Supporting Information for "Case studies in bias reduction and inference for electronic health record data with selection bias and phenotype misclassification" by

Lauren J. Beesley[*1,2] **and Bhramar Mukherjee**[1]
[1]University of Michigan, Department of Biostatistics
[2]Los Alamos National Laboratory, Information Systems and Modeling
*Corresponding Author: lvandervort@lanl.gov

Contents

A Michigan Genomics Initiative dataset at a glance

In the main paper, we provide an overview of The Michigan Genomics Initiative (MGI) dataset and provide detailed analyses. Here, we include some additional information referred to in the main text. **Supp. Figure A.1** provides a visual schematic of the data generation mechanisms in MGI along with the conceptual difference between source and target populations. **Supp. Table A.1** provides comparisons between MGI patients, people included in the National Health and Nutrition Examination Survey (NHANES) in 2017-2018, and the US adult population. **Supp. Figure A.2** shows the rates of observed disease by age in these three groups of people, and **Supp. Table A.2** provides sources for external summary information used for these comparisons.

In this paper, we consider several EHR-derived phenotypes. International Classification of Disease (ICD) codes were used to define disease status in MGI data. These codes were aggregated into a coding system known as phenotype codes or "phecodes" following the coding systems described elsewhere (Denny et al., 2010). Cancer diagnosis was defined as receipt of any phecode corresponding to a cancer diagnosis during follow-up in the Michigan Medicine EHR. Diabetes diagnosis was defined as receipt of phecode 250 ("diabetes mellitus"), which includes diabetes types I and II. Coronary artery disease (CAD) diagnosis was defined as receipt of phecode 411.4 ("coronary atherosclerosis"). Macular degeneration diagnosis was defined as receipt of phecode 362.2 ("degeneration of macula and posterior pole of retina") among patients at least 50 years old. Since our goal in case study (b) is to study age-related macular degeneration (AMD), we will use "macular degeneration" and "AMD" interchangeably. Body mass index (BMI) was defined as the median observed BMI value prior to any cancer diagnosis or bariatric surgery. For patients without BMI measurements before such diagnoses, the earliest observed BMI was chosen.

Figure A.1: Schematic of MGI data generation and desired data analysis

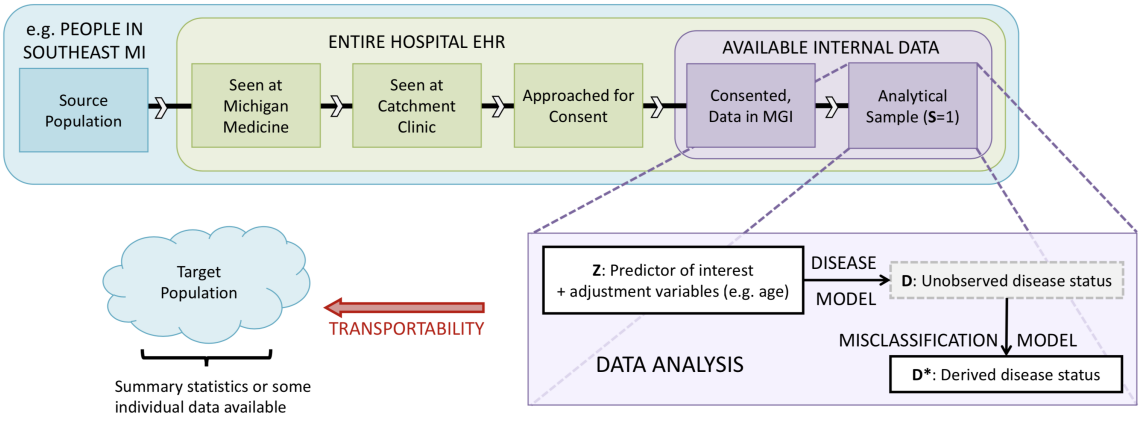


Table A.1: Comparison of disease and demographic characteristics between cohorts

Characteristic	**MGI**[1]	**NHANES, 2017-2018 (Interview and Examination, 18+)**	**US Adult Population**
Sample size	43339	5533	>200,000,000
Age, median	59.0	51	35.3[3]
Number of visits, mean (range)	84 (1 - 1425)	-	-
Length of follow-up (d = day, y = year)	8y (1d - 41.0y)	-	-
Body Mass Index (BMI), mean	29.9	29.7	27.6[4]
BMI category, n (%)			
Underweight, <18.5	473 (1.1)	99 (1.8)	11.9%[4]
Normal, [18.5, 25.0)	10338 (23.9)	1372 (24.8)	27.3%
Overweight, [25.0, 30.0)	14025 (32.4)	1727 (31.2)	26.3%
Obese, 30.0+	17880 (41.3)	2236 (40.4)	34.4%
Unknown	623 (1.4)	99 (1.8)	-
Female, n (%)	22710 (52.4)	2861 (51.7)	50.9%[3]
Smoking habits, n (%)			
Never	21575 (49.8)	3301 (59.7)	58.9%[4]
Former	14261 (32.9)	1260 (22.8)	24.0%
Current	7402 (17.1)	972 (17.6)	17.1%
Unknown	101 (0.2)	0 (0)	-
Lifetime disease prevalence, (%)			
Diabetes (Type I or II)	11838 (27.3)	838 (15.2)[2]	13.0[5]
Macular degeneration	1849 (4.3)	-	2.1%[6]
Coronary artery disease	6943 (16.0)	243 (4.6)	12.1%[7]
Cancer (any type)	23587 (54.4)	551 (10.5)	39.5%[8]
Race/Ethnicity (%)			
Non-Hispanic White	42387 (97.8)	1898 (34.3)	69.1%[3]
Hispanic/Other/Unknown	952 (2.2)	3635 (65.7)	30.9%

