EXTENDING INFERENCES FROM SAMPLE TO TARGET POPULATIONS: ON THE GENERALIZABILITY OF A REAL-WORLD CLINICO-GENOMIC DATABASE NON-SMALL CELL LUNG CANCER COHORT

Darren S. Thomas[1], Simon Collin[1], Luis Berrocal-Almanza[1], Heide Stirnadel-Farrant[1], Yiduo Zhang[1], Ping Sun[1]

[1] **Centre of Oncology Data Excellence, Oncology Business Unit, AstraZeneca**

Disclosure: All authors are current employees of AstraZeneca at the time of writing and may hold stock options

ABSTRACT

The representativeness of Real-World Data is generally assumed, but findings will rarely generalise to the target population when the potential outcomes under treatment are influenced by variables causative of selection into a study. Using a de-identified nationwide US Clinico-Genomic Database (CGDB) Non-Small Cell Lung Cancer (NSCLC) cohort as an example of collider bias, we assess its representativeness in relation to two target populations: a superset of all NSCLC patients in the Flatiron Health network and Surveillance, Epidemiology and End Results cancer registrations. Informed by causal Directed Acyclic Graphs, the CGDB cohort was weighted to be representative of the target populations and real-world overall survival (rwOS) was re-estimated. Despite Standardised Differences suggesting differences in individual covariates between sample and target populations, the conditional distributions of selection were alike, and indices of generalizability were very high (≥ 0.96 on a proportional scale of 0—1). Estimates of rwOS in a population weighted to be representative did not differ from naïve estimates in the unweighted cohort. The Tipton generalizability index provides a quantitative assessment of the generalizability of findings that can be used to determine the influence of selection biases.

BACKGROUND

Rarely is a study sample randomly drawn from the target population. Real-world data are oftentimes a convenience sample of patients receiving healthcare at centres using a proprietary Electronic Medical Record system, covered by a defined insurance policy or enrolled in a disease registry (1,2). When the mechanisms underlying the selection into a study sample affect the potential outcomes under treatment, the findings will rarely generalise to the target population as the Sample Average Treatment Effect (SATE) is expected to diverge from the Population Average Treatment Effect (PATE) (3–5).

Our motivating example is in the study of the Flatiron Health-Foundation Medicine Clinico-Genomic Database (CGDB) Non-Small Cell Lung Cancer (NSCLC) cohort (6). The database represents the intersection of NSCLC patients treated within the Flatiron Health Research Network who i) underwent human technology-assisted chart abstraction and ii) whose tumour biopsy was submitted for Next Generation Sequencing (NGS). NGS, at its introduction, was not widely adopted by clinical guidelines or covered by insurers, which could introduce a selection bias (7). A recent study suggested only slight differences in the distributions of patient characteristics between the CGDB and US cancer registrations (8), but judgements on the generalizability of study findings are necessarily subjective and do not quantify the potential bias introduced.

We outline a quantitative alternative to assess the representativeness of the sample population based on the Tipton generalizability index (9) and detail a solution to reweight samples using Inverse Probability Weighting of Marginal Structural Models where necessary (10). These methods have been used to extend randomised-controlled trial findings to a pragmatic setting (11,12). The Tipton generalizability Index is a measure of how closely a sample population approximates a sample randomly drawn from a defined target population (9). The metric is useful in providing a quantitative basis to assess the generalizability of inferences from a sample population to the target population and, where unlikeness exists, to indicate the extent to which propensity score-based adjustment may improve our inferences. Inverse Probability Weighting in the context of selection involves cloning selected subjects so that they account for themselves and their unselected counterparts with equal probability of selection (3,10). Within this weighted cohort, selection is independent of the outcome. For data on the underlying US target population, we used two sources: the Flatiron Health Machine Learning Extracted (ML-E) database, which includes the super-sample of all patients treated within the Flatiron Health network, not just a nested sample that have undergone human chart abstraction and NGS; and the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) cancer registrations, which cover 34.6% of the US population.

 # METHODS

Sample population

Flatiron Health-Foundation Medicine Clinico-Genomic Database (2011—2021)

The Flatiron Health-Foundation Medicine CGDB is a nationwide US de-identified database that links via deterministic matching Flatiron Health Research Network Electronic Medical Records from ~ 280 US cancer clinics (~800 sites of care) with genomic data derived from Foundation Medicine Comprehensive Genomic Profiling NGS tests (6). Retrospective longitudinal clinical data were curated via technology-enabled abstraction of variables from structured and unstructured data stored in Electronic Medical Records. Our sample was obtained from records covering the period 1 Jan 2011–31 Dec 2021. NSCLC case inclusion criteria were: a lung cancer diagnosis (ICD-9-CM 162.x or ICD-10-CM C34.x or C39.9); confirmation of NSCLC based on unstructured data; two visits on or after 1 Jan 2011; and a Foundation Medicine NGS test on a biopsy with pathologist-confirmed histology that is consistent with the abstracted tumour type and taken within 30 days before the chart-confirmed date of initial diagnosis (or anytime afterwards).

Target populations

Flatiron Health Machine Learning Extracted Database (2011—2021)

The Flatiron Health Machine Learning Extracted (ML-E) database is a de-identified database of all NSCLC patients receiving care within the Flatiron Health Research Network. The ML-E database uses Machine Learning to extract variables from unstructured clinical data, without confirmation by chart abstraction (13). The target population was obtained from records covering the period 1 Jan 2011–31 Dec 2021 NSCLC case inclusion criteria were the same as for the CGDB cohort except the requirement for a NGS test.

Surveillance, Epidemiology and End Results *Incidence Data (1975-2016)*

The Surveillance, Epidemiology and End Results (SEER) program records cross-sectional data on the demographic and clinical characteristics of cancer registrations collected by state-level population-based cancer registries covering 34.6% of the US population [https://seer.cancer.gov/data]. From publicly available de-identified patient-level data covering the period 1 Jan 2011-31 Dec 2016, we generated a target population of patients with histologically confirmed NSCLC defined by ICD codes (ICD-9-CM 162.x or ICD-10-CM C34x or C39.9) and relevant ICD-O-2 histology (Supplementary Table S1).

Statistical analyses

Variable selection

In the language of Directed Acyclic Graphs (14), a type-1 selection bias results from conditioning on a common effect of both the treatment (or cause of treatment) and the outcome (or cause of the outcome) (3,4). Conditioning on a collider variable by studying a restricted cohort can induce a spurious association between its parents (treatment and outcome) even if they are marginally independent. This collider bias can inflate, attenuate or even reverse-sign associations (15,16).

We shortlisted the following variables as causative of selection under the theorems of d-separation (Figure 1): age [at diagnosis] (continuous)), gender (Female, Male), race (Black, Other, White), stage (American Joint Committee on Cancer (AJCC) 7/8th edition I-IV for CGDB & ML-E or SEER stages (Localised, Regional, and Distant) for CGDB & SEER), and histological classification (Non-squamous carcinoma, Squamous carcinoma). While the AJCC stage was generally unrecorded for SEER patients, the SEER stages Localised, Regional, and Distant are broadly equivalent to AJCC stages I-II, III, and IV, respectively. Gender was unrecorded for SEER therefore we used biological sex as a proxy.

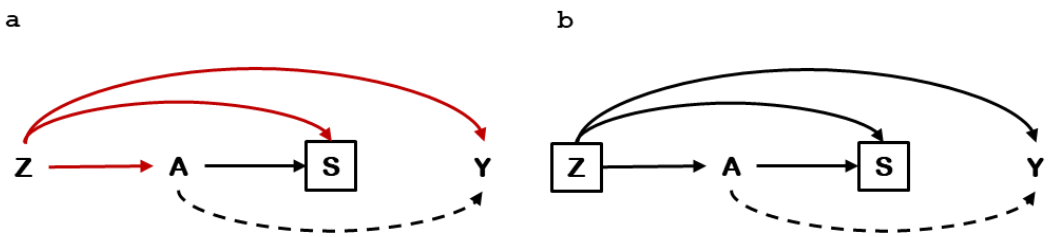


Figure 1: Directed Acyclic Graphs highlighting the open (red) and closed (black) backdoor paths in the (a) unweighted and (b) weighted, where A represents the exposure of interest (treatment or diagnosis), Y the outcome (survival), Z a set of baseline variables (age, gender, race, stage, histological classification), and S a selection node (having undergone NGS). A [box] represents conditioning, either through experimental design (restriction) or statistical analysis (weighting). Conditioning reverses the operation of a path (closed becomes open, open becomes closed). By default, confounding and mediating paths are open while collider paths are closed. The overall goal is to isolate the main causal effect of A > Y (dashed line) by closing all non-causal backdoor paths (A < Z > Y & A > [S] < Z > Y). By studying only patients who underwent NGS (*i.e.*, conditioning on S=1), we open these non-causal backdoor paths. Conditioning on [Z] (by weighting) closes these paths.

Missing data

We excluded all patients from the CGDB & ML-E/SEER who did not have complete data (Complete Case Analysis).

Inverse Probability Weighting

We estimated Stabilised Inverse Probability Weights of selection using a logistic binomial model on the outcome indicating selection into the study sample (1 for CGDB, 0 for ML-E or SEER (individually)), conditional on a set of baseline variables (Z) causative of selection. Age [at diagnosis] was modelled with a natural cubic spline with 3 knots. All other variables were categorical. Weights for $_i th$ CGDB patient (w_i) were calculated as the marginal probability of selection into the target population ($Pr[S = 1]$) divided by the probability of selection conditional on a set of variables Z ($Pr[S = 1 \mid Z = z]$) (Equation 1). We verified the positivity assumption that the conditional probability of being selected is > 0% for all levels of Z.

