Uncovering interpretable potential confounders in electronic medical records

Jiaming Zeng* Michael F. Gensheimer † Daniel L. Rubin† Susan Athey‡ Ross D. Shachter*

February 17, 2021

Abstract

In medicine, randomized clinical trials (RCT) are the gold standard for informing treatment decisions. Observational comparative effectiveness research (CER) is often plagued by selection bias, and expert-selected covariates may not be sufficient to adjust for confounding. We explore how the unstructured clinical text in electronic medical records (EMR) can be used to reduce selection bias and improve medical practice. We develop a method based on natural language processing to uncover interpretable potential confounders from the clinical text. We validate our method by comparing the hazard ratio (HR) from survival analysis with and without the confounders against the results from established RCTs. We apply our method to four study cohorts built from localized prostate and lung cancer datasets from the Stanford Cancer Institute Research Database and show that our method adjusts the HR estimate towards the RCT results. We further confirm that the uncovered terms can be interpreted by an oncologist as potential confounders. This research helps enable more credible causal inference using data from EMRs, offers a transparent way to improve the design of observational CER, and could inform high-stake medical decisions. Our method can also be applied to studies within and beyond medicine to extract important information from observational data to support decisions.

*Department of Management Science and Engineering, Stanford University. This research was supported by a seed grant from the Stanford Institute for Human-Centered Artificial Intelligence.

†School of Medicine, Stanford University

‡Graduate School of Business, Stanford University. Athey thanks the Sloan Foundation and the Office of Naval Research grant ONR N00014-17-1-2131 for generous support.

1 Introduction

As the number of highly targeted cancer treatments increase, it is increasingly difficult for oncologists to decide on optimal treatment practices. In the recent years, medicine has seen the reversal of 146 standard medical practices [1], and many unanswered questions remain on treatment decisions in oncology. The gold standard for assessing treatment effects is randomized clinical trials (RCT). However, RCTs can be very expensive, time-consuming, and limited by the lack of external validity [2, 3]. Hence, there has been a growing interest in using observational data to compare and evaluate the effectiveness of clinical interventions, also known as *comparative effectiveness research* (CER) [2].

Many studies have used large-scale observational registries such as the Surveillance, Epidemiology, and Ends Results (SEER) and National Cancer Data Base (NCDB) to perform CER. However, such studies may be unreliable due to the systemic bias present in observational data and the presence of unmeasured confounders [1, 2, 4]. Moreover, population-based CERs in oncology often also face small data challenges. Electronic medical records (EMRs) are another source of rich observational information on patient demographics and past medical history. We hypothesize that the more detailed unstructured data present in EMRs can be harnessed to reduce confounding compared to prior CER studies.

We study how EMRs, especially clinical text, can be used to reduce selection bias in observational CER studies and better inform treatment decisions in oncology. A *confounder* is a variable that is associated with both treatment assignment and the potential outcomes a subject would have under different treatment regimes. In the presence of confounders, the correlation between treatment assignment and outcomes cannot be interpreted as causal. One way that confounding may arise is when patients are selected for a treatment group on the basis of the severity of their illness. In such a case, failing to adjust for patient severity can lead to *selection bias* when attempting to estimate causal effects. For exam-

ple, surgery tends to be performed on younger or healthier patients; certain doctors or institutions may prefer one treatment over another, and this creates confounding if those doctors or institutions treat patients with systematically different severity. Studies based on a small set of covariates tend not to capture the important confounders and result in biased estimates [5, 6]. Observational studies are more reliable when we can better control for these confounders. EMR clinical text is a potential source of additional information about factors that might relate to both treatment assignment and prognosis.

We use natural language processing (NLP) to uncover interpretable potential confounders from the EMRs for treatment decisions. For high-stake settings such as cancer treatment decisions, it is important to design models that are interpretable for trust and understanding [7]. NLP can be used to process the unstructured clinical notes and create covariates. We then augment our dataset with covariates that impact both treatment assignment and patient outcomes, where attempting to estimate causal effects while omitting such variables leads to biased estimates [8, 9]. Finally, we use methods designed to estimate causal effects in observational studies with observed confounders to estimate treatment effects in our augmented data set. We show that controlling for these confounders appears to reduce selection bias when compared against the results from established RCTs and clinical judgement.

