Title: Extending Inferences From Sample To Target Populations: On The Generalizability Of A Real-World Clinico-Genomic Database Non-Small Cell Lung Cancer Cohort

Authors: Darren S. Thomas, Simon Collin, Luis C. Berrocal-Almanza, Heide Stirnadel-Farrant, Yiduo Zhang, Ping Sun

Correspondence Address: AstraZeneca, City House, 136 Hills Road, Cambridge, CB2 8PA, United Kingdom

Affiliations: Centre of Oncology Data Excellence, Oncology Business Unit, AstraZeneca (Darren S. Thomas, Simon Collin, Luis C. Berrocal-Almanza, Heide Stirnadel-Farrant, Yiduo Zhang, Ping Sun)

Funding: This work was supported by AstraZeneca

Data Availability Statement: The data that support the findings of this study have been originated by Flatiron Health, Inc. Requests for data sharing by license or by permission for the specific purpose of replicating results in this manuscript can be submitted to dataaccess@flatiron.com and cgdb-fmi@flatiron.com. Surveillance, Epidemiology and End Results (SEER) data is publicly available with a signed data-use agreement [https://seer.cancer.gov/data]

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

Conference presentation: N/A

Preprint Information: A pre-print was deposited on *medRxiv* [https://doi.org/10.1101/2023.06.15.23291372]

Disclaimer: This study was undertaken by AstraZeneca

Conflict of Interest: Authors, all employees of AstraZeneca at the time of writing, may report stock options.

Running Head: Extending inferences from sample to target populations

Key words: Real-world Data, Oncology, Causal Inference, Selection Bias, Generalizability

Abbreviations: AJCC, American Joint Committee on Cancer; ASD, Absolute Standardized Difference; CGDB, Clinico-Genomic Database [cohort]; IPW, Inverse Probability Weighting; ML-E, Machine Learning-Extracted [cohort]; NGS, Next Generation Sequencing; NSCLC, Non-Small Cell Lung Cancer; PATE, Population Average Treatment Effect; rwOS, Real-world Overall Survival; SATE, Sample Average Treatment Effect; SEER, Surveillance, Epidemiology, and End Results [cohort]

ABSTRACT

The representativeness of Real-world Data is 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. We assess the extent of selection biases in a de-identified nationwide US Clinico-Genomic Database Non-Small Cell Lung Cancer cohort through each process using two referent populations: a superset of all NSCLC patients in the Flatiron Health network and the National Cancer Institute's Surveillance, Epidemiology and End Results cancer registrations. Despite Standardised Differences suggesting differences in individual covariates between sample and referent populations, the conditional distributions of selection were alike, and indices suggest the results being generalizable (≥ 0.96 on a proportional scale of 0–1). Estimates of Real-world Overall Survival in a population weighted to be representative did not differ from naïve estimates in the unweighted cohort. We conclude with a counterfactual analysis highlighting how the Average Treatment Effect in the Sample and Population were concordant under an example having a Generalizability Index of 0.97. 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 underwent human technology-assisted chart abstraction and whose tumor biopsy was submitted for Next Generation Sequencing (NGS). The differential selection of patients into the CGDB can therefore be caused by the geographic sampling of Flatiron Health clinics across the US, or in the requirement for patients to meet further eligibility criteria. Though the selection of patients to undergo chart abstraction is random, NGS at its introduction was not widely adopted by clinical guidelines or covered by insurers, which could introduce 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 (IPW) 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

quantitative measure of how closely a sample population approximates a sample randomly drawn from a defined referent population [9]. IPW in the context of selection involves weighting 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.

Two referent datasets were identified to assess the selection of subgroups through each process. We define the referent population as the population preceding a selection process [4]. The National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) dataset includes cancer registrations for participating states covering 34.6% of the US population. In the absence of US-wide cancer registrations, we use this as the underlying target population to which we want to extend our inferences. The Flatiron Health Machine Learning Extracted (ML-E) database is a super-sample of all patients treated within the Flatiron Health network; not just a nested sample that have undergone human chart abstraction and NGS. Selection from SEER to ML-E therefore represents the influence of geographic sampling; from ML-E to CGDB due to the requirement for NGS; and from SEER to CGDB due to both processes combined. We demonstrate with an applied example on how these methods can be used to calculate the SATE and PATE as a causal contrast of the potential outcomes under a hypothetical binary treatment in the presence of confounding biases.

METHODS

Sample population

Flatiron Health-Foundation Medicine CGDB (2011–2021)

The 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 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).

Referent populations

Flatiron Health ML-E Database (2011–2021)

The 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 referent 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.

SEER Incidence Data (1975-2016)

The SEER program records cross-sectional data on the demographic and clinical characteristics of cancer registrations collected by participating state-level 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 1).

Statistical analyses

The analysis follows two parts. Part I is a descriptive analysis that compares a set of baseline variables at each stage of selection (ML-E vs. SEER, CGDB vs. ML-E, CGDB vs. SEER), followed by a comparison of marginal overall survival in the unweighted CGDB population relative to the CGDB population weighted to be represented of the ML-E or SEER referents. In part II we outline a counterfactual analysis in which we calculate the SATE and PATE. Part I uses only Inverse Probability of Selection Weights (*sw*), where part II uses the product terms of *sw* and Inverse Probability of Treatment Weights (*tw*).