Equation 1: Stabilised Inverse Probability Weights for the $_i th$ CGDB patient (w_i) where S denotes selection into the study sample (1 for the CGDB sample population, 0 for the target population (ML-E/SEER individually)) and Z a set of baseline variables (age, gender, race, stage, histological classification) causative of selection.

$$w_i = \frac{Pr[S = 1]}{Pr[S = 1 \mid Z = z]}$$

Baseline characteristics

We tabulated the set of baseline characteristics (Z) of the unweighted and weighted sample populations and calculated the Absolute Standardised Difference (ASD) as a metric to assess the differences in these distributions relative to those in the target population. A rule-of-thumb for conditional exchangeability is an ASD < 0.1 for all variables (17).

Generalizability

The Tipton generalizability Index is a measure of the degree of similarity in the conditional distributions of selection within the sample and target populations (9). The index (B) is calculated by integrating the product of the kernel densities for the conditional probabilities of selection in the sample ($f_s(s)$) and target populations ($f_p(s)$) (Equation 2). Importantly, the index requires no distributional assumptions. Bounded between 0 and 1 from distinctly unrepresentative to perfectly representative, an index of ≥ 0.90 is indicative of high generalizability (9). Indices lower than 0.9 may suggests the inappropriateness of generalisations without remedial action via weighting. We used the open-source implementation available as the R package {generalize} (18).

Equation 2: Calculation of the Tipton Generalizability Index (B) where $f_s(s)$ denotes the conditional distribution of selection for the sample, $f_p(s)$ the conditional distribution of selection for the target population, and ds a kernel density bandwidth defined using Silverman's rule-of-thumb.

$$B = \int_{-\infty}^{\infty} \sqrt{f_s(s) f_p(s)}\, ds$$

Real-world overall survival

We estimated real-world overall survival (rwOS) via Kaplan-Meier estimation of the unweighted and weighted CGDB cohorts (each for stages I—IV, stages I—II, stages III—IV). Time zero was the date of diagnosis for analyses of stages I—IV & I—II and the date of initiating first-line systemic therapy for the stages III—IV. The endpoint was mortality or right censorship on the latest evidence of clinical activity (oral medication, clinical visit or genetic report date). Because the day of mortality was withheld, the day was imputed as the 15th of the same month or, if there was evidence of clinical activity after this, to the last calendar day of that month. To circumvent a left-truncation bias caused by eligibility potentially occurring after time zero, entry into the analysis was delayed until the day of earliest eligibility (the latest of their second visit or on their genetic report date) relative to the index date (risk-set adjustment) (19,20). For the weighted cohorts we used robust variance estimation. A weighted log-rank test was used to compare survival times in the unweighted and weighted cohorts, using a prespecified α of 0.00833 (α of 0.05 Bonferroni adjusted for six comparisons).

Sensitivity analysis

We undertook a sensitivity analysis wherein we compared baseline characteristics of stage I—IV CGDB patients restricted to be in harmony with the study period of SEER cancer registrations (both diagnoses during 1 Jan 2011—31 Dec 2016).

Software

All analyses were undertaken with R version 4.1 (21). A JSON with all dependencies and a PDF with all package citations are made available as supplements.

RESULTS

Sample and target populations

There were 17,230, 199,278, and 240,943 patients with lung cancer histologically confirmed to be of non-small cell origin in the CGDB, ML-E, and SEER databases, respectively (Table 1). Notably, relative to SEER, Flatiron Health data (CGDB & ML-E) had higher rates of missingness in capturing race (9.0% for CGDB, 9.6% for ML-E, 0.3% for SEER) and stage (3.7% for CGDB, 5.9% for ML-E, 2.5% for SEER, whereas SEER had higher rates of missingness for histological classification (4.0% for CGDB, 4.9% for ML-E, 15.7% for SEER). SEER staging is notably less granular than the AJCC system, however. The respective data for cohorts of stage I—II & III-IV are presented in Supplementary Tables S2 & S3. After exclusion of missing data, there were 14,545, 162,577, and 198,741 remaining in the CGDB, ML-E, and SEER cohorts.

Table 1: Baseline characteristics of CGDB, ML-E, and SEER cohorts (stages I–IV), before exclusion of missing data

Characteristic	CGDB (n 17,230)	ML-E (n 199,278)	SEER (n 240,943)
Age at diagnosis			
Median (Interquartile range)	68.0 (61.0, 75.0)	70.0 (63.0, 76.0)	70.0 (62.0, 77.0)
Unknown	0 (0.0%)	1 (0.0%)	0 (0.0%)
Gender			
Female	8,793 (51.0%)	98,928 (49.6%)	114,973 (47.7%)
Male	8,437 (49.0%)	100,334 (50.3%)	125,970 (52.3%)
Unknown	0 (0.0%)	16 (0.0%)	0 (0.0%)
Race			
Black	1,140 (6.6%)	15,898 (8.0%)	28,263 (11.7%)
Other	2,944 (17.1%)	29,901 (15.0%)	18,023 (7.5%)
White	11,600 (67.3%)	134,421 (67.5%)	193,988 (80.5%)
Unknown	1,546 (9.0%)	19,058 (9.6%)	669 (0.3%)
AJCC stage			
Stage I	1,849 (10.7%)	44,729 (22.4%)	–
Stage II	1,290 (7.5%)	18,164 (9.1%)	–
Stage III	3,938 (22.9%)	43,747 (22.0%)	–
Stage IV	9,518 (55.2%)	80,883 (40.6%)	–
Unknown	635 (3.7%)	11,755 (5.9%)	–
SEER stage[1]			
Localized	3,139 (18.2%)	62,893 (31.6%)	55,754 (23.1%)
Regional	3,938 (22.9%)	43,747 (22.0%)	56,664 (23.5%)
Distant	9,518 (55.2%)	80,883 (40.6%)	122,410 (50.8%)
Unknown	635 (3.7%)	11,755 (5.9%)	6,115 (2.5%)
Histological classification			
Non-squamous cell carcinoma	12,814 (74.4%)	131,339 (65.9%)	144,502 (60.0%)
Squamous cell carcinoma	3,727 (21.6%)	58,273 (29.2%)	58,637 (24.3%)
Unknown	689 (4.0%)	9,666 (4.9%)	37,804 (15.7%)

Abbreviations: CGDB, Clinico-Genomic Database [cohort]; ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]; AJCC, American Joint Committee on Cancer

[1] SEER stage approximated for CGDB & ML-E cohorts

 # ML-E cohort as target population

The baseline characteristics of CGDB patients I the unweighted and weighted in relation to ML-E are presented in Table 2. Based on the ASDs (Figure 2), CGDB patients were younger (median 68 years (IQR 61—75) vs. 70 years (IQR 62—75) (ASD 0.13)) and had a different distributions of AJCC stage (ASD 0.38) and histological classification (ASD 0.18) than all patients who underwent ML-E. Variables in the weighted cohort had ASDs between 0.01—0.03. Within the subset of cancer stages I—II (Supplementary Table S4), CGDB patients differed in the distribution in AJCC stage (ASD 0.26) and histological classification (ASD 0.18); differences that were corrected via weighting (ASD 0.01). Within the subset of stages III—IV (Supplementary Table S5), CGDB patients differed in the distribution in age (median 67 years (IQR 60—74) vs. 69 years (IQR 61—75) (ASD 0.13)), AJCC stage (ASD 0.24), and histological classification (ASD 0.20), which was bought under closer alignment by weighting (all ASDs 0.00—0.03).

Table 2: Baseline characteristics of CGDB & ML-E cohorts (stages I—IV), in the unweighted and weighted

	Unweighted			Weighted		
Characteristic	**CGDB (n 14,545)**	**ML-E (n 162,577)**	**ASD**	**CGDB (n 14,525)**	**ML-E (n 162,577)**	**ASD**
Age at diagnosis[1]	68.0 (61.0, 75.0)	70.0 (62.0,75.0)	0.13	69.0 (62.0, 75.0)	70.0 (62.0,75.0)	0.02
Gender			0.03			0.01
Female	7,449 (51.2%)	81,106 (49.9%)		7,208 (49.6%)	81,106 (49.9%)	
Male	7,096 (48.8%)	81,471 (50.1%)		7,317 (50.4%)	81,471 (50.1%)	
Race			0.07			0.01
Black	1,053 (7.2%)	14,188 (8.7%)		1,248 (8.6%)	14,188 (8.7%)	
Other	2,713 (18.7%)	27,015 (16.6%)		2,467 (17.0%)	27,015 (16.6%)	
White	10,779 (74.1%)	121,374 (74.7%)		10,809 (74.4%)	121,374 (74.7%)	
AJCC stage			0.38			0.03
Stage I	1,658 (11.4%)	39,761 (24.5%)		3,384 (23.3%)	39,761 (24.5%)	
Stage II	1,168 (8.0%)	16,270 (10.0%)		1,415 (9.7%)	16,270 (10.0%)	
Stage III	3,474 (23.9%)	38,079 (23.4%)		3,401 (23.4%)	38,079 (23.4%)	
Stage IV	8,245 (56.7%)	68,467 (42.1%)		6,324 (43.5%)	68,467 (42.1%)	
Histological classification			0.18			0.02
Non-squamous cell carcinoma	11,335 (77.9%)	113,642 (69.9%)		10,272 (70.7%)	113,642 (69.9%)	
Squamous cell carcinoma	3,210 (22.1%)	48,935 (30.1%)		4,253 (29.3%)	48,935 (30.1%)	

Abbreviations: CGDB, Clinico-Genomic Database [cohort]; ML-E, Machine Learning-Extracted [cohort]; ASD, Absolute Standardized Difference; AJCC, American Joint Committee on Cancer

[1] Median (Interquartile range)

 # SEER cohort as target population

The baseline characteristics of CGDB patients and ASDs in the unweighted and weighted in relation to SEER are presented in Table 3. Based on the ASDs (Figure 2), CGDB patients were younger (median 68 years (IQR 61—75) vs. 70 years (IQR 62—77 (ASD 0.20)) and had different distributions of race (ASD 0.35), SEER stage (ASD 0.15), and histological classification (ASD 0.15) than SEER cancer registrations. These became aligned in the weighted population with ASDs of 0.00—0.06. Within the subset of stages I—II (Supplementary Table S6), differences in the distributions of age (ASD 0.16), race (ASD 0.36), and histological classification (ASD 0.10) in CGDB patients were balanced by weighting with ASDs ranging from 0.00—0.06. Stages III—IV (Supplementary Table S7) differed in age

(ASD 0.24), race (ASD 0.35), SEER stage (ASD 0.18), and histological classification (ASD 0.20) in the unweighted cohorts but were balanced in the weighted (ASDs 0.01—0.07).