We apply our method to localized prostate and lung cancer patients. Based on cohorts from established RCTs, we built four treatment groups for comparison. We uncovered interpretable potential confounders from clinical text and validated the potential confounders against the results from the RCTs. Simple NLP techniques (e.g. lemmatization, entity identification) were used to construct a bag-of-words representation of the frequently occurring terms. A Lasso model [10] was then used to select the terms that are predictive of both the treatment and survival outcome as potential confounders. Finally, we validated

our method by comparing the hazard ratio (HR) from survival analysis with and without the confounders.

Our main contribution is presenting an approach to uncover interpretable potential confounders from clinical text. Existing work in observational causal inference rarely employs unstructured data [11, 12, 13, 14], and most NLP studies on clinical text focus on prediction settings [15, 16, 17, 18]. Our paper is the first to uncover interpretable potential confounders from clinical notes for causal analysis on cancer therapies, and one of the few works that combines NLP and causal inference in a time-to-event setting. Our method allows researchers to extract and control for confounders that are not typically available. This appears to be a useful step for future observational CER studies to help reduce selection bias unique to that dataset. The research presented can help unlock the potential of clinical notes to help clinicians understand current clinical practice and support future medical decisions.

1.1 Related Work

In the past decade, there has been a growing interest in using observational data for clinical decision making and causal inference in oncology [2]. However, such studies are often unreliable, and many observational studies have been refuted by RCTs soon after [2, 4]. For example, Yeh et al. [6] performed a comparison of surgery vs. radiotherapy for oropharynx cancer and suggested that surgery may be superior to radiation for quality of life outcomes. A few years later, this claim was refuted by an RCT study Nichols et al. [19], which showed that radiation is in fact superior to surgery in terms of 1-year quality of life scores. A similar example is seen with prostate cancer. In 2016, Wallis et al. [5] showed through population-based studies that surgery is superior to radiation for early-stage prostate cancer for overall and prostate-cancer specific survival; a few months later,

the finding was refuted by Hamdy et al. [20], which showed that surgery and radiation are equivalent in terms of overall and prostate-cancer specific survival. Many other studies have shown the fallibility of population CERs that rely on expert-curated features to draw conclusions about treatment effects [2, 21].

Beyond clinical studies, there is a relatively large literature on performing causal inference from observational data. Various papers have explored how to correct for bias when evaluating average treatment effect (ATE) from observational studies with propensity score matching or weighting [22, 11, 12]; see [23] for a review. There is also a growing amount of literature that adapts machine learning models, such as random forest or regularized regression, for doubly-robust ATE estimation in high-dimensional settings [24, 25, 26, 27, 9, 28]. However, most of the methods do not include unstructured data.

Recent literature has shown the usefulness of conditioning on textual data to adjust for confounding [29, 30, 31, 32]. Roberts et al. [31] proposes text matching to employ textual data for causal inference. Mozer et al. [30] applies text matching to patient charts texts for a medical procedure evaluation; however, they focus on continuous outcomes and rely mostly on expert-curated terms from the clinical text. Veitch et al. [32] is another work that employ unstructured data for causal inference; however, they rely on black-box models that are not interpretable. Moreover, many existing causal inference methods are developed for continuous outcomes and do not transfer easily to the time-to-event outcomes for survival analysis used in oncology. Of the ones that perform causal inference on time-to-event outcomes for medical applications [13, 14], we did not find any that include unstructured data in a systematic way. Austin [14] presents methods for using propensity scores to reduce bias in observational studies with time-to-event outcomes. Our study leverages some of the ideas and methods in this literature to develop our approach for identifying and evaluating the potential confounders from the unstructured clinical notes.

There is also a growing literature that seeks to better employ EMRs for clinical tasks. Existing work has employed structured EMR data and unstructured clinical notes for survival prediction and analysis [17, 33], predicting metastatic recurrence [15, 16], clinical risk prediction [34], and prediction of multiple medical events [18]. However, most current work involving EMRs focuses on prediction tasks. In studies that include the unstructured notes, most use deep learning to produce context-rich embedding representations of words or documents [17, 18]. While these representations are highly accurate for prediction tasks, they are often black-box and very difficult to interpret for causal insights. Our approach differs in that we use simple NLP techniques (e.g. entity identification, bag-of-words) to generate matrix representations that can be easily mapped to specific words and phrases. This increases the interpretability of our method and allows us to explain our confounders to clinicians.

