and BMI (kg/m^2), respectively. The longitudinal data in the EHR was reduced to the participant's median value after removing extreme values (values outside 1.5x the interquartile range) for the corresponding variable. Age was considered the participant's age at the time of data pull (March 23, 2022). Sex (indicator for female) and race/ethnicity were obtained from EHR data. Multiple measurements of self-reported smoking status were recorded and recoded into a binary ever/never indicator variable.

Ma and colleagues previously calculated several PRS for 27 exposures in MGI participants.[43] In this paper, we selected the Lassosum PRS for BMI and the deterministic

Bayesian sparse linear mixed model PRS for glucose because they had the highest R^2 value for their respective traits in the published paper.[43] These PRS relied on publicly available GWAS summary statistics of UK Biobank data (Neale lab[315,316]). Both PRS's were predictive in MGI, with BMI PRS being much stronger (Pearson correlation between BMI and BMI PRS: 0.30; glucose and glucose PRS: 0.09; Supplementary Figure 5.2).

5.3.2.3 Estimated regression coefficient corresponding to BMI with glucose as the outcome

The target estimand of interest was the regression coefficient corresponding to BMI with glucose as the outcome. We conducted analyses among individuals 40 and older without a diabetes diagnosis in non-Hispanic White and non-Hispanic Black strata as well as in the full (i.e., unstratified) cohort. Unlike in the simulations, the selection weights in MGI were not known. Salvatore and colleagues estimated inverse probability selection weights to make MGI more representative of the US adult population using National Health Interview Survey data.[215] Using the same methods to calculate stratum-specific weights, we conducted weighted versions of each regression analysis. Using the non-Hispanic White and non-Hispanic Black samples with missing data (n=42,999 and 2,297, respectively), we performed multiple imputation with and without PRS (adjusting for age, sex, and smoking status). We also conducted a PRS-informed multiple imputation analysis where observations were restricted to only those with observed PRS (PRS-imputed (subset): n=25,520 and 1,240 for non-Hispanic Whites and non-Hispanic Blacks, respectively). We reported the estimated beta coefficients and 95% confidence intervals.

5.3.3 Software

Analyses were conducted using R version 4.3.3. The code used to conduct analyses in this paper is available at https://github.com/maxsal/exprs_imputation.

5.4 Results

5.4.1 Simulation study

Random sampling, exposure only missingness: When BMI missingness was MCAR (orange in Figure 5.3), estimated BMI coefficients for glucose maintained the nominal coverage rate (95% confidence level; panel A) and exhibited essentially no bias (panel C). When BMI missingness was MAR, complete case analysis was unable to retain the nominal coverage rate, which decreased as sample size increased, and was consistently biased (e.g., 8.86% for n=1,000, 7.80% for n=10,000). woPRS-imputed analyses were able to achieve a stable coverage rate (e.g., 0.931 for n=1,000, 0.924 for n=10,000) and reduced bias, which decreased as sample size increased (e.g., 2.89% for n=1,000, 0.10% for n=10,000). PRS-imputed analyses were able to full retain the nominal coverage rate at all sample sizes and had less bias, which decreased as sample size increased (e.g., 1.5% for n=1,000, 0.04% for n=10,000). When BMI missingness was MNAR, all analyses failed to retain the nominal coverage rate and exhibited substantial bias (>30%), with multiply imputed analyses performing better than complete case analysis. PRS-imputed analyses were able to achieve a slightly better coverage rate (e.g., for n=1,000: 0.637, PRS-imputed; 0.561, woPRS-imputed) and lower bias (e.g., for n=1,000: 31.95%, PRS-imputed; 36.12%, woPRS-imputed) than woPRS-imputed analyses.

Biased sampling, exposure only missingness: When BMI missingness was MCAR in biased sampling simulations (Figure 5.4), all unweighted analyses were unable to retain the nominal coverage rate and exhibited substantial bias (>29%), which worsened as sample size increased. When missingness was MAR or MNAR, unweighted complete case analyses demonstrated less bias than multiple imputation analyses. This is likely due to multiple imputation amplifying biases when imputing biased samples without considering sampling weights. However, weighted analyses improved association estimation properties, with multiple imputation approaches outperforming complete case analysis. For example, in the 10,000-observation sample with MAR missingness, the coverage rate was 0.784, 0.877, and 0.883 for complete case, woPRS-imputed, and PRS-imputed analysis, respectively. Coverage decreased and bias increased for all analyses when BMI was MNAR. Multiple imputation methods performed better than complete case analysis, with PRS-imputed analyses performing slightly better than woPRS-imputed analyses as sample size increased (e.g., coverage rate for n=10,000: 23.22%, complete case; 14.00%, woPRS-imputed; 12.66%, PRS-imputed).

Random sampling, exposure and outcome missingness: When BMI and glucose missingness were MCAR, all analyses maintained the nominal coverage rate and were unbiased (Figure 5.3). For example, the coverage rate of complete case analysis was 0.938, 0.956, 0.961, and 0.951 for MCAR sample sizes 1,000, 2,500, 5,000, and 10,000, respectively. As expected, complete case analyses with MAR and MNAR data could not maintain the nominal coverage rate, and the estimates were biased. The coverage rate decreased, and the percent bias remained stable as the sample size increased. For example, the coverage rate for complete case analyses with MAR data was 0.925, 0.914, 0.857, and 0.784, as the sample size increased from 1,000 to 10,000. For woPRS-imputed and PRS-imputed analyses of MAR data, the coverage rate returned to the nominal level (e.g., 10,000-observation sample: woPRS-imputed: 0.946; PRS-imputed: 0.947) and little-to-no bias was observed (e.g., 10,000-observation sample: woPRS-imputed: 1.32%; PRS-imputed: 0.97%). woPRS-imputed and PRS-imputed analyses were unable to achieve the nominal coverage rate and remained substantially biased in MNAR analyses. Notably, PRS-imputed analyses of MNAR data had better coverage rates (e.g., 1,000-observation sample: complete case: 0.000, woPRS-imputed: 0.296; PRS-imputed: 0.299) and percent bias (e.g., 1,000-observation sample: complete case: 62.27%; woPRS-imputed: 55.68%; PRS-imputed: 53.89%) properties than woPRS-imputed analyses but remained poor. Plots describing average 95% CI width and RMSE are shown in Supplementary Figure 5.3.

Biased sampling, exposure and outcome missingness: Because PRS-informed multiple imputation exhibited superior properties when analyzing missing data in random samples, we only comment on unweighted and weighted PRS-imputed analyses in biased/covariate-informed sampling here (results for other analyses shown in Figure 5.4). Unweighted PRS-imputed analyses of MCAR, MAR, and MNAR data were unable to recover the nominal coverage rate (e.g., 2,500-observation MAR sample: 0.080; Figure 5.4B, left) and demonstrated substantial bias (e.g., 2,500-observation MAR sample: 53.37%; Figure 5.4D, left) when selection weights were not considered. While analyses for all missingness mechanisms were similarly poor in terms of coverage rate, MCAR analyses performed best in terms of percent bias, followed by MAR and then MNAR, regardless of the missing data method. Notably, complete case analyses of unweighted data exhibited slightly better properties of MNAR data than the multiple imputation approaches (e.g., percent bias in 2,500-observation sample: complete case: 98.04%; woPRS-imputed: 135.57%; PRS-imputed: 122.88%).

Sampling weighted analyses told a different story (Figure 5.4, right side of panels). Multiple imputation analyses of MCAR data performed better than the complete case analysis, and the coverage rate decreased as the sample size increased. However, PRS-imputed analyses of MAR data saw improved coverage rates, exceeding those seen in MCAR analyses, and increased as sample size increased. For example, the PRS-imputed coverage rate for MCAR analyses of sample sizes 1,000, 2,500, 5,000, and 10,000 were 0.890, 0.896, 0.875, and 0.834, respectively. Meanwhile, the same analysis applied to MAR data resulted in coverage rates of 0.840, 0.860, 0.895, and 0.906, respectively. Regardless of the method for handling missing data, MNAR analyses had similarly poor coverage rates, which decreased with increasing sample size.

Similar trends were observed regarding percent bias (Figure 5.4D, right). PRS-imputed analyses of MAR data exhibited little bias (e.g., n=1,000: 3.67%; n=2,500: 5.99%; n=5,000: 7.54%; n=10,000: 7.38%). PRS-imputed analyses of MNAR data performed substantially worse in terms of percent bias (e.g., n=1,000: 32.40%; n=2,500: 21.94%; n=5,000: 21.07%; n=10,000: 23.19%), while still performing better than complete case and woPRS-imputed analyses (plots depicting average 95% CI width and RMSE are shown in Supplementary Figure 5.4 and Supplementary Figure 5.5).

5.4.2 Analysis in the Michigan Genomics Initiative (MGI)

5.4.2.1 Descriptive characteristics of the study population

We initially considered a cohort of 50,026 MGI participants aged 40 or older without diabetes, which were 54.5% female and 86.0% non-Hispanic White, with a mean (standard deviation (SD)) age of 62.9 (12.5), BMI of 29.1 (6.0), and glucose of 99.0 (14.1) mg/dL (Supplementary Table 5.8). However, because the literature suggested racial/ethnic heterogeneity,[317,318] we stratified our analysis into non-Hispanic White and non-Hispanic Black strata.

Of 42,999 non-Hispanic White MGI participants aged 40 or older without diabetes, 53.8% were female, with a mean (SD) age of 63.5 (12.5) years old, BMI of 29.1 (6.0), and glucose of 99.5 (14.2) mg/dL (). A subset of 25,520 participants had no missing data. Individuals with any missing data were more likely to be younger (mean age 62.6 vs. 64.1; p<0.001), more female (54.6% vs. 53.3%; p=0.006), less likely to have ever smoked (46.5% vs. 50.7%; p<0.001) and have lower glucose values (99.47 mg/dL vs 99.55 mg/dL; p=0.023) compared to complete cases. No statistical differences in BMI PRS (p=0.6) or glucose PRS (p=0.4) were observed between complete cases and participants with missing data.

Of 2,297 non-Hispanic Black MGI participants aged 40 or older without diabetes, 63.3% were female, with a mean (SD) age of 57.8 (11.4) years old, BMI of 30.8 (6.4) and glucose of 95.1 (12.5) mg/dL (Table 5.2). A subset of 1,240 participants had no missing data. Individuals with any missing data were less likely to have ever smoked (39.2% vs 44.7%; p=0.017) and have lower glucose values (94.0 mg/dL vs 95.8 mg/dL; p<0.001). Again, no statistical differences in BMI PRS (p=0.8) or glucose PRS (p=0.061) were observed between complete cases and participants with missing data.

Missing PRS values exist because some samples have not been genotyped yet. Subsets of 30,492 non-Hispanic Whites and 1,437 non-Hispanic Blacks with genotyped biospecimen were also analyzed with complete PRS data (last column, Table 5.1 and

Table 5.2). Smoking status and glucose exhibited fair missingness (i.e., was never recorded at any encounter), approximately 14% and 11% in the non-Hispanic White sample and 13% and 7% in the non-Hispanic Black sample. BMI was hardly missing (0.5% in non-Hispanic Whites and 1.1% in non-Hispanic Blacks).

5.4.2.2 Estimation of the coefficient for BMI with glucose as the outcome

In non-Hispanic Whites: Among non-Hispanic White individuals aged 40 years or older without a diabetes diagnosis in MGI, the unweighted, covariate-adjusted, complete case coefficient estimate was 0.288 (0.264, 0.312) (Figure 5.5). This differed from the benchmark range of (0.376, 0.423), which was derived from All of Us data using 2019 National Health Interview Survey (NHIS) data and sampling weights to make it representative of US adults aged 40 years or older without diabetes. Deriving similar stratum-specific sampling weights in MGI using NHIS data to reduce selection bias, the weighted, covariate-adjusted complete case estimate of 0.324 (0.283, 0.365) aligned more closely with the benchmark. woPRS-imputed and PRS-imputed, unweighted analyses aimed at reducing bias due to missing data saw a nominal increase in the coefficient estimate compared to the corresponding complete case analysis (woPRS-imputed: 0.300 (0.277, 0.323); PRS-imputed: 0.302 (0.280, 0.324)). Weighted woPRS-imputed and PRS-imputed analyses achieved estimates even closer to the benchmark range (woPRS-imputed: 0.331 (0.292, 0.371); PRS-imputed: 0.338 (0.299, 0.373)). We also considered an analysis using the subset of individuals in MGI with completely observed exposure and outcome PRS, which nominally decreased the estimate (0.329 (0.284, 0.375)) compared to the PRS-imputed analysis on the full data. In all cases, weighting and, to a lesser extent, multiple imputation of missing data resulted in coefficient estimates that more closely aligned with the benchmark range. Similar trends were seen in the corresponding unadjusted analyses in Figure 5.5A.

In non-Hispanic Blacks: Among non-Hispanic Black individuals aged 40 years or older without a diabetes diagnosis in MGI, the unweighted, covariate-adjusted, complete case coefficient estimate was 0.178 (0.097, 0.261) (Figure 5.5). This differed from the benchmark range of (0.196, 0.297), similarly derived from stratum-specific All of Us and NHIS derived sampling weights. The weighted, covariate-adjusted complete case point estimate of 0.202 (0.086, 0.317) fell within the benchmark range. woPRS-imputed and PRS-imputed, unweighted analyses saw nominal increases in the coefficient estimates, with point estimates also falling within the benchmark range (woPRS-imputed: 0.204 (0.119, 0.288); PRS-imputed: 0.203 (0.116, 0.290)). Weighted woPRS-imputed and PRS-imputed analyses also had point estimates in the benchmark range with virtually no change in the coefficient estimate (woPRS-imputed: 0.200 (0.082, 0.318); PRS-imputed: 0.214 (0.096, 0.332)). Weighting had a smaller impact than in the non-Hispanic White analyses, in part because point estimates were already in the benchmark range. Unlike the non-Hispanic White analyses, analyses on the subset of individuals with completely observed exposure and outcome PRS performed worse, possibly due to genotyping prioritization mechanisms not captured in the sample weight estimation process.

We present unstratified results on the full MGI adult cohort aged 40 years or older without diabetes in Supplementary Figure 5.6. Results largely mirrored those seen in the non-Hispanic White strata because the cohort was mostly non-Hispanic White (86%) (e.g., covariate-adjusted: unweighted, complete case: 0.277 (0.255, 0.299); weighted woPRS-imputed: 0.305 (0.269, 0.342); weighted PRS-imputed: 0.312 (0.274, 0.349; NHIS-weighted All of Us-derived benchmark range: (0.346, 0.386)).

5.5 Discussion

In this paper, we examined the combined impact of selection bias and missing data on association analyses using EHR data and demonstrated how biobanks with genetic data can use genetic summaries of exposures and outcomes to reduce bias due to these systematic sources of error. Genetic data, often largely observed in EHR-linked biobanks, provides a unique opportunity to further mitigate bias due to missing data. Li and colleagues demonstrated improved properties of data imputed using genetic information,[44] while Ma and colleagues calculated ExPRS predictive of several exposures that frequently have missing values in EHR-linked biobanks.[43] Building on these works, we used simulations and a case study to explore whether PRS-informed multiple imputation improves association estimation properties with missing exposure only and exposure and outcome data and how PRS-informed imputation and sample weighting jointly impact exposure-outcome association estimates. To our knowledge, this is the first paper to explore the joint impacts of genetic-informed multiple imputation and sample weighting methods in EHR-linked biobank data.

Findings from the simulation study: Our random sampling simulation revealed some key findings. Multiple imputation, as expected, improved association estimation of MAR data compared to complete case analyses.[112,306] However, PRS-informed multiple imputation exhibited superior properties than multiple imputation without PRS for MAR and MNAR data. While woPRS- and PRS-imputed analyses improved coverage rates relative to complete case analyses and improved as sample size increased, they failed to recover the nominal coverage rate. PRS-informed multiple imputation produced smaller confidence intervals than multiple imputation without PRS at all sample sizes (Supplementary Figure 5.4). PRS-imputed analyses also had the smallest RMSE of MAR analyses.

While all analyses of MCAR data were (and are expected to be) unbiased, MCAR assumptions are often not applicable to real-world EHR-linked biobank data.[76,243] Clinically informative observation processes due to health status (e.g., less healthy patients have more complete EHR), healthcare access (e.g., those with limited access might have gaps that correlate with health access), and referrals (e.g., patients referred for testing or treatments based on their symptoms and health conditions) violate the MCAR assumption.[23,107,211,312,319] Lack of interoperability and connectivity between EHR across different providers (e.g., changes in patient's insurance status or location, out-of-network referrals for specialized care; i.e., EHR fragmentation), as in the US, further complicates observation processes, providing incomplete snapshots of a patient's health history.[27,28,319–321] While PRS-informed multiple imputation estimates of MNAR had better properties than multiple imputation without PRS and complete case estimates, all methods performed poorly. The improvements in PRS-informed multiple imputation estimates could be attributable to the property that correlations between PRS mimic correlations between their traits' observed values.[43] Notably, the correlations between exposures and their PRS are not very strong (Supplementary Figure 5.2), and stronger correlates would likely improve multiple imputation.

EHR-linked biobank data are subject to selection bias, primarily because of healthy volunteer bias (as in the UK Biobank[4]) or non-random recruitment strategies like recruitment through specific clinics (as in MGI[45]) and oversampling groups historically underrepresented in biomedical research (as in the NIH All of Us Research Program[2]). We explored the joint roles of missing data and selection bias through simulation by oversampling older individuals, individuals with higher BMI, and individuals with higher glucose levels. As expected, all missing data methods for all missing data mechanisms could not capture the true value (e.g., woPRS-imputed coverage rate for n=1,000 MAR exposure and outcome missingness sample: 0.754; Figure 5.4) and exhibited substantial bias (all analyses had at least 21% bias) in unweighted analyses where selection bias was present. In weighted analyses, multiple imputation with and without PRS had better properties than complete case analysis when data were MAR (e.g., complete case, woPRS-imputed, and PRS-imputed coverage rate for exposure and outcome missingness with n=1,000: 0.682, 0.842, 0.840, respectively; percent bias: 35.89%, 1.04%, 3.67%, respectively). However, comparing diagnostics between woPRS- and PRS-imputed analyses showed they were virtually identical for MAR analyses (e.g., woPRS-imputed and PRS-imputed coverage rate for exposure and outcome missingness with n=1,000: 0.842, 0.840, respectively; n=10,000: 0.884, 0.906, respectively), while PRS-imputed analyses showed slightly improved percent bias and RMSE of MNAR data (e.g., woPRS-imputed and PRS-imputed percent bias for exposure and outcome missingness with n=1,000: 33.67% and 32.40%, respectively; n=10,000: 25.46% and 23.19%, respectively).

Findings from the case study: Our estimation of the BMI coefficient for glucose using MGI data demonstrated relatively small differences between complete case and multiple imputation estimates (e.g., weighted complete case, woPRS-imputed, and PRS-imputed coefficient (95% CI) among non-Hispanic Whites: 0.324 (0.283, 0.365), 0.331 (0.292, 0.371), and 0.338 (0.292, 0.371), respectively; Figure 5.5). These small changes are likely attributable to relatively low levels of missingness (e.g., in non-Hispanic Whites and non-Hispanic Blacks, respectively: glucose: 14% and 13%; BMI: 0.5% and 1.1%). We saw substantial changes in the coefficient estimate after accounting for selection bias (e.g., PRS-imputed estimate before weighting: 0.302 (0.280, 0.324; after weighting: 0.338 (0.302, 0.374)). These larger changes are attributable to MGI's recruitment mechanism (primarily at pre-/peri-operative appointments requiring anesthesia) and MGI being unrepresentative of the presumptive target population (US adults).[215] In this case study, bias due to selection played a more substantial role than bias due to missing data.

Guidance for practitioners: In a given situation, we may not know the missingness or selection mechanisms. While it is possible to test the plausibility of data being MCAR[322–325] (a strong and often unrealistic assumption in practice[32]), it is impossible to distinguish between MAR and MNAR data empirically.[300,326] Methods like multiple imputation and inverse probability weighting can reduce bias in MAR data. However, they can theoretically increase bias in MNAR data.[300] While results from our simulation analyses suggest that PRS-informed multiple imputation has preferable properties for the analysis of MAR data and, to a lesser extent, MNAR data, than multiple imputation without PRS, when data are suspected to be MNAR, bounds and sensitivity analyses are recommended.[327] For example, if, based on external information, we hypothesize that glucose values are more likely to be lower than those with observed glucose values, one can model alternative scenarios that fit the hypothesis and compare the robustness and variability in the resulting estimates. Despite well-documented problems with mishandling missing data, missing data mechanisms are not rigorously examined.[21,305,328–330] Researchers must use expert knowledge[113,331] and tools like m-graphs or m-DAGs[332–335] to interrogate not only missingness mechanisms but also identify variables related to missingness before employing analyses appropriate for their missing data. For most regression models, complete case analyses can give unbiased results when the probability of being a complete case is independent of the outcome after taking covariates into account, regardless of the missingness mechanism (Supplementary Table 5.9).[30,121,336] Correctly specifying the selection probability in EHR data is also a challenge.[215] While removing bias from these two sources is a daunting task, our findings suggest PRS-informed multiple imputation with sampling weights help with bias reduction and improving coverage.