Causal estimand

Akin to an intention-to-treat analog, the causal estimand of interest for the counterfactual analysis was the Average Treatment Effect at times *t* (ATE_t) as a difference in potential outcomes of survival under initiating a hypothetical binary treatment *A*, regardless of intercurrent treatment adherence or discontinuation. The ATE_t *therefore equals* $E[Y_t^{a=1} - Y_t^{a=0}]$, where the $E[Y^a]$ represents the potential outcome under treatment *A* (1 or 0). For a binary outcome of survival, the potential outcome $E[Y^a]$ equals the $Pr[Y^a = 0]$ — the probability that death did not occur. The hypothetical treatment was assigned to all stage III-IV CGDB patients by sampling from probability tables indicating initiation of two unnamed EGFR inhibitors as first line, conditional of a set of baseline variables Z, trained on a subset of ML-E patients. The ATE_t was calculated within the CGDB sample ($SATE_t = E[Y_t^{a=1} - Y_t^{=0} | S = 1]$, where *S* denotes selection into the sample) using *tw*, and extended to the target population ($PATE_t = E[Y_t^{a=1} - Y_t^{a=0} | S = 0]$) with the product terms of *tw* and *sw*.

We present our causal assumptions in the Supplementary Appendix, which depicts two non-causal paths: a confounding path ($A < Z > Y$) which we sever with *tw* and a collider-restriction (type-1 selection bias) path ($A > [S] < Z > Y$) which we reverse with *sw*. A set of baseline variables *Z*, shortlisted as causative of selection and treatment assignment under the theorems of d-separation, were used to estimate conditional probabilities during IPW: 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 unrecorded for SEER patients, the SEER stages Localised, Regional, and Distant are equivalent to AJCC stages I–II, III, and IV. Gender was unrecorded for SEER therefore we used biological sex as a proxy.

Missing data

All patients from the CGDB & ML-E/SEER who did not have complete data were excluded.

Inverse Probability of Selection Weights

We estimated Stabilised Inverse Probability of Selection Weights (sw) using a logistic binomial model on the outcome indicating selection into the study sample (1 for sample, 0 for referent population), conditional on a set of baseline variables (Z) causative of selection. Age was modelled with a natural cubic spline with 3 knots. All other variables were categorical indicator variables. Weights for $_i th$ CGDB patient (sw_i) were calculated as the marginal probability of selection into the sample 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 $0 < Pr[S = 1 \mid Z = z]$ for all levels of Z.

Equation 1: Stabilised Inverse Probability of Selection Weights for the $_i th$ CGDB patient (sw_i) where S denotes selection into the study sample (1 for the sample population, 0 for the referent population) and Z a set of baseline variables (age, gender, race, stage, histological classification) causative of selection.

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

Inverse Probability of Treatment Weights

Stabilized Inverse Probability Treatment Weights (*tw*) were calculated for CGDB patients as per Equation 2. The calculation was the same as for the calculation of *sw* with the exceptions of the outcome variable being binary treatment (*A*), the weights being calculated for both counterfactuals (*A = 1 & A = 0*), and the training data was restricted to the CGDB sample.

Equation 2: Stabilized Inverse Probability of Treatment Weights for the $_i th$ CGDB patient (tw_i) where *A* denotes a binary treatment (1 or 0), S a binary variable indicating selection into the CGDB sample, and *Z* a set of baseline variables causative of unrandom treatment initiation.

$$tw_i = \begin{cases} \dfrac{Pr[A = 1 \mid S = 1]}{Pr[A = 1 \mid S = 1,\ Z = z]} & \boldsymbol{if\ A = 1} \\ \dfrac{1 - Pr[A = 1 \mid S = 1]}{1 - Pr[A = 1 \mid S = 1,\ Z = z]} & \boldsymbol{if\ A = 0} \end{cases}$$

Baseline characteristics

We tabulated the set of baseline characteristics (*Z*) of the unweighted and *sw*-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 referent population. A rule-of-thumb for conditional exchangeability is an ASD < 0.1 for all variables [14].

Generalizability

The Tipton Generalizability Index (*β*) is a measure of the degree of similarity in the conditional distributions of selection ($Pr[S = 1 \mid Z = z]$) within the sample and referent populations [9]. *β* is calculated by integrating the product of the kernel densities for the conditional probabilities of selection in the sample ($f_s(s)$) and referent 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 generalizable findings [9]. We used the open-source implementation available as the R package {generalize} [15].

Equation 3: Calculation of the Tipton Generalizability Index (β) where $f_s(s)$ denotes the conditional distribution of selection for the sample, $f_p(s)$ the conditional distribution of selection for the referent population, and ds a kernel density bandwidth defined using Silverman's rule-of-thumb.

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

Real-world Overall Survival

We estimated Real-world Overall Survival (rwOS) via Kaplan-Meier estimation of the unweighted and *sw*-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). While the ascertainment of rwOS through medical records and external commercial & public data has high sensitivity relative to the National Death Index [16], right censorship on last known activity is necessary as patients lost to follow-up may not be at risk of a future event. The withheld 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) [17,18]. For the weighted cohorts we used robust variance estimation. For the descriptive analysis, a weighted log-rank test was used to compare survival times in the unweighted and *sw*-weighted CGDB cohorts, using a prespecified α of $P < 0.00833$ (α of 0.05 Bonferroni adjusted for six comparisons). *P* values are two-sided. For the counterfactual analysis, the SATE and PATE were calculated (*sw* using SEER as referent) at 12, 24, 36, and 48 months, with 95% confidence intervals for the difference in survival calculated from 500 bootstraps.

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 [19]. A JSON with all dependencies and a PDF with all package citations are made available as supplements.

RESULTS

Sample and referent 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 (Table 1). Relative to SEER, 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 2 & 3. After exclusion of missing data, there were 14,545, 162,577, and 198,741 remaining in the CGDB, ML-E, and SEER cohorts.