Table 3: Baseline characteristics of CGDB & SEER cohorts (stages I—IV), in the unweighted and weighted

	Unweighted			Weighted		
Characteristic	**CGDB (n 14,545)**	**SEER (n 198,741)**	**ASD**	**CGDB (n 14,461)**	**SEER (n 198,741)**	**ASD**
Age at diagnosis[1]	68.0 (61.0, 75.0)	70.0 (62.0, 77.0)	0.20	70.0 (63.0, 76.0)	70.0 (62.0, 77.0)	0.06
Gender			0.06			0.00
Female	7,449 (51.2%)	96,221 (48.4%)		7,007 (48.5%)	96,221 (48.4%)	
Male	7,096 (48.8%)	102,520 (51.6%)		7,454 (51.5%)	102,520 (51.6%)	
Race			0.35			0.04
Black	1,053 (7.2%)	23,042 (11.6%)		1,655 (11.4%)	23,042 (11.6%)	
Other	2,713 (18.7%)	15,311 (7.7%)		1,262 (8.7%)	15,311 (7.7%)	
White	10,779 (74.1%)	160,388 (80.7%)		11,544 (79.8%)	160,388 (80.7%)	
SEER stage			0.15			0.01
Localised[2]	2,826 (19.4%)	49,685 (25.0%)		3,569 (24.7%)	49,685 (25.0%)	
Regional[2]	3,474 (23.9%)	49,014 (24.7%)		3,522 (24.4%)	49,014 (24.7%)	
Distant[2]	8,245 (56.7%)	100,042 (50.3%)		7,370 (51.0%)	100,042 (50.3%)	
Histological classification			0.15			0.02
Non-squamous cell carcinoma	11,335 (77.9%)	141,547 (71.2%)		10,450 (72.3%)	141,547 (71.2%)	
Squamous cell carcinoma	3,210 (22.1%)	57,194 (28.8%)		4,011 (27.7%)	57,194 (28.8%)	

Abbreviations: CGDB, Clinico-Genomic Database [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]; ASD, Absolute Standardized Difference

[1] Median (Interquartile range)

[2] Localized broadly equivalent to AJCC stages I—II, Regional to stage III, and Distant to stage IV

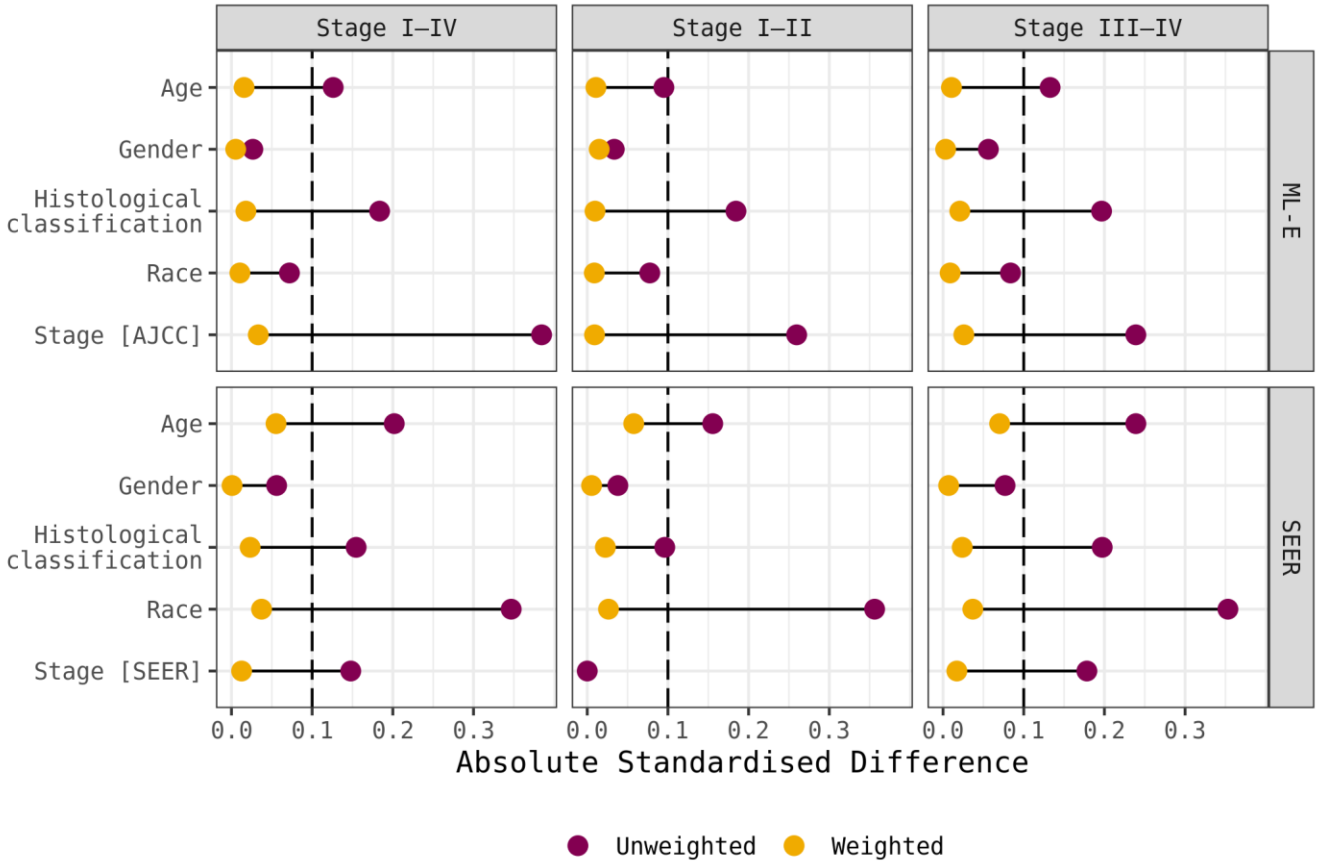


Figure 2: Absolute Standardised Differences summarising the differences in individual covariates between the sample and target populations. An Absolute Standardised Difference < 0.1 suggests conditional exchangeability. Abbreviations: ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]

 # Generalizability

Figure 3 shows the kernel densities for the conditional probabilities of selection in the sample and target populations, in the unweighted and weighted. These suggest no major differences in the distributions before weighting. More formally, the generalizability indices using the ML-E cohort as the target population were 0.98 for stages I–IV, 0.97 for stages I–II, and 0.99 for stages III–IV. Using SEER as the target population, the generalizability indices were 0.98 for stages I–IV, 0.97 for stages I–II, and 0.97 for stages III–IV. These results suggest that CGDB findings have very high generalizability without weighting, and accordingly we would not expect estimates of rwOS obtained from the weighted cohort to be substantially different.