5.5.1 Strengths and limitations

This study not only underscores the importance of considering missing data and selection bias in EHR-linked biobanks but also calls for specific actions from researchers. Our simulation studies, informed by real-world variance-covariance conditions, demonstrate the potential of multiple imputation to improve the handling of missing data. Additionally, we examined the combined impacts of missing data and selection bias through multiple imputation and weighting methods. By conducting weighted analyses on multiply imputed data and sharing our code, we promote reproducibility and transparency. We recommend that all researchers adopt these practices to enhance the reliability of their findings.

This study also had several limitations. First, our simulation study considered missingness in two settings: exposure only and exposure and outcome. In practice, multiple missingness patterns simultaneously impact exposures, outcomes, and covariates. Relatedly, this study considered only one level of missingness (~25% missingness). Again, in practice, multiple variables of interest will have varying levels of missingness. Future work should consider different levels and patterns of missingness, preferably those informed by levels and patterns seen in real-world data. Second, the recruitment strategies into EHR-linked biobanks vary greatly, meaning selection bias and its impact on analyses varies dataset-to-dataset. For example, MGI has substantial selection biases (relative to a US adult target population, as explored here). Other prominent biobanks, such as the NIH All of Us Research Program and the UK Biobank, are more representative of presumptive target populations. Thus, the relative impact of selection bias presented in this study may be more pronounced than for biobanks that are more representative of their target population. Third, the real-world case study examined a single association with a relatively small level of missingness compared to the simulations. Further, an established estimate for this association does not exist; thus, assessing changes in the estimates is based on simulation conclusions and expert knowledge rather than a preferable gold standard benchmark. Future studies should explore additional associations for which gold standard estimates exist for comparison. Fourth, this study looked at glucose (LOINC: 2345-7), which was collected primarily through basic or comprehensive metabolic panels and may or may not have been collected under fasting conditions. The uncontrolled collection of glucose compared to fasting glucose makes its interpretation challenging and of limited use. We explored alternative LOINC codes specifying fasting conditions for glucose collection, but they were rarely used. Fifth, this study considered the association between two continuous variables after collapsing longitudinal measurements; however, much clinical research uses binary outcome and longitudinal data available in EHR. Future work should consider association analyses in binary and longitudinal data. Sixth, clinically informative visiting processes increase the likelihood that missing data in EHR are MNAR. While we saw some improvements in association estimation properties of PRS-imputed analyses of MNAR data compared to woPRS-imputed analyses, future studies should incorporate targeted methods that model and account for visiting processes.[337–340]

5.6 Conclusion

Missing data is a pervasive issue in EHR-linked biobank data. In our study, we leveraged a unique aspect of biobanks – non-missing genetic data – to assess whether using PRS-informed multiple imputation of missing data could reduce bias in association estimation. Our simulation studies demonstrated a substantial reduction in bias and an improvement in the coverage rate for MAR data when multiple imputation incorporated genetic information. We also investigated the combined impacts of missing data and selection bias using real-world data from MGI. This case study showed that the impact of missing data was smaller relative to selection bias. Our findings call for future research to explore additional patterns and levels of missingness across several associations and cohorts. Our results indicate that biobanks should provide PRS for common exposures available as proxies to inform multiple imputation of missing data and offer sampling weights to address selection bias. This will enable researchers better to mitigate multiple biases in EHR-linked biobank association analyses, enhancing the reliability and validity of their findings.

5.7 Tables

Table 5.1 Comparison of demographic, health measurements, and polygenic risk score values overall and among non-Hispanic Whites 40 or older without diabetes, with and without any missing values in the Michigan Genomics Initiative.

		Incomplete	**Complete**		**Non-missing**
	Overall	**observations**	**observations**		**PRS**
Characteristic	N = 42,999[a]	N = 17,479[a]	N = 25,520[a]	**p-value**[b]	N = 30,942[a]
Age	63.5 (12.5)	62.6 (12.4)	64.1 (12.5)	<0.001	63.7 (12.5)
Female	53.8 (23,145)	54.6 (9,547)	53.3 (13,598)	0.006	53.4 (16,509)
Smoking status (ever)	49.4 (18,187)	46.5 (5,255)	50.7 (12,932)	<0.001	50.1 (14,320)
Missing	*6,178*	*6,178*	*0*		*2348*
BMI	29.1 (6.0)	29.1 (6.0)	29.0 (5.9)	0.3	29.0 (5.9)
Missing	*236*	*236*	*0*		*138*
Glucose	99.5 (14.2)	99.5 (14.7)	99.5 (14.0)	0.023	99.8 (14.3)
Missing	*4,836*	*4,836*	*0*		*3560*
BMI PRS[c]	0.000 (1.000)	0.004 (1.021)	-0.001 (0.996)	0.6	0.000 (1.000)
Missing	*12,057*	*12,057*	*0*		*0*
Glucose PRS[c]	0.000 (1.000)	0.013 (0.989)	-0.003 (1.002)	0.4	0.000 (1.000)
Missing	*12,057*	*12,057*	*0*		*0*

[a] continuous: mean (SD); dichotomous: % (n)

[b] Wilcoxon rank sum test; Pearson's Chi-squared test

[c] PRS were mean standardized

Abbreviations: BMI, body mass index; PRS, polygenic risk score

Table 5.2 Comparison of demographic, health measurements, and polygenic risk score values overall and among non-Hispanic Blacks 40 or older without diabetes, with and without any missing values in the Michigan Genomics Initiative.

		Incomplete	**Complete**		**Non-missing**
	Overall	**observations**	**observations**		**PRS**
Characteristic	N = 2,297[a]	N = 1,057[a]	N = 1,240[a]	**p-value**[b]	N = 1,437[a]
Age	57.8 (11.4)	57.3 (11.2)	58.2 (11.6)	0.069	57.7 (11.6)
Female	63.3 (1,454)	63.2 (668)	63.4 (786)	>0.9	63.3 (910)
Smoking status (ever)	42.6 (847)	39.2 (293)	44.7 (554)	0.017	44.3 (597)
Missing	*310*	*310*	*0*		*88*
BMI	30.8 (6.4)	30.6 (6.3)	31.0 (6.5)	0.2	31.0 (6.5)
Missing	*26*	*26*	*0*		*8*
Glucose	95.1 (12.5)	94.0 (12.6)	95.8 (12.4)	<0.001	95.8 (12.4)
Missing	*160*	*160*	*0*		*116*
BMI PRS[c]	0.000 (1.000)	-0.006 (0.977)	0.001 (1.004)	0.8	0.000 (1.000)
Missing	*860*	*860*	*0*		*0*
Glucose PRS[c]	0.000 (1.000)	0.158 (1.064)	-0.025 (0.988)	0.061	0.000 (1.000)
Missing	*860*	*860*	*0*		*0*

[a] continuous: mean (SD); dichotomous: % (n)

[b] Wilcoxon rank sum test; Pearson's Chi-squared test

[c] PRS were mean standardized

Abbreviations: BMI, body mass index; PRS, polygenic risk score

5.8 Figures

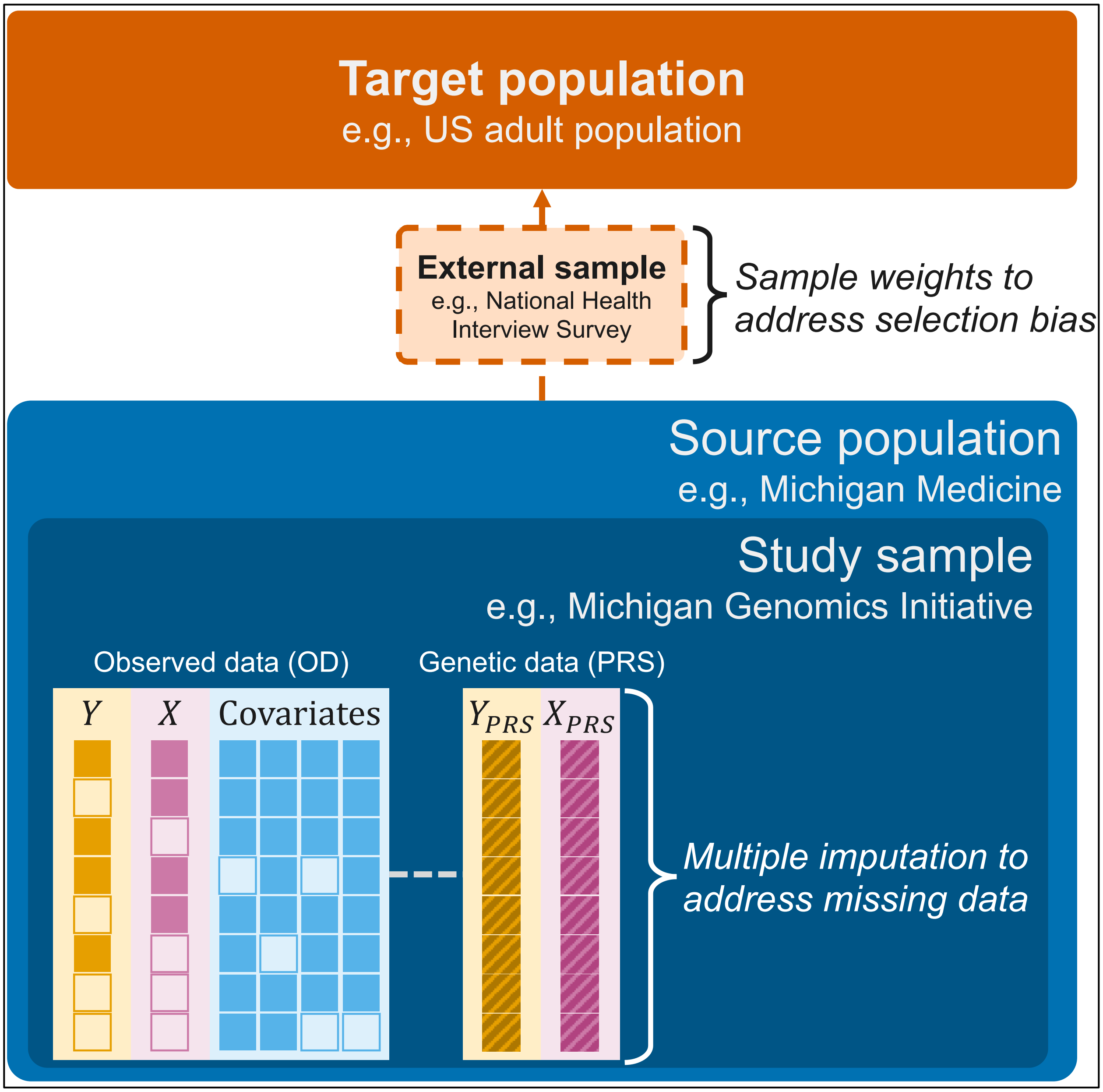

Target population
e.g., US adult population
External sample
e.g., National Health Interview Survey
Sample weights to address selection bias
Source population
e.g., Michigan Medicine
Study sample
e.g., Michigan Genomics Initiative
Observed data (OD)
Genetic data (PRS)
Y
X
Covariates
Y_{PRS}
X_{PRS}
Multiple imputation to address missing data

Figure 5.1 Schematic representation depicting multiple imputation and weighted analyses to jointly address missing data and selection bias. Y represents the outcome (e.g., glucose), X represents the exposure (e.g., body mass index) and covariates could include age, sex, race/ethnicity, and smoking status. The empty boxes represent missing data. Y_{PRS} and X_{PRS} are the polygenic risk scores (PRS) corresponding to the outcome and exposure, respectively.

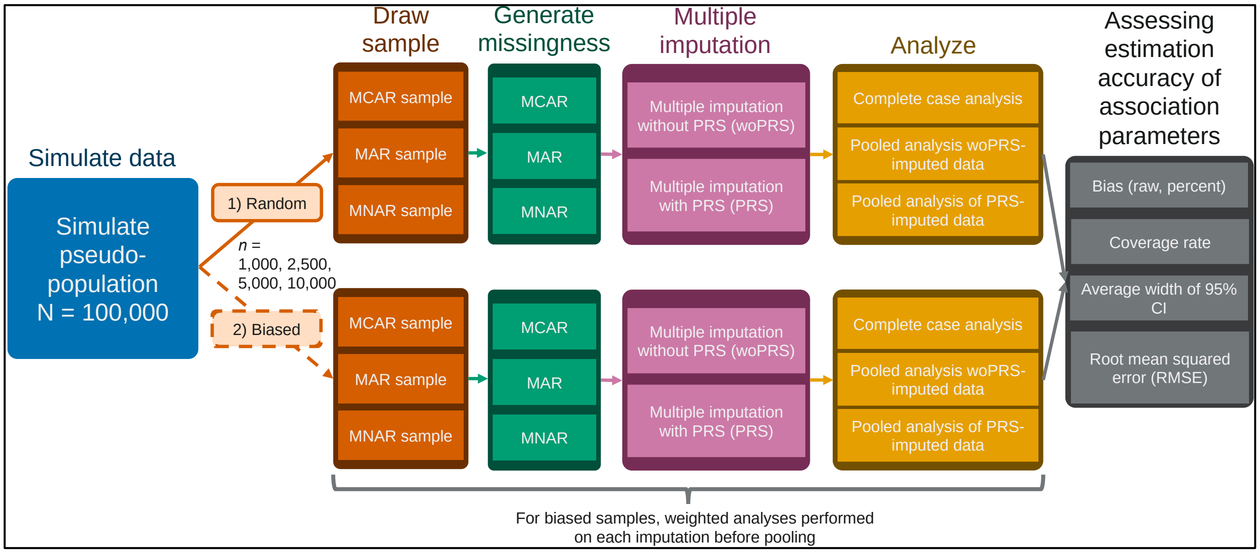

Simulate data
Simulate pseudo-population N = 100,000
1) Random
2) Biased
n = 1,000, 2,500, 5,000, 10,000
Draw sample
MCAR sample
MAR sample
MNAR sample
Generate missingness
MCAR
MAR
MNAR
Multiple imputation
Multiple imputation without PRS (woPRS)
Multiple imputation with PRS (PRS)
Analyze
Complete case analysis
Pooled analysis woPRS-imputed data
Pooled analysis of PRS-imputed data
Assessing estimation accuracy of association parameters
Bias (raw, percent)
Coverage rate
Average width of 95% CI
Root mean squared error (RMSE)
For biased samples, weighted analyses performed on each imputation before pooling

Figure 5.2 Schematic representation of random and biased sampling simulation analyses. Abbreviations: CI, confidence interval; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS, polygenic risk score; RMSE, root mean square error; woPRS, without polygenic risk score

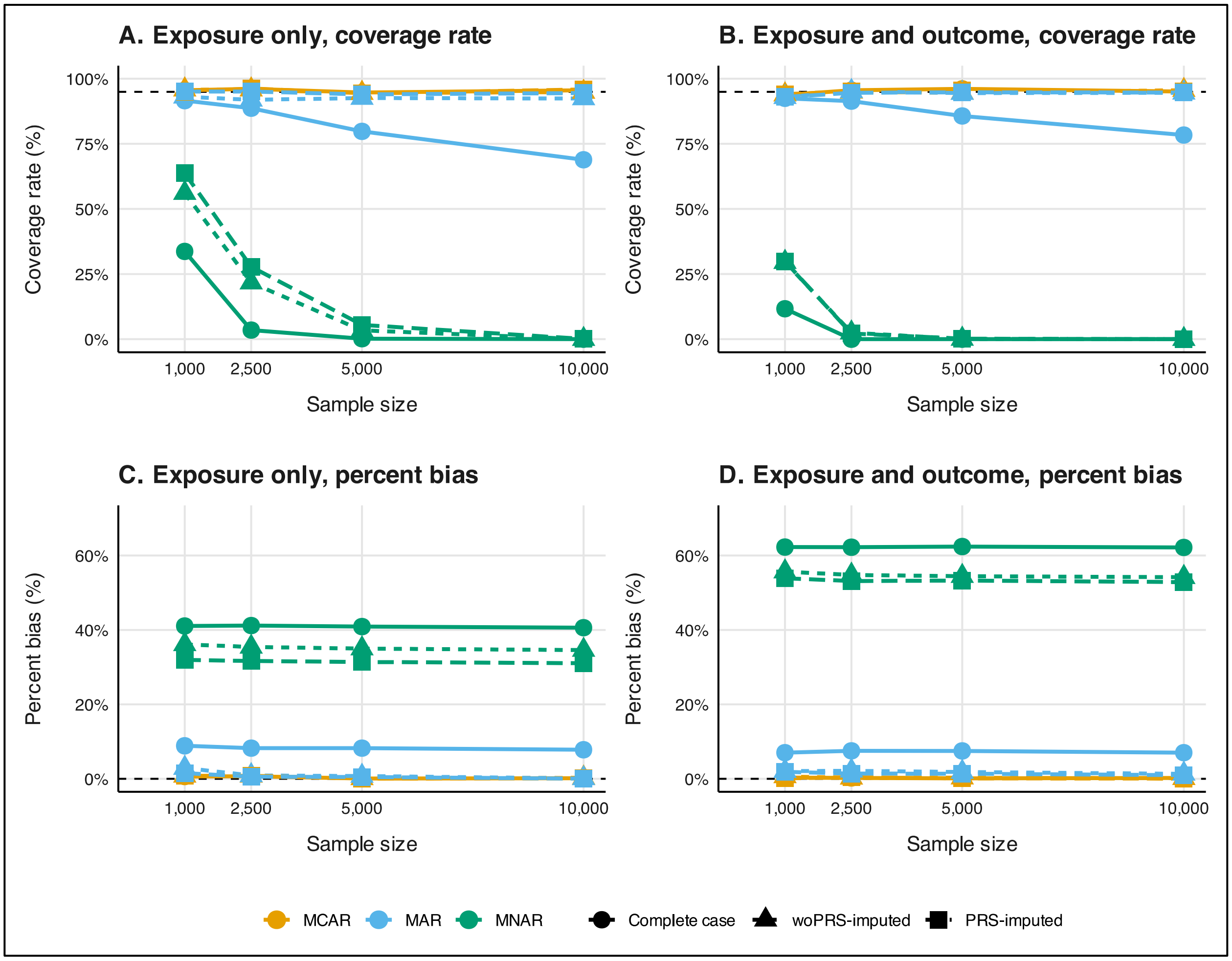


Figure 5.3 Coverage rate (panels A and B) and percent bias (panels C and D) diagnostics for exposure only (panels A and C) and exposure and outcome missingness (panels B and D) BMI coefficient for glucose by missing data mechanism and method and sample size under random sampling in a 1,000-iteration simulation. Analyses were adjusted for age, sex, non-Hispanic White, and smoking status (ever/never). Corresponding coverage rate, percent bias, average confidence interval width, and root mean squared error diagnostics are reported in Supplementary Table 5.2 and Supplementary Table 5.3. Abbreviations: MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS-imputed, polygenic risk score-informed multiple imputation; woPRS-imputed, multiple imputation without exposure and outcome PRS.

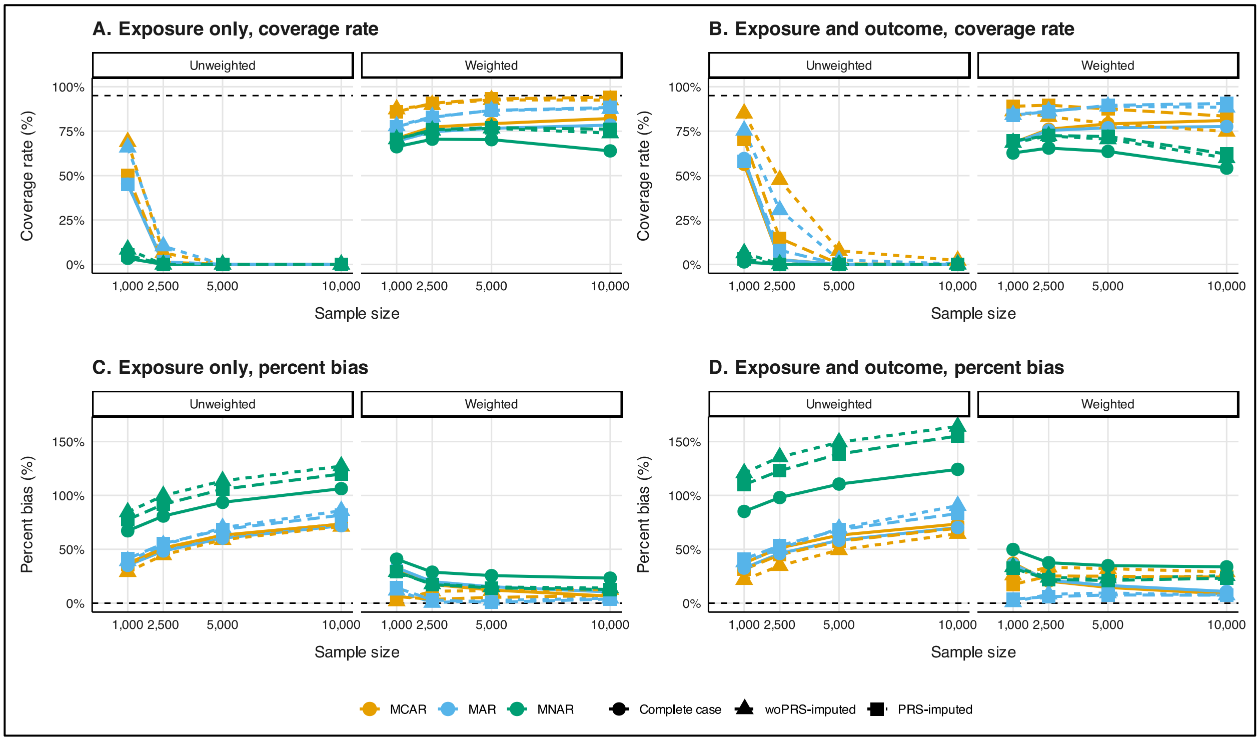


Figure 5.4 Coverage rate (panels A and B) and percent bias (panels C and D) diagnostics for unweighted (left) and weighted (right) BMI coefficient for glucose estimation by missing data mechanism and method and sample size under biased sampling and exposure only (panels A and C) and exposure and outcome missingness (panels B and D) in a 1,000-iteration simulation. For biased sampling simulations, unweighted and weighted diagnostics are reported in Supplementary Table 5.4, Supplementary Table 5.5, Supplementary Table 5.6, and Supplementary Table 5.7, respectively. Analyses were adjusted for age, sex, non-Hispanic White, and smoking status (ever/never).