ML-E in comparison to SEER

To understand the geographic sampling of the Flatiron Network from the target population, the baseline characteristics of the ML-E and SEER cohorts are presented for stages I–IV, I–I, and III–IV in Table 3 and Supplementary Tables 4 & 5. Based on ASDs, the ML-E cohort differed from SEER in the distributions of race (ASD 0.28) and stage (ASD 0.21). Cohorts defined by stages I–II differed in the composition of race (ASD 0.30) and in stages III-IV by age (ASD 0.17), race (ASD 0.28), and stage (ASD 0.18). The β indices for stages I-IV, I–II, and III–IV were all 0.98.

CGDB in comparison to ML-E

To understand the influence of selection based on requiring NGS, the baseline characteristics of CGDB patients in the unweighted and weighted in relation to ML-E are presented in Table 3. Based on the ASDs (Figure 1), 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 6), 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 7), 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).

CGDB in comparison to SEER

To understand the influence of both selections processes combined, the baseline characteristics of CGDB patients and ASDs in the unweighted and weighted in relation to SEER are presented in Table 4. Based on the ASDs (Figure 1), 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 8), 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 9) 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).

Generalizability

Figure 2 shows the kernel densities for the conditional probabilities of selection in the sample and referent populations, in the unweighted and weighted. These suggest no major differences in the distributions before weighting. More formally, the β using the ML-E cohort as the referent 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 referent population, the βwere 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 are generalizable without weighting.

Real-world Overall survival

Figure 3 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 5 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 5). **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 10). The β was 0.96.**Causal estimand**

Using the subset of stage III-IV CGDB patients that initiated systemic therapy and SEER as the referent population (β 0.97), the SATE and PATE was estimated as at 12-month intervals (Table 6). The corresponding counterfactual survival curves are presented in Figure 4. The SATE at times 12, 24, 36, and 48 months were 1.0, 0.6, -0.1, and 0.0 percentage points, whereas the PATE were 0.5, 0.0, -1.2, and -1.0. Inferring from the bootstrap intervals, the SATE and PATE were concordant.

DISCUSSION

We sought to quantify how representative the NSCLC CGDB cohort was of the underlying target population, through each stepwise selection process. While CGDB patients tended to differ from the referent populations in the distributions of individual characteristics, indices suggest that findings are generalizable. Accordingly, re-estimation of rwOS within samples weighted to be representative of the US NSCLC target population, or all NSCLC patients in the Flatiron Health network, caused only slight changes in point estimates. We also show, by estimating a causal estimand as a contrast of potential outcomes under a hypothetical binary treatment, that within a sample with a generalisability index of 0.97 in relation to the target population, the SATE and PATE were concordant.

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 [20] cohorts with 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 databases. 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. Based on simulations in the study of pedagogy trials, current rules-of-thumb suggests that indices ≥ 0.9 are indicative of findings being generalizable [9]. By this threshold, the influence of selection due to geographic sampling of Flatiron Health from the underlying cancer population (β 0.98), on the requirement for a NGS test ($\beta \geq 0.97$), or the effect of both processes combined ($\beta \geq 0.97$) was uninfluential. The utility of this index as a measure of generalizability is further corroborated in our counterfactual example wherein the SATE and PATE were concordant within a sample of stage III–IV CGDB patients that had a β index of 0.97 relative to the SEER population. When the conditional distributions are alike, concordance between unweighted & weighted estimates are guaranteed. Future work should attempt establish interpretation rules relevant to the study of pharmacoepidemiology and outcomes of survival, which would aid researchers in assessing the generalizability of findings.

The sample population of 14,545 CGDB patients was considerably larger than the typical real-world study cohort, enabling precise variance estimation of rwOS. Nevertheless, it is important to acknowledge the increased susceptibility of smaller sample sizes to selection biases because there is a greater expectation for these to be unrepresentative by chance alone. While IPW can address the issue of representativeness, it does not automatically correct for the increased variance inherit to smaller samples; and in some circumstances may increase random error [22]. Accordingly, one must be mindful of the necessity for precise variance estimation using appropriate power and sample-size calculations[22], inclusion of auxiliary variables influential of the outcome, and using robust or bootstrap variance estimation [23].

We used the graphical rules of Directed Acyclic Graphs to inform the selection of variables based on our understanding of the data-generating mechanism [3,24]. The causal assumption of exchangeability requires that no variables causative of selection or treatment are omitted from the respective propensity models [25,26]. This also holds true for the generalizability index; a mis-specified selection model through omission of variables — measured or unmeasured — would still return a high index if the distributions of selection conditional on those variables specified showed a high degree of overlap between the sample and referent populations. The presence of residual confounding in such circumstances could be triangulated with negative controls [27] and measured with Quantitative Bias Analysis [28]. Nonetheless, exchangeability remains a strong yet untestable assumption in observational studies. However, it is important to avoid the substantial bias that can arise from controlling for all variables without careful consideration of causal pathways [29–32].

Further assumptions required to endow these estimates with causal interpretation include random delayed entry and random loss to follow-up. If there exists any violations to these whereby progression of disease after diagnosis is causative of either ordering NGS leading to a dependent left-truncation bias [33] or informative censoring on last clinical activity after which death cannot be ascertained [3], further weights should be derived with time-varying clinical parameters measured post baseline. Estimation of these weights would mirror how one would account for treatment adherence in the case of time-varying treatments and time-varying confounding [34], while weights would be combined as the product terms as we have described herein.

Our study has some limitations. We used two data sources to understand the processes underlying each selection process: the ML-E superset of all Flatiron Health patients who underwent Machine Learning chart extraction and SEER cancer registrations. Data for the Flatiron Health ML-E database coincided with the CGDB study period, but SEER data represented an earlier period due to the lag period

required for their curation. 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. Defining the target population can too be problematic, particularly in the US where healthcare and surveillance systems are fragmented. In the absence of population-wide cancer registration in the US, we make the strong assumption that SEER cancer registrations for participating states covering 34.6% of the US population is a valid representation of the target population. Nonetheless, there may be circumstances in which the research question requires the study of the treated population as represented by the ML-E cohort as opposed to the diagnosed population as represented by cancer registrations. An advantage of using SEER is that this data is publicly available, and so the methods described herein can be readily applied.