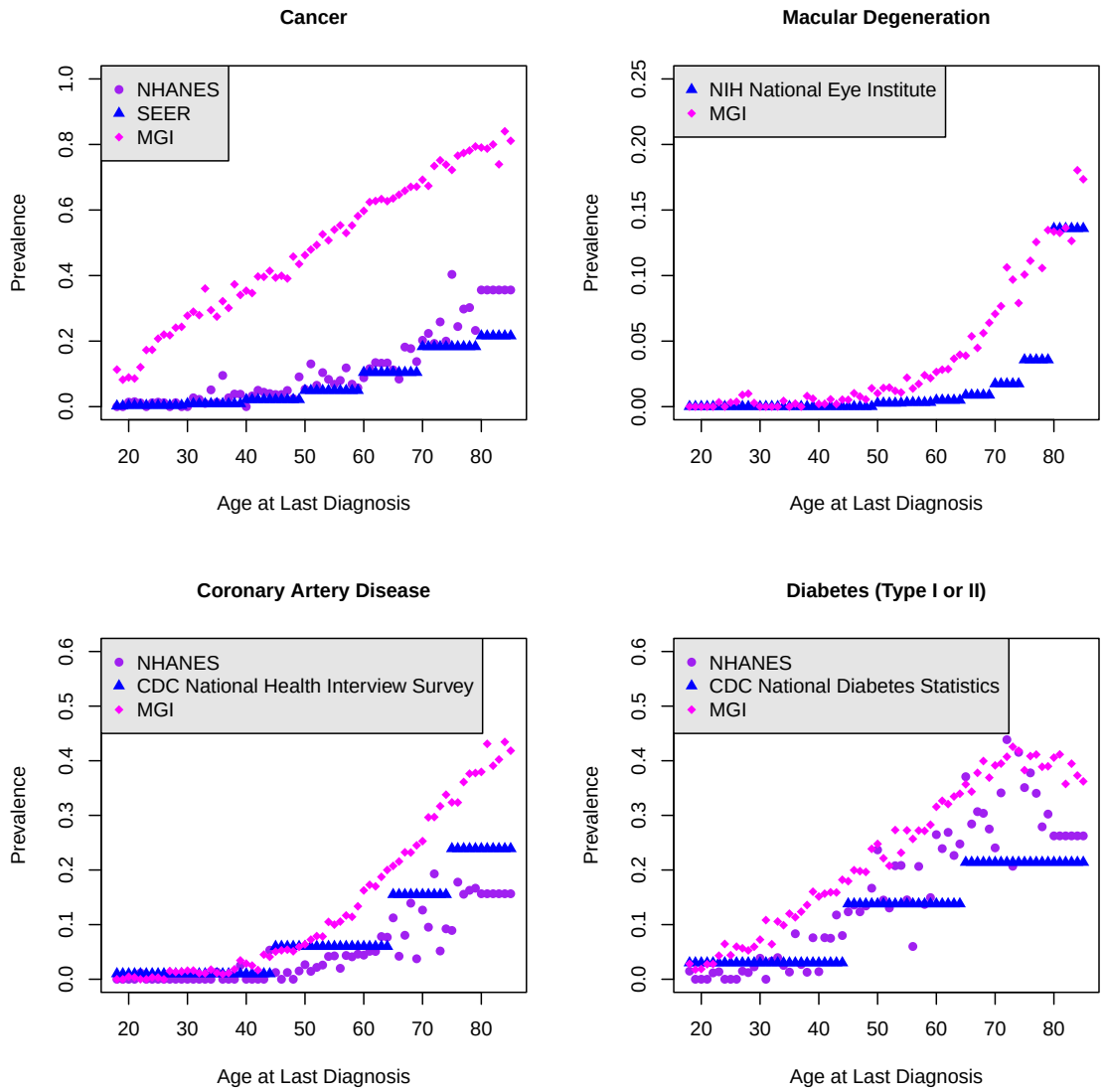
[1] age represents age of last diagnosis in EHR. Summaries provided for entire MGI cohort used in case study (a). Case study (b) uses a subset of unrelated MGI patients of recent European decent aged 50+.
[2] NHANES diagnosis status was missing for 4, 284, and 270 patients for diabetes, coronary artery disease, and cancer (any type), respectively. Macular degeneration diagnosis was not collected for NHANES in 2017-2018. Lifetime disease prevalence was calculated for NHANES using patients with observed disease status.
[3] Source: US Census, 2000. Adult-only median age is between 40 and 44.
[4] Source: National Health and Nutrition Examination Survey (NHANES) with selection weighting, 2017-2018
[5] Source: Centers for Disease Control and Prevention (CDC), National Diabetes Statistics Report, 2013-2016
[6] Source: NIH National Eye Institute Statistics, 2010 prevalence for ages 50+.
[7] Source: CDC, National Center for Health Statistics, National Health Interview Survey, 2017-2018
[8] Source: NIH National Cancer Institute; Surveillance, Epidemiology and End Results Program (SEER), 2015-2017

Table A.2: External data sources used for bias correction

Data Source	Quantity	Link
US Census, 2000	age distribution	https://www.census.gov
NIH National Cancer Institute; Surveillance, Epidemiology and End Results Program (SEER)	cancer prevalence[1] by age (2016)	https://seer.cancer.gov/data/
	lifetime risk of developing cancer (all sites, 2015-2017)	https://seer.cancer.gov/data/
	lifetime cancer risk by gender (all sites, 2008-2016)	https://seer.cancer.gov/csr/previous.html
NIH National Eye Institute Statistics, 2010	age-related macular degeneration prevalence by age (and overall for ages 50+)	https://www.nei.nih.gov
Centers for Disease Control and Prevention (CDC), National Diabetes Statistics Report, 2013-2016	diabetes prevalence by age	https://www.cdc.gov
CDC, National Center for Health Statistics, National Health Interview Survey, 2017-2018	coronary artery disease prevalence by age	https://www.cdc.gov/nchs/index.htm
National Health and Nutrition Examination Survey (NHANES), 2017-2018	cancer, diabetes, CAD, AMD, BMI, age, and smoking joint distribution	https://www.cdc.gov/nchs/...

[1] invasive cancers only, limited to cancers occurring in the previous 24 years.

Figure A.2: Disease prevalence by age in MGI, NHANES, and the US adult population[1]



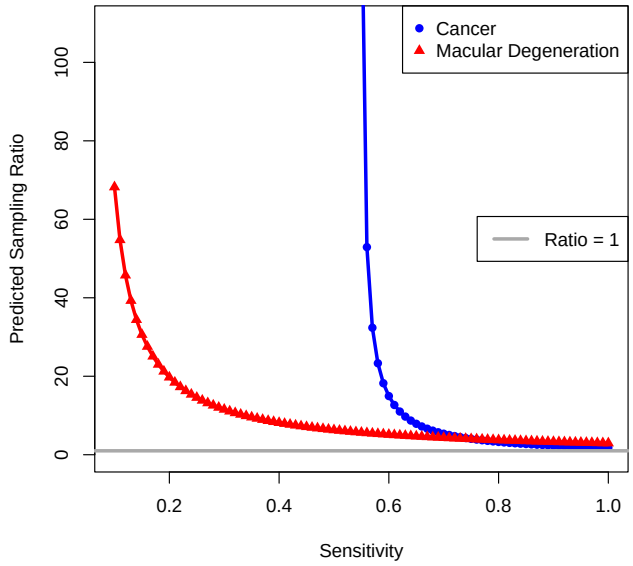
[1] Prevalence by age in NHANES was calculated using data from 2017-2018. Data on macular degeneration diagnosis are not available for NHANES for this time period. Cancer prevalence by age is for invasive cancers only.

B Case studies: additional figures and tables

Step 1: Fixing reasonable values for $\tilde{r}$

We can use the population disease rates for cancer and macular degeneration reported in **Supp. Table A.1** along with the observed EHR-derived disease status rates in MGI to obtain estimates of the marginal sampling ratio, $\tilde{r}$, as a function of potential values for the marginal sensitivity, $\tilde{c}$. **Supp. Figure B.1** shows these predicted values. We cannot use these data alone to determine the "true" value for $\tilde{r}$. Instead, we can use this plot to guide reasonable choices for $\tilde{r}$ for further analysis. For both outcomes, we consider values between 1 (no disease-related selection) and 100 (patients with disease 100x more likely to be included).

Figure B.1: Marginal sampling ratio as a function of marginal sensitivity



[1] These relationships were estimated using assumed disease prevalences in **Supp. Table A.1**. For the cancer outcome, some values of $\tilde{c}$ were incompatible with the data (estimated $\tilde{r} < 0$), and $\tilde{r}$ is not plotted.

Step 2: Estimating sensitivity

We then estimate sensitivity $c_{true}(X)$ as a function of covariates X using Method 2b in **Figure 1**. This approach requires specification of $P(D = 1|X)$. This distribution is not known, but we do know the marginal disease prevalence, $P(D = 1)$, and the relationship between disease diagnosis and age, $P(D = 1|\text{Age})$. We estimate $c_{true}(X)$ first assuming $P(D = 1|X) = P(D = 1)$ and then assuming $P(D = 1|X) = P(D = 1|\text{Age})$. We present results assuming $P(D = 1|X) = P(D = 1|\text{Age})$ in the main paper. **Supp. Figure B.2** presents the distributions of estimated $c_{true}(X)$ across MGI participants. This figure demonstrates that the choice $P(D = 1|X)$ can have a strong impact in the estimation of $c_{true}(X)$. For the cancer outcome, where diagnosis is expected to be associated with more doctors visits and longer follow-up, neither of our specifications for $P(D = 1|X)$ is very reasonable. We showed in Beesley and Mukherjee (2022), however, that downstream estimation of θ is only weakly impacted by the specification of $P(D = 1|X)$, so we are not very concerned with this assumption violation in practice. **Supp. Figure B.3** shows the estimated β log-odds ratios for covariates in the sensitivity model. Interestingly, the estimated odds ratios tend to be similar for the two outcomes, with a slightly stronger estimated association with longer follow-up time for the macular degeneration outcome. Higher sensitivity in both outcomes is associated with longer follow-up and more visits per follow-up time.