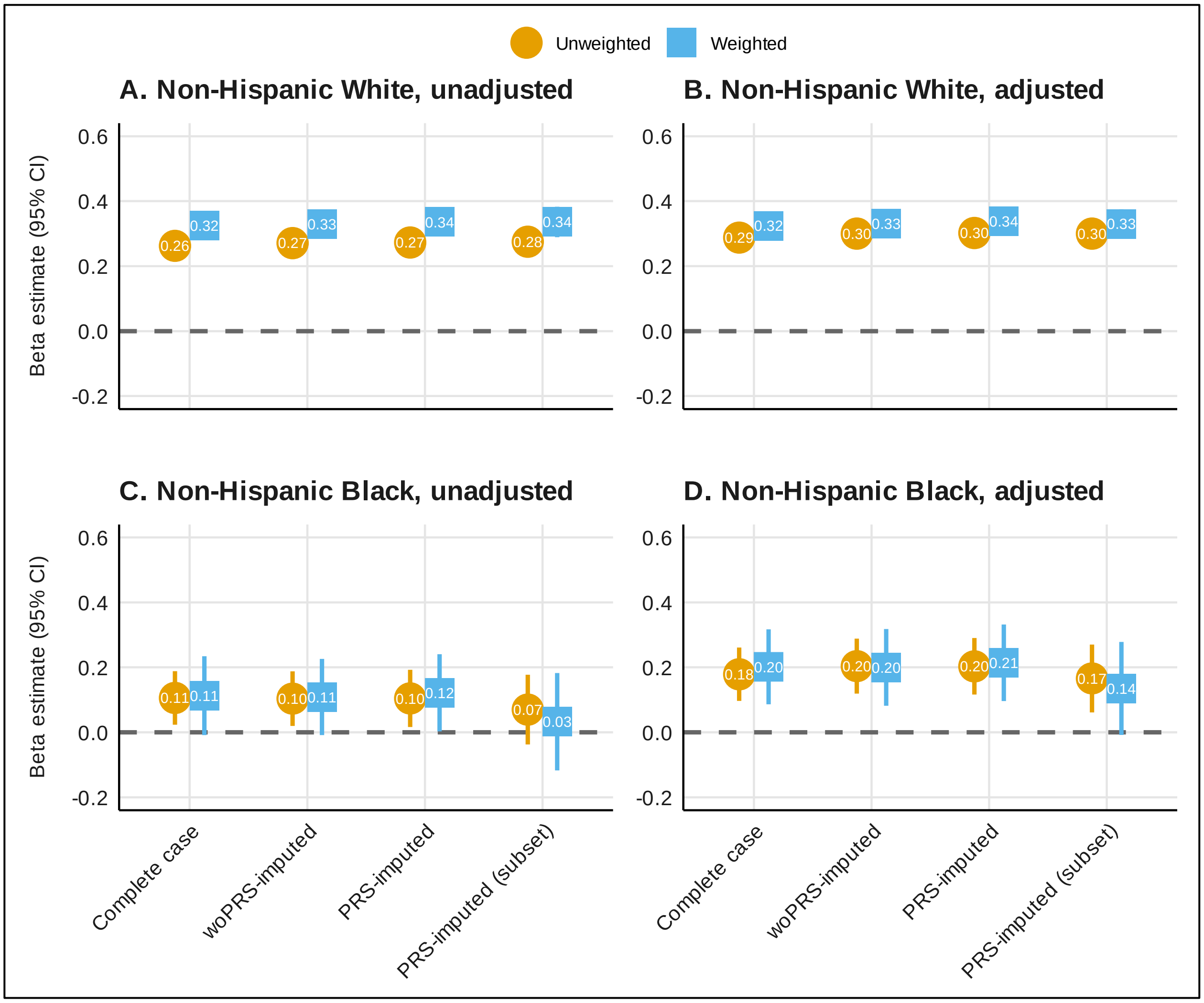
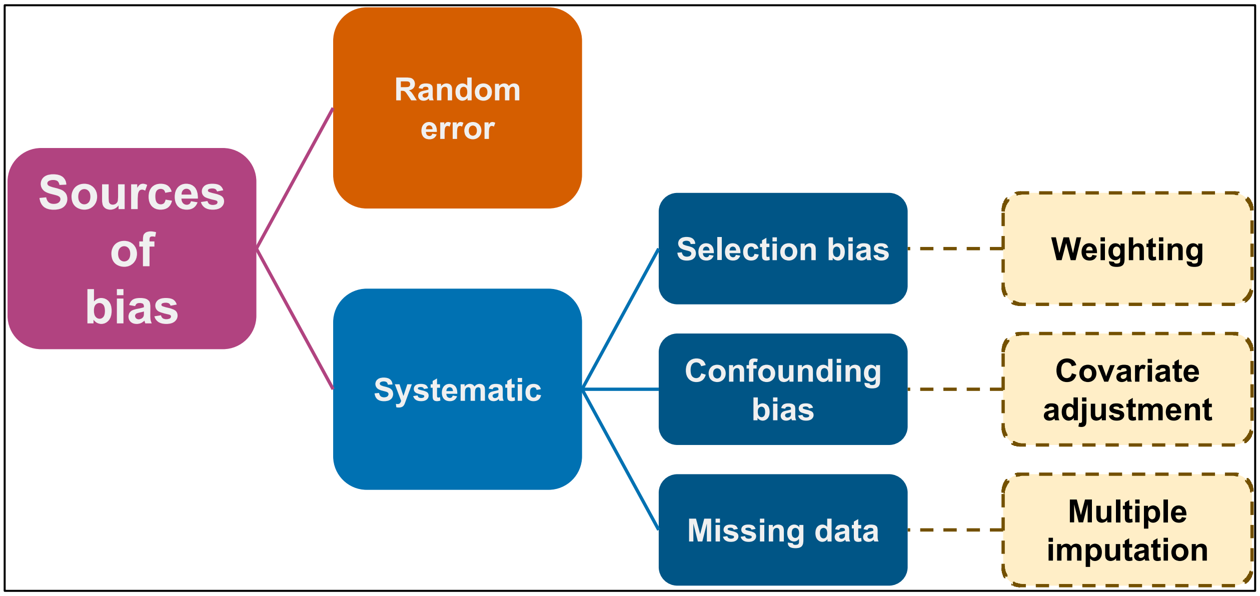
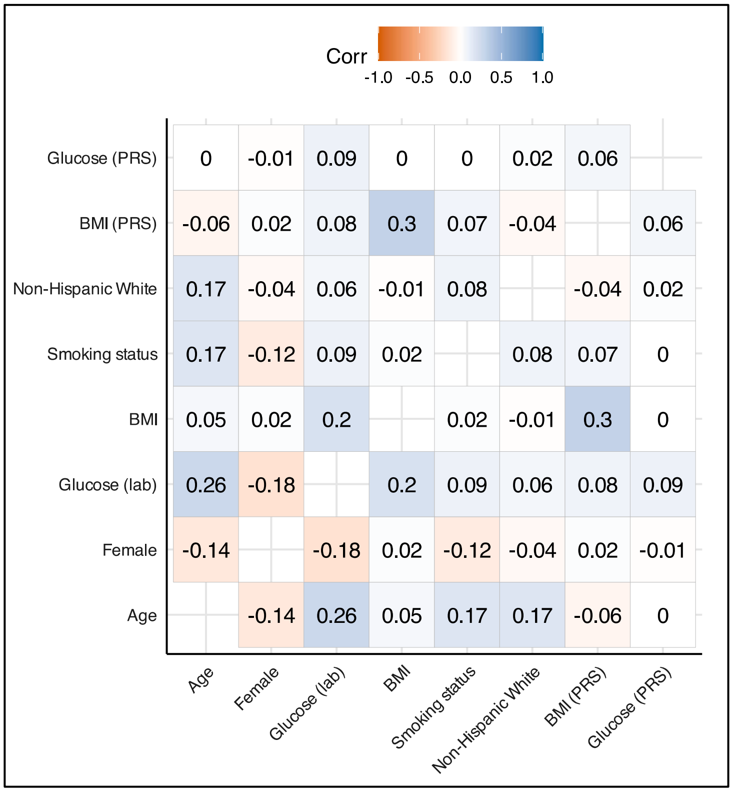


Figure 5.5 Estimation of the coefficient for BMI with glucose as the outcome by missing data method and weighting approach among non-Hispanic Whites (n=42,999; panels A and B) and non-Hispanic Blacks (n=2,297; panels C and D) in all MGI adults age 40 or older without diabetes. The PRS-imputed subset sample (n=30,942 for non-Hispanic Whites; n=1,437 for non-Hispanic Blacks) was restricted to individuals with non-missing genotype data before multiple imputation. Analyses were adjusted for age, sex, and smoking status (ever/never). Gray shaded regions represent corresponding 95% confidence interval from National Health Interview Survey-weighted All of Us data where weights are calculated separately for non-Hispanic Whites and non-Hispanic Blacks to make All of Us data for each of these groups more representative of their corresponding US population (target population). Results for the full, unstratified cohort are shown in Supplementary Figure 5.6. Abbreviations: PRS, polygenic risk score.

5.9 Supplementary materials



Supplementary Figure 5.1 Schematic representation of sources of bias. While missing data is not classically considered a source of systematic bias, we have here considered it a child of systematic error because of its ability to induce selection bias and misclassification. The yellow boxes with dashed outlines indicate the analytic methods applied in this manuscript to address each bias.



	Age	Female	Glucose (lab)	BMI	Smoking status	Non-Hispanic White	BMI (PRS)	Glucose (PRS)
Glucose (PRS)	0	-0.01	0.09	0	0	0.02	0.06	
BMI (PRS)	-0.06	0.02	0.08	0.3	0.07	-0.04		0.06
Non-Hispanic White	0.17	-0.04	0.06	-0.01	0.08		-0.04	0.02
Smoking status	0.17	-0.12	0.09	0.02		0.08	0.07	0
BMI	0.05	0.02	0.2		0.02	-0.01	0.3	0
Glucose (lab)	0.26	-0.18		0.2	0.09	0.06	0.08	0.09
Female	-0.14		-0.18	0.02	-0.12	-0.04	0.02	-0.01
Age		-0.14	0.26	0.05	0.17	0.17	-0.06	0

Supplementary Figure 5.2 Pairwise complete observation correlation matrix of relevant variables observed in MGI.

Supplementary Table 5.1 Intercept values for exposure only and exposure and outcome missingness generation models by missing data mechanism.

		Exposure and outcome		**Exposure only**
Mechanisms	**n**	Exposure	Outcome	Exposure
MAR	1,000	-5.53	-4.09	-5.58
	2,500	-5.36	-4.02	-5.36
	5,000	-5.15	-3.90	-5.15
	10,000	-4.88	-3.79	-4.88
	n	Exposure	Outcome	Exposure
MNAR	1,000	-7.36	-7.36	-7.41
	2,500	-7.05	-7.05	-7.04
	5,000	-6.74	-6.74	-6.72
	10,000	-6.32	-6.32	-6.31

Supplementary Table 5.2 Performance of missing data methods for estimating covariate-adjusted BMI coefficient for glucose by missing data mechanism, metric, and sample size in random sampling simulations with exposure only missingness.

			Sample size			
Mechanism	**Metric**	**Method**	**1,000**	**2,500**	**5,000**	**10,000**
MCAR	Percent bias	Complete case	0.613	0.668	0.054	0.157
		woPRS-imputed	0.773	0.806	0.095	0.254
		PRS-imputed	0.807	0.811	0.071	0.125
	Coverage rate	Complete case	0.956	0.961	0.947	0.956
		woPRS-imputed	0.958	0.958	0.945	0.949
		PRS-imputed	0.949	0.963	0.943	0.959
	Average width	Complete case	0.135	0.085	0.060	0.042
		woPRS-imputed	0.139	0.088	0.062	0.044
		PRS-imputed	0.138	0.087	0.061	0.043
	RMSE	Complete case	0.033	0.021	0.016	0.011
		woPRS-imputed	0.033	0.022	0.016	0.011
		PRS-imputed	0.033	0.022	0.016	0.011
MAR	Percent bias	Complete case	8.861	8.231	8.240	7.801
		woPRS-imputed	2.892	0.904	0.736	0.100
		PRS-imputed	1.538	0.555	0.465	0.041
	Coverage rate	Complete case	0.916	0.887	0.798	0.689
		woPRS-imputed	0.931	0.919	0.926	0.924
		PRS-imputed	0.951	0.951	0.940	0.945
	Average width	Complete case	0.129	0.081	0.058	0.041
		woPRS-imputed	0.147	0.093	0.065	0.045
		PRS-imputed	0.143	0.091	0.064	0.045
	RMSE	Complete case	0.038	0.026	0.022	0.019
		woPRS-imputed	0.040	0.025	0.017	0.012
		PRS-imputed	0.036	0.022	0.016	0.011
MNAR	Percent bias	Complete case	41.080	41.174	40.905	40.615
		woPRS-imputed	36.115	35.437	34.986	34.598
		PRS-imputed	31.946	31.670	31.391	31.074
	Coverage rate	Complete case	0.337	0.035	0.002	0.000
		woPRS-imputed	0.561	0.217	0.034	0.000
		PRS-imputed	0.637	0.278	0.055	0.001
	Average width	Complete case	0.136	0.086	0.061	0.043
		woPRS-imputed	0.159	0.101	0.071	0.050
		PRS-imputed	0.156	0.098	0.069	0.049
	RMSE	Complete case	0.089	0.084	0.083	0.081
		woPRS-imputed	0.082	0.074	0.072	0.070
		PRS-imputed	0.074	0.067	0.065	0.063

Abbreviations: BMI, body mass index; CC, complete case; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS, polygenic risk score; RMSE, root mean square error; woPRS, without polygenic risk score

Supplementary Table 5.3 Performance of missing data methods for estimating covariate-adjusted BMI coefficient for glucose by missing data mechanism, metric, and sample size in random sampling simulations with exposure and outcome missingness.

			Sample size			
Mechanism	**Metric**	**Method**	**1,000**	**2,500**	**5,000**	**10,000**
MCAR	Percent bias	Complete case	0.320	0.234	0.125	0.217
		woPRS-imputed	0.628	0.256	0.286	0.151
		PRS-imputed	0.109	0.429	0.120	0.043
	Coverage rate	Complete case	0.938	0.956	0.961	0.951
		woPRS-imputed	0.939	0.950	0.946	0.956
		PRS-imputed	0.941	0.951	0.956	0.951
	Average width	Complete case	0.156	0.098	0.069	0.049
		woPRS-imputed	0.170	0.107	0.076	0.054
		PRS-imputed	0.167	0.107	0.074	0.053
	RMSE	Complete case	0.042	0.025	0.017	0.012
		woPRS-imputed	0.043	0.026	0.018	0.013
		PRS-imputed	0.042	0.026	0.018	0.012
MAR	Percent bias	Complete case	7.021	7.506	7.489	7.009
		woPRS-imputed	2.061	2.140	1.828	1.319
		PRS-imputed	1.877	1.395	1.364	0.968
	Coverage rate	Complete case	0.925	0.914	0.857	0.784
		woPRS-imputed	0.929	0.952	0.945	0.946
		PRS-imputed	0.930	0.946	0.948	0.947
	Average width	Complete case	0.145	0.091	0.064	0.046
		woPRS-imputed	0.180	0.113	0.078	0.056
		PRS-imputed	0.166	0.104	0.075	0.053
	RMSE	Complete case	0.041	0.028	0.022	0.018
		woPRS-imputed	0.044	0.027	0.019	0.013
		PRS-imputed	0.042	0.026	0.018	0.013
MNAR	Percent bias	Complete case	62.273	62.237	62.420	62.177
		woPRS-imputed	55.679	54.781	54.450	54.180
		PRS-imputed	53.889	53.148	53.282	52.867
	Coverage rate	Complete case	0.117	0.000	0.000	0.000
		woPRS-imputed	0.296	0.025	0.001	0.000
		PRS-imputed	0.299	0.022	0.002	0.000
	Average width	Complete case	0.149	0.094	0.066	0.047
		woPRS-imputed	0.172	0.109	0.076	0.053
		PRS-imputed	0.167	0.107	0.074	0.052
	RMSE	Complete case	0.130	0.126	0.125	0.124
		woPRS-imputed	0.118	0.112	0.109	0.108
		PRS-imputed	0.115	0.108	0.107	0.106

Abbreviations: BMI, body mass index; CC, complete case; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS, polygenic risk score; RMSE, root mean square error; woPRS, without polygenic risk score

Supplementary Table 5.4 Biased/covariate-informed sampling simulation performance of missing data methods for estimating unweighted and covariate-adjusted BMI coefficient for glucose by missing data mechanism, metric, and sample size with exposure only missingness.

			Sample size			
Mechanism	**Metric**	**Method**	**1,000**	**2,500**	**5,000**	**10,000**
MCAR	Percent bias	Complete case	36.873	51.226	63.133	73.374
		woPRS-imputed	29.057	44.652	58.862	71.132
		PRS-imputed	35.796	50.593	62.758	73.435
	Coverage rate	Complete case	0.456	0.005	0.000	0.000
		woPRS-imputed	0.690	0.065	0.001	0.000
		PRS-imputed	0.500	0.013	0.000	0.000
	Average width	Complete case	0.137	0.087	0.062	0.043
		woPRS-imputed	0.155	0.099	0.071	0.050
		PRS-imputed	0.142	0.090	0.064	0.046
	RMSE	Complete case	0.081	0.104	0.126	0.146
		woPRS-imputed	0.070	0.092	0.118	0.142
		PRS-imputed	0.080	0.103	0.125	0.146
MAR	Percent bias	Complete case	34.998	48.445	60.272	71.844
		woPRS-imputed	39.082	54.004	69.995	85.814
		PRS-imputed	41.165	55.207	68.433	81.637
	Coverage rate	Complete case	0.452	0.005	0.000	0.000
		woPRS-imputed	0.659	0.103	0.000	0.000
		PRS-imputed	0.450	0.014	0.001	0.000
	Average width	Complete case	0.132	0.083	0.059	0.042
		woPRS-imputed	0.186	0.121	0.083	0.058
		PRS-imputed	0.154	0.097	0.070	0.050
	RMSE	Complete case	0.077	0.098	0.120	0.143
		woPRS-imputed	0.092	0.113	0.142	0.171
		PRS-imputed	0.091	0.113	0.137	0.163
MNAR	Percent bias	Complete case	67.098	80.842	93.566	106.255
		woPRS-imputed	84.533	99.765	113.517	127.188
		PRS-imputed	77.451	91.727	105.719	119.905
	Coverage rate	Complete case	0.034	0.000	0.000	0.000
		woPRS-imputed	0.085	0.000	0.000	0.000
		PRS-imputed	0.052	0.000	0.000	0.000
	Average width	Complete case	0.138	0.087	0.062	0.044
		woPRS-imputed	0.177	0.109	0.075	0.052
		PRS-imputed	0.160	0.101	0.072	0.050
	RMSE	Complete case	0.138	0.162	0.186	0.211
		woPRS-imputed	0.175	0.200	0.226	0.253
		PRS-imputed	0.160	0.184	0.211	0.238

Abbreviations: BMI, body mass index; CC, complete case; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS, polygenic risk score; RMSE, root mean square error; woPRS, without polygenic risk score

Supplementary Table 5.5 Biased/covariate-informed sampling simulation performance of missing data methods for estimating weighted and covariate-adjusted BMI coefficient for glucose by missing data mechanism, metric, and sample size with exposure only missingness.