For simplicity, we conducted a Complete Case Analysis, excluding missing data in the study sample and referent 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 referent 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 referent populations in the distribution of individual characteristics, but indices suggest findings being generalizable. Estimates of rwOS in a population weighted to be representative did not differ from naïve estimates in the unweighted cohort, and estimates of the SATE and PATE in a counterfactual example were concordant. 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.

REFERENCES

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

2. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clin Pharmacol Ther*. 2019;105(4):867-877. doi:10.1002/cpt.1351

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

4. Lu H, Cole SR, Howe CJ, Westreich D. Toward a clearer definition of selection bias when estimating causal effects. *Epidemiology*. 2022;33(5):699-706. doi:10.1097/EDE.0000000000001516

5. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: A potential outcomes perspective. *Epidemiology*. 2017;28(4):553-561. doi: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*. 2019;321(14):1391-1399. doi:10.1001/jama.2019.3241

7. Sheinson DM, Wong WB, Flores C, Ogale S, Gross CP. Association between medicare's national coverage determination and utilization of next-generation sequencing. *JCO Oncol Pract*. 2021;17(11):e1774-e1784. doi: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*. Published online January 5, 2023:2023.01.03.22283682. doi:10.1101/2023.01.03.22283682

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

10. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656-664. doi:10.1093/aje/kwn164

11. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. Extending inferences from a randomized trial to a new target population. *Stat Med*. 2020;39(14):1999-2014. doi: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*. 2010;172(1):107-115. doi: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*. Published online March 6, 2023:2023.03.02.23286522. doi:10.1101/2023.03.02.23286522

14. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28(25):3083-3107. doi:10.1002/sim.3697

15. Ackerman B, Schmid I, Rudolph KE, et al. Implementing statistical methods for generalizing randomized trial findings to a target population. *Addict Behav*. 2019;94:124-132. doi:10.1016/j.addbeh.2018.10.033

16. Curtis MD, Griffith SD, Tucker M, et al. Development and Validation of a High-Quality Composite Real-World Mortality Endpoint. *Health Serv Res*. 2018;53(6):4460-4476. doi:10.1111/1475-6773.12872

17. Brown S, Lavery JA, Shen R, et al. Implications of selection bias due to delayed study entry in clinical genomic studies. *JAMA Oncol*. 2022;8(2):287-291. doi:10.1001/jamaoncol.2021.5153

18. 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*. 2022;31(6):1195-1201. doi:10.1158/1055-9965.EPI-21-0876

19. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2019. https://www.R-project.org/

20. Ma X, Long L, Moon S, Adamson BJS, Baxi SS. Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron Health, SEER, and NPCR. *bioRxiv*. Published online June 7, 2023. doi:10.1101/2020.03.16.20037143

21. Tipton E, Hallberg K, Hedges LV, Chan W. Implications of small samples for generalization: Adjustments and rules of thumb. *Eval Rev*. 2017;41(5):472-505. doi:10.1177/0193841X16655665

22. Austin PC. Informing power and sample size calculations when using inverse probability of treatment weighting using the propensity score. *Stat Med*. 2021;40(27):6150-6163. doi:10.1002/sim.9176

23. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med*. 2016;35(30):5642-5655. doi:10.1002/sim.7084

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

25. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60(7):578-586. doi:10.1136/jech.2004.029496

26. Hernan MA, Robins JM. *Causal Inference*. CRC Press; 2023.

27. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*. 2010;21(3):383-388. doi:10.1097/EDE.0b013e3181d61eeb

28. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol*. 1996;25(6):1107-1116. doi:10.1093/ije/25.6.1107-a

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

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

31. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol*. 2013;177(4):292-298. doi:10.1093/aje/kws412

32. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*. 2009;20(4):488-495. doi:10.1097/ede.0b013e3181a819a1

33. Sondhi A. Estimating survival parameters under conditionally independent left truncation. *Pharm Stat*. 2022;21(5):895-906. doi:10.1002/pst.2202

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

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

Table 2: Baseline characteristics of stages I–IV ML-E & SEER cohorts, in the unweighted

Characteristic	ML-E (n 162,577)	SEER (n 198,741)	ASD
Age at diagnosis[1]	70.0 (62.0, 75.0)	70.0 (62.0, 77.0)	0.09
Gender			0.03
Female	81,106 (49.9%)	96,221 (48.4%)	
Male	81,471 (50.1%)	102,520 (51.6%)	
Race			0.28
Black	14,188 (8.7%)	23,042 (11.6%)	
Other	27,015 (16.6%)	15,311 (7.7%)	
White	121,374 (74.7%)	160,388 (80.7%)	
SEER stage			0.21
Localized	56,031 (34.5%)	49,685 (25.0%)	
Regional	38,079 (23.4%)	49,014 (24.7%)	
Distant	68,467 (42.1%)	100,042 (50.3%)	
Histological classification			0.03
Non-squamous cell carcinoma	113,642 (69.9%)	141,547 (71.2%)	
Squamous cell carcinoma	48,935 (30.1%)	57,194 (28.8%)	

Abbreviations: ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]; ASD, Absolute Standardized Difference

[1] Median (Interquartile range)

Table 3: 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)

Table 4: 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

Table 5: 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 referent 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 referent 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]

[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

Table 6: Calculation of Average Treatment Effects in the Sample and extended to the Population