Figure B.2: Estimated patient-varying sensitivities $c_{true}(X)$ as a function of marginal sampling ratio using method 2b. $P(D = 1|X)$ was assumed to equal either $P(D = 1)$ or $P(D = 1|\text{Age})$.

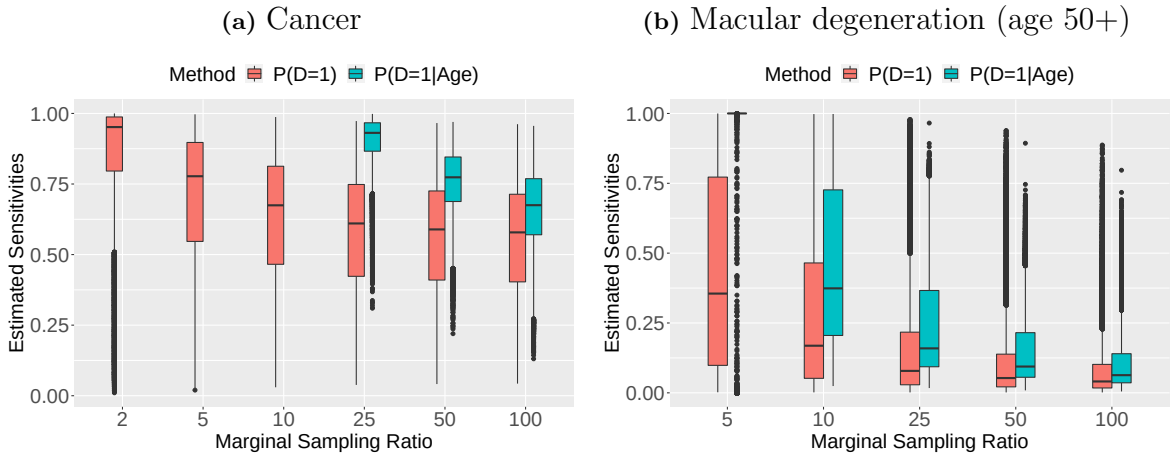


Figure B.3: Estimated sensitivity model parameters β as a function of marginal sampling ratio. Sensitivity was estimated using the method 2b and setting $P(D = 1|X) = P(D = 1|\text{Age})$

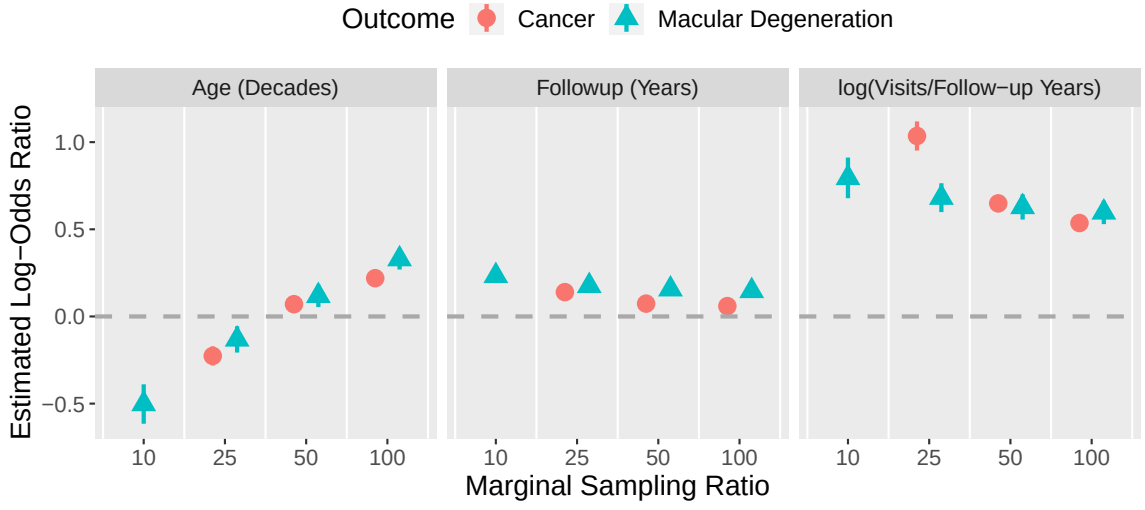
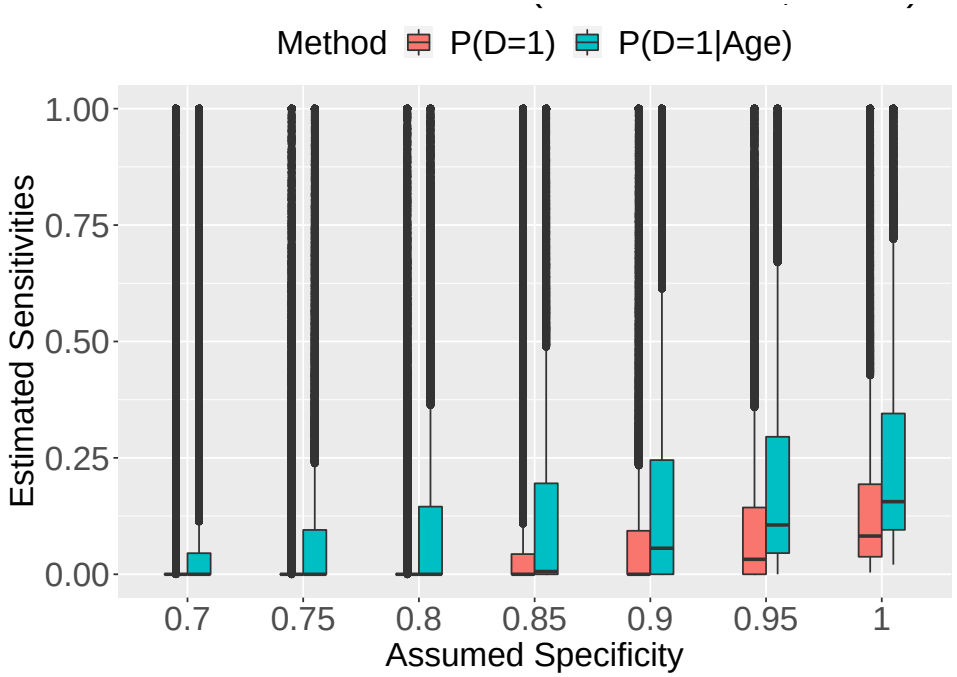


Figure B.4: Estimated patient-varying sensitivities $c_{true}(X)$ as a function of specificity for Macular degeneration (age 50+), assuming $\tilde{r} = 25$. $P(D = 1|X)$ was approximated by either $P(D = 1)$ or $P(D = 1|\text{Age})$.



Step 3: Estimating weights for selection bias adjustment

In estimating weights for selection bias adjustment using NHANES data, we fit regression models for (1) the probability of selection into NHANES among the US adult population and (2) the probability of inclusion in MGI given inclusion in MGI or NHANES. Parameter estimates for these regression model fits are provided in **Supp. Table B.1**.