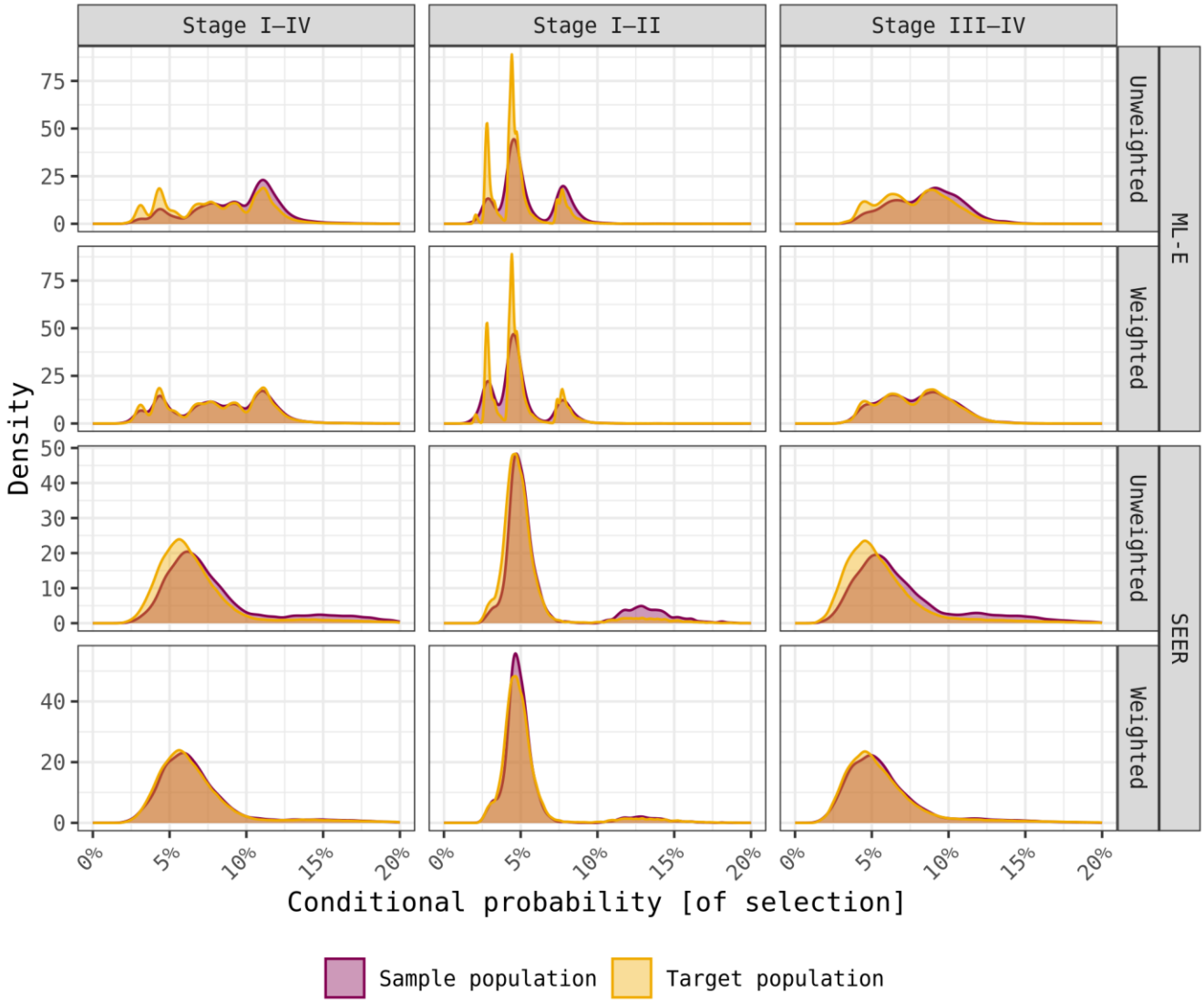


Figure 3: Kernel densities for the estimated probabilities of sample selection conditional on a set of baseline variables, before and after weighting, using the ML-E or SEER as the target population. Abbreviations: ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]

 # Real-world Overall survival

Figure 4 shows the Kaplan-Meier estimation of time-conditional rwOS estimates in the unweighted and weighted cohorts. The corresponding estimates of median survival presented in Table 4 shows that weighting only slightly changed the point estimates and confidence intervals. There were no statistical differences in the distributions of event times for the unweighted and weighted at a prespecified adjusted α of 0.00833 (Table 4).

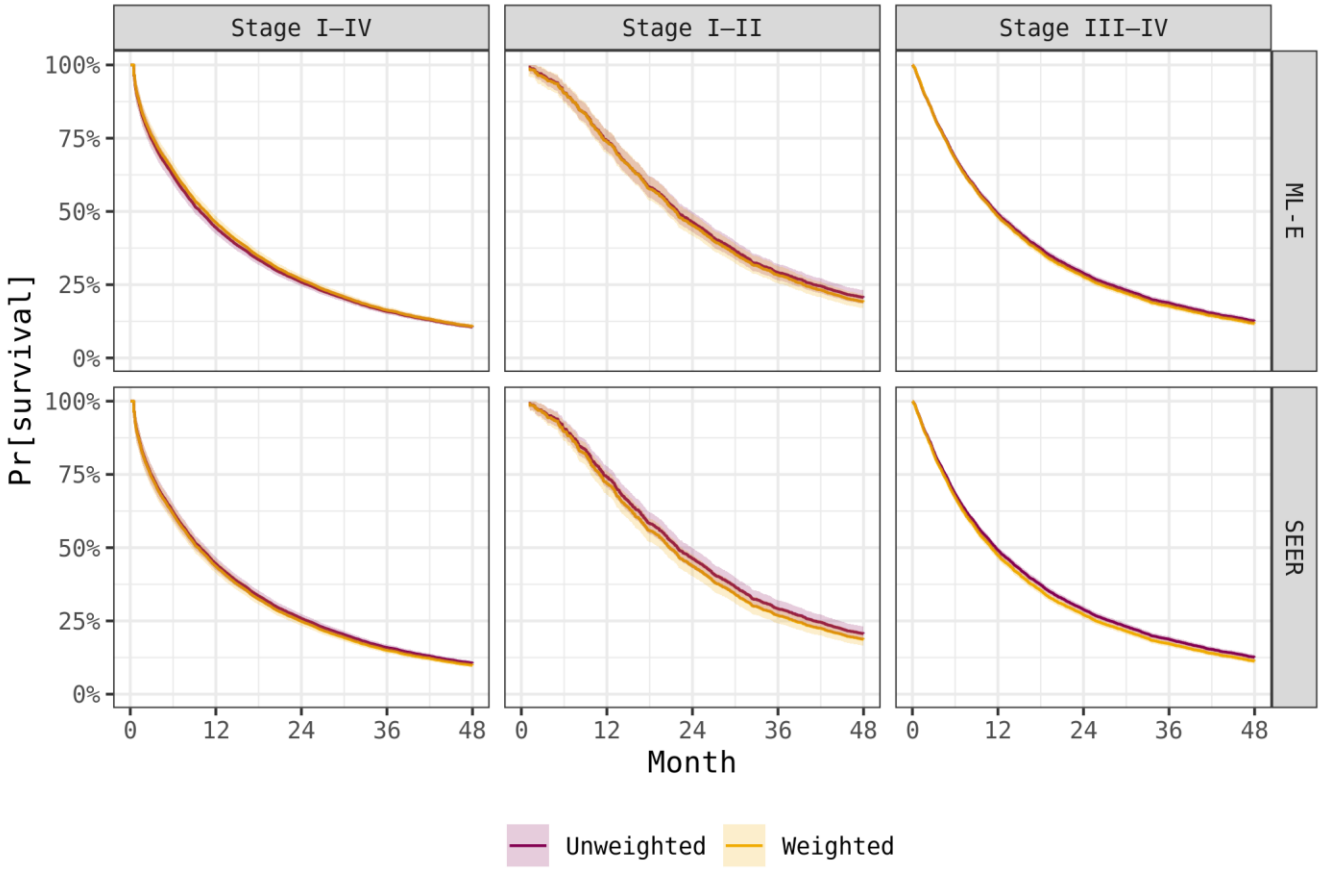


Figure 4: Kaplan-Meier estimation of rwOS, with delayed entry until relative date of eligibility. Time zero was the date of diagnosis for stages I–IV & I–II and the date of initiating first-line systemic therapy for the stages III–IV. Abbreviations: rwOS, Real-world Overall Survival; ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]

Table 4: Estimates of median rwOS and comparisons of event-time distributions

	Median rwOS (95% confidence intervals)		Event times
Cohort	**Unweighted**	**Weighted**	**Log-rank p value[1]**
ML-E as target population			
Stages I-IV[2]	9.8 (8.9–10.8)	10.5 (9.7–11.4)	0.830
Stages I–II[2]	21.9 (20.5–23.8)	21.6 (20.1–23.4)	0.268
Stages III–IV[3]	11.7 (11.2–12.2)	11.4 (11.0–11.9)	0.146
SEER as target population			
Stages I-IV[2]	9.8 (8.9–10.8)	9.5 (8.6–10.5)	0.049
Stages I–II[2]	21.9 (20.5–23.8)	20.8 (19.4–22.6)	0.181
Stages III–IV[3]	11.7 (11.2–12.2)	11.1 (10.6–11.6)	0.017

Abbreviations: rwOS, Real-world Overall Survival; ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]; ASD, Absolute Standardised Difference

[1] prespecified α of 0.00833 (0.05 Bonferroni-adjusted for six comparisons)
[2] from date of diagnosis
[3] from date of initiating first-line systemic therapy

Sensitivity analysis

In a CGDB sample restricted to diagnoses during the same time period as SEER cancer registrations (2011–2016), CGDB patients differed in the distributions of age (ASD 0.41), gender (ASD 0.10), race (ASD 0.33), and histological

classification (ASD 0.24) (Supplementary Table S8). The Tipton generalizability index was 0.96.

DISCUSSION

We sought to quantify how representative the Flatiron Health Foundation Medicine NSCLC CGDB cohort was of the underlying target population. While CGDB patients tended to differ from the target populations in the distributions of individual characteristics, indices of generalizability were very high. Accordingly, re-estimation of rwOS within a sample weighted to be representative of the target population caused only slight changes in point estimates.

Our findings are consistent with the few studies on the representativeness of Real-World Data. Flatiron Health have, for example, compared their CGDB (8) & Enhanced Database (22) and cohorts with aggregated cancer registrations, finding only differences in individual covariates, most notably in the clinical stage and geographic distribution. For NSCLC, these differences included an under-representation of stage I and over-representation of stage IV in both the CGDB and Enhanced Database. These assessments, however, are necessarily subjective and only compare individual covariates. The Tipton index has been used for assessing the generalizability of randomised trials (9). We have shown that it can be applied to Real-World Data whose representativeness should not be assumed, especially when subject to selection processes. Our findings also highlight the necessity of comparing the conditional distributions of selection over individual covariates because, whereas standardised differences can suggest differences in individual covariates, these would only lead to biased estimates if the distributions differed in a meaningful way.

The sample population of 14,545 CGDB patients was considerably larger than the typical real-world study cohort, enabling precise variance estimation of rwOS which showed that the sample rwOS did not differ statistically from the population rwOS. Smaller sample sizes could theoretically benefit greater from weighting because there is a greater expectation for these to be unrepresentative by chance alone (23); but one must also be mindful of precise variance estimation using appropriate power and sample-size calculations (24).

We used the graphical rules of Directed Acyclic Graphs to inform the selection of variables based on our understanding of the data-generating mechanism (3,14). Causal interpretation is dependent on correctly specifying the causal process and recording all variables necessary for conditional exchangeability (25,26). As with causal inference in observational studies, exchangeability remains a strong yet untestable assumption. However, it is important to avoid the substantial bias that can arise from controlling for all variables without careful consideration of causal pathways (15,27–29). Use of Directed Acyclic Graphs ensures transparency in reporting our assumptions and indicates the actions required to produce unbiased estimates. Our clinical vignettes have only one potential outcome; estimating either the effect of diagnosis or systemic therapy on survival. This approach can be extended to estimate the comparative effectiveness of two counterfactuals (medicine A vs. medicine B, for example). Weights can also be stacked as product terms should we want to account for treatment adherence in the case of time-varying treatments (30), informative censoring caused by non-random loss to follow-up (3), or dependent left-truncation bias caused by non-random delayed entry (31).