	Sample			Population		
Month *t*	$Y_t^{a=1}$	$Y_t^{a=0}$	**SATE[1]** $Y_t^{a=1} - Y_t^{a=0}$	$Y_t^{a=1}$	$Y_t^{a=0}$	**PATE[1]** $Y_t^{a=1} - Y_t^{a=0}$
12	50.9%	49.9%	1.0 PP (-2.9–5.0)	49.1%	48.6%	0.5 PP (-5.7–6.8)
24	30.9%	30.3%	0.6 PP (-3.3–4.0)	28.7%	28.7%	0.0 PP (-5.4–4.4)
36	20.1%	20.2%	-0.1 PP (-2.9–2.7)	18.2%	19.4%	-1.2 PP (-5.4–2.4)
48	13.7%	13.8%	0.0 PP (-2.5–2.1)	12.1%	13.1%	-1.0 PP (-4.2–1.8)

Abbreviations: $Y_t^{a=}$, potential outcome of Real-world Overall Survival under treatment *A* 1 or 0; SATE, Sample Average Treatment Effect; PATE, Population Average Treatment Effect; PP, Percentage Points

[1] 95% confidence intervals calculated from 500 bootstraps

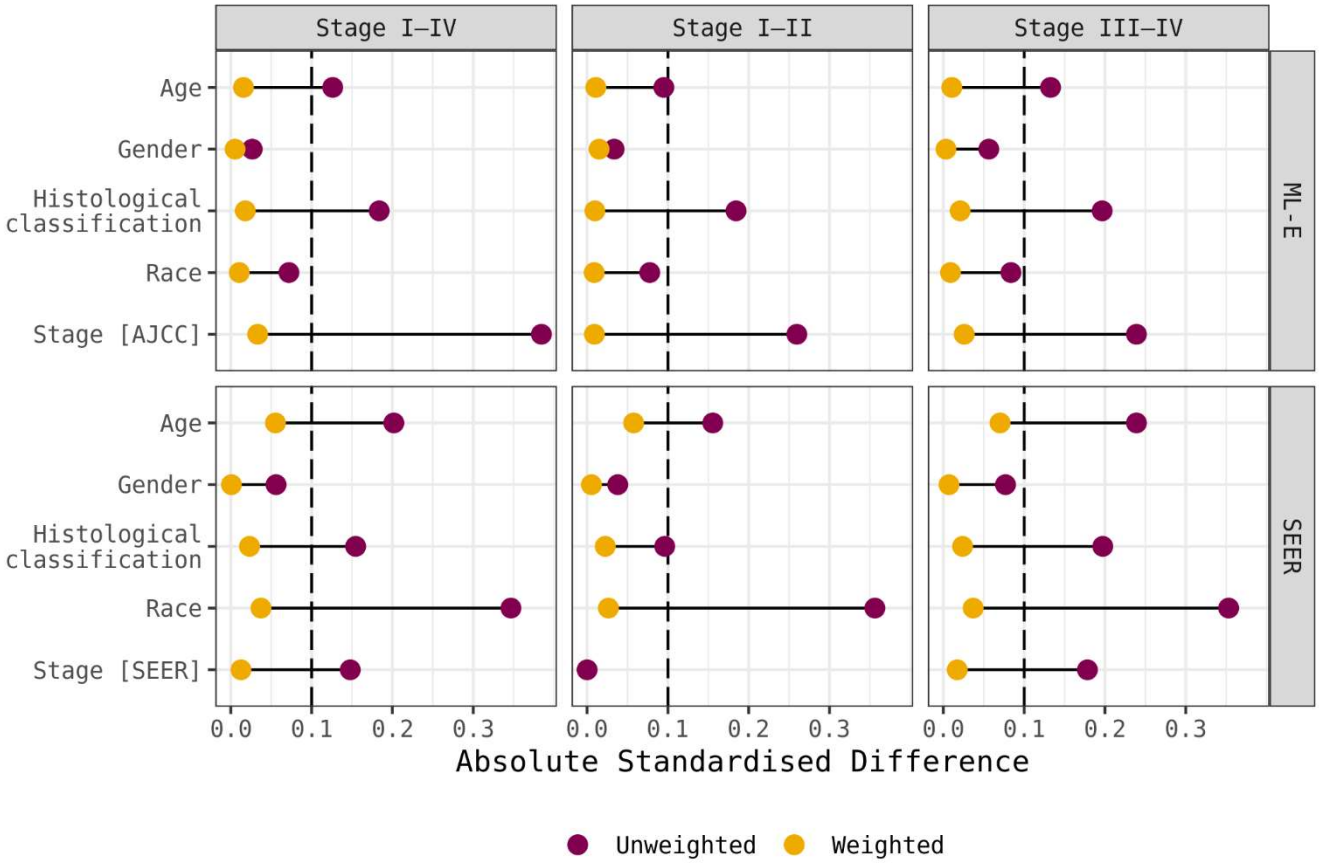


Figure 1: Absolute Standardised Differences summarising the differences in individual covariates between the sample and referent populations. An Absolute Standardised Difference < 0.1 suggests conditional exchangeability. Abbreviations: ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]