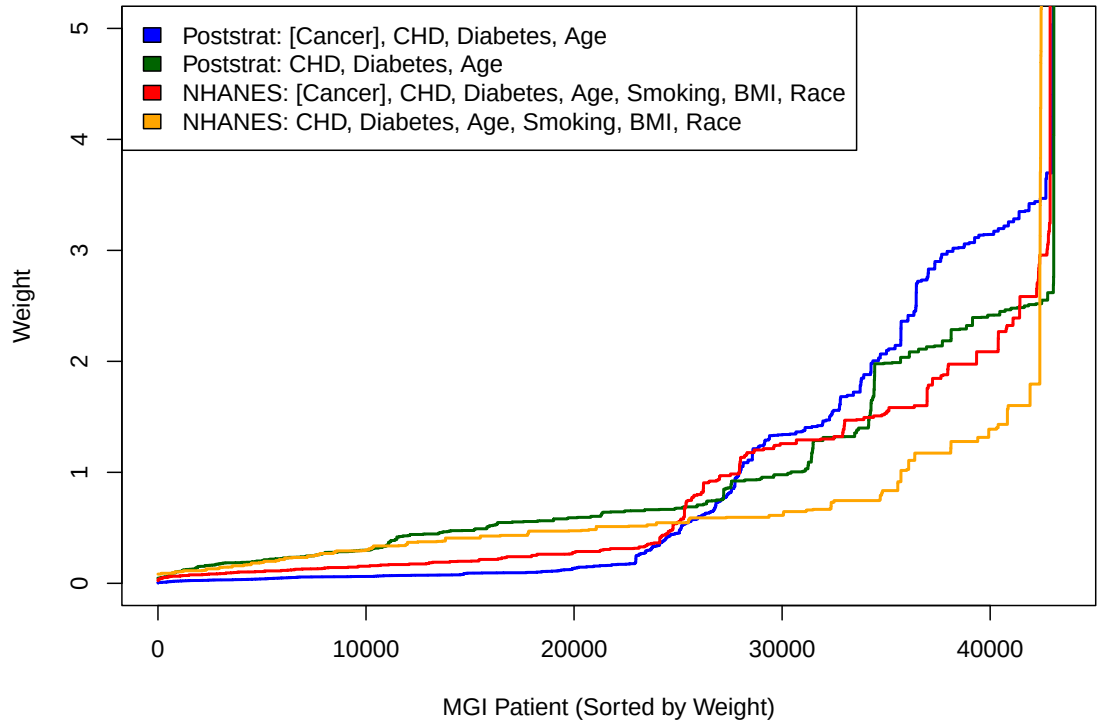
Supp. Figure B.5 shows the estimated individual-level selection weights for the MGI patients included in case study (a), where each type of weight is sorted by increasing value along the x-axis. All weights are trimmed above by 10 to ensure no one patient dominates the estimation. Although not shown, poststratification weights with and without the AMD outcome are extremely similar, indicating that AMD status may not appreciably differ between MGI and the US populations after accounting for differences due to age and other diseases. In contrast, poststratification weights with and without the cancer outcome differ substantially, reflecting the clear enrichment of MGI in terms of cancer outcomes relative to the US adult population. This same phenomenon occurs for weights estimated using NHANES.

Table B.1: Beta regression of NHANES sampling probabilities (relative to US adult population) and logistic regression for including in MGI given inclusion in NHANES or MGI

	Beta regression for sampling probabilities[1] Log-Odds Ratio, (95% CI)	**Logistic regression for inclusion in MGI** Log-Odds Ratio, (95% CI)
Age		
<40	reference	reference
40-59	0.21 (0.17, 0.25)	0.56 (0.45, 0.66)
60+	0.66 (0.62, 0.69)	0.34 (0.24, 0.45)
Diabetes diagnosis		
No	reference	reference
Yes	0.08 (0.04, 0.12)	0.48 (0.37, 0.58)
CAD diagnosis		
No	reference	reference
Yes	-0.04 (-0.02, 0.11)	0.90 (0.74, 1.07)
BMI category		
Underweight (<= 18.5)	0.15 (0.03, 0.26)	-0.32 (-0.66, 0.03)
Normal (18.6 − 24.9)	reference	reference
Overweight (25.0 − 29.9)	-0.06 (-0.10, -0.03)	-0.02 (-0.12, 0.09)
Obese (30+)	-0.06 (-0.10, -0.03)	-0.25 (-0.35, -0.14)
Smoking habits		
Never	reference	reference
Current or Former	0.03 (-0.02, 0.05)	-
Current	-	-0.14 (-0.25, -0.03)
Former	-	0.12 (0.02, 0.21)
Race/ethnicity		
Hispanic/Other/Multi	reference	reference
Non-Hispanic White	-0.91 (-0.95, -0.88)	4.40 (4.31, 4.49)

[1] Implemented with a logit link function for mean of beta distribution

Figure B.5: Estimated selection bias adjustment weights (not correcting for misclassification of disease phenotypes) for case study (a)



[1] "Poststrat" indicates poststratification weights estimated using population summary statistics. "NHANES" indicates IPW weights estimated using NHANES data. Estimated weights are shown along the y-axis, and each weight is sorted by increasing value across the MGI patients along the x-axis. For case study (b), we compute similar poststratification weights focusing on subset of unrelated MGI patients aged 50+ of recent European ancestry. Brackets in legend labels correspond to outcome variables for case studies (a) and (b).

Step 4, case study (a): Association between cancer and gender

Table B.2: Log-odds ratio point estimates and width of 95% confidence intervals for cancer-gender associations (reference = male) [Case study (a)] [1]

	Log-odds ratio	**Width of 95% confidence intervals**
Uncorrected analysis	-0.10	0.076
Selection weighting only		
Poststratification: without cancer	0.14	0.106
Poststratification: with cancer (uncorrected)	-0.18	0.093
NHANES IPW: without cancer (**from Elliot 2009**)	0.01	0.154
NHANES IPW: with cancer (**from Elliot 2009**)	-0.18	0.112
Matching only		
Matching with BMI, smoking status	-0.05	0.082
Matching without BMI, smoking status	-0.06	0.082
Approx. $D^*\|Z$ method [method 4a]		
No weighting (**from Duffy et al. (2004)**)	-0.15	0.115
Poststratification: without cancer	0.17	0.132
Poststratification: with cancer (uncorrected)	-0.18	0.094
Poststratification: with cancer (corrected)	-0.17	0.093
NHANES IPW: without cancer	0.02	0.206
NHANES IPW: with cancer (uncorrected)	-0.19	0.122
NHANES IPW: with cancer (corrected)	-0.17	0.120
Non-logistic link method [method 4c]		
No weighting (**extension of Sinnott et al. (2014)**)	-0.15	0.115
Poststratification: without cancer	0.17	0.132
Poststratification: with cancer (uncorrected)	-0.18	0.094
Poststratification: with cancer (corrected)	-0.17	0.093
NHANES IPW: without cancer	0.02	0.206
NHANES IPW: with cancer (uncorrected)	-0.19	0.122
NHANES IPW: with cancer (corrected)	-0.17	0.120

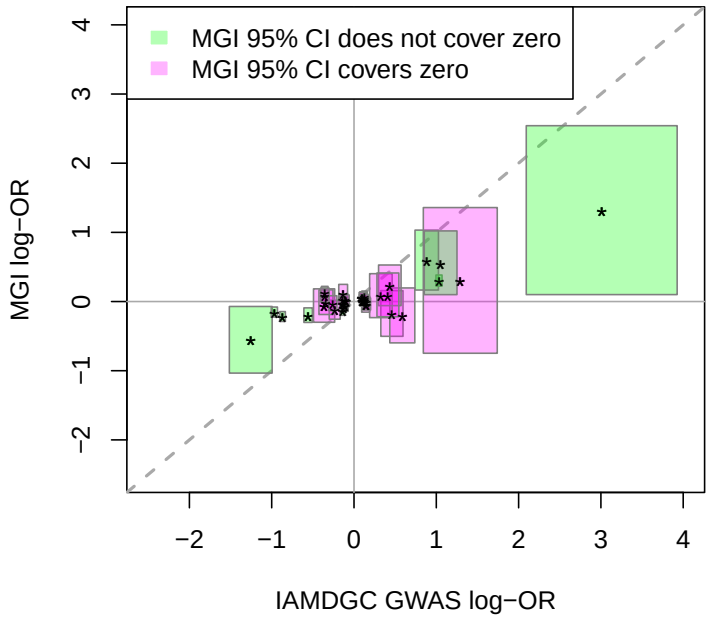
[1] For Approx. $D^*|Z$ and Non-logistic link function methods, sensitivity is estimated assuming $\tilde{r} = 25$. The former method without weighting is equivalent to the approach in Duffy et al. (2004), and the non-logistic link function implementation without weighting is a generalization of the method in Sinnott et al. (2014) to allow for covariate-related sensitivity.