			Sample size			
Mechanism	**Metric**	**Method**	**1,000**	**2,500**	**5,000**	**10,000**
MCAR	Percent bias	Complete case	32.506	17.254	12.133	6.518
		woPRS-imputed	1.842	11.046	11.583	12.559
		PRS-imputed	6.144	3.518	5.327	7.055
	Coverage rate	Complete case	0.709	0.773	0.792	0.821
		woPRS-imputed	0.874	0.899	0.926	0.925
		PRS-imputed	0.859	0.907	0.931	0.941
	Average width	Complete case	0.356	0.286	0.229	0.177
		woPRS-imputed	0.346	0.266	0.208	0.163
		PRS-imputed	0.349	0.270	0.213	0.165
	RMSE	Complete case	0.157	0.108	0.084	0.061
		woPRS-imputed	0.112	0.084	0.065	0.052
		PRS-imputed	0.114	0.081	0.063	0.048
MAR	Percent bias	Complete case	32.936	19.958	15.178	10.345
		woPRS-imputed	11.881	0.514	2.535	4.511
		PRS-imputed	14.274	2.992	0.715	3.800
	Coverage rate	Complete case	0.693	0.747	0.766	0.784
		woPRS-imputed	0.779	0.831	0.866	0.877
		PRS-imputed	0.774	0.828	0.867	0.883
	Average width	Complete case	0.343	0.274	0.219	0.173
		woPRS-imputed	0.327	0.258	0.205	0.160
		PRS-imputed	0.328	0.259	0.206	0.161
	RMSE	Complete case	0.151	0.107	0.082	0.063
		woPRS-imputed	0.129	0.093	0.071	0.056
		PRS-imputed	0.130	0.093	0.070	0.055
MNAR	Percent bias	Complete case	40.836	28.779	25.531	23.222
		woPRS-imputed	29.135	17.697	14.954	14.002
		PRS-imputed	29.067	17.204	14.151	12.657
	Coverage rate	Complete case	0.662	0.705	0.702	0.639
		woPRS-imputed	0.705	0.754	0.766	0.739
		PRS-imputed	0.705	0.758	0.770	0.761
	Average width	Complete case	0.347	0.278	0.224	0.179
		woPRS-imputed	0.337	0.269	0.215	0.170
		PRS-imputed	0.336	0.268	0.214	0.170
	RMSE	Complete case	0.160	0.117	0.093	0.077
		woPRS-imputed	0.145	0.103	0.080	0.065
		PRS-imputed	0.145	0.103	0.079	0.063

Abbreviations: BMI, body mass index; CC, complete case; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS, polygenic risk score; RMSE, root mean square error; woPRS, without polygenic risk score

Supplementary Table 5.6 Biased/covariate-informed sampling simulation performance of missing data methods for estimating unweighted and covariate-adjusted BMI coefficient for glucose by missing data mechanism, metric, and sample size with exposure and outcome missingness.

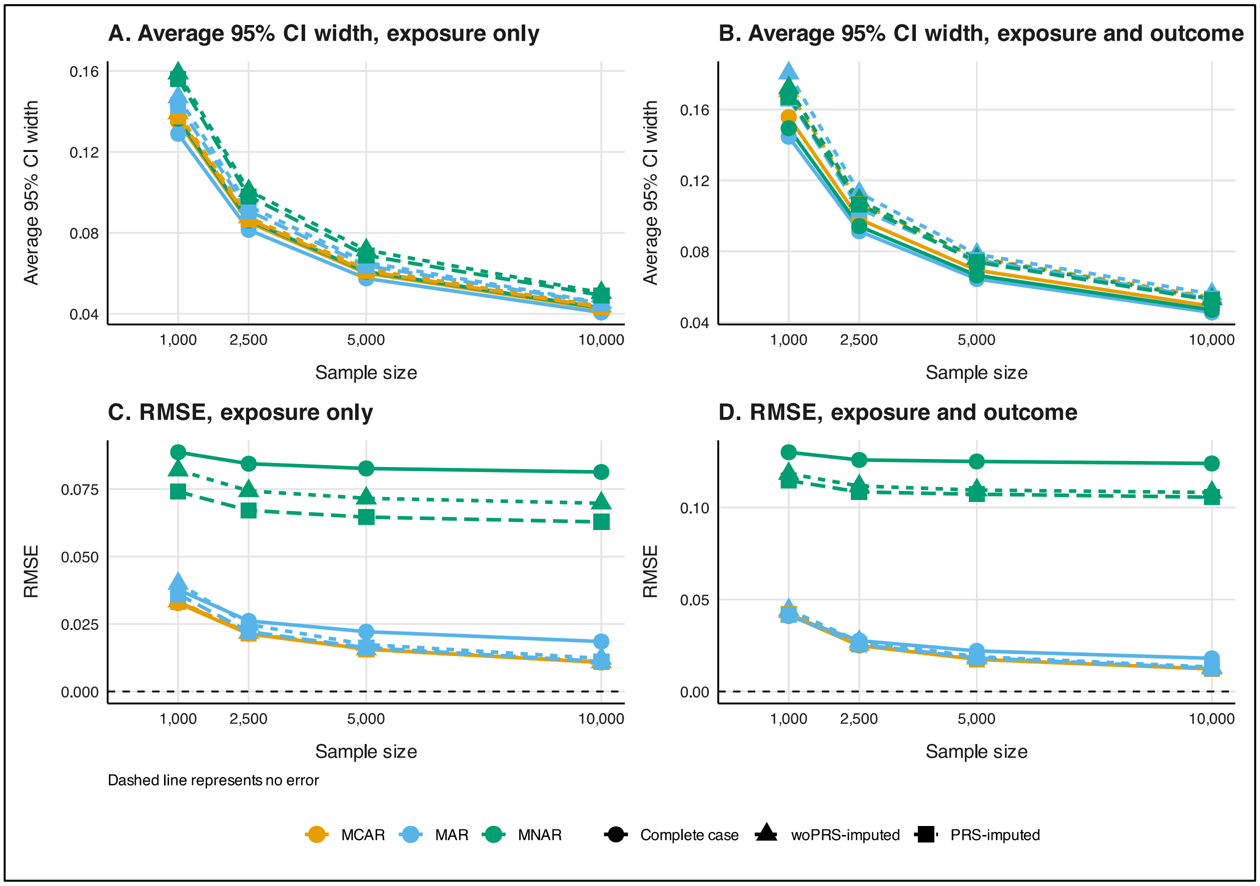
			Sample size			
Mechanism	**Metric**	**Method**	**1,000**	**2,500**	**5,000**	**10,000**
MCAR	Percent bias	Complete case	37.159	51.197	63.088	73.468
		woPRS-imputed	21.696	34.714	49.350	64.485
		PRS-imputed	31.311	44.849	57.490	69.527
	Coverage rate	Complete case	0.563	0.027	0.000	0.000
		woPRS-imputed	0.850	0.476	0.077	0.022
		PRS-imputed	0.703	0.147	0.003	0.000
	Average width	Complete case	0.159	0.100	0.071	0.050
		woPRS-imputed	0.196	0.132	0.107	0.092
		PRS-imputed	0.180	0.114	0.084	0.061
	RMSE	Complete case	0.084	0.105	0.126	0.146
		woPRS-imputed	0.068	0.077	0.103	0.132
		PRS-imputed	0.078	0.094	0.116	0.139
MAR	Percent bias	Complete case	32.508	46.041	58.330	70.157
		woPRS-imputed	38.110	50.608	69.346	90.519
		PRS-imputed	40.978	53.365	68.150	83.145
	Coverage rate	Complete case	0.596	0.026	0.000	0.000
		woPRS-imputed	0.754	0.306	0.025	0.003
		PRS-imputed	0.579	0.080	0.000	0.000
	Average width	Complete case	0.148	0.093	0.066	0.047
		woPRS-imputed	0.223	0.149	0.114	0.083
		PRS-imputed	0.185	0.117	0.087	0.062
	RMSE	Complete case	0.075	0.094	0.117	0.140
		woPRS-imputed	0.102	0.111	0.145	0.184
		PRS-imputed	0.098	0.111	0.138	0.167
MNAR	Percent bias	Complete case	85.115	98.036	110.564	124.209
		woPRS-imputed	120.797	135.571	149.429	164.005
		PRS-imputed	109.921	122.884	138.468	155.203
	Coverage rate	Complete case	0.014	0.000	0.000	0.000
		woPRS-imputed	0.064	0.002	0.000	0.000
		PRS-imputed	0.032	0.000	0.000	0.000
	Average width	Complete case	0.152	0.096	0.068	0.048
		woPRS-imputed	0.217	0.141	0.094	0.063
		PRS-imputed	0.195	0.122	0.086	0.060
	RMSE	Complete case	0.173	0.196	0.220	0.247
		woPRS-imputed	0.249	0.272	0.298	0.326
		PRS-imputed	0.227	0.247	0.277	0.309

Abbreviations: BMI, body mass index; CC, complete case; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS, polygenic risk score; RMSE, root mean square error; woPRS, without polygenic risk score

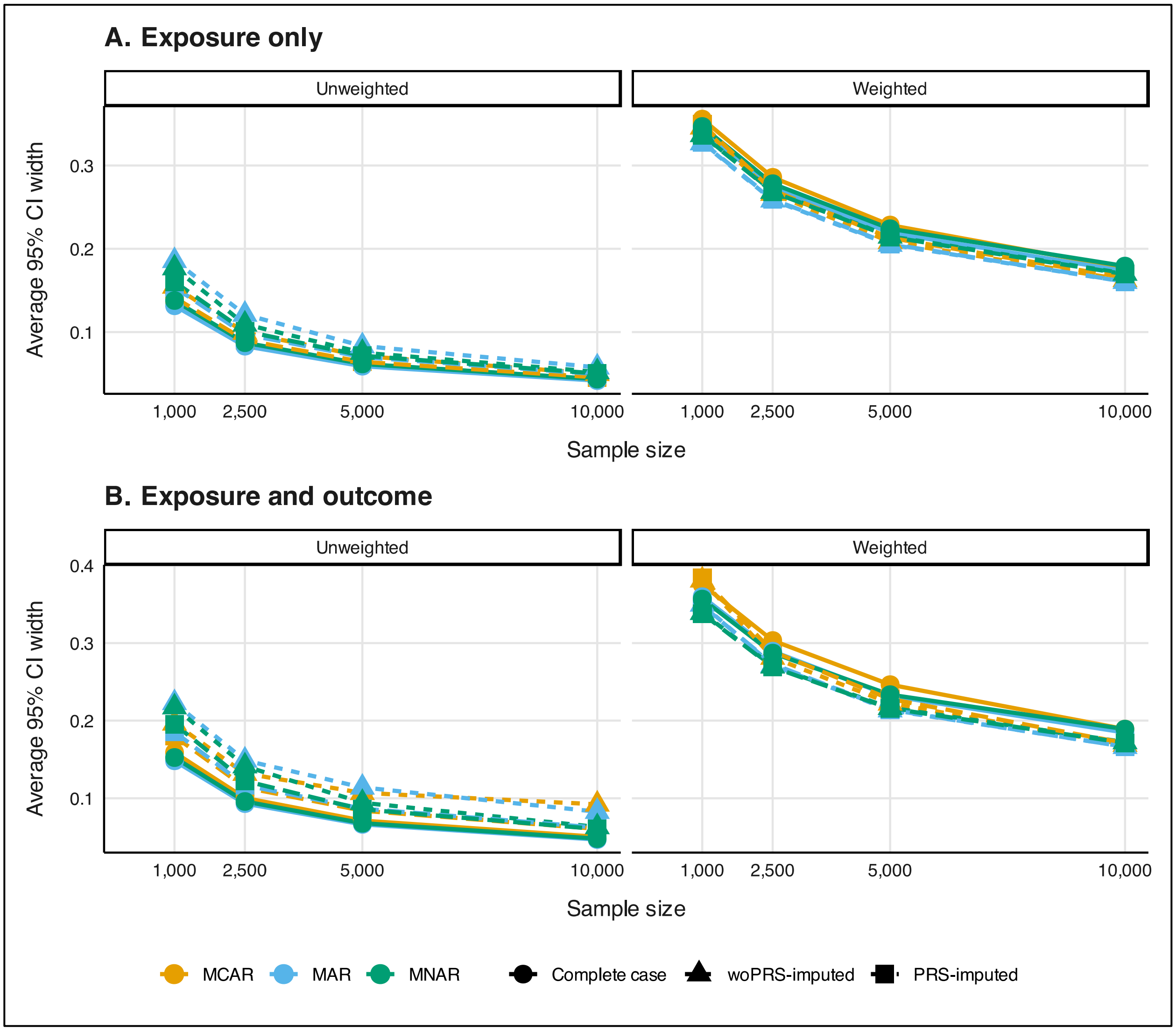
Supplementary Table 5.7 Biased/covariate-informed sampling simulation performance of missing data methods for estimating weighted and covariate-adjusted BMI coefficient for glucose by missing data mechanism, metric, and sample size with exposure and outcome missingness.

			Sample size			
Mechanism	**Metric**	**Method**	**1,000**	**2,500**	**5,000**	**10,000**
MCAR	Percent bias	Complete case	37.174	20.417	14.551	8.517
		woPRS-imputed	25.938	33.332	31.954	28.911
		PRS-imputed	17.280	25.151	24.982	24.115
	Coverage rate	Complete case	0.684	0.761	0.791	0.810
		woPRS-imputed	0.863	0.831	0.793	0.748
		PRS-imputed	0.890	0.896	0.875	0.834
	Average width	Complete case	0.375	0.303	0.246	0.189
		woPRS-imputed	0.381	0.281	0.222	0.169
		PRS-imputed	0.384	0.289	0.228	0.171
	RMSE	Complete case	0.172	0.119	0.090	0.066
		woPRS-imputed	0.123	0.104	0.087	0.073
		PRS-imputed	0.115	0.093	0.078	0.065
MAR	Percent bias	Complete case	35.885	21.217	16.427	10.861
		woPRS-imputed	1.039	8.231	9.369	7.404
		PRS-imputed	3.671	5.991	7.541	7.380
	Coverage rate	Complete case	0.682	0.754	0.769	0.776
		woPRS-imputed	0.842	0.863	0.890	0.884
		PRS-imputed	0.840	0.860	0.895	0.906
	Average width	Complete case	0.360	0.289	0.232	0.185
		woPRS-imputed	0.350	0.271	0.213	0.166
		PRS-imputed	0.348	0.273	0.214	0.166
	RMSE	Complete case	0.160	0.114	0.087	0.068
		woPRS-imputed	0.125	0.092	0.072	0.058
		PRS-imputed	0.124	0.092	0.069	0.055
MNAR	Percent bias	Complete case	49.811	37.496	34.837	33.700
		woPRS-imputed	33.674	23.788	23.089	25.463
		PRS-imputed	32.404	21.942	21.067	23.194
	Coverage rate	Complete case	0.627	0.655	0.636	0.542
		woPRS-imputed	0.685	0.723	0.705	0.599
		PRS-imputed	0.695	0.725	0.720	0.621
	Average width	Complete case	0.357	0.287	0.233	0.189
		woPRS-imputed	0.339	0.270	0.217	0.172
		PRS-imputed	0.338	0.269	0.216	0.171
	RMSE	Complete case	0.174	0.129	0.108	0.094
		woPRS-imputed	0.151	0.110	0.089	0.079
		PRS-imputed	0.149	0.108	0.087	0.076

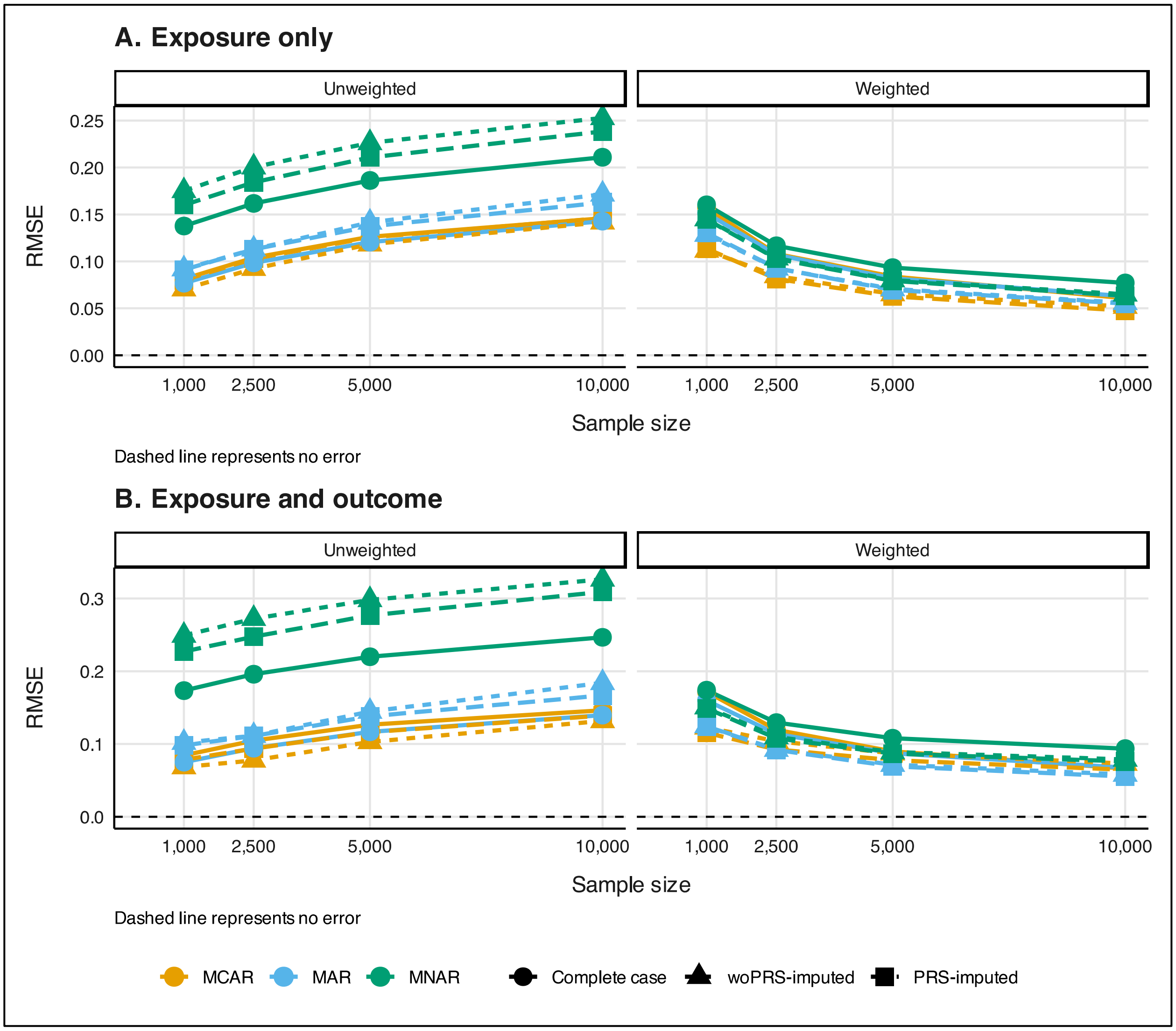
Abbreviations: BMI, body mass index; CC, complete case; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS, polygenic risk score; RMSE, root mean square error; woPRS, without polygenic risk score



Supplementary Figure 5.3 Average 95% CI width (panels A and B) and RMSE (panels C and D) diagnostics for BMI coefficient for glucose by missing data mechanism and method and sample size under random sampling with exposure only (panels A and C) and exposure and outcome missingness (panels B and D). Analyses were adjusted for age, sex, non-Hispanic White, and smoking status (ever/never). Corresponding coverage rate, percent bias, average confidence interval width, and root mean squared error diagnostics are reported in Supplementary Table 5.2 and Supplementary Table 5.3. Abbreviations: MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS, polygenic risk score; RMSE, root mean square error; woPRS, without polygenic risk score.



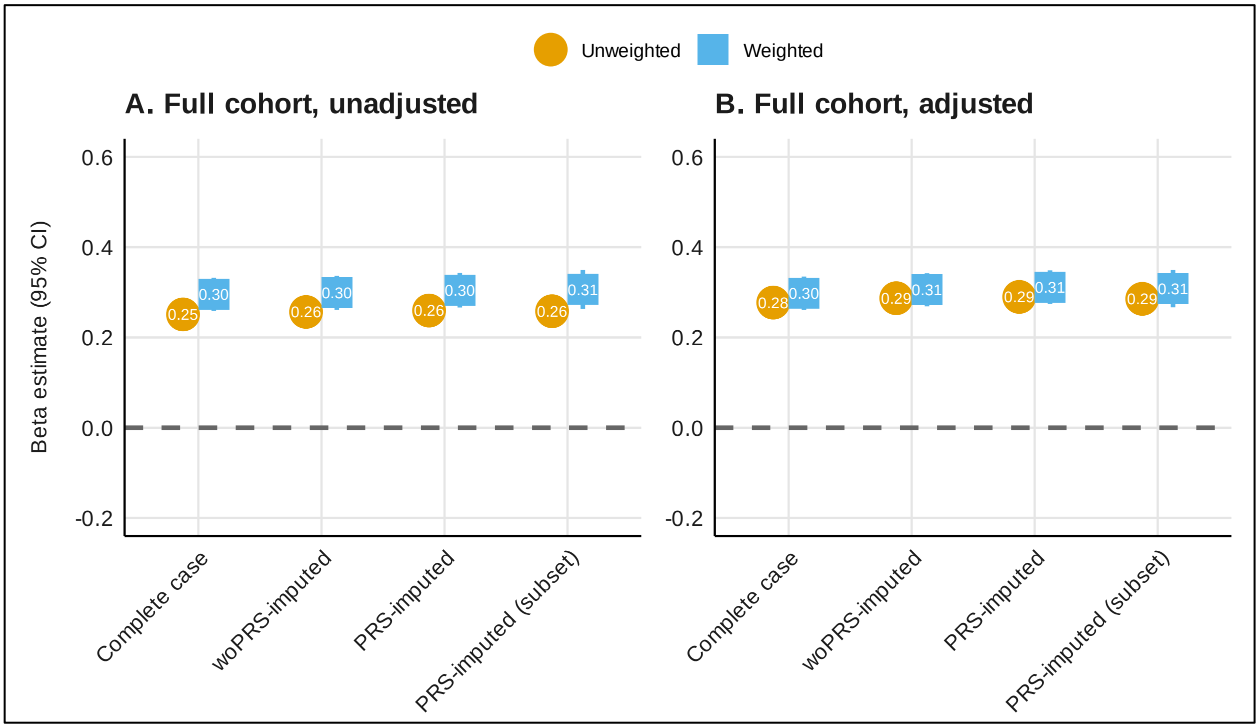
Supplementary Figure 5.4 Average 95% CI width for unweighted (left) and weighted (right) BMI coefficient for glucose by missing data mechanism and method and sample size under biased sampling with exposure only (panel A) and exposure and outcome missingness (panel B). For biased sampling simulations, unweighted and weighted diagnostics are reported in Supplementary Table 5.4, Supplementary Table 5.5, Supplementary Table 5.6, and Supplementary Table 5.7. Analyses were adjusted for age, sex, non-Hispanic White, and smoking status (ever/never). Abbreviations: MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS, polygenic risk score; RMSE, root mean square error; woPRS, without polygenic risk score.