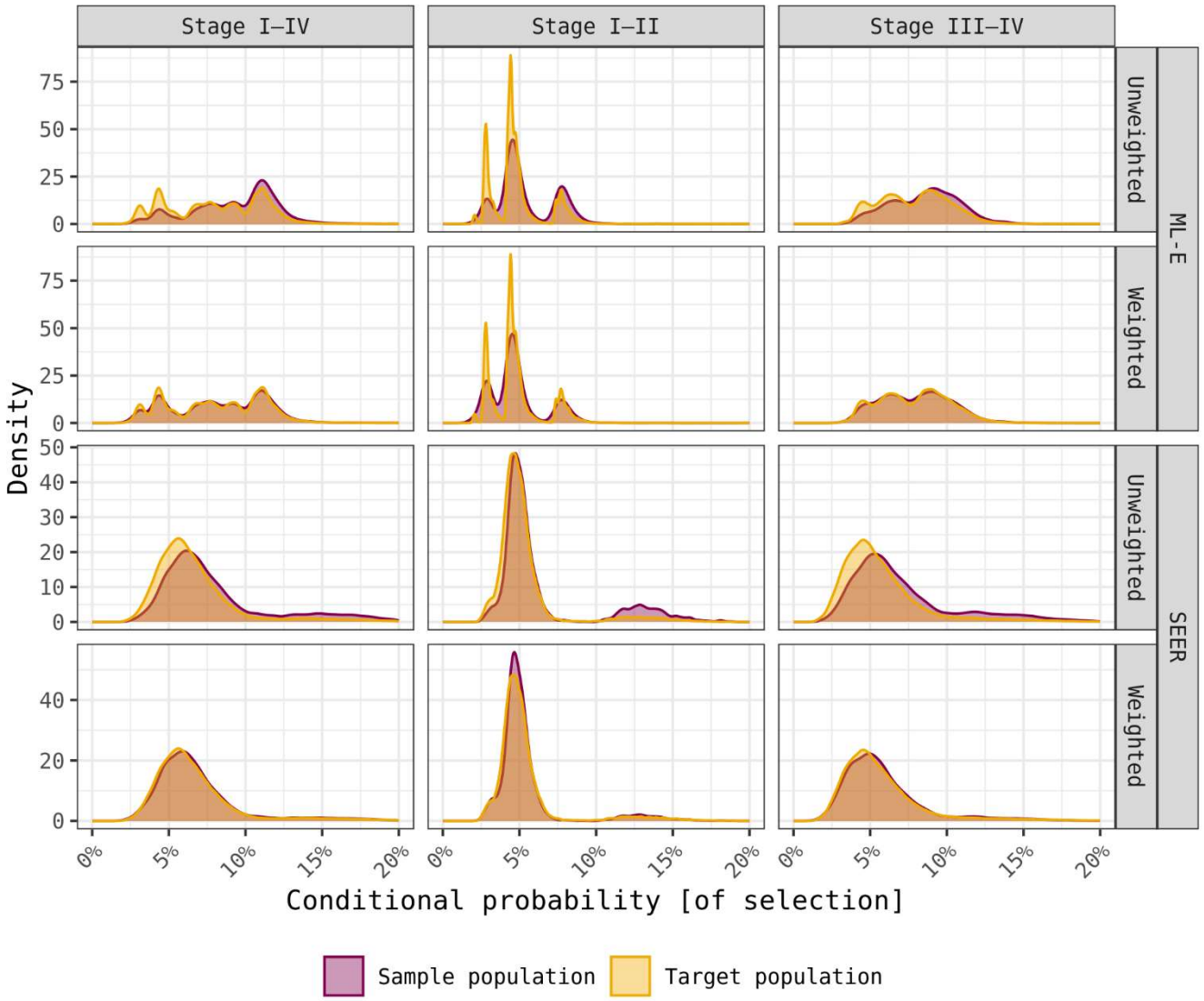


Figure 2: Kernel densities for the estimated probabilities of sample selection conditional on a set of baseline variables, before and after weighting, using the ML-E or SEER as the referent population. Abbreviations: ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]