Selection bias adjustment weights with and without correction for phenotype misclassification are presented in this table. The IPW weighting method without misclassification correction is equivalent to the method proposed in Elliot (2009).

Step 4, case study (b): Association between AMD and genetic loci

Supp. Figure B.6 compares estimated associations between macular degeneration diagnosis and 43 genetic loci in MGI and IAMDGC. The box around each point corresponds to the 95% confidence interval in either dataset. We observe smaller estimated effects in MGI compared to IAMDGC estimates. There are many possible explanations for this phenomenon, three of which seem most likely. Firstly, the 43 genetic loci were chosen as the top hits in the IAMDGC GWAS, so the resulting point estimates may be over-estimated following the "winner's" curse. Secondly, GWAS results for *advanced* AMD were obtained for IAMDGC data, and resulting genetic associations may be stronger for this outcome than for the MGI macular degeneration outcome including less advanced cases in addition to advanced ones. Thirdly, MGI results may be attenuated as a result of misclassification and/or selection bias.

Figure B.6: Estimated AMD log-odds for 43 SNPs using MGI and IAMDGC data. 95% confidence intervals in MGI and IAMDGC data are shown as shaded boxes.



In **Supp. Table B.3**, we provide summary metrics for the performance of the bias-correction strategies in **Supp. Figure 1** to reduce potential bias due to phenotype misclassification and selection. An abridged version of this table is included and discussed in the main paper. **Supp. Figure B.8** shows the 43 estimated log-odds ratio associations in IAMDGC and MGI (uncorrected and corrected using method 4a without selection weighting). Method 4a does not uniformly map the uncorrected MGI point estimates to the IAMDGC GWAS estimates. This plot does demonstrate, however, that the point estimates and confidence intervals can sometimes differ substantially between the various data analysis methods for a given genetic locus, and these differences here are more pronounced for extreme values of the IAMDGC GWAS θ (far left and far right values).

Table B.3: Bias-adjusted AMD log-odds ratios across 43 genetic loci [Case study (b)] [1]

	Avg. Absolute Deviation	Lin's Concordance Correlation	MAPE	Avg. Relative Standard Error
IAMDGC GWAS	0	1	0	1
Uncorrected Analysis	0.30	0.61	**0.81**	2.2
No misclassification adjustment				
Weighting without AMD	**0.25**	**0.82**	**0.85**	4.2
Weighting with AMD (Uncorrected)	0.27	0.78	0.93	6.4
Approx. $D^*\|Z$ method [method 4a]				
No weighting (**from Duffy et al. (2004)**)	**0.26**	0.76	**0.75**	3.1
Weighting without AMD	**0.26**	**0.85**	0.90	5.6
Weighting with AMD (Uncorrected)	0.28	**0.80**	0.97	7.1
Weighting with AMD (Marg. Corrected)	0.27	**0.80**	0.90	5.6
Non-logistic link method [method 4c]				
No weighting (**ext. of Sinnott et al. (2014)**)	0.28	0.73	**0.79**	3.1
Weighting without AMD	0.29	0.61	1.02	5.6
Weighting with AMD (Uncorrected)	0.28	**0.84**	1.04	7.3
Weighting with AMD (Corrected)	0.29	0.61	1.03	5.6

[1] For Approx. $D^*|Z$ and Non-logistic link function methods, sensitivity is estimated assuming $\tilde{r} = 25$. The former method without weighting is equivalent to the approach in Duffy et al. (2004), and the non-logistic link function implementation without weighting is a extension of the method in Sinnott et al. (2014) to allow for covariate-related sensitivity. Bolded values indicate the best performing methods.

Average absolute deviation = average absolute difference between MGI and IAMDGC point estimates (lower is better)

Definitions: Average absolute deviation = average absolute difference between MGI and IAMDGC point estimates (lower is better); Lin's concordance correlation = estimated concordance between MGI and IAMDGC point estimates (higher is better); MAPE (mean absolute percentage error) = average absolute difference between 1 and the ratio of MGI and IAMDGC point estimates (lower is better); Avg. relative standard error = ratio of standard errors for MGI and IAMDGC point estimates.

Figure B.7: Bias-adjusted AMD log-odds ratios across 43 genetic loci as a function of specificity [Case study (b)]

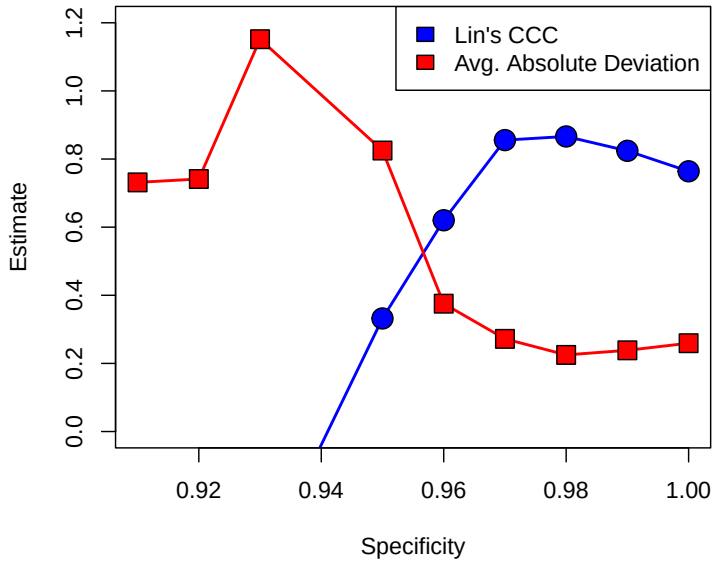
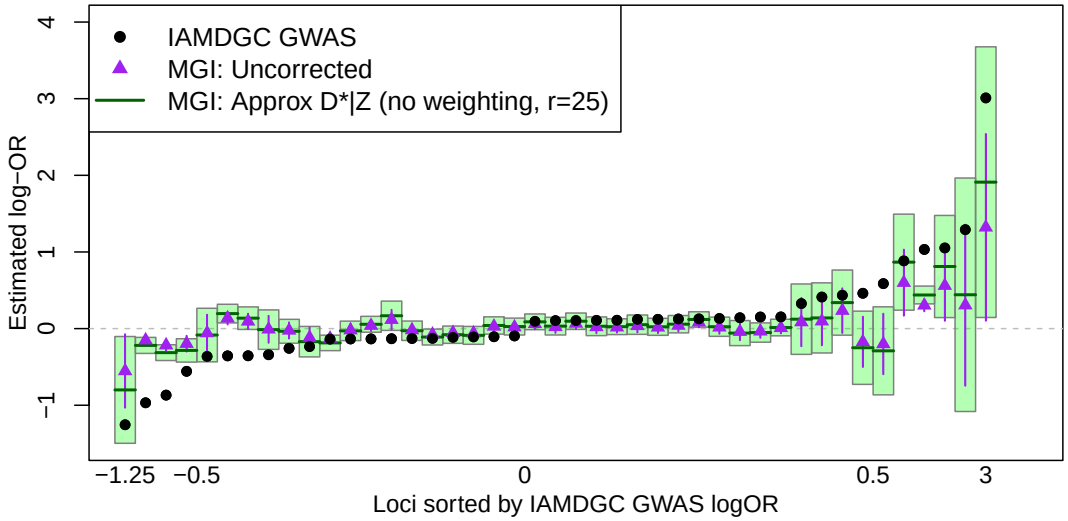


Figure B.8: Bias-adjusted AMD log-odds for estimates for 43 SNPs



C On the question of transportability

We suppose our goal is to use the internal EHR data ($S = 1$) to make inference about some well-defined target population. Suppose first that our goal is to generalize inference to some subset of the study's source population (Dahabreh and Hernán, 2019). For MGI, we might want to learn about the source population of people living in southeast Michigan as illustrated in **Figure A.1**, or we might be interested in some subset of the source population such as people living in this area with pre-existing comorbidities. More commonly, however, our goal is will be to *transport* inference based on the study "internal data" to some external target population containing some members not included in the source population. For example, we may want to use MGI data to make statements about the US adult population. In order to use data sampled non-probabilistically from the source population to learn about some other target population, additional assumptions are needed relating the source and target populations. In this section, we clarify some of the assumptions needed to transport our inference to the desired target population and provide some crude strategies for evaluation.