Our study advances both the clinical and causal inference literature by using NLP to perform causal inference on clinical text in time-to-event settings. We hope this will inform clinical practice and improve patient outcomes.

2 Results

We apply our methods to localized prostate and stage I non-small cell lung cancer (NSCLC) patients and compare the results against established RCTs. We select these diseases due to data availability and having established clinical RCTs for validation. After filtering and assignment, we include 1,822 patients for prostate cancer, with 988 surgery patients, 385 radiation patients, and 449 active monitoring patients; the average follow-up time is 4.11 years. For stage I NSCLC, we include 749 patients, with 492 surgery patients and 257 radiation patients; the average follow-up time is 4.96 years. The patient characteristic descriptions of the prostate cancer cohort are shown in Table 1 and the NSCLC cohort are

shown in Table 2.

We use the findings from established RCTs and clinical judgement as a benchmark for evaluating our results. For localized prostate cancer, Hamdy et al. [20] compared active monitoring, radical prostatectomy, and external-beam radiotherapy. A total of 1,643 patients were included in the study, with 553 men assigned to surgery, 545 men assigned to radiotherapy, and 545 men to active monitoring. They observed no significant difference among the groups for prostate-cancer or all-cause mortality ($P = 0.48$ and $P = 0.87$ respectively). Similarly, a recent study showed that difference in treatment effects for surgery vs. radiation observed from observational studies is entirely due to treatment selection bias [35]. For stage I NSCLC, The Chang et al. [36] study is a pooled study comparing stereotactic ablative radiotherapy (SABR) to surgery. A total of 58 patients were included, with 31 patients assigned to SABR and 27 to surgery. The study observed that SABR had slightly better overall survival than surgery ($P = 0.037$), but claims to be consistent with the clinical judgement that surgery is equipoise to radiation.

Following the design of Hamdy et al. [20] and Chang et al. [36], we evaluate our results for the following four treatment groups for an outcome of all-cause mortality:

- *surgery* vs. *radiation* for prostate cancer
- *surgery* vs. *monitoring* for prostate cancer
- *radiation* vs. *monitoring* for prostate cancer
- *surgery* vs. *radiation* for stage I NSCLC

We do not analyze other treatment groups for lung cancer due to patient count constraints.

Our approach identifies covariates that are likely potential confounders in this particular dataset from the high-dimensional and high-noise EMR data. These covariates are interpretable as they are represented by structured data or words from a bag-of-words

matrix. To evaluate the effectiveness of the potential confounders selected in the model, we use these potential confounders to perform survival analysis for the treatment groups for prostate and stage I NSCLC. We compare the results of various methods for time-to-event analysis in terms of HR. Although we cannot know what the true HR is, we suggest that using medical notes improves on the traditional covariates. We compare our results against existing RCTs to evaluate how the confounders we have uncovered can help correct for selection bias. The overall workflow is shown in Figure 1.