Our study has some limitations. Defining the target population can be problematic, particularly in the US where healthcare and surveillance systems are fragmented. We used two data sources: a superset of all Flatiron Health

patients who underwent Machine Learning chart extraction and SEER cancer registrations. We make the strong assumptions that either all patients in the Flatiron Health network, or SEER cancer registrations covering 34.6% of the population, are valid representations of the target population. Data for the Flatiron Health ML-E database coincided with the CGDB study period, but SEER data represented an earlier period. Temporal shifts in demographics — delayed diagnosis due to the indirect effects of the COVID-19 pandemic, for example — could explain some of the observed differences between target and study populations not owed to selection biases; however, our sensitivity analysis suggests that differences between CGDB and SEER remained even while restricting to the same study period. An additional consideration when defining the target population is determining whether the research question requires the study of the treated population as represented by the ML-E cohort or the diagnosed population as represented by cancer registrations. For simplicity, we conducted a Complete Case Analysis, excluding missing data in the study sample and target populations. This is justified only under the assumption of Missing Completely At Random. If data were Missing At Random — poor prognosis leading to unnecessary diagnostic work-up in ascertaining stage, histological classification, for example — then this exclusion would skew study and target population characteristics toward earlier disease by disproportionally excluding late-stage cancers. Multiple Imputation of missing data would be justified under this circumstance.

In summary, Flatiron Health CGDB patients tended to differ from the target populations in the distribution of individual characteristics, but indices of generalizability were very high. Estimates of rwOS in a population weighted to be representative did not differ from naïve estimates in the unweighted cohort. The Tipton generalizability index provides a quantitative assessment of the generalizability of findings which, together with Directed Acyclic Graphs, can be used to determine the influence of selection biases.

 # ACKNOWLEDGMENTS

The authors thank Miguel Miranda (AstraZeneca) for his involvement in discussions on statical analyses.

 # REFERENCES

1. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J. Clin. Epidemiol.* [electronic article]. 2005;58(4):323–337. (http://dx.doi.org/10.1016/j.jclinepi.2004.10.012)

2. Franklin JM, Glynn RJ, Martin D, et al. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clin. Pharmacol. Ther.* [electronic article]. 2019;105(4):867–877. (http://dx.doi.org/10.1002/cpt.1351)

3. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* [electronic article]. 2004;15(5):615–625. (http://dx.doi.org/10.1097/01.ede.0000135174.63482.43)

4. Lu H, Cole SR, Howe CJ, et al. Toward a clearer definition of selection bias when estimating causal effects. *Epidemiology* [electronic article]. 2022;33(5):699–706. (http://dx.doi.org/10.1097/EDE.0000000000001516)

5. Lesko CR, Buchanan AL, Westreich D, et al. Generalizing study results: A potential outcomes perspective. *Epidemiology* [electronic article]. 2017;28(4):553–561. (http://dx.doi.org/10.1097/EDE.0000000000000664)

6. Singal G, Miller PG, Agarwala V, et al. Association of patient characteristics and tumor genomics with clinical outcomes among patients with non-small cell lung cancer using a clinicogenomic database. *JAMA* [electronic article]. 2019;321(14):1391–1399. (http://dx.doi.org/10.1001/jama.2019.3241)

7. Sheinson DM, Wong WB, Flores C, et al. Association between medicare's national coverage determination and utilization of next-generation sequencing. *JCO Oncol. Pract.* [electronic article]. 2021;17(11):e1774–e1784. (http://dx.doi.org/10.1200/OP.20.01023)

8. Snow T, Snider J, Comment L, et al. Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron health-Foundation Medicine Clinico-Genomic Databases, Flatiron Health Research Databases, and the National Cancer Institute SEER population-based cancer registry. *medRxiv*. 2023;2023.01.03.22283682. (https://www.medrxiv.org/content/10.1101/2023.01.03.22283682v1). (Accessed February 1, 2023)

9. Tipton E. How Generalizable Is Your Experiment? An Index for Comparing Experimental Samples and Populations. *J. Educ. Behav. Stat.* [electronic article]. 2014;39(6):478–501. (https://doi.org/10.3102/1076998614558486)

10. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am. J. Epidemiol.* [electronic article]. 2008;168(6):656–664. (http://dx.doi.org/10.1093/aje/kwn164)

11. Dahabreh IJ, Robertson SE, Steingrimsson JA, et al. Extending inferences from a randomized trial to a new target population. *Stat. Med.* [electronic article]. 2020;39(14):1999–2014. (http://dx.doi.org/10.1002/sim.8426)

12. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am. J. Epidemiol.* [electronic article]. 2010;172(1):107–115. (http://dx.doi.org/10.1093/aje/kwq084)

13. Adamson B, Waskom M, Blarre A, et al. Approach to Machine Learning for Extraction of Real-World Data Variables from Electronic Health Records. *medRxiv*. 2023;2023.03.02.23286522. (https://www.medrxiv.org/content/10.1101/2023.03.02.23286522v1.abstract). (Accessed March 28, 2023)

14. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* [electronic article]. 1999;10(1):37–48. (https://www.ncbi.nlm.nih.gov/pubmed/9888278)

15. Munafò MR, Tilling K, Taylor AE, et al. Collider scope: when selection bias can substantially influence observed associations. *Int. J. Epidemiol.* [electronic article]. 2018;47(1):226–235. (http://dx.doi.org/10.1093/ije/dyx206)

16. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int. J. Epidemiol.* [electronic article]. 2010;39(2):417–420. (http://dx.doi.org/10.1093/ije/dyp334)

17. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* [electronic article]. 2009;28(25):3083–3107. (http://dx.doi.org/10.1002/sim.3697)

18. Ackerman B, Schmid I, Rudolph KE, et al. Implementing statistical methods for generalizing randomized trial findings to a target population. *Addict. Behav.* [electronic article]. 2019;94:124–132. (http://dx.doi.org/10.1016/j.addbeh.2018.10.033)

19. Brown S, Lavery JA, Shen R, et al. Implications of selection bias due to delayed study entry in clinical genomic studies. *JAMA Oncol.* [electronic article]. 2022;8(2):287–291. (http://dx.doi.org/10.1001/jamaoncol.2021.5153)

20. Backenroth D, Snider J, Shen R, et al. Accounting for Delayed Entry in Analyses of Overall Survival in Clinico-Genomic Databases. *Cancer Epidemiol. Biomarkers Prev.* [electronic article]. 2022;31(6):1195–1201. (http://dx.doi.org/10.1158/1055-9965.EPI-21-0876)

21. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.(https://www.R-project.org/)

22. Ma X, Long L, Moon S, et al. Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron Health, SEER, and NPCR. *medRxiv*. 2023;(https://www.medrxiv.org/content/10.1101/2020.03.16.20037143v3)

23. Tipton E, Hallberg K, Hedges LV, et al. Implications of small samples for generalization: Adjustments and rules of thumb. *Eval. Rev.* [electronic article]. 2017;41(5):472–505. (http://dx.doi.org/10.1177/0193841X16655665)

24. Austin PC. Informing power and sample size calculations when using inverse probability of treatment weighting using the propensity score. *Stat. Med.* [electronic article]. 2021;40(27):6150–6163. (http://dx.doi.org/10.1002/sim.9176)

25. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* [electronic article]. 2006;60(7):578–586. (http://dx.doi.org/10.1136/jech.2004.029496)

26. Hernan MA, Robins JM. Causal Inference. Boca Raton, FL: CRC Press; 2023 312 p.

27. Griffith GJ, Morris TT, Tudball MJ, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat. Commun.* [electronic article]. 2020;11(1):5749. (http://dx.doi.org/10.1038/s41467-020-19478-2)

28. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am. J. Epidemiol.* [electronic article]. 2013;177(4):292–298. (http://dx.doi.org/10.1093/aje/kws412)

29. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* [electronic article]. 2009;20(4):488–495. (http://dx.doi.org/10.1097/ede.0b013e3181a819a1)

30. Murray EJ, Hernán MA. Adherence adjustment in the Coronary Drug Project: A call for better per-protocol effect estimates in randomized trials. *Clin. Trials* [electronic article]. 2016;13(4):372–378. (http://dx.doi.org/10.1177/1740774516634335)

31. Sondhi A. Estimating survival parameters under conditionally independent left truncation. *Pharm. Stat.* [electronic article]. 2022;21(5):895–906. (http://dx.doi.org/10.1002/pst.2202)

SUPPLEMENTARY TABLES

Table S1: ICD-O-2 codes used for histological confirmation and classification of Non-Small Cell Lung Cancer	
Histological classification	**ICD-O-2 codes**
Non-squamous cell carcinoma	**Defined as adenocarcinomas, large-cell carcinomas or other specified carcinomas** **Adenocarcinomas:** 8015, 8050, 8140, 8143, 8147, 8190, 8201, 8211, 8250, 8260, 8290, 8310, 8320, 8323, 8333, 8401, 8440, 8470, 8480, 8490, 8503, 8507, 8550, 8570, 8574, 8576 **Large-cell carcinomas:** 8012, 8021, 8034, 8082 **Other specified carcinomas:** 8003, 8022, 8030, 8035, 8200, 8240, 8243, 8249, 8430, 8525, 8560, 8562, 8575
Squamous cell carcinoma	8051, 8070, 8078, 8083, 8090, 8094, 8120, 8123

Table S2: Baseline characteristics of stages I—II CGDB, ML-E, and SEER cohorts, before exclusion of missing data