Supplementary Figure 5.5 RMSE for unweighted (left) and weighted (right) BMI coefficient for glucose by missing data mechanism and method and sample size under biased sampling with exposure only (panel A) and exposure and outcome missingness (panel B). For biased sampling simulations, unweighted and weighted diagnostics are reported in Supplementary Table 5.4, Supplementary Table 5.5, Supplementary Table 5.6, and Supplementary Table 5.7. Analyses were adjusted for age, sex, non-Hispanic White, and smoking status (ever/never). Abbreviations: MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS, polygenic risk score; woPRS, without polygenic risk score.

Supplementary Table 5.8 Comparison of demographic, health measurements, and polygenic risk score values overall and among adults 40 or older without diabetes, with and without any missing values in the Michigan Genomics Initiative.

Characteristic	**Overall** N = 50,026[a]	**Incomplete observations** N = 20,799[a]	**Complete observations** N = 29,227[a]	**p-value[b]**	**Non-missing PRS** N = 35,353[a]
Age	62.9 (12.5)	61.9 (12.4)	63.5 (12.5)	<0.001	63.1 (12.6)
Female	54.5 (27,261)	55.1 (11,465)	54.0 (15,796)	0.017	54.0 (19,100)
Non-Hispanic White	86.0 (42,999)	84.0 (17,479)	87.3 (25,520)	<0.001	87.5 (30,942)
Smoking status (ever)	48.1 (20,561)	44.8 (6,071)	49.6 (14,490)	<0.001	49.1 (16,022)
Missing	*7,238*	*7,238*	*0*		*2691*
BMI	29.1 (6.0)	29.1 (6.0)	29.0 (6.0)	0.7	29.0 (6.0)
Missing	*292*	*292*	*0*		*156*
Glucose	99.0 (14.1)	98.8 (14.6)	99.1 (13.9)	<0.001	99.3 (14.2)
Missing	*5,467*	*5,467*	*0*		*3991*
BMI PRS[c]	0.000 (1.000)	0.008 (1.023)	-0.002 (0.995)	0.3	0.000 (1.000)
Missing	*14,673*	*14,673*	*0*		*0*
Glucose PRS[c]	0.000 (1.000)	0.013 (0.996)	-0.003 (1.001)	0.2	0.000 (1.000)
Missing	*14,673*	*14,673*	*0*		*0*



Supplementary Figure 5.6 Estimation of the coefficient for BMI with glucose as the outcome by missing data method and weighting approach among all MGI adults age 40 or older without diabetes (i.e., not stratified by race/ethnicity; n=50,026). The PRS-imputed subset sample (n=35,353) was restricted to individuals with non-missing genotype data before multiple imputation. Analyses were adjusted for age, sex, smoking status (ever/never), and a non-Hispanic White indicator. Gray shaded regions represent corresponding 95% confidence interval from National Health Interview Survey-weighted All of Us data to make All of Us data more representative of the US population (target population). Abbreviations: PRS, polygenic risk score.

Supplementary Table 5.9 Potential bias in intercept, exposure, and confounder regression coefficients in complete case analysis of linear and logistic regression by reason for missing data.

	Linear regression coefficient			Logistic regression coefficient		
Variables missingness is dependent upon	Intercept	Exposure	Confounder	Intercept	Exposure	Confounder
None (e.g., missing completely at random)	Unbiased	Unbiased	Unbiased	Unbiased	Unbiased	Unbiased
Outcome (Y) only	Biased	Biased[a]	Biased[a]	Biased	Unbiased	Unbiased
Exposure (X) and/or other covariates (C)	Unbiased	Unbiased	Unbiased	Unbiased	Unbiased	Unbiased
Outcome (Y) and confounders (C)	Biased	Biased	Biased	Biased	Unbiased	Biased
Outcome (Y), exposure (X), and possible confounders (C)	Biased	Biased	Biased	Biased	Biased[b]	Biased

[a] Biased in general, except when in truth there is no association between the outcome and the exposure or confounding in question (i.e., the true value of the regression coefficient is zero)

[b] Biased in general, except when missingness depends on the outcome and exposure independently

Adapted from Supplementary Table 1 from Hughes and colleagues (2019, doi: 10.1093/ije/dyz032) and Table 1 from Bartlett and colleagues (2015; doi: 10.1093/aje/kwv114)

Chapter 6 Conclusion

6.1 Summary of the dissertation

6.1.1 Aim 1: To weight or not to weight? The effect of selection bias in three large EHR-linked biobanks and recommendations for practice

EHR-linked biobanks can often be subject to different forms of selection bias such as healthy volunteer bias (as in UKB[4]), have recruitment strategies such as oversampling groups historically underrepresented in biomedical research (as in AOU[2]) or recruiting patients at pre-/peri-operative appointments requiring anesthesia (as in MGI[45]). Such non-probability sampling makes them unrepresentative of presumptive target populations. This lack of representativeness threatens the generalizability of results from EHR-linked biobank analyses. Weighting-based approaches like IP and PS weights, which rely on estimating the probability of someone in the analytic sample representing someone in the target population, are commonly used to address selection bias. Van Alten and colleagues estimated "selection weights" in the UKB.[8] However, determining what factors go into weighting models and how vital weighting is in different types of EHR-linked biobank analyses are open questions.

Aim 1 of this dissertation sought to make recommendations on whether and when to use weights to reduce selection bias in three EHR-linked biobanks with different recruitment strategies: AOU, MGI, and UKB. We estimated IPW and PS weights in AOU (n=244,071) and MGI (n=81,243) using the 2019 National Health Interview Survey to make the cohorts more similar to the US adult population. Using these weights, alongside those previously calculated in UKB (n=401,167),[8] we compared the impact of weighting on four common analyses: prevalence, dimensionality, and association estimation and large-scale hypothesis testing.

Estimated phecode prevalences decreased in AOU (weighted-to-unweighted median prevalence ratio [MPR]: 0.82) and MGI (0.61) and increased in UKB (1.06). Prevalence ratios less than 1 indicate that phecodes are overrepresented in the sample compared to the target population. Weighting minimally impacted latent phenome dimensionality estimation, which has implications for determining the number of independent tests used in Bonferroni multiple testing corrected p-values. Weighting impacted targeted association estimation, aligning coefficient estimates more closely with national registry-based estimates in MGI. Weighted analyses of AOU data could not recover benchmark estimates, likely due to significant racial/ethnic heterogeneity, highlighting the need for expert knowledge and data exploration during the analytic process. Large-scale hypothesis testing, captured by a phenome-wide association study (PheWAS), demonstrated considerable overlap of significant hits between weighted and unweighted PheWAS.

Our findings show that researchers should use weight analyses to reduce bias in prevalence and association estimation. Weights should be curated when conducting association estimation. On the other hand, weighting is less crucial for dimensionality estimation and large-scale hypothesis testing, where specific signals should be followed up by weighted analysis when effect size estimation is of interest. EHR-linked biobanks should report selection mechanisms and make selection weights available for researchers, who should carefully consider their analytic goals and target populations and weight analyses accordingly. This manuscript has been published in JAMIA.[215]

6.1.2 Aim 2: The impact of sample-weighting on risk prediction and risk stratification properties of prediction models trained in one EHR-linked biobank when applied to another biobank with a different recruitment strategy: A case study in the United States

Risk prediction models are classical tools in public health and precision medicine. EHR-linked biobanks are multi-modal data sources that link EHR and genetic information with other linkable data like cancer and vital status registries, neighborhood-level environmental exposures, and complementary survey data, presenting an ideal environment for more holistic risk prediction and stratification. Salvatore and colleagues developed a framework for summarizing diagnosis history into a single-number risk score called a phenotype risk score (PheRS).[18] However, should weights be considered in risk prediction/stratification models (e.g., PheRS) when the external cohort has a different sampling strategy than the internal training sample? Because EHR-linked biobanks have sampling mechanisms like recruiting patients awaiting surgery (MGI)[45] and oversampling groups historically underrepresented in biomedical research (AOU),[2] they are not representative of presumptive target populations. When there are differences in data distributions between the sample used to develop the model and the sample to which the model is applied, a reduction in prediction performance called lack of transferability can occur. Iparragirre and colleagues developed a framework for tuning model hyperparameters in weighted settings,[17] allowing us to explore weighting-based methods for addressing transferability.

In Aim 2, we developed PheRS for esophageal, liver, and pancreatic cancers in MGI (n=76,757), a cohort enriched for cancer, and evaluated their performance in the external AOU cohort (n=226,764). Our goals were to (a) compare different PheRS construction approaches, (b) determine whether using weights improved PheRS performance, and (c) contrast PheRS performance with other data domains. First, we estimated poststratification (PS) weights to make MGI more like AOU (PS_{BASIC} accounting for age, sex, race/ethnicity; PS_{FULL} additionally accounting for smoking, alcohol consumption, BMI, depression, hypertension, and the Charlson Comorbidity Index). Then, adapting the framework from Iparragirre and colleagues,[17] we tuned model hyperparameters (λ and α for regularized regression; the number of randomly sampled features as each split and minimum number of observations in a leaf node in the random forest) using these weights. Next, we developed weighted and unweighted one- (lasso, ridge, and elastic net regression and random forest) and two-step (analogous to pruning-and-thresholding) PheRS for esophageal, liver, and pancreatic cancer outcomes, restricting data to 0, 1, 2, and 5 years prior to the cancer diagnosis. Finally, we assessed PheRS performance in terms of risk stratification, discriminatory ability, accuracy, and calibration alongside and in combination with other domains: basic sociodemographic covariates, risk factors, and presenting symptoms.

We found that no single PheRS construction approach uniformly performed better in terms of risk stratification or discrimination, though elastic net and random forest tended to exhibit good properties. We also observed that using weights in model development did not consistently or meaningfully improve PheRS risk stratification performance. Health history summarized as PheRS appeared to be the most important domain in risk stratification compared to the other three domains.

Our findings show that PheRS contributes to risk stratification and discriminatory ability alongside demographic covariates, risk factors, and a presenting symptom. PheRS should be considered when developing risk prediction models using EHR-linked biobank data. We also found that weighting-based approaches to PheRS construction do not improve model performance in an external cohort. Other methods to address transferability may be more worthwhile (e.g., transfer learning). In addition, we expanded on Iparragirre and colleagues'[17] framework to perform hyperparameter tuning for regularized regression and random forest in weighted settings. We further contributed to the adoption of these methods by providing R code.

6.1.3 Aim 3: Impact of polygenic risk score-informed multiple imputation and sample weighting for handling missing data and selection bias on association estimation in EHR-linked biobanks

EHR-linked biobanks are subject to multiple cooccurring biases, including those due to missing data and selection bias. Regarding missing data, Bell and colleagues found that 95% of RCTs published in top-tier journals in late 2013 reported missing outcome data.[305] Understanding why the data are missing (the missing data mechanism) and whether the data are MCAR, MAR, or MNAR has implications for analysis. For example, complete case analyses are expected to give unbiased estimates when data are MCAR. However, the MCAR assumption is strong and often not reasonable in practice. MAR is a less stringent assumption that is more common in real-life data, where complete case analyses can result in biased parameter estimation, leading to invalid conclusions. Unfortunately, complete case analyses of missing data are the most common in RCTs[305] and epidemiologic studies.[21] Multiple imputation is the most common method for handling missing data. It fills in missing data several times, informed by relationships between observed data. EHR-linked biobanks contain missing data, but they also contain non-missing genetic data. Li and colleagues leveraged genetic information to improve imputation of missing cardiovascular data in EHR-linked biobanks.[44] However, whether the use of genetic information in multiple imputation can improve association estimation is an open question. Simultaneously, EHR-linked biobanks are subject to selection bias because of recruitment mechanisms (e.g., recruitment through specific clinics[45]) and participant-driven factors (e.g., healthy volunteers[4]).

In Aim 3, we explored (a) whether PRS-informed multiple imputation reduces bias due to missing data and (b) the joint impact of PRS-informed multiple imputation and sample weighting on exposure-outcome association estimation when both missing data and selection bias are at play. First, we curated an analytic sample containing demographic, anthropometric, lifestyle, and genetic information in MGI to estimate the BMI coefficient for glucose. We simulated 100,000 observation multivariate-normal datasets using observed variance-covariance information and induced ~25% missingness in (a) BMI only and (b) BMI and glucose under MCAR, MAR, and MNAR mechanisms. We estimated the unadjusted and covariate-adjusted BMI coefficient for glucose using complete case, woPRS-imputed, and PRS-imputed analyses on random and biased samples of sizes 1,000, 2,500, 5,000, and 10,000. After 1,000 iterations, we calculated the coverage rate, percent bias, average 95% confidence interval width, and RMSE. For our case study, we estimated unadjusted and covariate-adjusted unweighted and weighted BMI coefficients for glucose in non-Hispanic White and non-Hispanic Black adults (aged 40 or older) without diabetes in MGI, calculating stratum-specific weights using the approach described in Salvatore and colleagues.[215] To our knowledge, this is the first study to consider missing data and selection bias jointly.

We found that PRS-imputed analyses exhibited better properties than woPRS-imputed in simulations of MAR data in random samples. This observation was also true in weighted analyses in our simulations of MAR data in biased samples. In our MGI case study, we found only small changes in coefficient estimates between complete, woPRS-imputed, and PRS-imputed analyses. However, performing weighted analyses changed the estimates considerably.

Our findings highlight the utility of genetic information in reducing bias due to missing data in EHR-linked biobanks. Researchers must carefully apply appropriate missing data methods alongside approaches to reduce other simultaneous biases like selection bias.

6.2 Public health relevance

As EHR-linked biobanks grow in size, number, and use for research, researchers must grapple with fundamental data issues to obtain impactful results that translate to specific groups. This dissertation addresses two issues – selection bias and missing data – and provides users with valuable guidance in performing principled analyses.

In the first aim, I explored the impact of selection weights on several common analyses conducted in EHR-linked biobanks. In doing so, I discovered that weights are crucial for prevalence and association estimation. Improving prevalence estimation, which describes the burden of diseases and informs areas needing public health attention, increases the utility of EHR-linked biobanks as a real-time surveillance tool. Addressing selection bias in association estimation, which captures the strength of the relationship between an exposure and outcome (or two diseases), is paramount because it can have indeterminate effects and lead to invalid conclusions. I also constructed selection weights for two US-based cohorts that are commonly used but in which weights are not available. Moreover, code for reproducing these weights or calculating weights in other cohorts is shared to promote the adoption of weighted analyses. Now more than ever, it is critical to think about the representativeness of samples and their intended target populations and apply methods to reduce selection bias as EHR-linked biobank data are reaching a broader range of researchers.

In my second aim, I found no single approach to summarizing health history in EHR data as a phenotype risk score (PheRS) that results in uniformly better risk stratification. Critically, as EHR-linked biobanks are used to develop risk prediction models for use in populations with different data distributions, weighting-based approaches to PheRS development do not improve stratification performance. However, diagnosis health history summarized as PheRS was the most important data domain in risk stratification compared to demographic covariates, risk factors, and presenting symptoms. Additionally, PheRS contributed to risk stratification alongside these domains, meaning researchers should consider the breadth of data domains in EHR-linked biobanks to improve risk prediction and stratification models. Resulting models can better identify individuals with higher-than-average risk for whom primary prevention, screening, or potentially invasive diagnostic procedures should be considered.

In my third aim, PRS-informed multiple imputation improved association estimation in EHR-linked biobanks with missing data. Missing data is a chronic problem in public health and EHR-linked biobank research.[21,300,305] My results demonstrate that leveraging a unique feature of EHR-linked biobanks – non-missing genetic information – reduces percent bias and improves and maintains the nominal coverage rate of association estimation relative to multiple imputation without PRS. This finding suggests EHR-linked biobanks should make PRS for common exposures available to researchers to reduce biases due to missing data. My case study, which also employed weighted analyses for reducing selection bias, suggested that bias due to missing data plays a relatively smaller role. However, this result highlights the importance of jointly considering and addressing the impacts of multiple simultaneous biases.

My aims address fundamental issues in EHR-linked biobank data and provide recommendations to achieve less biased and potentially more generalizable results, enhancing translation and improving public health impact. My work will be instrumental in leading to higher-quality EHR-linked biobank analyses.

6.3 Recommendations for future studies

There are natural continuations and extensions for future studies resulting from the findings in this dissertation. The first aim introduces methodological and substantive works. Methodologically, the weights we developed used factors that the cohorts explicitly stated influenced recruitment. However, these factors may not capture participant-driven factors impacting selection. Future work can use data-driven approaches, like that employed by van Alten and colleagues,[8] and can be compared with weights based on stated recruitment factors. Substantively, the first aim ignored analyses focused on or combined with genetic data. While Schoeler and colleagues explore the impact of selection weights on genetic analyses,[97] future research should explore the effect of selection bias on the genome-by-phenome landscape.

Future studies should also investigate non-weighting-based methods to enhance PheRS risk stratification transferability to explicit external populations. Other model-building approaches, like neural networks, support vector machines, and SuperLearner,[153] can be considered alongside methods to improve generalizability, including transfer, semi-supervised, and federated learning.[173,174,265,267] Again, while PheRS was evaluated alongside demographic covariates, risk factors, and presenting symptoms, we did not consider genetic predictors. We previously demonstrated that PheRS and PRS independently improve risk prediction,[18] meaning incorporation of genetic information in future work is warranted.

Finally, in Aim 3, we began to explore the joint impacts of missing data and selection bias in EHR-linked biobank data. Beesley and Mukherjee developed methods for jointly addressing selection bias and outcome misclassification.[7,11] Future studies should explore the relative impacts of multiple simultaneous biases more thoroughly. This should include evaluations across several exposure-outcome associations with gold standard estimates, in more missing data settings (e.g., multivariate missingness), and in more cohorts, and to ultimately develop recommendations for dealing with multiple biases jointly in practice.

My PhD dissertation has equipped me with invaluable skills and experience in designing and applying methods-based EHR-linked biobank data analysis, a frontier in epidemiology. But more than that, it holds the promise of a significant role in advancing the translation into practice of results from EHR-linked biobank cohorts. These cohorts, often needing to be more representative of the intended target population, can benefit greatly from the insights of this dissertation. This work can inspire progress in our field by paving the way for more meaningful impacts in healthcare and medical research tailored to the specific needs and characteristics of the intended groups.

References

1. Beesley LJ, Salvatore M, Fritsche LG, et al. The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Stat Med*. 2020;39(6):773-800. doi:10.1002/sim.8445

2. All of Us Research Program Investigators, Denny JC, Rutter JL, et al. The “All of Us” Research Program. *N Engl J Med*. 2019;381(7):668-676. doi:10.1056/NEJMsr1809937

3. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779

4. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol*. 2017;186(9):1026-1034. doi:10.1093/aje/kwx246

5. Hernán MA, Robins JM. Selection bias. In: *Causal Inference: What If*. Chapman & Hall/CRC; 2020:103-118. Accessed October 12, 2023. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

6. Lash TL, Rothman KJ. Selection Bias and Generalizability. In: *Modern Epidemiology*. 4th ed. Wolters Kluwer; 2021:315-331.