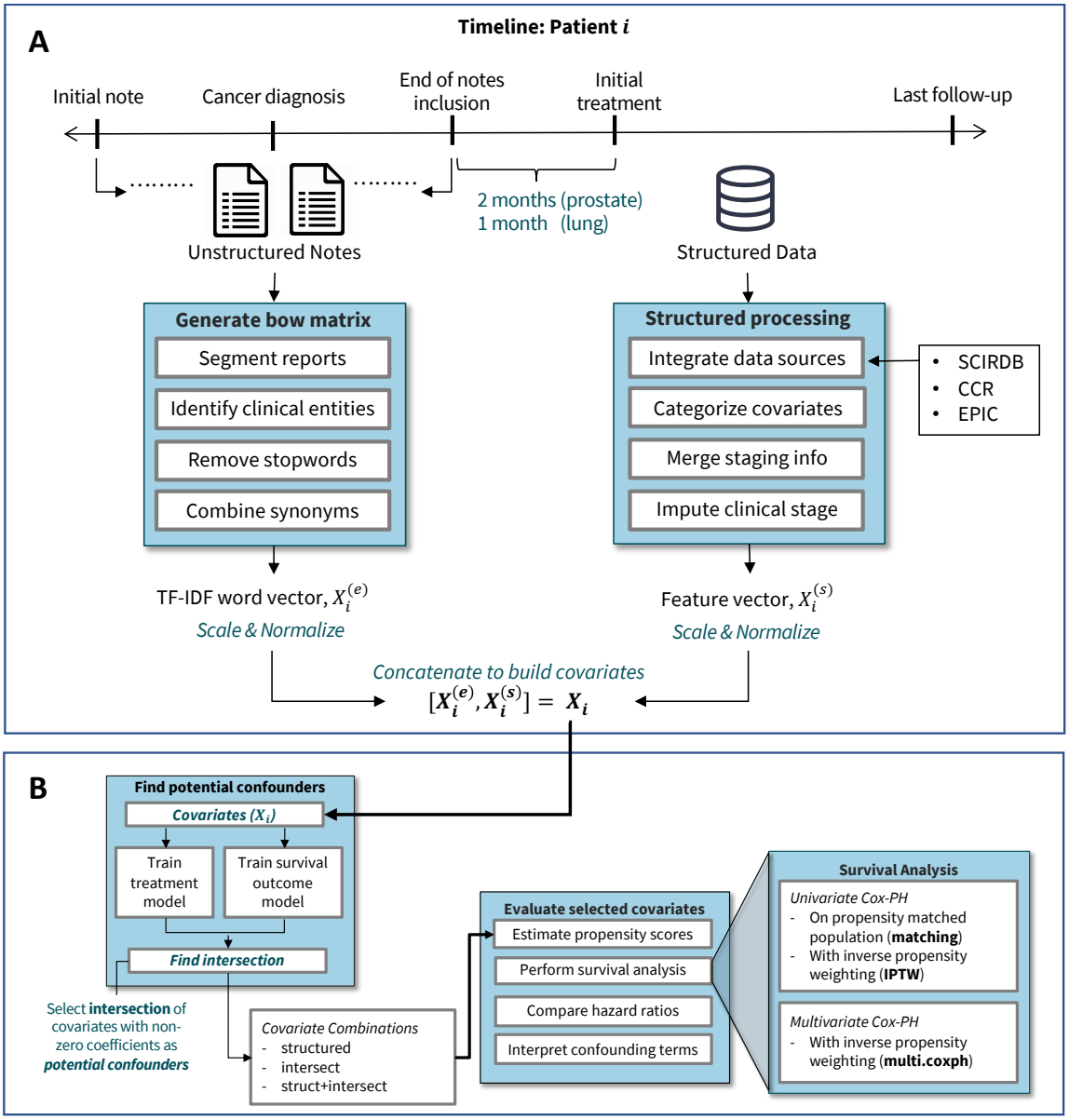


Figure 1: Pictorial overview for uncovering potential confounders. **A** *Data processing for each patient.* We preprocess and concatenate the structured and unstructured covariates before applying our method. **B** *Workflow for identifying how potential confounders affect survival analysis for each treatment group.* We uncover covariates that are predictive of both the treatment and outcome as potential confounders. We then perform survival analysis on different combinations of the selected covariates.

Table 1: Characteristics of the localized prostate cancer patients.

	Treatment Groups		
Features	*Surgery* ($n = 988$)	*Radiation* ($n = 385$)	*Monitoring* ($n = 449$)
Age, mean (std)	64.04 (7.8)	70.18 (7.6)	66.22 (8.2)
Race, no. (%)			
white	709 (71.8%)	221 (57.4%)	292 (65.0%)
black	32 (3.2%)	15 (3.9%)	14 (3.1%)
asian	94 (9.5%)	41 (10.6%)	42 (9.4%)
unknown	153 (15.4%)	108 (28.1%)	101 (22.5%)
Ethnicity, no. (%)			
hispanic	71 (7.1%)	20 (5.2%)	23 (5.1%)
non-hispanic	890 (90.1%)	348 (90.4%)	393 (87.5%)
unknown	27 (2.7%)	17 (4.4%))	33 (7.3%)
Clinical Stage, no. (%)			
stage I	219 (22.2%)	36 (9.4%)	227 (50.6%)
stage II	750 (75.9%)	289 (75.1%)	217 (48.3%)
stage III	12 (1.2%)	38 (9.9%)	3 (0.7%)
stage IV	7 (0.7%)	22 (5.7%)	2 (0.4%)
Tumor Grade, no. (%)			
grade 1	66 (66.8%)	33 (8.6%)	157 ((35.0%)
grade 2	429 (43.4%)	132 (34.3%)	205 (45.7%)
grade 3	474 (48.0%)	208 (54.0%)	62 (13.8%)
grade 4	3 (0.3%)	2 (0.5%)	0 (0%)
unknown	16 (1.6%)	10 (2.6%)	25 (5.6%)
No. notes/patient, mean (std)	24.96 (44.4)	53.93 (105.7)	54.48 (93.4)
Days of survival, mean (std)	1,564.90 (979.4)	1424.76 (1,031.6)	1,403.72 (921.2)
Death, no. (%)	70 (7.1%)	19 (4.9%)	17 (3.8%)