Characteristic	CGDB (n 3,139)	ML-E (n 62,893)	SEER (n 55,754)
Age at diagnosis			
Median (Interquartile range)	70.0 (64.0, 76.0)	71.0 (65.0, 76.0)	71.0 (64.0, 78.0)
Unknown	0 (0.0%)	0 (0.0%)	0 (0.0%)
Gender			
Female	1,754 (55.9%)	34,154 (54.3%)	29,956 (53.7%)
Male	1,385 (44.1%)	28,736 (45.7%)	25,798 (46.3%)
Unknown	0 (0.0%)	3 (0.0%)	0 (0.0%)
Race			
Black	166 (5.3%)	4,389 (7.0%)	5,196 (9.3%)
Other	500 (15.9%)	9,088 (14.4%)	3,478 (6.2%)
White	2,222 (70.8%)	44,429 (70.6%)	46,875 (84.1%)
Unknown	251 (8.0%)	4,987 (7.9%)	205 (0.4%)
AJCC stage			
Stage I	1,849 (58.9%)	44,729 (71.1%)	—
Stage II	1,290 (41.1%)	18,164 (28.9%)	—
SEER stage[1]			
Localized	3,139 (100.0%)	62,893 (100.0%)	55,754 (100.0%)
Histological classification			
Non-squamous cell carcinoma	2,331 (74.3%)	41,101 (65.4%)	35,725 (64.1%)
Squamous cell carcinoma	741 (23.6%)	19,722 (31.4%)	14,137 (25.4%)
Unknown	67 (2.1%)	2,070 (3.3%)	5,892 (10.6%)

Abbreviations: CGDB, Clinico-Genomic Database [cohort]; ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]; AJCC, American Joint Committee on Cancer

[1] SEER stage approximated for CGDB & ML-E

Table S3: Baseline characteristics of stages III—IV CGDB, ML-E, and SEER cohorts, before exclusion of missing data

Characteristic	CGDB (n 13,456)[1]	ML-E (n 124,630)	SEER (n 179,074)
Age at diagnosis			
Median (Interquartile range)	68.0 (60.0, 75.0)	69.0 (61.0, 75.0)	69.0 (61.0, 77.0)
Unknown	0 (0.0%)	0 (0.0%)	0 (0.0%)
Gender			
Female	6,709 (49.9%)	58,969 (47.3%)	82,093 (45.8%)
Male	6,747 (50.1%)	65,650 (52.7%)	96,981 (54.2%)
Unknown	0 (0.0%)	11 (0.0%)	0 (0.0%)
Race			
Black	931 (6.9%)	10,484 (8.4%)	22,415 (12.5%)
Other	2,315 (17.2%)	19,022 (15.3%)	13,953 (7.8%)
White	8,986 (66.8%	82,601 (66.3%)	142,347 (79.5%)
Unknown	1,224 (9.1%)	12,523 (10.0%)	359 (0.2%)
AJCC stage			
Stage III	3,938 (29.3%)	43,747 (35.1%)	—
Stage IV	9,518 (70.7%)	80,883 (64.9%)	—
SEER stage[2]			
Regional	3,938 (29.3%)	43,747 (35.1%)	56,664 (31.6%)
Distant	9,518 (70.7%)	80,883 (64.9%)	122,410 (68.4%)
Histological classification			
Non-squamous cell carcinoma	10,097 (75.0%)	84,169 (67.5%)	106,210 (59.3%)
Squamous cell carcinoma	2,783 (20.7%)	34,138 (27.4%)	43,147 (24.1%)
Unknown	576 (4.3%)	6,323 (5.1%)	29,717 (16.6%)

Abbreviations: CGDB, Clinico-Genomic Database [cohort]; ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]; AJCC, American Joint Committee on Cancer

[1] Stage III—IV patients that also initiated systemic therapy
[2] SEER stage approximated for CGDB & ML-E

Table S4: Baseline characteristics of stages I—II CGDB & ML-E cohorts, in the unweighted and weighted

	Unweighted			Weighted		
Characteristic	**CGDB (n 2,826)**	**ML-E (n 56,031)**	**ASD**	**CGDB (n 2,832)**	**ML-E (n 56,031)**	**ASD**
Age at diagnosis[1]	70.0 (64.0, 76.0)	71.0 (65.0, 76.0)	0.09	70.0 (65.0, 76.0)	71.0 (65.0, 76.0)	0.01
Gender			0.03			0.01
Female	1,583 (56.0%)	30,453 (54.4%)		1,518 (53.6%)	30,453 (54.4%)	
Male	1,243 (44.0%)	25,578 (45.6%)		1,314 (46.4%)	25,578 (45.6%)	
Race			0.08			0.01
Black	165 (5.8%)	4,250 (7.6%)		216 (7.6%)	4,250 (7.6%)	
Other	489 (17.3%)	8,827 (15.8%)		455 (16.1%)	8,827 (15.8%)	
White	2,172 (76.9%)	42,954 (76.7%)		2,161 (76.3%)	42,954 (76.7%)	
AJCC stage			0.26			0.01
Stage I	1,658 (58.7%)	39,761 (71.0%)		1,998 (70.6%)	39,761 (71.0%)	
Stage II	1,168 (41.3%)	16,270 (29.0%)		834 (29.4%)	16,270 (29.0%)	
Histological classification			0.18			0.01
Non-squamous cell carcinoma	2,143 (75.8%)	37,856 (67.6%)		1,901 (67.1%)	37,856 (67.6%)	
Squamous cell carcinoma	683 (24.2%)	18,175 (32.4%)		932 (32.9%)	18,175 (32.4%)	

Abbreviations: CGDB, Clinico-Genomic Database [cohort]; ML-E, Machine Learning-Extracted [cohort]; ASD, Absolute Standardized Difference; AJCC, American Joint Committee on Cancer

[1] Median (Interquartile range)

Table S5: Baseline characteristics of stages III—IV CGDB & ML-E cohorts, in the unweighted and weighted

	Unweighted			Weighted		
Characteristic	**CGDB (n 9,172)[1]**	**ML-E (n 106,546)**	**ASD**	**CGDB (n 9,164)**	**ML-E (n 106,546)**	**ASD**
Age at diagnosis[2]	67.0 (60.0, 74.0)	69.0 (61.0, 75.0)	0.13	68.0 (61.0, 75.0)	69.0 (61.0, 75.0)	0.01
Gender			0.06			0.00
Female	4,618 (50.3%)	50,653 (47.5%)		4,342 (47.4%)	50,653 (47.5%)	
Male	4,554 (49.7%)	55,893 (52.5%)		4,822 (52.6%)	55,893 (52.5%)	
Race			0.08			0.01
Black	686 (7.5%)	9,938 (9.3%)		849 (9.3%)	9,938 (9.3%)	
Other	1,776 (19.4%)	18,188 (17.1%)		1,594 (17.4%)	18,188 (17.1%)	
White	6,710 (73.2%)	78,420 (73.6%)		6,721 (73.3%)	78,420 (73.6%)	
AJCC stage			0.24			0.03
Stage III	2,278 (24.8%)	38,079 (35.7%)		3,162 (34.5%)	38,079 (35.7%)	
Stage IV	6,894 (75.2%)	68,467 (64.3%)		6,001 (65.5%)	68,467 (64.3%)	
Histological classification			0.20			0.02
Non-squamous cell carcinoma	7,297 (79.6%)	75,786 (71.1%)		6,604 (72.1%)	75,786 (71.1%)	
Squamous cell carcinoma	1,875 (20.4%)	30,760 (28.9%)		2,560 (27.9%)	30,760 (28.9%)	

Abbreviations: CGDB, Clinico-Genomic Database [cohort]; ML-E, Machine Learning-Extracted [cohort]; ASD, Absolute Standardized Difference; AJCC, American Joint Committee on Cancer

[1] Stage III—IV patients that also initiated systemic therapy

[2] Median (Interquartile range)

Table S6: Baseline characteristics of stages I—II CGDB & SEER cohorts, in the unweighted and weighted

	Unweighted			Weighted		
Characteristic	**CGDB (n 2,826)**	**SEER (n 49,685)**	**ASD**	**CGDB (n 2,814)**	**SEER (n 49,685)**	**ASD**
Age at diagnosis[1]	70.0 (64.0, 76.0)	71.0 (64.0, 78.0)	0.16	71.0 (65.0, 76.0)	71.0 (64.0, 78.0)	0.06
Gender			0.04			0.01
Female	1,583 (56.0%)	26,894 (54.1%)		1,531 (54.4%)	26,894 (54.1%)	
Male	1,243 (44.0%)	22,791 (45.9%)		1,283 (45.6%)	22,791 (45.9%)	
Race			0.36			0.03
Black	165 (5.8%)	4,637 (9.3%)		260 (9.2%)	4,637 (9.3%)	
Other	489 (17.3%)	3,185 (6.4%)		199 (7.1%)	3,185 (6.4%)	
White	2,172 (76.9%)	41,863 (84.3%)		2,355 (83.7%)	41,863 (84.3%)	
SEER stage			0.00			0.00
Localized	2,826 (100.0%)	49,685 (100.0%)		2,814 (100.0%)	49,685 (100.0%)	
Histological classification			0.10			0.02
Non-squamous cell carcinoma	2,143 (75.8%)	35,578 (71.6%)		2,043 (72.6%)	35,578 (71.6%)	
Squamous cell carcinoma	683 (24.2%)	14,107 (28.4%)		771 (27.4%)	14,107 (28.4%)	

Abbreviations: CGDB, Clinico-Genomic Database [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]; ASD, Absolute Standardized Difference

[1] Median (Interquartile range)

Table S7: Baseline characteristics of stages III–IV CGDB & SEER cohorts, in the unweighted and weighted