7. Beesley LJ, Mukherjee B. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics*. Published online December 3, 2020:biom.13400. doi:10.1111/biom.13400

8. Van Alten S, Domingue BW, Faul J, Galama T, Marees AT. Reweighting UK Biobank corrects for pervasive selection bias due to volunteering. *Int J Epidemiol*. 2024;53(3):dyae054. doi:10.1093/ije/dyae054

9. Brayne C, Moffitt TE. The limitations of large-scale volunteer databases to address inequalities and global challenges in health and aging. *Nat Aging*. 2022;2(9):775-783. doi:10.1038/s43587-022-00277-x

10. Bishop CD, Leite WL, Snyder PA. Using Propensity Score Weighting to Reduce Selection Bias in Large-Scale Data Sets. *J Early Interv*. 2018;40(4):347-362. doi:10.1177/1053815118793430

11. Beesley LJ, Mukherjee B. Case studies in bias reduction and inference for electronic health record data with selection bias and phenotype misclassification. *Stat Med*. 2022;41(28):5501-5516. doi:10.1002/sim.9579

12. Farzadfar F. Cardiovascular disease risk prediction models: challenges and perspectives. *Lancet Glob Health*. 2019;7(10):e1288-e1289. doi:10.1016/S2214-109X(19)30365-1

13. Hippisley-Cox J, Mei W, Fitzgerald R, Coupland C. Development and validation of a novel risk prediction algorithm to estimate 10-year risk of oesophageal cancer in primary care: prospective cohort study and evaluation of performance against two other risk prediction models. *Lancet Reg Health - Eur*. 2023;32:100700. doi:10.1016/j.lanepe.2023.100700

14. Wilson PWF, D’Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*. 1998;97(18):1837-1847. doi:10.1161/01.CIR.97.18.1837

15. Aleksandrova K, Reichmann R, Kaaks R, et al. Development and validation of a lifestyle-based model for colorectal cancer risk prediction: the LiFeCRC score. *BMC Med*. 2021;19(1):1. doi:10.1186/s12916-020-01826-0

16. Fröhlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. *BMC Med*. 2018;16(1):150. doi:10.1186/s12916-018-1122-7

17. Iparragirre A, Lumley T, Barrio I, Arostegui I. Variable selection with LASSO regression for complex survey data. *Stat*. 2023;12(1):e578. doi:10.1002/sta4.578

18. Salvatore M, Beesley LJ, Fritsche LG, et al. Phenotype risk scores (PheRS) for pancreatic cancer using time-stamped electronic health record data: Discovery and validation in two large biobanks. *J Biomed Inform*. Published online December 2020:103652. doi:10.1016/j.jbi.2020.103652

19. Little RJ. In Praise of Simplicity not Mathematistry! Ten Simple Powerful Ideas for the Statistical Scientist. *J Am Stat Assoc*. 2013;108(502):359-369. doi:10.1080/01621459.2013.787932

20. Pedersen A, Mikkelsen E, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol*. 2017;Volume 9:157-166. doi:10.2147/CLEP.S129785

21. Eekhout I, De Boer RM, Twisk JWR, De Vet HCW, Heymans MW. Missing Data: A Systematic Review of How They Are Reported and Handled. *Epidemiology*. 2012;23(5):729-732. doi:10.1097/EDE.0b013e3182576cdb

22. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592. doi:10.1093/biomet/63.3.581

23. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annu Symp Proc AMIA Symp*. 2013;2013:1472-1477.

24. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inform Decis Mak*. 2014;14(1):51. doi:10.1186/1472-6947-14-51

25. Harton J, Mitra N, Hubbard RA. Informative presence bias in analyses of electronic health records-derived data: a cautionary note. *J Am Med Inform Assoc*. Published online April 19, 2022:ocac050. doi:10.1093/jamia/ocac050

26. McGee G, Haneuse S, Coull BA, Weisskopf MG, Rotem RS. On the Nature of Informative Presence Bias in Analyses of Electronic Health Records. *Epidemiology*. 2022;33(1):105-113. doi:10.1097/EDE.0000000000001432

27. Bourgeois FC. Patients Treated at Multiple Acute Health Care Facilities: Quantifying Information Fragmentation. *Arch Intern Med*. 2010;170(22):1989. doi:10.1001/archinternmed.2010.439

28. Wei WQ, Leibson CL, Ransom JE, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc*. 2012;19(2):219-224. doi:10.1136/amiajnl-2011-000597

29. Gianfrancesco MA, McCulloch CE, Trupin L, Graf J, Schmajuk G, Yazdany J. Reweighting to address nonparticipation and missing data bias in a longitudinal electronic health record study. *Ann Epidemiol*. 2020;50:48-51.e2. doi:10.1016/j.annepidem.2020.06.008

30. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol*. 2019;48(4):1294-1304. doi:10.1093/ije/dyz032

31. Li L, Shen C, Li X, Robins JM. On weighting approaches for missing data. *Stat Methods Med Res*. 2013;22(1):14-30. doi:10.1177/0962280211403597

32. van Buuren S. *Flexible Imputation of Missing Data*. Accessed February 6, 2024. https://stefvanbuuren.name/fimd/

33. Klebanoff MA, Cole SR. Use of Multiple Imputation in the Epidemiologic Literature. *Am J Epidemiol*. 2008;168(4):355-357. doi:10.1093/aje/kwn071

34. Little RJ, Carpenter JR, Lee KJ. A Comparison of Three Popular Methods for Handling Missing Data: Complete-Case Analysis, Inverse Probability Weighting, and Multiple Imputation. *Sociol Methods Res*. Published online August 5, 2022:004912412211138. doi:10.1177/00491241221113873

35. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*. 2011;45(3). doi:10.18637/jss.v045.i03

36. Jazayeri A, Liang OS, Yang CC. Imputation of Missing Data in Electronic Health Records Based on Patients' Similarities. *J Healthc Inform Res*. 2020;4(3):295-307. doi:10.1007/s41666-020-00073-5

37. Getz K, Hubbard RA, Linn KA. Performance of Multiple Imputation Using Modern Machine Learning Methods in Electronic Health Records Data. *Epidemiology*. 2023;34(2):206-215. doi:10.1097/EDE.0000000000001578

38. Li J, Yan XS, Chaudhary D, et al. Imputation of missing values for electronic health record laboratory data. *Npj Digit Med*. 2021;4(1):147. doi:10.1038/s41746-021-00518-0

39. Ge Y, Li Z, Zhang J. A simulation study on missing data imputation for dichotomous variables using statistical and machine learning methods. *Sci Rep*. 2023;13(1):9432. doi:10.1038/s41598-023-36509-2

40. Fritsche LG, Patil S, Beesley LJ, et al. Cancer PRSweb: An Online Repository with Polygenic Risk Scores for Major Cancer Traits and Their Evaluation in Two Independent Biobanks. *Am J Hum Genet*. 2020;107(5):815-836. doi:10.1016/j.ajhg.2020.08.025

41. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. 2018;19(9):581-590. doi:10.1038/s41576-018-0018-x

42. Li R, Chen Y, Ritchie MD, Moore JH. Electronic health records and polygenic risk scores for predicting disease risk. *Nat Rev Genet*. 2020;21(8):493-502. doi:10.1038/s41576-020-0224-1

43. Ma Y, Patil S, Zhou X, Mukherjee B, Fritsche LG. ExPRSweb: An online repository with polygenic risk scores for common health-related exposures. *Am J Hum Genet*. 2022;109(10):1742-1760. doi:10.1016/j.ajhg.2022.09.001

44. Li R, Chen Y, Moore JH. Integration of genetic and clinical information to improve imputation of data missing from electronic health records. *J Am Med Inform Assoc*. 2019;26(10):1056-1063. doi:10.1093/jamia/ocz041

45. Zawistowski M, Fritsche LG, Pandit A, et al. The Michigan Genomics Initiative: A biobank linking genotypes and electronic clinical records in Michigan Medicine patients. *Cell Genomics*. Published online January 2023:100257. doi:10.1016/j.xgen.2023.100257

46. De Souza YG, Greenspan JS. Biobanking past, present and future: responsibilities and benefits. *AIDS*. 2013;27(3):303-312. doi:10.1097/QAD.0b013e32835c1244

47. Zhou W, Kanai M, Wu KHH, et al. Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics*. 2022;2(10):100192. doi:10.1016/j.xgen.2022.100192

48. Streeter AJ, Lin NX, Crathorne L, et al. Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *J Clin Epidemiol*. 2017;87:23-34. doi:10.1016/j.jclinepi.2017.04.022

49. Zhang L, Wang Y, Schuemie MJ, Blei DM, Hripcsak G. Adjusting for indirectly measured confounding using large-scale propensity score. *J Biomed Inform*. 2022;134:104204. doi:10.1016/j.jbi.2022.104204

50. Hubbard RA, Tong J, Duan R, Chen Y. Reducing Bias Due to Outcome Misclassification for Epidemiologic Studies Using EHR-derived Probabilistic Phenotypes. *Epidemiology*. 2020;31(4):542-550. doi:10.1097/EDE.0000000000001193

51. Tong J, Huang J, Chubak J, et al. An augmented estimation procedure for EHR-based association studies accounting for differential misclassification. *J Am Med Inform Assoc JAMIA*. 2020;27(2):244-253. doi:10.1093/jamia/ocz180

52. Li MX, Yeung JMY, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum Genet*. 2012;131(5):747-756. doi:10.1007/s00439-011-1118-2

53. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol*. 2008;32(4):361-369. doi:10.1002/gepi.20310

54. Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. Target Validity and the Hierarchy of Study Designs. *Am J Epidemiol*. 2019;188(2):438-443. doi:10.1093/aje/kwy228

55. 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;184(11):847-855. doi:10.1093/aje/kww112

56. Phelan M, Bhavsar N, Goldstein BA. Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System

Can Impact Inference. *EGEMs Gener Evid Methods Improve Patient Outcomes*. 2017;5(1):22. doi:10.5334/egems.243

57. Data Standardization – OHDSI. Accessed July 31, 2023. https://www.ohdsi.org/data-standardization/

58. SNOMED CT. Accessed July 31, 2023. https://www.nlm.nih.gov/healthit/snomedct/index.html

59. Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: an international classification of diseases for the twenty-first century. *BMC Med Inform Decis Mak*. 2021;21(6):206. doi:10.1186/s12911-021-01534-6

60. International Classification of Diseases (ICD). Accessed February 18, 2022. https://www.who.int/standards/classifications/classification-of-diseases

61. Shuey M, Stead W, Aka I, et al. Next-Generation Phenotyping: Introducing PhecodeX for Enhanced Discovery Research in Medical Phenomics. *Bioinformatics*. Published online November 1, 2023. doi:10.1101/2023.06.18.23291088

62. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinforma Oxf Engl*. 2010;26(9):1205-1210. doi:10.1093/bioinformatics/btq126

63. Bastarache L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu Rev Biomed Data Sci*. 2021;4(1):1-19. doi:10.1146/annurev-biodatasci-122320-112352

64. Yu S, Ma Y, Gronsbell J, et al. Enabling phenotypic big data with PheNorm. *J Am Med Inform Assoc*. 2018;25(1):54-60. doi:10.1093/jamia/ocx111

65. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat Protoc*. 2019;14(12):3426-3444. doi:10.1038/s41596-019-0227-6

66. Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc*. 2016;23(6):1046-1052. doi:10.1093/jamia/ocv202

67. Salvatore M, Clark-Boucher D, Fritsche LG, et al. Cohort profile: Epidemiologic Questionnaire (EPI-Q) – a scalable, app-based health survey linked to electronic health record and genotype data. *Epidemiol Health*. Published online August 8, 2023:e2023074. doi:10.4178/epih.e2023074

68. Guo A, Khan YM, Langabeer JR, Foraker RE. Discovering disease–disease associations using electronic health records in The Guideline Advantage (TGA) dataset. *Sci Rep*. 2021;11(1):20969. doi:10.1038/s41598-021-00345-z

69. Appelbaum L, Cambronero JP, Stevens JP, et al. Development and validation of a pancreatic cancer risk model for the general population using electronic health records: An observational study. *Eur J Cancer*. 2021;143:19-30. doi:10.1016/j.ejca.2020.10.019

70. Ananthakrishnan AN, Cagan A, Cai T, et al. Identification of Nonresponse to Treatment Using Narrative Data in an Electronic Health Record Inflammatory Bowel Disease Cohort: *Inflamm Bowel Dis*. 2016;22(1):151-158. doi:10.1097/MIB.0000000000000580

71. Wu P, Zeng D, Wang Y. Matched Learning for Optimizing Individualized Treatment Strategies Using Electronic Health Records. *J Am Stat Assoc*. 2020;115(529):380-392. doi:10.1080/01621459.2018.1549050

72. Rosenstrom E, Meshkinfam S, Ivy JS, et al. Optimizing the First Response to Sepsis: An Electronic Health Record-Based Markov Decision Process Model. *Decis Anal*. 2022;19(4):265-296. doi:10.1287/deca.2022.0455

73. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and Machine Learning for Heart Failure Survival Analysis. *Stud Health Technol Inform*. 2015;216:40-44.

74. Yuan Q, Cai T, Hong C, et al. Performance of a Machine Learning Algorithm Using Electronic Health Record Data to Identify and Estimate Survival in a Longitudinal Cohort of Patients With Lung Cancer. *JAMA Netw Open*. 2021;4(7):e2114723. doi:10.1001/jamanetworkopen.2021.14723

75. Samad MD, Ulloa A, Wehner GJ, et al. Predicting Survival From Large Echocardiography and Electronic Health Record Datasets. *JACC Cardiovasc Imaging*. 2019;12(4):681-689. doi:10.1016/j.jcmg.2018.04.026

76. Haneuse S, Daniels M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data are Observed and Why? *EGEMs Gener Evid Methods Improve Patient Outcomes*. 2016;4(1):16. doi:10.13063/2327-9214.1203

77. 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;30(10):2221-2238. doi:10.1177/09622802211027601

78. Weiskopf NG, Dorr DA, Jackson C, Lehmann HP, Thompson CA. Healthcare utilization is a collider: an introduction to collider bias in EHR data reuse. *J Am Med Inform Assoc*. Published online February 8, 2023:ocad013. doi:10.1093/jamia/ocad013

79. Kundu R, Shi X, Morrison J, Barrett J, Mukherjee B. A framework for understanding selection bias in real-world healthcare data. *J R Stat Soc Ser A Stat Soc*. Published online May 2, 2024:qnae039. doi:10.1093/jrsssa/qnae039

80. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol*. 2017;186(9):1026-1034. doi:10.1093/aje/kwx246

81. Lu H, Cole SR, Howe CJ, Westreich D. Toward a Clearer Definition of Selection Bias When Estimating Causal Effects. *Epidemiology*. 2022;33(5):699-706. doi:10.1097/EDE.0000000000001516

82. Degtiar I, Rose S. A Review of Generalizability and Transportability. *Annu Rev Stat Its Appl*. 2023;10(1):501-524. doi:10.1146/annurev-statistics-042522-103837

83. Hernán MA, Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias: *Epidemiology*. 2004;15(5):615-625. doi:10.1097/01.ede.0000135174.63482.43

84. Elliott MR, Valliant R. Inference for Nonprobability Samples. *Stat Sci*. 2017;32(2). doi:10.1214/16-STS598

85. Jager J, Putnick DL, Bornstein MH. II. MORE THAN JUST CONVENIENT: THE SCIENTIFIC MERITS OF HOMOGENEOUS CONVENIENCE SAMPLES. *Monogr Soc Res Child Dev*. 2017;82(2):13-30. doi:10.1111/mono.12296

86. Government of Canada SC. 3.2.3 Non-probability sampling. Published September 2, 2021. Accessed October 13, 2023. https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch13/nonprob/5214898-eng.htm

87. Odgaard-Jensen J, Vist GE, Timmer A, et al. Randomisation to protect against selection bias in healthcare trials. Cochrane Methodology Review Group, ed. *Cochrane Database Syst Rev*. 2011;2015(4). doi:10.1002/14651858.MR000012.pub3

88. Kaplan RM, Chambers DA, Glasgow RE. Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. *Clin Transl Sci*. 2014;7(4):342-346. doi:10.1111/cts.12178

89. Msaouel P. The Big Data Paradox in Clinical Practice. *Cancer Invest*. 2022;40(7):567-576. doi:10.1080/07357907.2022.2084621

90. Suissa S. Immortal Time Bias in Pharmacoepidemiology. *Am J Epidemiol*. 2008;167(4):492-499. doi:10.1093/aje/kwm324

91. Yadav K, Lewis RJ. Immortal Time Bias in Observational Studies. *JAMA*. 2021;325(7):686. doi:10.1001/jama.2020.9151

92. Ebrahim S, Davey Smith G. Commentary: Should we always deliberately be non-representative? *Int J Epidemiol*. 2013;42(4):1022-1026. doi:10.1093/ije/dyt105

93. Smith GD. The Wright Stuff: Genes in the Interrogation of Correlation and Causation. *Eur J Personal*. 2012;26(4):391-413. doi:10.1002/per.1865

94. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol*. 2018;47(1):226-235. doi:10.1093/ije/dyx206

95. Swanson SA. A Practical Guide to Selection Bias in Instrumental Variable Analyses. *Epidemiology*. 2019;30(3):345-349. doi:10.1097/EDE.0000000000000973

96. Gkatzionis A, Burgess S. Contextualizing selection bias in Mendelian randomization: how bad is it likely to be? *Int J Epidemiol*. 2019;48(3):691-701. doi:10.1093/ije/dyy202

97. Schoeler T, Speed D, Porcu E, Pirastu N, Pingault JB, Kutalik Z. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat Hum Behav*. Published online April 27, 2023. doi:10.1038/s41562-023-01579-9

98. Szklo M, Nieto FJ. Understanding Lack of Validity: Bias. In: *Epidemiology: Beyong the Basics*. 3rd ed. Jones & Bartlett Learning; 2014:139.

99. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol*. 2014;43(6):1969-1985. doi:10.1093/ije/dyu149

100. Nohr EA, Liew Z. How to investigate and adjust for selection bias in cohort studies. *Acta Obstet Gynecol Scand*. 2018;97(4):407-416. doi:10.1111/aogs.13319

101. Carry PM, Vanderlinden LA, Dong F, et al. Inverse probability weighting is an effective method to address selection bias during the analysis of high dimensional data. *Genet Epidemiol*. 2021;45(6):593-603. doi:10.1002/gepi.22418

102. Geneletti S, Richardson S, Best N. Adjusting for selection bias in retrospective, case-control studies. *Biostatistics*. 2008;10(1):17-31. doi:10.1093/biostatistics/kxn010

103. NHIS - National Health Interview Survey. Accessed August 9, 2023. https://www.cdc.gov/nchs/nhis/index.htm

104. Lumley T. Post-stratification, raking, and calibration. In: *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Soncs; 2010:135-156.

105. Wiśniowski A, Sakshaug JW, Perez Ruiz DA, Blom AG. Integrating Probability and Nonprobability Samples for Survey Inference. *J Surv Stat Methodol*. 2020;8(1):120-147. doi:10.1093/jssam/smz051

106. Rueda MDM, Pasadas-del-Amo S, Rodríguez BC, Castro-Martín L, Ferri-García R. Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain. *Biom J*. 2023;65(2):2200035. doi:10.1002/bimj.202200035

107. Tan ALM, Getzen EJ, Hutch MR, et al. Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? *J Biomed Inform*. 2023;139:104306. doi:10.1016/j.jbi.2023.104306

108. Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Can Med Assoc J*. 2012;184(11):1265-1269. doi:10.1503/cmaj.110977

109. McClatchey KD, ed. *Clinical Laboratory Medicine*. 2nd ed. Lippincott Wiliams & Wilkins; 2002.

110. Banerjee D, Chung S, Wong EC, Wang EJ, Stafford RS, Palaniappan LP. Underdiagnosis of Hypertension Using Electronic Health Records. *Am J Hypertens*. 2012;25(1):97-102. doi:10.1038/ajh.2011.179

111. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2014;21(2):221-230. doi:10.1136/amiajnl-2013-001935

112. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience; 2004.

113. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338(jun29 1):b2393-b2393. doi:10.1136/bmj.b2393

114. Li P, Stuart EA, Allison DB. Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA*. 2015;314(18):1966. doi:10.1001/jama.2015.15281

115. Buuren S van, Groothuis-Oudshoorn K, Vink G, et al. mice: Multivariate Imputation by Chained Equations. Published online November 24, 2021. Accessed July 6, 2022. https://CRAN.R-project.org/package=mice

116. Enders CK. *Applied Missing Data Analysis*. Guilford Press; 2010.

117. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*. 2013;22(3):278-295. doi:10.1177/0962280210395740

118. Enders CK. A Primer on Maximum Likelihood Algorithms Available for Use With Missing Data. *Struct Equ Model Multidiscip J*. 2001;8(1):128-141. doi:10.1207/S15328007SEM0801_7

119. Enders CK. The Performance of the Full Information Maximum Likelihood Estimator in Multiple Regression Models with Missing Data. *Educ Psychol Meas*. 2001;61(5):713-740. doi:10.1177/0013164401615001

120. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. 2013;64(5):402. doi:10.4097/kjae.2013.64.5.402

121. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 1st ed. Wiley; 2002. doi:10.1002/9781119013563

122. Petersen I, Welch CA, Nazareth I, et al. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clin Epidemiol*. 2019;Volume 11:157-167. doi:10.2147/CLEP.S191437

123. Haneuse S, Bogart A, Jazic I, et al. Learning About Missing Data Mechanisms in Electronic Health Records-based Research: A Survey-based Approach. *Epidemiology*. 2016;27(1):82-90. doi:10.1097/EDE.0000000000000393

124. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20(1):144-151. doi:10.1136/amiajnl-2011-000681

125. Farmer R, Mathur R, Bhaskaran K, Eastwood SV, Chaturvedi N, Smeeth L. Promises and pitfalls of electronic health record analysis. *Diabetologia*. 2018;61(6):1241-1248. doi:10.1007/s00125-017-4518-6

126. Bots SH, Groenwold RHH, Dekkers OM. Using electronic health record data for clinical research: a quick guide. *Eur J Endocrinol*. 2022;186(4):E1-E6. doi:10.1530/EJE-21-1088

127. Callahan A, Shah NH, Chen JH. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. *Ann Intern Med*. 2020;172(11_Supplement):S79-S84. doi:10.7326/M19-0873

128. Cyganek B, Graña M, Krawczyk B, et al. A Survey of Big Data Issues in Electronic Health Record Analysis. *Appl Artif Intell*. 2016;30(6):497-520. doi:10.1080/08839514.2016.1193714

129. Wei WQ, Bastarache LA, Carroll RJ, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. Rzhetsky A, ed. *PLOS ONE*. 2017;12(7):e0175508. doi:10.1371/journal.pone.0175508

130. Liu X, Chubak J, Hubbard RA, Chen Y. SAT: a Surrogate-Assisted Two-wave case boosting sampling method, with application to EHR-based association studies. *J Am Med Inform Assoc*. 2022;29(5):918-927. doi:10.1093/jamia/ocab267

131. Yin Z, Tong J, Chen Y, Hubbard RA, Tang CY. A cost-effective chart review sampling design to account for phenotyping error in electronic health records (EHR) data. *J Am Med Inform Assoc*. 2021;29(1):52-61. doi:10.1093/jamia/ocab222

132. Teixeira PL, Wei WQ, Cronin RM, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *J Am Med Inform Assoc*. 2017;24(1):162-171. doi:10.1093/jamia/ocw071

133. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*. Published online April 30, 2018:k1479. doi:10.1136/bmj.k1479

134. Sisk R, Lin L, Sperrin M, et al. Informative presence and observation in routine health data: A review of methodology for clinical risk prediction. *J Am Med Inform Assoc*. 2021;28(1):155-166. doi:10.1093/jamia/ocaa242

135. Meng XL. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann Appl Stat*. 2018;12(2). doi:10.1214/18-AOAS1161SF

136. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer International Publishing; 2019. doi:10.1007/978-3-030-16399-0

137. US Preventive Services Task Force, Curry SJ, Krist AH, et al. Risk Assessment for Cardiovascular Disease With Nontraditional Risk Factors: US Preventive Services Task Force Recommendation Statement. *JAMA*. 2018;320(3):272. doi:10.1001/jama.2018.8359

138. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. Published online May 16, 2016:i2416. doi:10.1136/bmj.i2416

139. Lloyd-Jones DM. Cardiovascular Risk Prediction: Basic Concepts, Current Status, and Future Directions. *Circulation*. 2010;121(15):1768-1777. doi:10.1161/CIRCULATIONAHA.109.849166

140. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*. 1989;81(24):1879-1886. doi:10.1093/jnci/81.24.1879

141. Antoniou AC, Pharoah PPD, Smith P, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *Br J Cancer*. 2004;91(8):1580-1590. doi:10.1038/sj.bjc.6602175

142. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med*. 2004;23(7):1111-1130. doi:10.1002/sim.1668

143. Wells BJ, Kattan MW, Cooper GS, Jackson L, Koroukian S. ColoRectal Cancer Predicted Risk Online (CRC-PRO) Calculator Using Data from the Multi-Ethnic Cohort Study. *J Am Board Fam Med*. 2014;27(1):42-55. doi:10.3122/jabfm.2014.01.130040

144. Zhang YD, Hurson AN, Zhang H, et al. Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nat Commun*. 2020;11(1):3353. doi:10.1038/s41467-020-16483-3

145. Usher-Smith JA, Emery J, Kassianos AP, Walter FM. Risk Prediction Models for Melanoma: A Systematic Review. *Cancer Epidemiol Prev Biomark*. 2014;23(8):1450-1463. doi:10.1158/1055-9965.EPI-14-0295

146. Berhane S, Toyoda H, Tada T, et al. Role of the GALAD and BALAD-2 Serologic Models in Diagnosis of Hepatocellular Carcinoma and Prediction of Survival in Patients. *Clin Gastroenterol Hepatol*. 2016;14(6):875-886.e6. doi:10.1016/j.cgh.2015.12.042

147. Liu Y, Zhang J, Wang W, Li G. Development and validation of a risk prediction model for incident liver cancer. *Front Public Health*. 2022;10:955287. doi:10.3389/fpubh.2022.955287

148. Friedrich S, Groll A, Ickstadt K, et al. Regularization approaches in clinical biostatistics: A review of methods and their applications. *Stat Methods Med Res*. 2023;32(2):425-440. doi:10.1177/09622802221133557

149. Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Stat Sci*. 2001;16(3). doi:10.1214/ss/1009213726

150. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123-140. doi:10.1007/BF00058655

151. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

152. Schapire RE. The Boosting Approach to Machine Learning: An Overview. In: Denison DD, Hansen MH, Holmes CC, Mallick B, Yu B, eds. *Nonlinear Estimation and Classification*. Vol 171. Lecture Notes in Statistics. Springer New York; 2003:149-171. doi:10.1007/978-0-387-21579-2_9

153. Van Der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol*. 2007;6(1). doi:10.2202/1544-6115.1309

154. Bannick MS, McGaughey M, Flaxman AD. Ensemble modelling in descriptive epidemiology: burden of disease estimation. *Int J Epidemiol*. 2021;49(6):2065-2073. doi:10.1093/ije/dyz223

155. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, Van Der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. Aalto-Setala K, ed. *PLOS ONE*. 2019;14(5):e0213653. doi:10.1371/journal.pone.0213653

156. Douville NJ, Surakka I, Leis A, et al. Use of a Polygenic Risk Score Improves Prediction of Myocardial Injury After Non-Cardiac Surgery. *Circ Genomic Precis Med*. 2020;13(4):e002817. doi:10.1161/CIRCGEN.119.002817

157. Kesar A, Baluch A, Barber O, et al. Actionable absolute risk prediction of atherosclerotic cardiovascular disease based on the UK Biobank. Gadekallu TR, ed. *PLOS ONE*. 2022;17(2):e0263940. doi:10.1371/journal.pone.0263940

158. Muller DC, Johansson M, Brennan P. Lung Cancer Risk Prediction Model Incorporating Lung Function: Development and Validation in the UK Biobank Prospective Cohort Study. *J Clin Oncol*. 2017;35(8):861-869. doi:10.1200/JCO.2016.69.2467

159. Smith T, Gunter MJ, Tzoulaki I, Muller DC. The added value of genetic information in colorectal cancer risk prediction models: development and evaluation in the UK Biobank prospective cohort study. *Br J Cancer*. 2018;119(8):1036-1039. doi:10.1038/s41416-018-0282-8

160. Wong KCY, Xiang Y, Yin L, So HC. Uncovering Clinical Risk Factors and Predicting Severe COVID-19 Cases Using UK Biobank Data: Machine Learning Approach. *JMIR Public Health Surveill*. 2021;7(9):e29544. doi:10.2196/29544

161. Dabbah MA, Reed AB, Booth ATC, et al. Machine learning approach to dynamic risk modeling of mortality in COVID-19: a UK Biobank study. *Sci Rep*. 2021;11(1):16936. doi:10.1038/s41598-021-95136-x

162. Willette AA, Willette SA, Wang Q, et al. Using machine learning to predict COVID-19 infection and severity risk among 4510 aged adults: a UK Biobank cohort study. *Sci Rep*. 2022;12(1):7736. doi:10.1038/s41598-022-07307-z

163. Grady C. Institutional Review Boards. *Chest*. 2015;148(5):1148-1155. doi:10.1378/chest.15-0706

164. Nosowsky R, Giordano TJ. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule: Implications for Clinical Research. *Annu Rev Med*. 2006;57(1):575-590. doi:10.1146/annurev.med.57.121304.131257

165. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated Learning for Healthcare Informatics. *J Healthc Inform Res*. 2021;5(1):1-19. doi:10.1007/s41666-020-00082-4

166. Crowson MG, Moukheiber D, Arévalo AR, et al. A systematic review of federated learning applications for biomedical data. Mordaunt DA, ed. *PLOS Digit Health*. 2022;1(5):e0000033. doi:10.1371/journal.pdig.0000033

167. Luo C, Islam MdN, Sheils NE, et al. Lossless Distributed Linear Mixed Model with Application to Integration of Heterogeneous Healthcare Data. Published online November 18, 2020. doi:10.1101/2020.11.16.20230730

168. Duan R, Boland MR, Moore JH, Chen Y. ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *Pac Symp Biocomput Pac Symp Biocomput*. 2019;24:30-41.

169. Tong J, Duan R, Li R, Scheuemie MJ, Moore JH, Chen Y. Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data. *Pac Symp Biocomput Pac Symp Biocomput*. 2020;25:695-706.

170. Duan R, Boland MR, Liu Z, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *J Am Med Inform Assoc*. 2020;27(3):376-385. doi:10.1093/jamia/ocz199

171. Duan R, Luo C, Schuemie MJ, et al. Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. *J Am Med Inform Assoc JAMIA*. 2020;27(7):1028-1036. doi:10.1093/jamia/ocaa044

172. Luo C, Duan R, Naj AC, Kranzler HR, Bian J, Chen Y. ODACH: a one-shot distributed algorithm for Cox model with heterogeneous multi-center data. *Sci Rep*. 2022;12(1):6627. doi:10.1038/s41598-022-09069-0

173. Pan SJ, Yang Q. A Survey on Transfer Learning. *IEEE Trans Knowl Data Eng*. 2010;22(10):1345-1359. doi:10.1109/TKDE.2009.191

174. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data*. 2016;3(1):9. doi:10.1186/s40537-016-0043-6

175. Bickel S, Brückner M, Scheffer T. Discriminative learning for differing training and test distributions. In: *Proceedings of the 24th International Conference on Machine Learning*. ACM; 2007:81-88. doi:10.1145/1273496.1273507

176. Sugiyama M, Krauledat M, Müller KR. Covariate Shift Adaptation by Importance Weighted Cross Validation. *J Mach Learn Res*. 2007;8:985-1005.

177. Datta A, Fiksel J, Amouzou A, Zeger SL. Regularized Bayesian transfer learning for population-level etiological distributions. *Biostatistics*. 2021;22(4):836-857. doi:10.1093/biostatistics/kxaa001

178. Vergouwe Y, Moons KGM, Steyerberg EW. External Validity of Risk Models: Use of Benchmark Values to Disentangle a Case-Mix Effect From Incorrect Coefficients. *Am J Epidemiol*. 2010;172(8):971-980. doi:10.1093/aje/kwq223

179. Kouw WM, Loog M. An introduction to domain adaptation and transfer learning. Published online 2018. doi:10.48550/ARXIV.1812.11806

180. Steingrimsson JA, Gatsonis C, Li B, Dahabreh IJ. Transporting a Prediction Model for Use in a New Target Population. *Am J Epidemiol*. 2023;192(2):296-304. doi:10.1093/aje/kwac128

181. SEER Cancer Stat Facts. SEER. Accessed December 1, 2021. https://seer.cancer.gov/statfacts/index.html

182. Yang HI, Yuen MF, Chan HLY, et al. Risk estimation for hepatocellular carcinoma in chronic hepatitis B (REACH-B): development and validation of a predictive score. *Lancet Oncol*. 2011;12(6):568-574. doi:10.1016/S1470-2045(11)70077-8

183. Papatheodoridis G, Dalekos G, Sypsa V, et al. PAGE-B predicts the risk of developing hepatocellular carcinoma in Caucasians with chronic hepatitis B on 5-year antiviral therapy. *J Hepatol*. 2016;64(4):800-806. doi:10.1016/j.jhep.2015.11.035

184. Wong VWS, Chan SL, Mo F, et al. Clinical Scoring System to Predict Hepatocellular Carcinoma in Chronic Hepatitis B Carriers. *J Clin Oncol*. 2010;28(10):1660-1665. doi:10.1200/JCO.2009.26.2675

185. Yuen MF, Tanaka Y, Fong DYT, et al. Independent risk factors and predictive score for the development of hepatocellular carcinoma in chronic hepatitis B. *J Hepatol*. 2009;50(1):80-88. doi:10.1016/j.jhep.2008.07.023

186. Kierans AS, Makkar J, Guniganti P, et al. Validation of Liver Imaging Reporting and Data System 2017 (LI-RADS) Criteria for Imaging Diagnosis of Hepatocellular Carcinoma. *J Magn Reson Imaging*. 2019;49(7). doi:10.1002/jmri.26329

187. Sharma SA, Kowgier M, Hansen BE, et al. Toronto HCC risk index: A validated scoring system to predict 10-year risk of HCC in patients with cirrhosis. *J Hepatol*. 2018;68(1):92-99. doi:10.1016/j.jhep.2017.07.033

188. Fujiwara N, Kubota N, Crouchet E, et al. Molecular signatures of long-term hepatocellular carcinoma risk in nonalcoholic fatty liver disease. *Sci Transl Med*. 2022;14(650):eabo4474. doi:10.1126/scitranslmed.abo4474

189. Sharma A, Kandlakunta H, Nagpal SJS, et al. Model to Determine Risk of Pancreatic Cancer in Patients With New-Onset Diabetes. *Gastroenterology*. 2018;155(3):730-739.e3. doi:10.1053/j.gastro.2018.05.023

190. Clift AK, Tan PS, Patone M, et al. Predicting the risk of pancreatic cancer in adults with new-onset diabetes: development and internal–external validation of a clinical risk prediction model. *Br J Cancer*. 2024;130(12):1969-1978. doi:10.1038/s41416-024-02693-9

191. Wenker TN, Rubenstein JH, Thrift AP, Singh H, El-Serag HB. Development and Validation of the Houston-BEST, a Barrett's Esophagus Risk Prediction Model Adaptable to Electronic Health Records. *Clin Gastroenterol Hepatol*. 2023;21(9):2424-2426.e0. doi:10.1016/j.cgh.2022.08.007

192. Rubenstein JH, McConnell D, Waljee AK, et al. Validation and Comparison of Tools for Selecting Individuals to Screen for Barrett's Esophagus and Early Neoplasia. *Gastroenterology*. 2020;158(8):2082-2092. doi:10.1053/j.gastro.2020.02.037

193. Johnson PJ, Pirrie SJ, Cox TF, et al. The Detection of Hepatocellular Carcinoma Using a Prospectively Developed and Validated Model Based on Serological Biomarkers. *Cancer Epidemiol Biomarkers Prev*. 2014;23(1):144-153. doi:10.1158/1055-9965.EPI-13-0870

194. Jia K, Kundrot S, Palchuk MB, et al. A pancreatic cancer risk prediction model (Prism) developed and validated on large-scale US clinical data. *eBioMedicine*. 2023;98:104888. doi:10.1016/j.ebiom.2023.104888

195. Liu H, Li K, Xia J, et al. Prediction of esophageal cancer risk based on genetic variants and environmental risk factors in Chinese population. *BMC Cancer*. 2024;24(1):598. doi:10.1186/s12885-024-12370-y

196. Llovet JM, Kelley RK, Villanueva A, et al. Hepatocellular carcinoma. *Nat Rev Dis Primer*. 2021;7(1):6. doi:10.1038/s41572-020-00240-3

197. Klein AP. Pancreatic cancer epidemiology: understanding the role of lifestyle and inherited risk factors. *Nat Rev Gastroenterol Hepatol*. 2021;18(7):493-502. doi:10.1038/s41575-021-00457-x

198. Zhang HZ, Jin GF, Shen HB. Epidemiologic differences in esophageal cancer between Asian and Western populations. *Chin J Cancer*. 2012;31(6):281-286. doi:10.5732/cjc.011.10390

199. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2017;24(1):198-208. doi:10.1093/jamia/ocw042

200. Pusztai L, Hatzis C, Andre F. Reproducibility of research and preclinical validation: problems and solutions. *Nat Rev Clin Oncol*. 2013;10(12):720-724. doi:10.1038/nrclinonc.2013.171

201. Rountree L, Lin YT, Liu C, et al. Reporting of Fairness Metrics in Clinical Risk Prediction Models: A Call for Change. Published online March 18, 2024. doi:10.1101/2024.03.16.24304390

202. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *J Biomed Inform*. 2021;113:103621. doi:10.1016/j.jbi.2020.103621

203. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *Npj Digit Med*. 2020;3(1):99. doi:10.1038/s41746-020-0304-9

204. Huddar V, Desiraju BK, Rajan V, Bhattacharya S, Roy S, Reddy CK. Predicting Complications in Critical Care Using Heterogeneous Clinical Data. *IEEE Access*. 2016;4:7988-8001. doi:10.1109/ACCESS.2016.2618775

205. Steyerberg EW, Nieboer D, Debray TPA, Van Houwelingen HC. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. *Stat Med*. 2019;38(22):4290-4309. doi:10.1002/sim.8296

206. Chan WX, Wong L. Obstacles to effective model deployment in healthcare. *J Bioinform Comput Biol*. 2023;21(02):2371001. doi:10.1142/S0219720023710014

207. Sharma V, Ali I, Veer S van der, Martin G, Ainsworth J, Augustine T. Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. *BMJ Health Care Inform*. 2021;28(1):e100253. doi:10.1136/bmjhci-2020-100253

208. Watson J, Hutyra CA, Clancy SM, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? *JAMIA Open*. 2020;3(2):167-172. doi:10.1093/jamiaopen/ooz046

209. Giddings R, Joseph A, Callender T, et al. Factors influencing clinician and patient interaction with machine learning-based risk prediction models: a systematic review. *Lancet Digit Health*. 2024;6(2):e131-e144. doi:10.1016/S2589-7500(23)00241-8

210. University of California, San Francisco, Adler-Milstein J, Aggarwal N, et al. Meeting the Moment: Addressing Barriers and Facilitating Clinical Adoption of Artificial Intelligence in Medical Diagnosis. *NAM Perspect*. 2022;22(9). doi:10.31478/202209c

211. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagn Progn Res*. 2020;4(1):8. doi:10.1186/s41512-020-00077-0

212. Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless…. *J Am Med Inform Assoc*. 2019;26(12):1645-1650. doi:10.1093/jamia/ocz145

213. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc*. 2017;24(6):1052-1061. doi:10.1093/jamia/ocx030

214. Minne L, Eslami S, De Keizer N, De Jonge E, De Rooij SE, Abu-Hanna A. Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. *Intensive Care Med*. 2012;38(1):40-46. doi:10.1007/s00134-011-2390-2

215. Salvatore M, Kundu R, Shi X, et al. To weight or not to weight? The effect of selection bias in 3 large electronic health record-linked biobanks and recommendations for practice. *J Am Med Inform Assoc*. Published online May 14, 2024:ocae098. doi:10.1093/jamia/ocae098

216. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*. 2016;538(7624):161-164. doi:10.1038/538161a

217. Goldstein JA, Weinstock JS, Bastarache LA, et al. LabWAS: Novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. *PLOS Genet*. 2020;16(11):e1009077. doi:10.1371/journal.pgen.1009077

218. Tsuo K, Zhou W, Wang Y, et al. Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity. *Cell Genomics*. 2022;2(12):100212. doi:10.1016/j.xgen.2022.100212

219. Wu KHH, Douville NJ, Konerman MC, et al. *Polygenic Risk Score from a Multi-Ancestry GWAS Uncovers Susceptibility of Heart Failure*. Cardiovascular Medicine; 2021. doi:10.1101/2021.12.06.21267389

220. Surakka I, Wu KH, Hornsby W, et al. *Multi-Ancestry Meta-Analysis Identifies 2 Novel Loci Associated with Ischemic Stroke and Reveals Heterogeneity of Effects between Sexes and Ancestries*. Genetic and Genomic Medicine; 2022. doi:10.1101/2022.02.28.22271647

221. Chen Y, Li P, Wu C. Doubly Robust Inference With Nonprobability Survey Samples. *J Am Stat Assoc*. 2020;115(532):2011-2021. doi:10.1080/01621459.2019.1677241

222. Ramirez AH, Sulieman L, Schlueter DJ, et al. The All of Us Research Program: Data quality, utility, and diversity. *Patterns*. 2022;3(8):100570. doi:10.1016/j.patter.2022.100570

223. University of Michigan Precision Health. Michigan Genomics Initiative. Accessed February 18, 2022. https://precisionhealth.umich.edu/our-research/michigangenomics/

224. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z

225. Horvitz DG, Thompson DJ. A Generalization of Sampling Without Replacement from a Finite Universe. *J Am Stat Assoc*. 1952;47(260):663-685. doi:10.1080/01621459.1952.10483446

226. Pfeffermann D. The Role of Sampling Weights When Modeling Survey Data. *Int Stat Rev Rev Int Stat*. 1993;61(2):317. doi:10.2307/1403631

227. PheWAS/PhecodeX. Accessed August 11, 2023. https://github.com/PheWAS/PhecodeX

228. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinforma Oxf Engl*. 2014;30(16):2375-2376. doi:10.1093/bioinformatics/btu197

229. Surveillance Research Program, National Cancer Institute. Recent trends in SEER Age-adjusted incidence rates, 2000-2020: Colon and rectum. SEER*Explorer: An interactive website for SEER cancer statistics. Published November 16, 2023. Accessed January 15, 2024. https://seer.cancer.gov/statistics-network/explorer/

230. White A, Ironmonger L, Steele RJC, Ormiston-Smith N, Crawford C, Seims A. A review of sex-related differences in colorectal cancer incidence, screening uptake, routes to diagnosis, cancer stage and survival in the UK. *BMC Cancer*. 2018;18(1):906. doi:10.1186/s12885-018-4786-7

231. Schwarzer G. meta: General Package for Meta-Analysis. Published online June 7, 2023. Accessed July 23, 2023. https://cran.r-project.org/web/packages/meta/index.html

232. Gao X. Multiple testing corrections for imputed SNPs. *Genet Epidemiol*. 2011;35(3):154-158. doi:10.1002/gepi.20563

233. Meng XL. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann Appl Stat*. 2018;12(2). doi:10.1214/18-AOAS1161SF

234. Rice K, Higgins JPT, Lumley T. A Re-Evaluation of Fixed Effect(s) Meta-Analysis. *J R Stat Soc Ser A Stat Soc*. 2018;181(1):205-227. doi:10.1111/rssa.12275

235. Getzen E, Ungar L, Mowery D, Jiang X, Long Q. Mining for equitable health: Assessing the impact of missing data in electronic health records. *J Biomed Inform*. 2023;139:104269. doi:10.1016/j.jbi.2022.104269