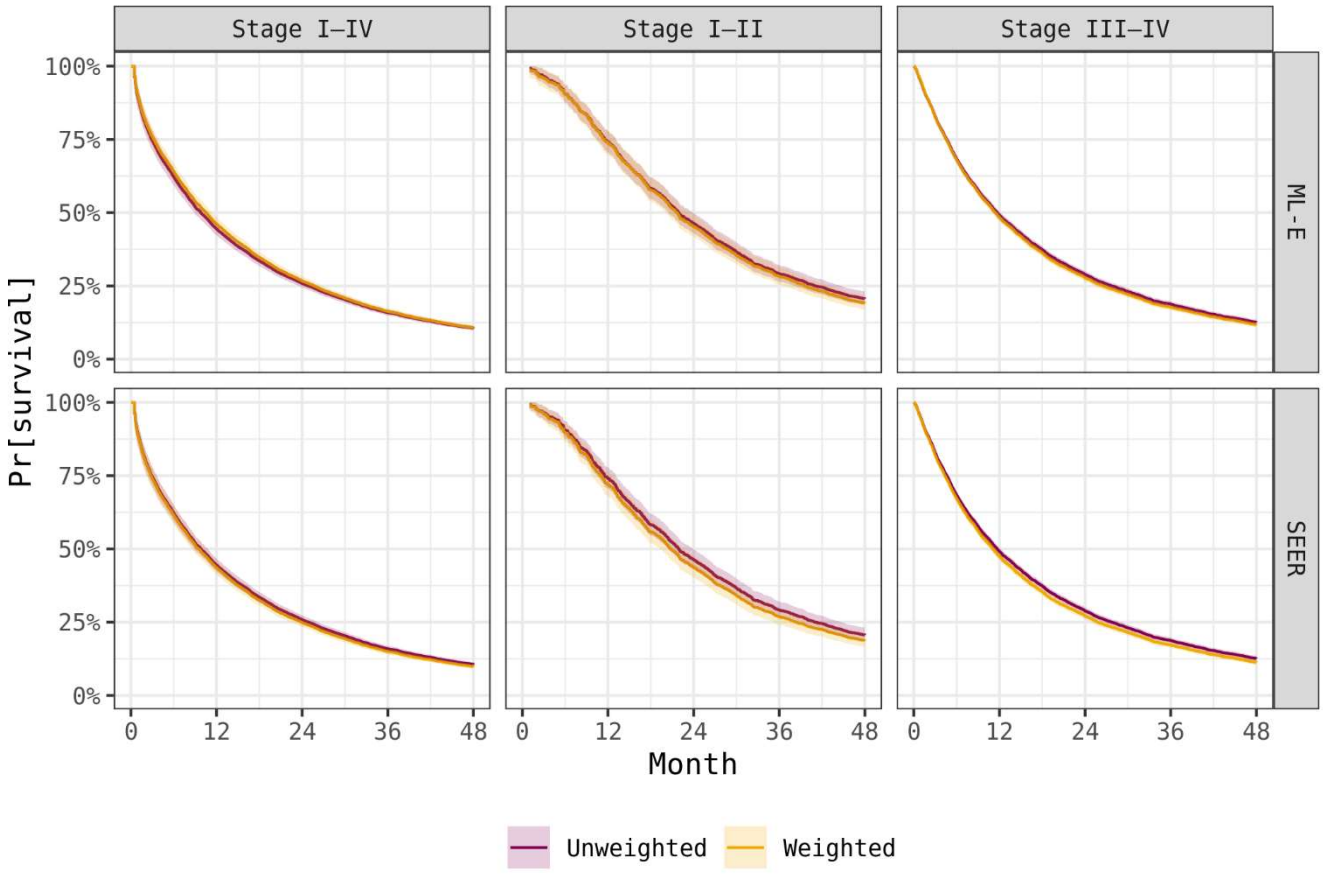


Figure 3: Kaplan-Meier estimation of rwOS for CGDB patients as an unweighted sample and extended to the population within a CGDB sample weighted to be representative of the SEER or ML-E referent populations (facet labels). Time zero was the date of diagnosis for stages I–IV & I–II and the date of initiating 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]

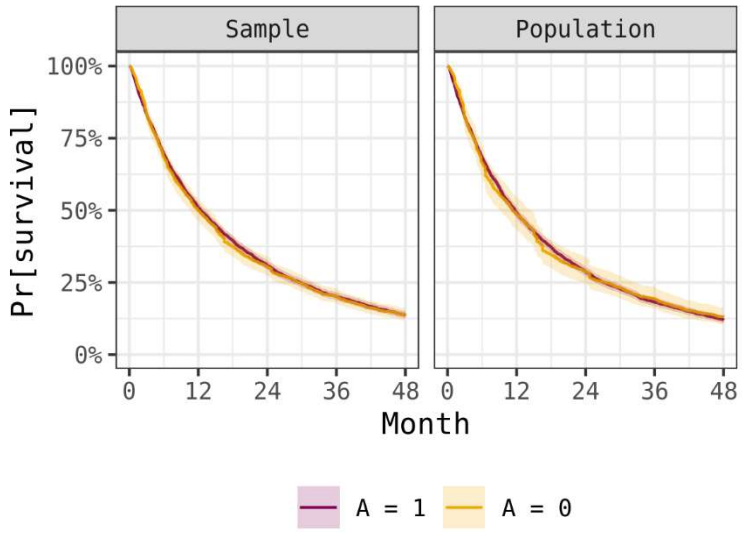


Figure 4: Counterfactual rwOS curves under a hypothetical binary treatment *A* for stage III-IV CGDB patients calculated within the selected sample and extended to the target population. Curves for the sample were weighted with treatment weights and were extended to the population with the product-terms of selection & treatment weights. Abbreviations: rwOS, Real-world Overall Survival

Supplementary Table 1: 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

Supplementary Table 2: 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

Supplementary Table 3: 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

Supplementary Table 4: Baseline characteristics of stages I–II ML-E & SEER cohorts, in the unweighted and weighted

Characteristic	ML-E (n 56,031)	SEER (n 49,685)	ASD
Age at diagnosis[1]	71.0 (65.0, 76.0)	71.0 (64.0, 78.0)	0.07
Gender			0.00
Female	30,453 (54.4%)	26,894 (54.1%)	
Male	25,578 (45.6%)	22,791 (45.9%)	
Race			0.30
Black	4,250 (7.6%)	4,637 (9.3%)	
Other	8,827 (15.8%)	3,185 (6.4%)	
White	42,954 (76.7%)	41,863 (84.3%)	
SEER stage			0.00
Localized	56,031 (100.0%)	49,685 (100.0%)	
Histological classification			0.09
Non-squamous cell carcinoma	37,856 (67.6%)	35,578 (71.6%)	
Squamous cell carcinoma	18,175 (32.4%)	14,107 (28.4%)	

Abbreviations: ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]; ASD, Absolute Standardized Difference

[1] Median (Interquartile range)

Supplementary Table 5: Baseline characteristics of stages III–IV ML-E & SEER cohorts, in the unweighted and weighted

Characteristic	ML-E (n 65,580)	SEER (n 149,056)	ASD
Age at diagnosis[1]	68.0 (61.0, 75.0)	69.0 (61.0, 77.0)	0.17
Gender			0.02
Female	31,170 (47.5%)	69,327 (46.5%)	
Male	34,410 (52.5%)	79,729 (53.5%)	
Race			0.28
Black	6,326 (9.6%)	18,405 (12.3%)	
Other	11,390 (17.4%)	12,126 (8.1%)	
White	47,864 (73.0%)	118,525 (79.5%)	
SEER stage			0.18
Regional	16,248 (24.8%)	49,014 (32.9%)	
Distant	49,332 (75.2%)	100,042 (67.1%)	
Histological classification			0.07
Non-squamous cell carcinoma	48,592 (74.1%)	105,969 (71.1%)	
Squamous cell carcinoma	16,988 (25.9%)	43,087 (28.9%)	