First, we clarify some notation. Let Population A be the study source population, and let Population B represent the target population. By definition, the study sample is some subset of Population A. Our interest is in clarifying when we can make statements about Population B using data from the EHR sample. We will assume the following:

Assumption T1: $P(D = 1|Z, Pop = A) = P(D = 1|Z, Pop = B)$.

In other words, the target relationship between D and Z is the same in the source and target populations. This is a strong assumption, but it is key. We also note that we can achieve this relationship if there exists a (possibly empty) covariate set U not containing Z such that $P(D = 1|Z, U, Pop = A) = P(D = 1|Z, U, Pop = B)$ and if $f(U|Z, Pop = A) = f(U|Z, Pop = B)$. This implies that $P(D = 1|Z, Pop = A) = P(D = 1|Z, Pop = B)$, so **Assumption T1** is satisfied. In general, we will not know whether **Assumption T1** is satisfied. This would require us to know the association between D and Z in Population B. If this were known, there would be little reason to perform the current analysis at all, since the whole goal is to make inference about this distribution. Therefore, we will view this is a key untestable assumption for transportability to be viable. We also note that we not assume that the distribution of Z is the same in the two populations. Reframed using terms from the transfer learning literature

as in Kouw and Loog (2019), we will allow for a covariate shift between the two populations in terms of Z, but we do not allow for a concept shift (i.e., the conditional distribution of interest must be the same in the two populations).

We now focus on the particular setting with no misclassification, so D is measured directly in the study sample. We construct a theoretical random indicator T defined for everyone in Populations A and B such that $P(T = 1|W, D)$ follows the same selection pattern as $P(S = 1|W, D, Pop = A)$, the true selection mechanism relating the EHR sample to the source population. We further assume that $P(T = 1|W, D, Pop = A) > 0$ for all W and D (**Assumption T2**). Let $W^\dagger$ denote the variables in W that are not included in Z (i.e., the additional predictors related to selection that are not adjusted-for in the disease model). We have that

$$P(T = 1|D, Z, Pop = A) = \int P(T = 1|D, Z, W^\dagger, Pop = A) f(W^\dagger|D, Z, Pop = A) dW^\dagger.$$

We also have that

$$\begin{aligned} &f(D|Z, T = 1, Pop = A) \propto P(T = 1|D, Z, Pop = A) f(D|Z, Pop = A) \\ &= \left[\int P(T = 1|D, Z, W^\dagger, Pop = A) f(W^\dagger|D, Z, Pop = A) dW^\dagger \right] f(D|Z, Pop = B). \end{aligned}$$

Suppose further that

$$\textbf{Assumption T3: } f(W^\dagger|D, Z, Pop = A) = f(W^\dagger|D, Z, Pop = B).$$

Then, we have that

$$\begin{aligned} &f(D|Z, T = 1, Pop = A) \propto P(T = 1|D, Z, Pop = A) f(D|Z, Pop = A) \\ &= P(T = 1|D, Z, Pop = B) f(D|Z, Pop = B). \end{aligned}$$

This in turn implies that we can re-weight a model for $D|Z, T = 1, Pop = A$ with weights defined by the inverse of the distribution of $T|D, Z, Pop = B$ to make inference about the distribution for $D|Z, Pop = B$. Rephrased, we can use the internal data to transport inference to Population B by re-weighting the data based on $f(T|D, Z, Pop = B)$.

Of course, we will not know $f(T|D, Z, Pop = B)$ in practice. We can apply various data strategies discussed in the main paper to estimate weights for selection bias adjustment as a function of Z, D, and W. Therefore, we will require a final, less rigid assumption:

> **Assumption T4:** Constructed weights are "good enough" to recover the target population θ_Z through weighted analysis of the internal sample

What is "good enough" will necessarily depend on the problem being studied, the scientific context, and the investigator's bias tolerance. Later on, we discuss some diagnostics for evaluating the plausibility of **Assumption 4**, and additional information can be found elsewhere, e.g., Degtiar and Rose (2021).

As an alternative to **Assumptions 3-4**, we could instead make the stronger assumption that $D \perp T|Z, Pop = A$, which directly implies that $f(D|Z, T = 1, Pop = A) = f(D|Z, Pop = B)$. In this case, we can transport results to the target population directly using the observed data. However, this assumption is extremely strong and unlikely to occur in practice. Therefore, we will ignore this trivial case.

To summarize, we have transportability of inference if (1) the distribution of $D|Z$ is the same in the target and source populations, (2) $P(S = 1|W, D, Pop = A) > 0$ for all W and D, (3) the distribution of un-adjusted-for factors related to selection from the source population, $W^\dagger$, given D and Z is the same in Populations A and B, and (4) the weights constructed for selection bias adjustment are "good enough" to account for systematic differences between the internal data and the target population in terms of the $D|Z$ log-odds ratio. These last two assumptions are not required if we can make the stronger assumption that selection from the

source population is independent of D given Z (but not $W^\dagger$).

We make a distinction here where our goal is to study the *association* between D and Z in Population B. Given the current specification, however, we are not attempting to make *causal* statements about the relationship between D and Z. This allows us to transport results with somewhat looser assumptions than are required in usual causal inference settings (Degtiar and Rose, 2021). However, **Assumptions T1-T3** can be loosely framed in terms of more familiar concepts in causal inference as discussed in Degtiar and Rose (2021). **Assumption T1** relates to exchangeability, where the relationship between D and Z is assumed to be the same in both populations. **Assumption T2** is related to positivity of selection, where any individual with characteristics D and W in the target population (Population B) would have non-zero probability of selection based only on D and W if they had been a member of Population A. The other assumptions relate to unmeasured confounding of the association between D and Z in terms of associations between D and $W^\dagger$. If this association given Z is the same in the two populations, then the potential impact of excluding $W^\dagger$ from the model for D is the same in the two populations. In practice, we may be able to loosen this assumption further when our primary interest is in log-odds ratios for logistic regression models. In this setting, the association between D and Z given W may be only mildly impacted by excluding $W^\dagger$ from analysis if Z and $W^\dagger$ are independent given D (Neuhaus and Jewell, 1993). In other words, we may still be able to transport our results to Population B when the conditional distribution of $W^\dagger$ is different from Population A if we have that $W^\dagger \perp Z|D$ in both populations.

C.1 Assessment of reasonableness of transportability assumptions for case studies (a) and (b)

Although the assumptions described above are not easily tested using the observed data, it is possible to crudely evaluate the reasonableness of **Assumption 4** by comparing weighted inference using the study data to summary statistics for the target population (Degtiar and Rose, 2021). We may also compare the propensity weights calculated for internal and external individual-level data.