	Unweighted			Weighted		
Characteristic	**CGDB (n 9,172)**	**SEER (n 149,056)**	**ASD**	**CGDB (n 9,092)**	**SEER (n 149,056)**	**ASD**
Age at diagnosis[1]	67.0 (60.0, 74.0)	69.0 (61.0, 77.0)	0.24	69.0 (62.0, 76.0)	69.0 (61.0, 77.0)	0.07
Gender			0.08			0.01
Female	4,618 (50.3%)	69,327 (46.5%)		4,198 (46.2%)	69,327 (46.5%)	
Male	4,554 (49.7%)	79,729 (53.5%)		4,895 (53.8%)	79,729 (53.5%)	
Race			0.35			0.04
Black	686 (7.5%)	18,405 (12.3%)		1,118 (12.3%)	18,405 (12.3%)	
Other	1,776 (19.4%)	12,126 (8.1%)		834 (9.2%)	12,126 (8.1%)	
White	6,710 (73.2%)	118,525 (79.5%)		7,141 (78.5%)	118,525 (79.5%)	
SEER stage			0.18			0.02
Regional	2,278 (24.8%)	49,014 (32.9%)		2,917 (32.1%)	49,014 (32.9%)	
Distant	6,894 (75.2%)	100,042 (67.1%)		6,175 (67.9%)	100,042 (67.1%)	
Histological classification			0.20			0.02
Non-squamous cell carcinoma	7,297 (79.6%)	105,969 (71.1%)		6,562 (72.2%)	105,969 (71.1%)	
Squamous cell carcinoma	1,875 (20.4%)	43,087 (28.9%)		2,531 (27.8%)	43,087 (28.9%)	

Abbreviations: CGDB, Clinico-Genomic Database [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]; ASD, Absolute Standardized Difference

[1] Median (Interquartile range)

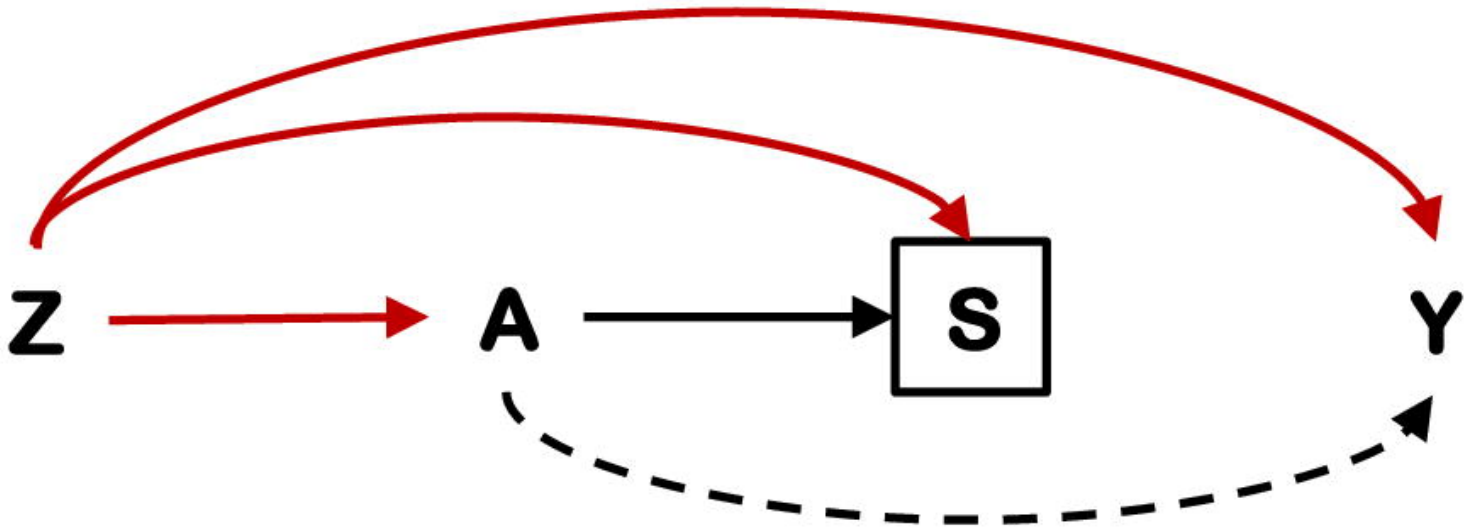
Table S8: Sensitivity analysis comparing the baseline characteristics of stages I—IV CGDB patients and SEER cancer registrations, harmonized to the same study period (2011—2016)

	Unweighted		
Characteristic	**CGDB (n 5,249)**	**SEER (n 198,741)**	**ASD**
Age at diagnosis[1]	66.0 (58.0, 73.0)	70.0 (62.0, 77.0)	0.41
Gender			0.10
Female	2,813 (53.6%)	96,221 (48.4%)	
Male	2,436 (46.4%)	102,520 (51.6%)	
Race			0.33
Black	325 (6.2%)	23,042 (11.6%)	
Other	902 (17.2%)	15,311 (7.7%)	
White	4,022 (76.6%)	160,388 (80.7%)	
SEER stage			0.02
Localized	1,353 (25.8%)	49,685 (25.0%)	
Regional	1,293 (24.6%)	49,014 (24.7%)	
Distant	2,603 (49.6%)	100,042 (50.3%)	
Histological classification			0.24
Non-squamous cell carcinoma	4,279 (81.5%)	141,547 (71.2%)	
Squamous cell carcinoma	970 (18.5%)	57,194 (28.8%)	

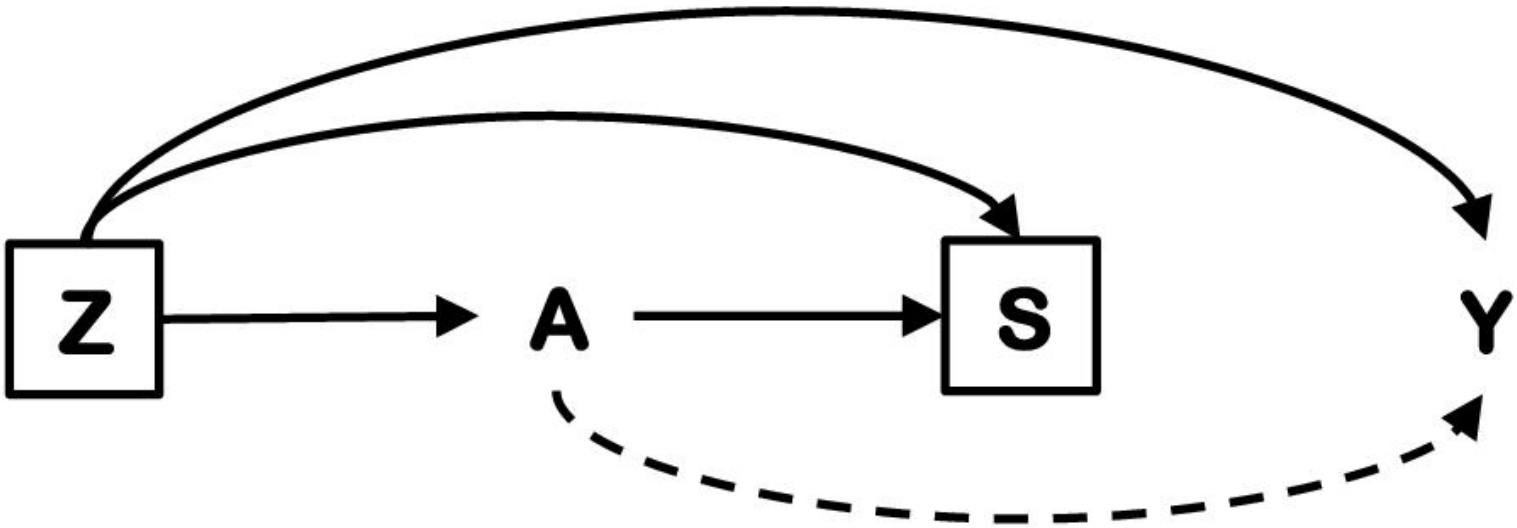
Abbreviations: CGDB, Clinico-Genomic Database [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]; ASD, Absolute Standardized Difference

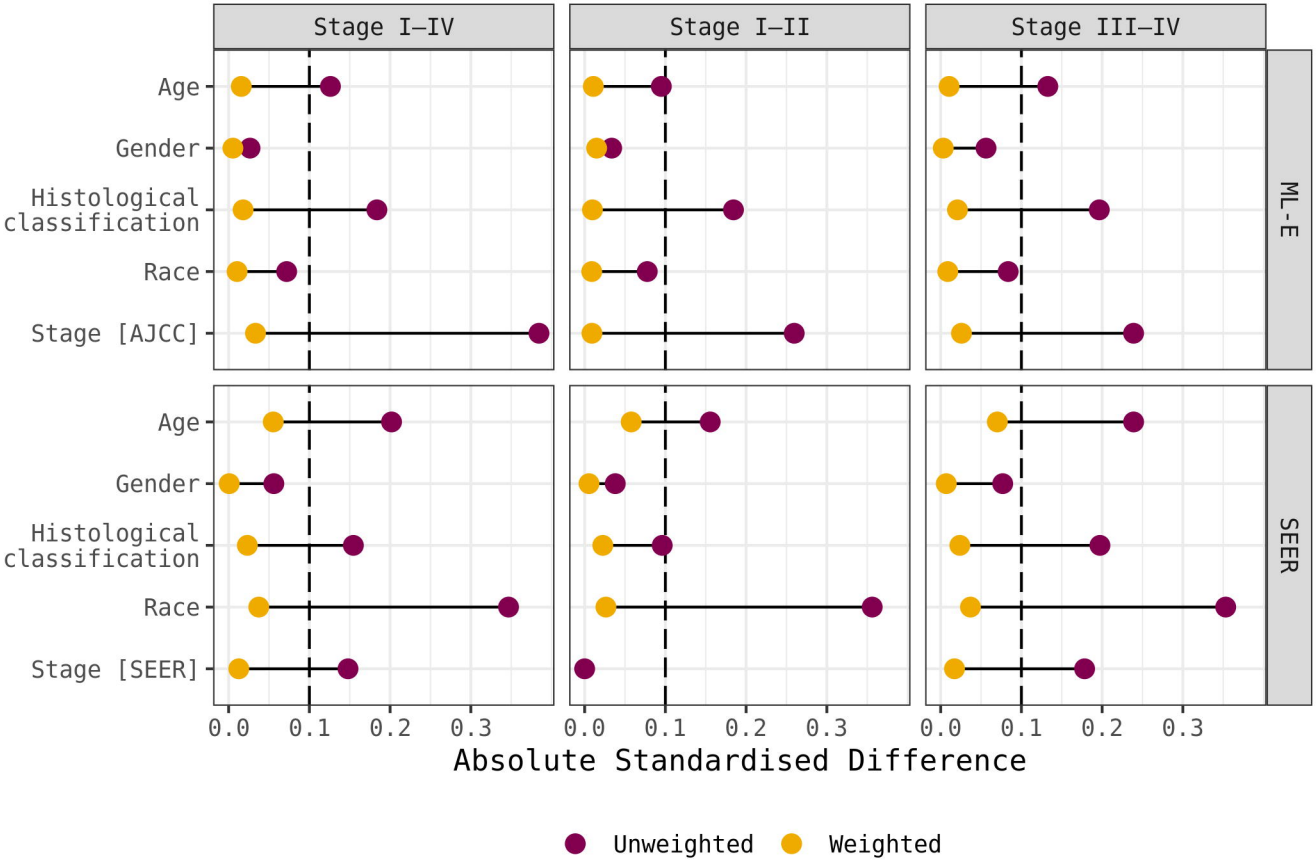
[1] Median (Interquartile range)