Diagnosis Year: 2008-2017; Avg. follow up: 4.11 years

Table 2: Characteristics of the stage I lung cancer patients.

	Treatment Groups	
Features	*Surgery* ($n = 484$)	*Radiation* ($n = 224$)
Age, mean (std)	68.05 (10.7)	74.60 (9.1)
Gender, no. (%)		
female	299 (62.0%)	87 (41.2%)
male	185 (38.0%)	137 (58.8%)
Race, no. (%)		
white	293 (60.8%)	152 (66.5%)
black	12 (2.2%)	5 (3.5%)
asian and pacific islander	99 (20.1%)	18 (9.7%)
unknown	80 (16.9%)	49 (20.2%)
Ethnicity, no. (%)		
hispanic	23 (4.9%)	10 (3.9%)
non-hispanic	411 (84.3%)	178 (81.7%)
unknown	50 (10.8%)	36 (14.4%)
No. notes/patient, mean (std)	57.49 (101.2)	57.73 (134.9)
Days of survival, mean (std)	2,060.13 (1,207.5)	1,350.29 (914.1)
Death, no. (%)	120 (24.8%)	126 (53.3%)

Diagnosis Year: 2000-2017; Avg. follow up: 4.96 years

2.1 Potential Confounders

We show that our methods uncover terms that are predictive of both the treatment and survival outcome. Hence, these are potential confounders that should be controlled for in observational CERs to reduce selection bias.

We select the intersection covariates from our treatment and outcome prediction models as the potential confounders. We base this idea on the selection of union variables to reduce confounding when performing causal inference on observational data in the case of continuous outcomes [9]. However, in survival analysis, it is recommended that the covariates analyzed be constrained by the statistical 1 in 10/20 rule of thumb with respect to the the event count [37, 38]. In our high-dimensional setting, the union of covariates that are predictive of treatment and outcome yield too many potential confounders relative to the sample size. Hence, we use the intersect as a heuristic to focus on the most important confounders.

In Figure 2, we illustrate the unpenalized coefficients of covariates from two models, the treatment assignment model and the survival outcome model. For each covariate, the x-axis plots the coefficient from the treatment prediction model while the y-axis plots the coefficient from the survival outcome model. Each covariate is labeled by the text next to it. The intersection covariates, `intersect`, are shown in blue; these are the covariates that have strong effects in both models. For the structured covariates, we illustrate in black the coefficients for the covariates that were not selected; these coefficients are closer to at least one of the axes in the figure. We do not illustrate the coefficients for unstructured covariates that are not selected, as there are a large number of these covariates. The axes are labeled to indicate which treatment the coefficient predicts and whether the coefficient is indicative of a good or bad survival prognosis. For example, in the treatment model, patients with a high "bladder" word-occurrence have a higher likelihood of receiving surgery; in the outcome

model, patients with a high "bladder" occurrence have a lower likelihood of survival.

In Supplementary A.2, we show the R^2 correlation among all the selected covariates for each treatment group.

Figure 2: For each treatment group, we show the unpenalized coefficients for the struct+intersect covariates. Blue text indicates the intersect covariates that have been selected as potential confounders by our method. Black text indicates the structured covariates that have not been selected. For the treatment model, these are the coefficients to a linear model. For the survival outcome model, these are the β for the Cox-PH model. The dotted lines are the axis, denote a coefficient value of 0.

2.2 Evaluation of Potential Confounders

We evaluate these potential confounders by comparing the results on 3 covariate combinations:

- structured: Using only the structured covariates. We use this as a baseline because

these are covariates that are typically used in retrospective oncology studies and are readily available in the structured data [5].