For case study (a), we constructed many weights for selection bias adjustment. Here, we will consider four weight formulations as shown in **Figure B.5**, where poststratification weights are constructed using summary statistics from the US census and SEER. For case study (b), we constructed poststratification weights using NIH National Eye Institute and US census summary statistics. For these two case studies, we are interested in exploring how the disease prevalence by age calculated using the weighted study data compare to the target population for each study. For this analysis, we ignore the potential impact of phenotype misclassification.

C.1.1 Positivity assumption: comparison of IPW scores calculated for internal (MGI) and external (NHANES) samples for case study (a)

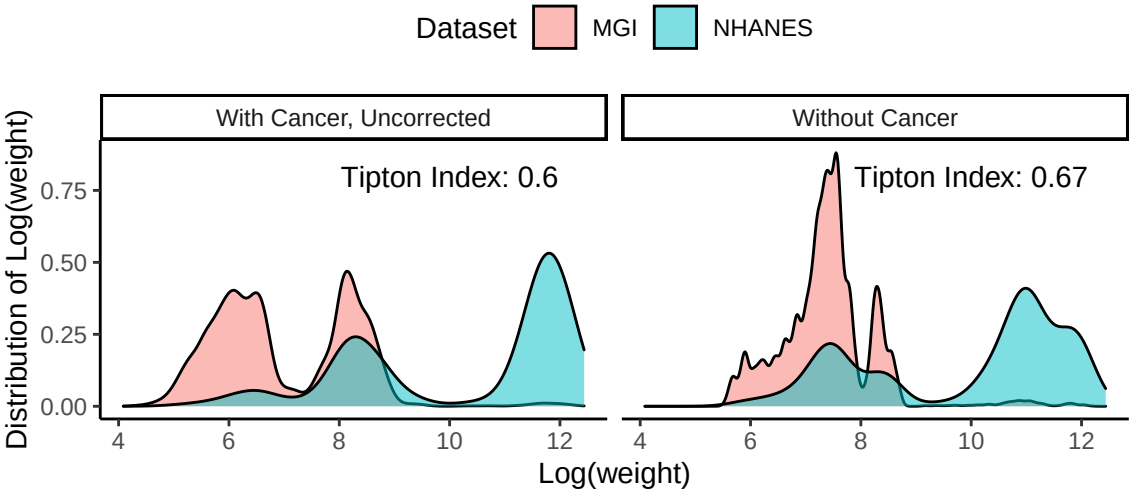
When a probability sample from the target population is available, one strategy for evaluating the reasonableness of transportability is to calculate the value of the IPW weights for both the internal and external samples (Degtiar and Rose, 2021). In our setting, IPW weights were constructed for MGI using data from external probability sample NHANES. While our target population is the US adult population and not the NHANES sample, the similarity in propensity scores between these two samples may give some insight into the reasonableness of transportability between related populations.

We calculate the IPW propensity weights for both the MGI and NHANES samples and plot the distribution of resulting (before sum-to-one scaling) weights for each sample in **Figure C.1**. We find that the propensity scores tend to be higher for NHANES than for MGI individuals. This supports the **reasonableness of the positivity assumption**, where all NHANES individuals have a nonzero probability of being included in MGI based on the characteristics used

for weight construction.

Tipton (2014) proposed a metric for comparing the similarity of the distributions of propensity scores. This index is defined as $\textit{Tipton Index} = \sum_b \sqrt{p_{internal,b} \times p_{external,b}}$, where b indexes bins of possible propensity score values and $p_{internal,b}$ and $p_{external,b}$ denote the proportion of internal and external sample scores that fall in each bin, respectively. This index ranges between 0 and 1, where values of 1 indicate strong potential for generalizability between populations. Using fine bins of width 5 between 0 and 300,000, the Tipton indices calculated for the IPW weights with and without including cancer diagnosis are 0.60 and 0.67, respectively.

Figure C.1: Distributions of unscaled IPW propensity scores (without correction for phenotype misclassification) for case study (a) in MGI and NHANES [1]



[1] Tipton Indices represent the similarity in these propensity scores between MGI and NHANES. Scores range between 0 and 1, with 1 indicating strong generalizability. Indices were calculated using the empirical distribution of (unscaled) propensity scores with bins defined between 0 and 300,000 with bin widths of 5.

C.1.2 Comparing weighted and unweighted estimates: Comparison of weighted and unweighted summary statistics in MGI with target population values for both case studies

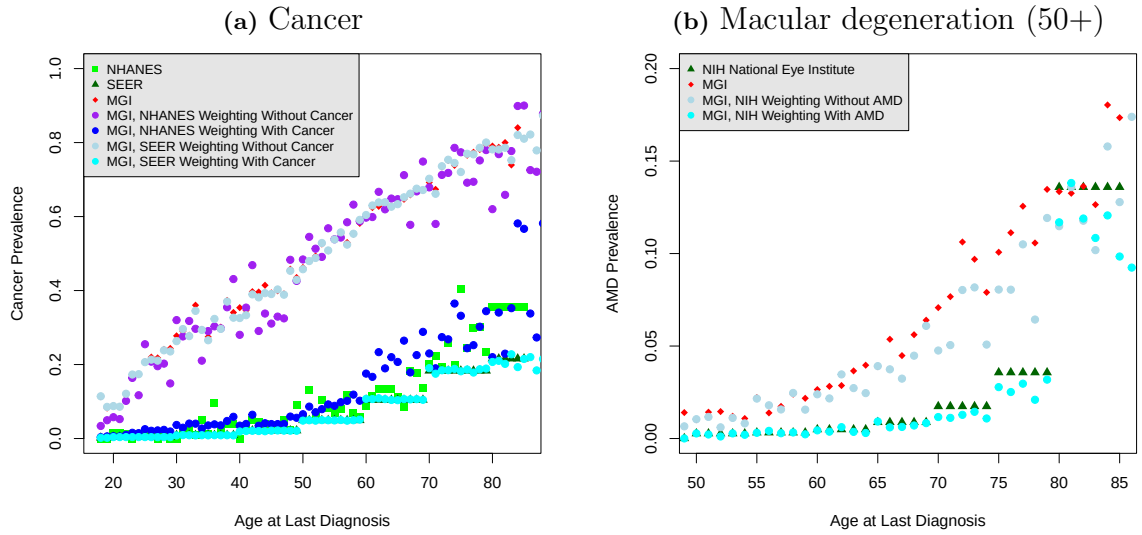
Figure C.2 provides the estimated disease prevalence by age using MGI data and calculated using various weight construction strategies. For (adjusted) cancer prevalence, we see strong bias in estimates of the population cancer prevalence by age using raw MGI data and using weighted MGI data with weights constructed ignoring the relationship between selection and cancer. When we re-weight the MGI data using weights that do condition on cancer diagnosis, the resulting estimates for the cancer prevalence by age closely resemble the truth in the target population. This is not surprising in the case of weights based on SEER data, since the disease prevalence by age in the target population was directly used to construct the weights. For (adjusted) AMD prevalence, we see comparatively less difference between MGI data results and the target disease rates. However, weights constructed adjusted for both age and AMD status better recover the target population characteristics.

We can also compare other weighted and unweighted estimands based on MGI data with those calculated using NHANES and those based on summary statistics compiled for the US adult population. Results are shown in **Table C.1**. Generally, IPW weighted estimates based on MGI data tended to be closer to the target population and NHANES estimates than estimates from crude analysis. Estimates using poststratification weights, however, often did a poor job at recovering population quantities. Two major factors may contribute. Firstly, the

IPW weights were constructed using individual-level data that allowed the joint distribution of key variables of interest to be estimated. In contrast, the poststratification weights constructed in this study relied only on marginal distributions of summary statistics. Secondly, the IPW weights included additional predictors including BMI, smoking status, and race/ethnicity in the modeling, while these variables were not accounted for directly in the poststratification weight construction. Since smoking status and BMI are both related to cancer diagnosis even after adjusting for age, we may expect inference based on the NHANES IPW weights to be more trustworthy than inference based on the poststratified weights.