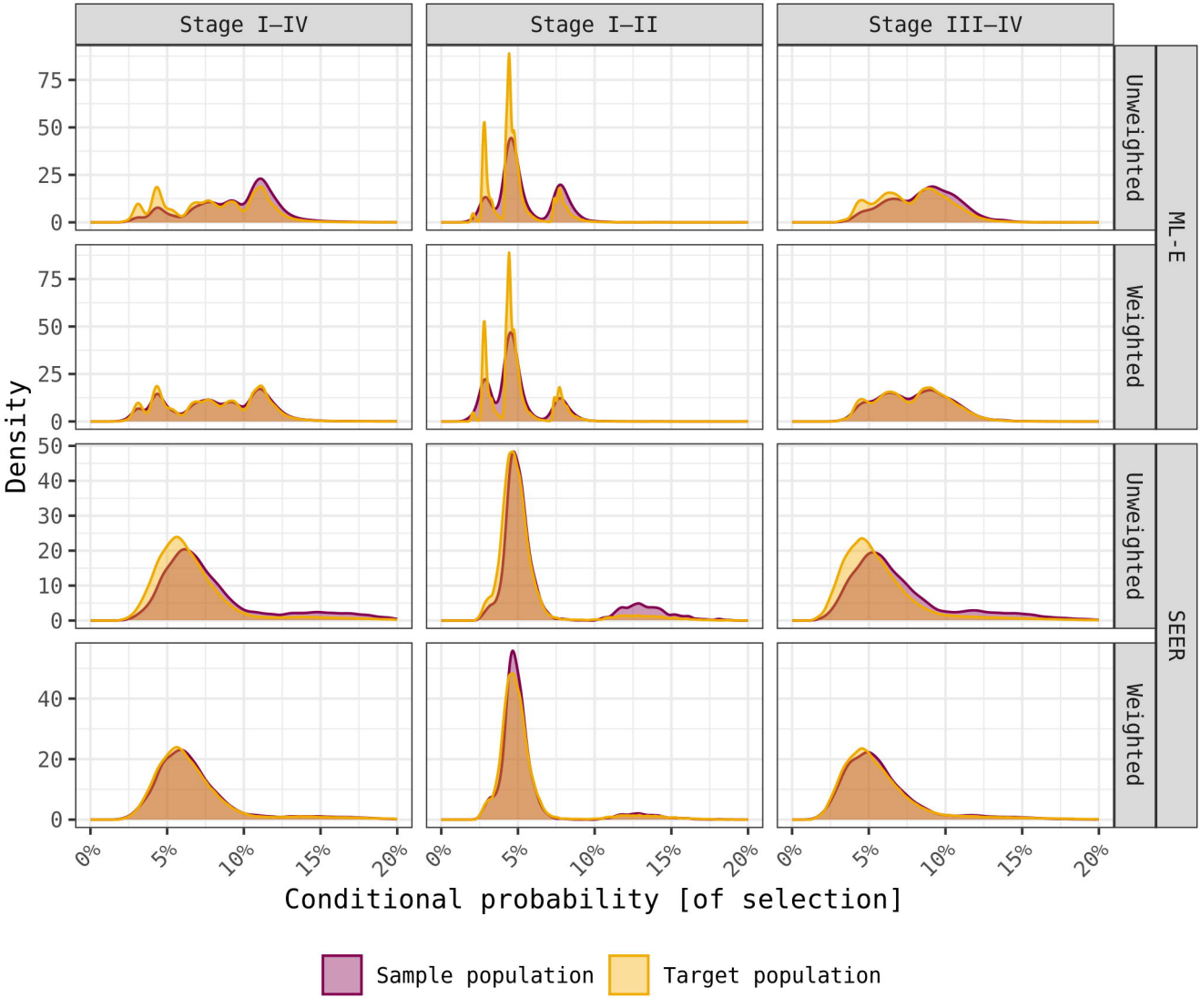
A)

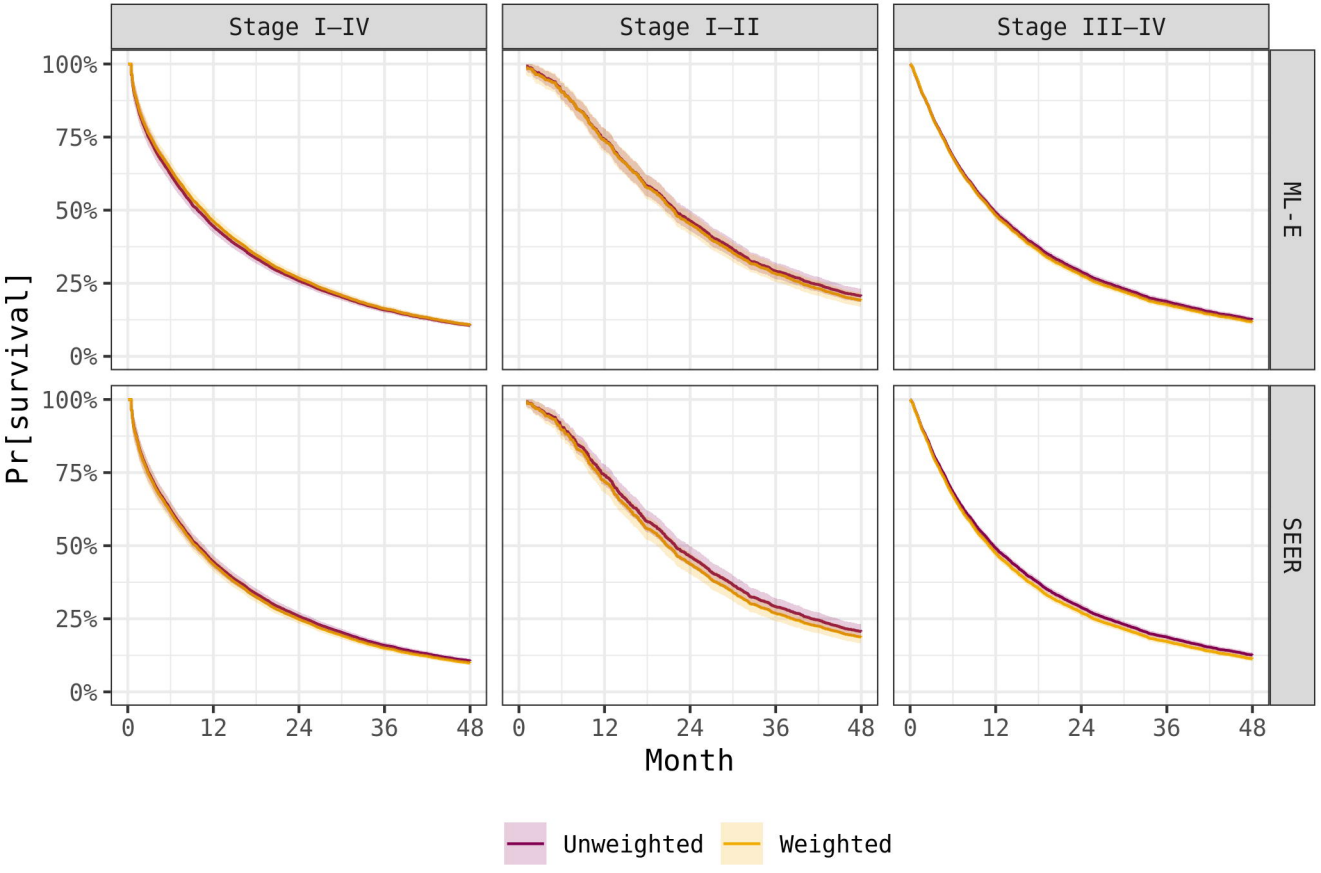


B)




Stage I–IV
Stage I–II
Stage III–IV
ML-E
SEER
Age
Gender
Histological classification
Race
Stage [AJCC]
Stage [SEER]
0.0
0.1
0.2
0.3
Absolute Standardised Difference
Unweighted
Weighted


Stage I–IV
Stage I–II
Stage III–IV
ML-E
SEER
Unweighted
Weighted
Unweighted
Weighted
Density
Conditional probability [of selection]
0%
5%
10%
15%
20%
Sample population
Target population


Stage I–IV
Stage I–II
Stage III–IV
ML-E
SEER
Pr[survival]
100%
75%
50%
25%
0%
0
12
24
36
48
Month
Unweighted
Weighted