Abbreviations: ML-E, Machine Learning-Extracted [cohort]; SEER, Surveillance, Epidemiology, and End Results [cohort]; ASD, Absolute Standardized Difference

[1] Median (Interquartile range)

Supplementary Table 6: 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)

Supplementary Table 7: 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)

Supplementary Table 8: 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)

Supplementary Table 9: 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)

Supplementary Table 10: Sensitivity analysis comparing the baseline characteristics of stages I–IV CGDB patients and SEER cancer registrations, harmonized to the same study period (2011–2016)

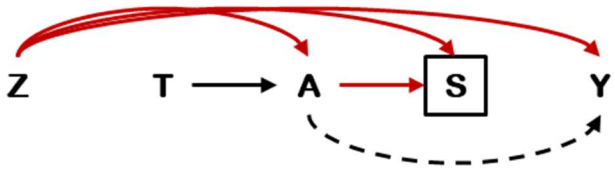
	Unweighted		
Characteristic	**CGDB (n 5,249)**	**SEER (n 198,741)**	**ASD**
Age at diagnosis[1]	66.0 (58.0, 73.0)	70.0 (62.0, 77.0)	0.41
Gender			0.10
Female	2,813 (53.6%)	96,221 (48.4%)	
Male	2,436 (46.4%)	102,520 (51.6%)	
Race			0.33
Black	325 (6.2%)	23,042 (11.6%)	
Other	902 (17.2%)	15,311 (7.7%)	
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)

SUPPLEMENTARY APPENDIX

We present our causal assumptions in Supplementary Figure 1, where *A* represents a hypothetical binary treatment, *Y* the outcome of Real-world Overall Survival, *Z* a set of baseline variables (age, gender, race, stage, histological classification), *T* a binary node indicating having undergone Next Generation Sequencing (NGS), and *S* a node indicating selection into the sample [vs. SEER target population]. A [box] depicts conditioning through experimental design (restriction where *S=1*). The causal and non-causal paths are depicted as dashed & red lines, respectively.



Supplementary Figure 1: Directed Acyclic Graph showing the causal assumptions underlying the study of a convenience CGDB sample.

Selection into the sample from the SEER target population was assumed *a priori* to be a product of geographic sampling and further eligibility criteria. Note that while in real-world clinical practice a set of baseline variables *Z* can cause the ordering of NGS *T*, in our example *T* is defined only by the population. SEER patients by definition cannot undergo NGS. Nevertheless, any differential selection of patient subgroups through the requirement of NGS is corrected downstream by weighting on the Inverse Probability of Selection. Separately, undergoing NGS *T* only affects the outcome *Y* through directing treatment *A* based on the tumor genotype.

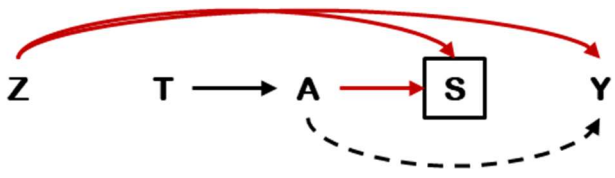
In the language of causal Directed Acyclic Graphs [1], a type-1 selection bias (collider-restriction bias) results from conditioning on a common effect of both the treatment (or cause of treatment) and the outcome (or cause of the outcome)

[2,3]. 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 [4,5]. Where type-1 selection biases is a result of conditioning on a common effect of A & Y (A > [Z] < Y), confounding biases are defined by open backdoor paths dealing with common causes of A & Y (A < Z > Y).

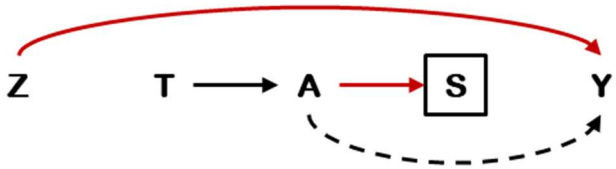
The graph above illustrates three paths:

- a causal path (*A* > *Y*) which we wish to isolate;
- a non-causal confounding path (*A* < *Z* > *Y*);
- a non-causal collider-restriction path (*A* > [*S*] < *Z* > *Y*), invoked by restricting on a collider variable by restricting on levels of *S* — a common effect of *Z* and treatment *A* (i.e. type-1 selection bias)

When use Inverse Probability of Treatment Weights, we weight on the conditional probability of *A* given *Z* (*Pr[A|Z]*), thereby severing the path *Z* > *A*. When use Inverse Probability of Selection Weights, we weight on the conditional probability of *S* given *Z* (*Pr[S|Z]*), and therefore sever the path *Z* > *S*. In the estimation of the Sample Average Treatment Effect, only the confounding path is blocked (with Inverse Probability of Treatment Weights), and there remains open the collider-restriction path as shown in Supplementary Figure 2. When we wish to extend the SATE to the PATE , we further weight the sample on the product terms of Inverse Probability of Treatment & Selection Weights as shown in Supplementary Figure 3.



Supplementary Figure 2: Identification of the Sample Average Treatment Effect (SATE). Note the presence of a backdoor collider-restriction path.



Supplementary Figure 3: Identification of the Sample Average Treatment Effect (SATE). Note the closure of all backdoor paths.

REFERENCES

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

2. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5):615-625. doi:10.1097/01.ede.0000135174.63482.43

3. Lu H, Cole SR, Howe CJ, Westreich D. Toward a clearer definition of selection bias when estimating causal effects. *Epidemiology*. 2022;33(5):699-706. doi:10.1097/EDE.0000000000001516

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

5. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol*. 2010;39(2):417-420. doi:10.1093/ije/dyp334