We may also use individual-level data on cancer diagnosis and other factors available through NHANES to calculate the association between cancer and many patient characteristics in the target population (through weighted analysis of NHANES data using provided NHANES weights). We can then compare estimates for the target population with estimates from weighted and unweighted MGI data. **Figure C.3** shows the estimated probability of having a cancer diagnosis adjusting for age, BMI, race/ethnicity, smoking status, diabetes diagnosis, CAD diagnosis, and gender in the target population and using MGI data. These predictions are calculated for each MGI patient. Probabilities using raw MGI data are far from the target population probabilities, and results using weights constructed ignoring cancer status do not correct these results. Weighted analysis of MGI data based on weights conditioning on cancer status perform much better, where analysis with weights constructed using NHANES data unsurprisingly do a good job at reproducing "true" associations also calculated using NHANES.

Figure C.2: Weighted and unweighted disease prevalences in study sample and target populations[1]



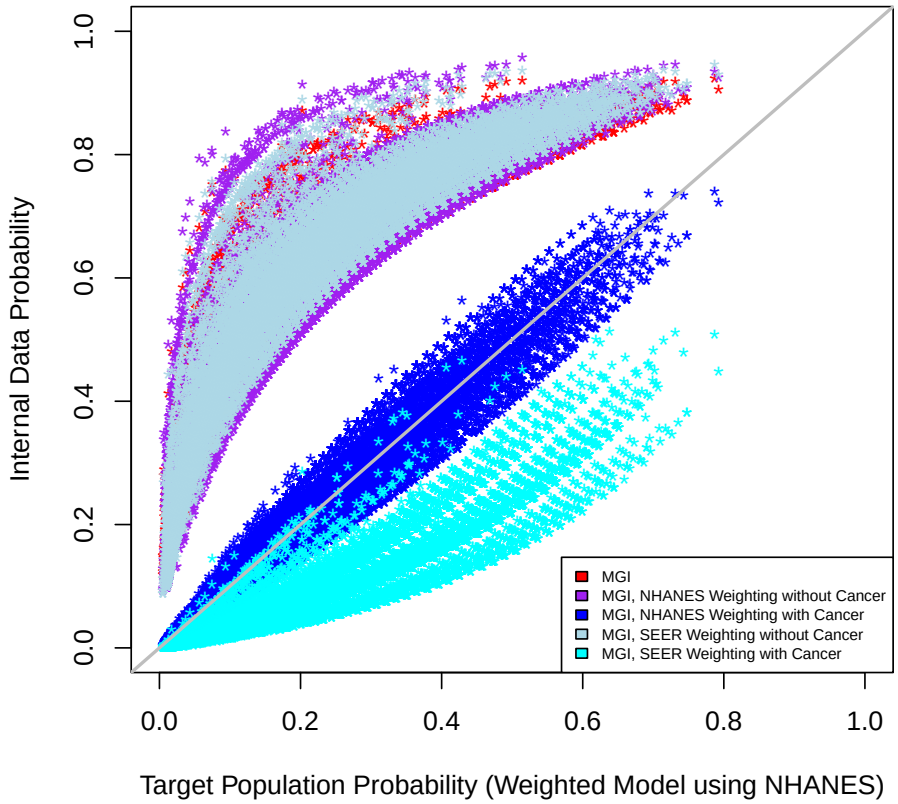
[1] For case study (b), we define our target population as the subset of the US adult population aged 50+ of recent European descent. For corresponding weight construction, however, we used summary statistics for the entire US adult population aged 50+. An underlying assumption is that the summary statistics used to develop the poststratification weights (i.e., the age distribution and the prevalence of AMD by age) do not differ appreciably by ancestry.

Table C.1: Comparison of weighted and unweighted MGI summary statistics with target population estimates

Analysis Type	**Crude**	**MGI Weighted Analysis**			**Target Population**	
Analysis Data	**MGI**	**MGI**	**MGI**	**MGI**	**NHANES**[1]	**US Adults**[2]
Type of Weights	**None**	**Case (a) IPW**	**Case (a) Poststrat.**	**Case (b) Poststrat.**	**Provided Weights**	**None**
Age, Mean	56.8	47.2	44.0	44.0	47.3	40-44
Female, %	52.4	53.6	55.9	54.7	51.8	50.9
Prior Cancer Diagnosis, %	54.4	10.1	39.6	4.5	11.0	39.5[3]
Prior CHD Diagnosis, %	16.0	4.3	4.5	4.7	4.1	12.1[3]
Prior AMD Diagnosis, %	4.2	1.9	2.0	0.71	-	2.1[3]
Obese, %	41.3	41.5	33.6	34.2	41.9	34.4
Current Smokers, %	17.1	21.7	19.0	18.3	17.1	17.1
Non-Hispanic White, %	97.8	56.8	97.6	28.8	62.0	69.1

[1]NHANES results have been weighted using the weights provided by NHANES, while the results for MGI with IPW weighting use constructed IPW weights or poststratification weights (constructed for either case study (a) or (b)) that include each disease diagnosis without correcting for phenotype misclassification. Age-related macular degeneration information were not collected for NHANES in the years under study.
[2] See **Table A.2** for sources of US adult summary statistics.
[3] US Adult summary statistics correspond to lifetime disease prevalence, not proportion of adults with prior diagnosis. Therefore, they are an upper bound on reasonable values for percent of adults with prior diagnosis.

Figure C.3: Weighted and unweighted probability of having cancer estimated using MGI and NHANES Data [1]



[1] Predicted probabilities from models adjusting for age, BMI, race/ethnicity, smoking status, diabetes diagnosis, CAD diagnosis, and gender. Results using NHANES data use a model weighted with the provided NHANES weights and can be interpreted as representative of the target population of interest.

References

Lauren J Beesley and Bhramar Mukherjee. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics*, 78(1):214–226, 2022.

Issa J Dahabreh and Miguel A Hernán. Extending inferences from a randomized trial to a target population. *European Journal of Epidemiology*, 34(8):719–722, 2019. ISSN 1573-7284. doi: 10.1007/s10654-019-00533-2. URL `https://doi.org/10.1007/s10654-019-00533-2`.

Irina Degtiar and Sherri Rose. A Review of Generalizability and Transportability. *arXiv*, pages 1–30, 2021.

Joshua C. Denny, Marylyn D. Ritchie, Melissa A. Basford, Jill M. Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R. Masys, Dan M. Roden, and Dana C. Crawford. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

S. W. Duffy, J. Warwick, A. R.W. Williams, H. Keshavarz, F. Kaffashian, T. E. Rohan, F. Nili, and A. Sadeghi-Hassanabadi. A simple model for potential use with a misclassified binary outcome in epidemiology. *Journal of Epidemiology and Community Health*, 58(8):712–717, 2004.

Michael R Elliot. Combining Data from Probability and Non- Probability Samples Using Pseudo-Weights. *Survey Practice*, 2(3):1–7, 2009.

Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv*, pages 1–42, 2019.

John M Neuhaus and Nicholas P Jewell. A Geometric Approach to Assess Bias Due to Omitted Covariates in Generalized Linear Models Author. *Biometrika*, 80(4):807–815, 1993.

Jennifer A. Sinnott, Wei Dai, Katherine P. Liao, Stanley Y. Shaw, Ashwin N. Ananthakrishnan, Vivian S. Gainer, Elizabeth W. Karlson, Susanne Churchill, Peter Szolovits, Shawn Murphy, Isaac Kohane, Robert Plenge, and Tianxi Cai. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Human Genetics*, 133(11):1369–1382, 2014.

Elizabeth Tipton. How Generalizable Is Your Experiment? An Index for Comparing Experimental Samples and Populations. *Journal of Educational and Behavioral Statistics*, 39(6): 478–501, 2014.