236. Zhou D, Gan Z, Shi X, et al. Multiview Incomplete Knowledge Graph Integration with application to cross-institutional EHR data harmonization. *J Biomed Inform*. 2022;133:104147. doi:10.1016/j.jbi.2022.104147

237. Robertson SE, Leith A, Schmid CH, Dahabreh IJ. Assessing Heterogeneity of Treatment Effects in Observational Studies. *Am J Epidemiol*. 2021;190(6):1088-1100. doi:10.1093/aje/kwaa235

238. Robins JM, Rotnitzky A. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *J Am Stat Assoc*. 1995;90(429):122-129. doi:10.1080/01621459.1995.10476494

239. Zivich PN, Breskin A. Machine Learning for Causal Inference: On the Use of Cross-fit Estimators. *Epidemiology*. 2021;32(3):393-401. doi:10.1097/EDE.0000000000001332

240. Zhong Y, Kennedy EH, Bodnar LM, Naimi AI. AIPW: An R Package for Augmented Inverse Probability–Weighted Estimation of Average Causal Effects. *Am J Epidemiol*. 2021;190(12):2690-2699. doi:10.1093/aje/kwab207

241. Kim S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Commun Stat Appl Methods*. 2015;22(6):665-674. doi:10.5351/CSAM.2015.22.6.665

242. Lumley T. CRAN - Package survey. Published 2023. Accessed August 11, 2023. https://cran.r-project.org/web/packages/survey/index.html

243. Haneuse S, Arterburn D, Daniels MJ. Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task. *JAMA Netw Open*. 2021;4(2):e210184. doi:10.1001/jamanetworkopen.2021.0184

244. Li L, Shen C, Li X, Robins JM. On weighting approaches for missing data. *Stat Methods Med Res*. 2013;22(1):14-30. doi:10.1177/0962280211403597

245. O'Neill TJ, Nguemo JD, Tynan AM, Burchell AN, Antoniou T. Risk of Colorectal Cancer and Associated Mortality in HIV: A Systematic Review and Meta-Analysis. *JAIDS J Acquir Immune Defic Syndr*. 2017;75(4):439-447. doi:10.1097/QAI.0000000000001433

246. Coghill AE, Engels EA, Schymura MJ, Mahale P, Shiels MS. Risk of Breast, Prostate, and Colorectal Cancer Diagnoses Among HIV-Infected Individuals in the United States. *JNCI J Natl Cancer Inst*. 2018;110(9):959-966. doi:10.1093/jnci/djy010

247. PheWAS. Published online May 5, 2023. Accessed May 5, 2023. https://github.com/PheWAS/PheWAS

248. Dandapani SV, Eaton M, Thomas CR, Pagnini PG. HIV- positive anal cancer: an update for the clinician. *J Gastrointest Oncol*. 2010;1(1):34-44. doi:10.3978/j.issn.2078-6891.2010.005

249. van Walraven C, Dhalla IA, Bell C, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *CMAJ Can Med Assoc J J Assoc Medicale Can*. 2010;182(6):551-557. doi:10.1503/cmaj.091117

250. Graff RE, Cavazos TB, Thai KK, et al. Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. *Nat Commun*. 2021;12(1):970. doi:10.1038/s41467-021-21288-z

251. Sassano M, Mariani M, Quaranta G, Pastorino R, Boccia S. Polygenic risk prediction models for colorectal cancer: a systematic review. *BMC Cancer*. 2022;22(1):65. doi:10.1186/s12885-021-09143-2

252. Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk Prediction Models for Colorectal Cancer: A Systematic Review. *Cancer Prev Res (Phila Pa)*. 2016;9(1):13-26. doi:10.1158/1940-6207.CAPR-15-0274

253. Li H, Sun D, Cao M, et al. Risk prediction models for esophageal cancer: A systematic review and critical appraisal. *Cancer Med*. 2021;10(20):7265-7276. doi:10.1002/cam4.4226

254. Chen R, Zheng R, Zhou J, et al. Risk Prediction Model for Esophageal Cancer Among General Population: A Systematic Review. *Front Public Health*. 2021;9:680967. doi:10.3389/fpubh.2021.680967

255. Yang JD, Addissie BD, Mara KC, et al. GALAD Score for Hepatocellular Carcinoma Detection in Comparison with Liver Ultrasound and Proposal of GALADUS Score. *Cancer Epidemiol Biomarkers Prev*. 2019;28(3):531-538. doi:10.1158/1055-9965.EPI-18-0281

256. Liang CW, Yang HC, Islam MM, et al. Predicting Hepatocellular Carcinoma With Minimal Features From Electronic Health Records: Development of a Deep Learning Model. *JMIR Cancer*. 2021;7(4):e19812. doi:10.2196/19812

257. Wang W, Chen S, Brune KA, Hruban RH, Parmigiani G, Klein AP. PancPRO: Risk Assessment for Individuals With a Family History of Pancreatic Cancer. *J Clin Oncol*. 2007;25(11):1417-1422. doi:10.1200/JCO.2006.09.2452

258. Placido D, Yuan B, Hjaltelin JX, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nat Med*. 2023;29(5):1113-1122. doi:10.1038/s41591-023-02332-5

259. Gao XR, Chiariglione M, Qin K, et al. Explainable machine learning aggregates polygenic risk scores and electronic health records for Alzheimer's disease prediction. *Sci Rep*. 2023;13(1):450. doi:10.1038/s41598-023-27551-1

260. Barnado A, Wheless L, Camai A, et al. Phenotype Risk Score but Not Genetic Risk Score Aids in Identifying Individuals With Systemic Lupus Erythematosus in the Electronic Health Record. *Arthritis Rheumatol*. 2023;75(9):1532-1541. doi:10.1002/art.42544

261. Takahashi P, Cerhan J, Ryu E, et al. Health behaviors and quality of life predictors for risk of hospitalization in an electronic health record-linked biobank. *Int J Gen Med*. Published online August 2015:247. doi:10.2147/IJGM.S85473

262. Petrazzini BO, Chaudhary K, Márquez-Luna C, et al. Coronary Risk Estimation Based on Clinical Data in Electronic Health Records. *J Am Coll Cardiol*. 2022;79(12):1155-1166. doi:10.1016/j.jacc.2022.01.021

263. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2017;24(1):198-208. doi:10.1093/jamia/ocw042

264. Chekroud AM, Hawrilenko M, Loho H, et al. Illusory generalizability of clinical prediction models. *Science*. 2024;383(6679):164-167. doi:10.1126/science.adg8538

265. Chi S, Li X, Tian Y, et al. Semi-supervised learning to improve generalizability of risk prediction models. *J Biomed Inform*. 2019;92:103117. doi:10.1016/j.jbi.2019.103117

266. Naderalvojoud B, Curtin CM, Yanover C, et al. Towards global model generalizability: independent cross-site feature evaluation for patient-level risk prediction models using the OHDSI network. *J Am Med Inform Assoc*. Published online February 27, 2024:ocae028. doi:10.1093/jamia/ocae028

267. Gu T, Lee PH, Duan R. COMMUTE: Communication-efficient transfer learning for multi-site risk prediction. *J Biomed Inform*. 2023;137:104243. doi:10.1016/j.jbi.2022.104243

268. Zhao Z, Fritsche LG, Smith JA, Mukherjee B, Lee S. The construction of cross-population polygenic risk scores using transfer learning. *Am J Hum Genet*. 2022;109(11):1998-2008. doi:10.1016/j.ajhg.2022.09.010

269. Salvatore M, Jeon J, Meza R. Changing trends in liver cancer incidence by race/ethnicity and sex in the US: 1992–2016. *Cancer Causes Control*. 2019;30(12):1377-1388. doi:10.1007/s10552-019-01237-4

270. Pereira SP, Oldfield L, Ney A, et al. Early detection of pancreatic cancer. *Lancet Gastroenterol Hepatol*. 2020;5(7):698-710. doi:10.1016/S2468-1253(19)30416-9

271. Singhi AD, Koay EJ, Chari ST, Maitra A. Early Detection of Pancreatic Cancer: Opportunities and Challenges. *Gastroenterology*. 2019;156(7):2024-2040. doi:10.1053/j.gastro.2019.01.259

272. Klein AP, Lindström S, Mendelsohn JB, et al. An Absolute Risk Model to Identify Individuals at Elevated Risk for Pancreatic Cancer in the General Population. Real FX, ed. *PLoS ONE*. 2013;8(9):e72311. doi:10.1371/journal.pone.0072311

273. Dong J, Buas MF, Gharahkhani P, et al. Determining Risk of Barrett's Esophagus and Esophageal Adenocarcinoma Based on Epidemiologic Factors and Genetic Variants. *Gastroenterology*. 2018;154(5):1273-1281.e3. doi:10.1053/j.gastro.2017.12.003

274. Freedman AN, Slattery ML, Ballard-Barbash R, et al. Colorectal Cancer Risk Prediction Tool for White Men and Women Without Known Susceptibility. *J Clin Oncol*. 2009;27(5):686-693. doi:10.1200/JCO.2008.17.4797

275. Rubenstein JH, Morgenstern H, Appelman H, et al. Prediction of Barrett's Esophagus Among Men. *Am J Gastroenterol*. 2013;108(3):353-362. doi:10.1038/ajg.2012.446

276. Baldwin-Hunter BL, Knotts RM, Leeds SD, Rubenstein JH, Lightdale CJ, Abrams JA. Use of the Electronic Health Record to Target Patients for Non-endoscopic Barrett's Esophagus Screening. *Dig Dis Sci*. 2019;64(12):3463-3470. doi:10.1007/s10620-019-05707-2

277. Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open*. 2015;5(3):e007825-e007825. doi:10.1136/bmjopen-2015-007825

278. Wang QL, Ness-Jensen E, Santoni G, Xie SH, Lagergren J. Development and Validation of a Risk Prediction Model for Esophageal Squamous Cell Carcinoma Using Cohort Studies. *Am J Gastroenterol*. 2021;116(4):683-691. doi:10.14309/ajg.0000000000001094

279. Zhang Y. Epidemiology of esophageal cancer. *World J Gastroenterol*. 2013;19(34):5598. doi:10.3748/wjg.v19.i34.5598

280. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. *J Am Med Inform Assoc*. 2021;29(1):187-196. doi:10.1093/jamia/ocab199

281. Sulieman L, Cronin RM, Carroll RJ, et al. Comparing medical history data derived from electronic health records and survey answers in the All of Us Research

Program. *J Am Med Inform Assoc*. 2022;29(7):1131-1141. doi:10.1093/jamia/ocac046

282. Bosch FX, Ribes J, Díaz M, Cléries R. Primary liver cancer: Worldwide incidence and trends. *Gastroenterology*. 2004;127(5):S5-S16. doi:10.1053/j.gastro.2004.09.011

283. Muhammad W, Hart GR, Nartowt B, et al. Pancreatic Cancer Prediction Through an Artificial Neural Network. *Front Artif Intell*. 2019;2:2. doi:10.3389/frai.2019.00002

284. Ahmed FE. Artificial neural networks for diagnosis and survival prediction in colon cancer. *Mol Cancer*. 2005;4(1):29. doi:10.1186/1476-4598-4-29

285. Zhao D, Liu H, Zheng Y, He Y, Lu D, Lyu C. A reliable method for colorectal cancer prediction based on feature selection and support vector machine. *Med Biol Eng Comput*. 2019;57(4):901-912. doi:10.1007/s11517-018-1930-0

286. Wang H, Zheng B, Yoon SW, Ko HS. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *Eur J Oper Res*. 2018;267(2):687-699. doi:10.1016/j.ejor.2017.12.001

287. Martini A, Stefanelli D, Biasiolo A, et al. New algorithm to identify patients at higher risk to develop hepatocellular carcinoma, based on machine learning approach. *Dig Liver Dis*. 2022;54:S166. doi:10.1016/j.dld.2022.08.003

288. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31(12):1102-1111. doi:10.1038/nbt.2749

289. Arnold M, Abnet CC, Neale RE, et al. Global Burden of 5 Major Types of Gastrointestinal Cancer. *Gastroenterology*. 2020;159(1):335-349.e15. doi:10.1053/j.gastro.2020.02.068

290. Quan H, Sundararajan V, Halfon P, et al. Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data: *Med Care*. 2005;43(11):1130-1139. doi:10.1097/01.mlr.0000182534.19832.83

291. Gasparini A, Salmasian H, Williman J, Chia SY, Teo E. comorbidity: Computing Comorbidity Scores. Published online May 1, 2023. Accessed February 15, 2024. https://cran.r-project.org/web/packages/comorbidity/index.html

292. Young JC, Conover MM, Jonsson Funk M. Measurement Error and Misclassification in Electronic Medical Records: Methods to Mitigate Bias. *Curr Epidemiol Rep*. 2018;5(4):343-356. doi:10.1007/s40471-018-0164-x

293. Lu CY. Observational studies: a review of study designs, challenges and strategies to reduce confounding. *Int J Clin Pract*. 2009;63(5):691-697. doi:10.1111/j.1742-1241.2009.02056.x

294. Goldstein BA, Phelan M, Pagidipati NJ, Peskoe SB. How and when informative visit processes can bias inference when using electronic health records data for clinical research. *J Am Med Inform Assoc*. 2019;26(12):1609-1617. doi:10.1093/jamia/ocz148

295. Tyrer F, Bhaskaran K, Rutherford MJ. Immortal time bias for life-long conditions in retrospective observational studies using electronic health records. *BMC Med Res Methodol*. 2022;22(1):86. doi:10.1186/s12874-022-01581-1

296. Prada-Ramallal G, Takkouche B, Figueiras A. Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. *BMC Med Res Methodol*. 2019;19(1):53. doi:10.1186/s12874-019-0695-y

297. Fu S, Leung LY, Raulli AO, et al. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. *BMC Med Inform Decis Mak*. 2020;20(1):60. doi:10.1186/s12911-020-1072-9

298. Glynn EF, Hoffman MA. Heterogeneity introduced by EHR system implementation in a de-identified data resource from 100 non-affiliated organizations. *JAMIA Open*. 2019;2(4):554-561. doi:10.1093/jamiaopen/ooz035

299. Japec L, Kreuter F, Berg M, et al. Big Data in Survey Research: AAPOR Task Force Report. *Public Opin Q*. 2015;79(4):839-880. doi:10.1093/poq/nfv039

300. Cole SR, Zivich PN, Edwards JK, et al. Missing Outcome Data in Epidemiologic Studies. *Am J Epidemiol*. 2023;192(1):6-10. doi:10.1093/aje/kwac179

301. Howe CJ, Cain LE, Hogan JW. Are All Biases Missing Data Problems? *Curr Epidemiol Rep*. 2015;2(3):162-171. doi:10.1007/s40471-015-0050-8

302. Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. *Int J Epidemiol*. 2015;44(4):1452-1459. doi:10.1093/ije/dyu272

303. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials*. 2004;1(4):368-376. doi:10.1191/1740774504cn032oa

304. Little RJ, D’Agostino R, Cohen ML, et al. The Prevention and Treatment of Missing Data in Clinical Trials. *N Engl J Med*. 2012;367(14):1355-1360. doi:10.1056/NEJMsr1203730

305. Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. *BMC Med Res Methodol*. 2014;14(1):118. doi:10.1186/1471-2288-14-118

306. Perkins NJ, Cole SR, Harel O, et al. Principled Approaches to Missing Data in Epidemiologic Studies. *Am J Epidemiol*. 2018;187(3):568-575. doi:10.1093/aje/kwx348

307. Seaman SR, White IR, Copas AJ, Li L. Combining Multiple Imputation and Inverse-Probability Weighting. *Biometrics*. 2012;68(1):129-137. doi:10.1111/j.1541-0420.2011.01666.x

308. Lee T, Shi D. A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychol Methods*. 2021;26(4):466-485. doi:10.1037/met0000381

309. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*. 2001;6(4):330-351.

310. Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for Handling Missing Data in Electronic Health Record Derived Data. *EGEMs Gener Evid Methods Improve Patient Outcomes*. 2013;1(3):7. doi:10.13063/2327-9214.1035

311. Harel O, Mitchell EM, Perkins NJ, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. *Am J Epidemiol*. 2018;187(3):576-584. doi:10.1093/aje/kwx349

312. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. *JMIR Med Inform*. 2018;6(1):e11. doi:10.2196/medinform.8960

313. Zeng C, Schlueter DJ, Tran TC, et al. Comparison of phenomic profiles in the All of Us Research Program against the US general population and the UK Biobank. *J Am Med Inform Assoc*. 2024;31(4):846-854. doi:10.1093/jamia/ocad260

314. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol*. 2009;9(1):57. doi:10.1186/1471-2288-9-57

315. Howrigan DP, Abbott L, rkwalters, Palmer D, Francioli L, Hammerbacher J. Nealelab/UK_Biobank_GWAS: v2. Published online June 6, 2023. doi:10.5281/zenodo.8011558

316. UK Biobank. Neale lab. Accessed April 24, 2024. http://www.nealelab.is/uk-biobank

317. Carson AP, Muntner P, Selvin E, et al. Do glycemic marker levels vary by race? Differing results from a cross-sectional analysis of individuals with and without diagnosed diabetes. *BMJ Open Diabetes Res Care*. 2016;4(1):e000213. doi:10.1136/bmjdrc-2016-000213

318. Zhu Y, Sidell MA, Arterburn D, et al. Racial/Ethnic Disparities in the Prevalence of Diabetes and Prediabetes by BMI: Patient Outcomes Research To Advance Learning (PORTAL) Multisite Cohort of Adults in the U.S. *Diabetes Care*. 2019;42(12):2211-2219. doi:10.2337/dc19-0532

319. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Med Care*. 2013;51(Supplement 8Suppl 3):S30-S37. doi:10.1097/MLR.0b013e31829b1dbd

320. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in Using Electronic Health Record Data for CER: Experience of 4 Learning Organizations and Solutions Applied. *Med Care*. 2013;51:S80-S86.

321. Samal L, Dykes PC, Greenberg JO, et al. Care coordination gaps due to lack of interoperability in the United States: a qualitative study and literature review. *BMC Health Serv Res*. 2016;16(1):143. doi:10.1186/s12913-016-1373-y

322. Dixon WJ, University of California, Los Angeles, eds. *BMDP Statistical Software: 1981*. 1981 ed. University of California Press; 1981.

323. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *J Am Stat Assoc*. 1988;83(404):1198-1202. doi:10.1080/01621459.1988.10478722

324. Jamshidian M, Jalal S. Tests of Homoscedasticity, Normality, and Missing Completely at Random for Incomplete Multivariate Data. *Psychometrika*. 2010;75(4):649-674. doi:10.1007/s11336-010-9175-3

325. Rouzinov S, Berchtold A. Regression-Based Approach to Test Missing Data Mechanisms. *Data*. 2022;7(2):16. doi:10.3390/data7020016

326. Austin PC, White IR, Lee DS, Van Buuren S. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Can J Cardiol*. 2021;37(9):1322-1331. doi:10.1016/j.cjca.2020.11.010

327. Robins JM, Rotnitzky A, Scharfstein DO. Sensitivity Analysis for Selection bias and unmeasured Confounding in missing Data and Causal inference models. In: Halloran ME, Berry D, eds. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. Vol 116. The IMA Volumes in Mathematics and its Applications. Springer New York; 2000:1-94. doi:10.1007/978-1-4612-1284-3_1

328. Díaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clin Trials*. 2014;11(5):590-600. doi:10.1177/1740774514537136

329. Sullivan TR, Yelland LN, Lee KJ, Ryan P, Salter AB. Treatment of missing data in follow-up studies of randomised controlled trials: A systematic review of the literature. *Clin Trials*. 2017;14(4):387-395. doi:10.1177/1740774517703319

330. Rombach I, Rivero-Arias O, Gray AM, Jenkinson C, Burke Ó. The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. *Qual Life Res*. 2016;25(7):1613-1623. doi:10.1007/s11136-015-1206-1

331. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7(2):147-177.

332. Lee KJ, Carlin JB, Simpson JA, Moreno-Betancur M. Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification. *Int J Epidemiol*. 2023;52(4):1268-1275. doi:10.1093/ije/dyad008

333. Mohan K, Pearl J, Tian J. Graphical models for inference with Missing data. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS’13. Curran Associates Inc.; 2013:1277-1285.

334. Mohan K, Pearl J. Graphical Models for Processing Missing Data. *J Am Stat Assoc*. 2021;116(534):1023-1037. doi:10.1080/01621459.2021.1874961

335. Thoemmes F, Mohan K. Graphical Representation of Missing Data Problems. *Struct Equ Model Multidiscip J*. 2015;22(4):631-642. doi:10.1080/10705511.2014.937378

336. Bartlett JW, Harel O, Carpenter JR. Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. *Am J Epidemiol*. 2015;182(8):730-736. doi:10.1093/aje/kwv114

337. Liang Y, Lu W, Ying Z. Joint Modeling and Analysis of Longitudinal Data with Informative Observation Times. *Biometrics*. 2009;65(2):377-384. doi:10.1111/j.1541-0420.2008.01104.x

338. Lin DY, Ying Z. Semiparametric and Nonparametric Regression Analysis of Longitudinal Data. *J Am Stat Assoc*. 2001;96(453):103-126. doi:10.1198/016214501750333018

339. Sun J, Park DH, Sun L, Zhao X. Semiparametric Regression Analysis of Longitudinal Data With Informative Observation Times. *J Am Stat Assoc*. 2005;100(471):882-889. doi:10.1198/016214505000000060

340. Bůržková P, Lumley T. Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *Can J Stat*. 2007;35(4):485-500. doi:10.1002/cjs.5550350402