- intersect: Using only the intersection covariates identified as confounders.
- struct+intersect: Using the union of the structured and intersection variables.

In Figure 3, we show the hazard ratio (HR) of the effect of treatment for each study cohort when the selected covariates are included in analysis. An HR below 1 indicates that patients with the second treatment are more likely to survive than those with the first treatments. An HR above 1 indicates the opposite, and an HR equal to 1 indicates that the two treatments are equipose.

We observe that with the additional covariates, we are able to shift the estimate of the HR towards the direction of the RCT for an outcome of all-cause mortality. We also compare the covariate-specific HR of each of the selected covariates in terms of univariate and multivariate Cox-PH analysis for an all-cause mortality outcome in Tables 3-6.

In Figure 3A and Table 3, we show the results with *surgery* vs. *radiation* for prostate cancer. The RCT reports no significant difference between *surgery* vs. *radiation* for localized prostate cancer [20]. With structured, we observe a significant effect that radiation is superior to surgery, a result that disagrees with most retrospective studies [5]. We observe a significant shift in the HR towards equipoise with the additional identified confounders for intersection and struct+intersect. For structured, we observe an HR of 2.51 with 95% CI (2.39-4.55) and p-value of 0.002 with **multi.coxph**. For struct+intersect, we estimate an HR of 1.54 with 95% CI (0.78-3.03) and p-value of 0.214 with **multi.coxph**. We shift the HR point estimate by 0.97, or 38.6%, towards equipose.

In Figure 3B and Table 4, we show the results of *surgery* vs. *active monitoring* for prostate cancer. Hamdy et al. [20], the RCT, reports the HR for *surgery* vs. *active*

monitoring as 0.93 with 95% CI (0.65, 1.35) and p-value of 0.92. With structured, we again have a significant effect that active monitoring is superior to surgery; this disagrees with most retrospective studies [5] and Hamdy et al. [20]. We again observe a significant shift in the HR towards equipoise with the additional identified confounders. For structured, we observe an HR of 2.71 with 95% CI (1.55-4.75) and p-value < 0.001 with **multi.coxph**. For struct+intersect, we estimate an HR of 1.10 with 95% CI (0.55-2.21) and p-value of 0.781 with **multi.coxph**. We shift the HR point estimate by 1.61, or 59.1%, towards equipose.

In Figure 3C and Table 5, we show the results of *radiation* vs. *active monitoring* for prostate cancer. We do not see as significant a shift with *radiation* vs. *active monitoring*. Hamdy et al. [20] records the HR for *radiation* vs. *active monitoring* as 0.94 with 95% CI of (0.65, 1.36) and p-value of 0.92. We observe that the **matching** results are not very far from the RCT results. All results with intersect and struct+intersect shift the HR estimate slightly towards equipose, with the most shift of 0.32 by intersect and **IPTW**. We suspect this may be due to the smaller dataset available for *radiation* vs. *active monitoring* or the confounding not being observable within the text.

In Figure 3D and Table 6, we show the results with *surgery* vs. *radiation* for stage I NSCLC. With *structured*, we observe a significant effect that surgery is superior to radiation. The results from Chang et al. [36] and clinical judgement tells us that surgery and radiation should be about equipose for stage I NSCLC. The shift is not as significant as with prostate cancer, but we also note that the established clinical standard for lung cancer is not as well studied. We do observe a more significant shift with **multi.coxph**; there is a slight shift with **IPTW** and **matching**. For structured, we observe an HR of 0.39 with 95% CI (0.30-0.51) and p-value < 0.001 with **multi.coxph**. For struct+intersect, we estimate an HR of 0.54 with 95% CI (0.40-0.53) and p-value < 0.001 with **multi.coxph**. We shift the HR point estimate by 0.15 towards equipose. We suspect the small changes with **IPTW** and

matching are again due to the even smaller data size of stage I NSCLC. The doubly-robust method of **multi.coxph** seem to perform better under these settings.

Overall, our methods uncover several potential confounders that can reduce selection bias in observational data. Although our method cannot uncover all potential confounders, we are able to uncover confounders that are not usually included in expert-selected covariates. Supplementary analysis of propensity scores and covariate balance plots for each analysis are seen in Supplementary A.1.

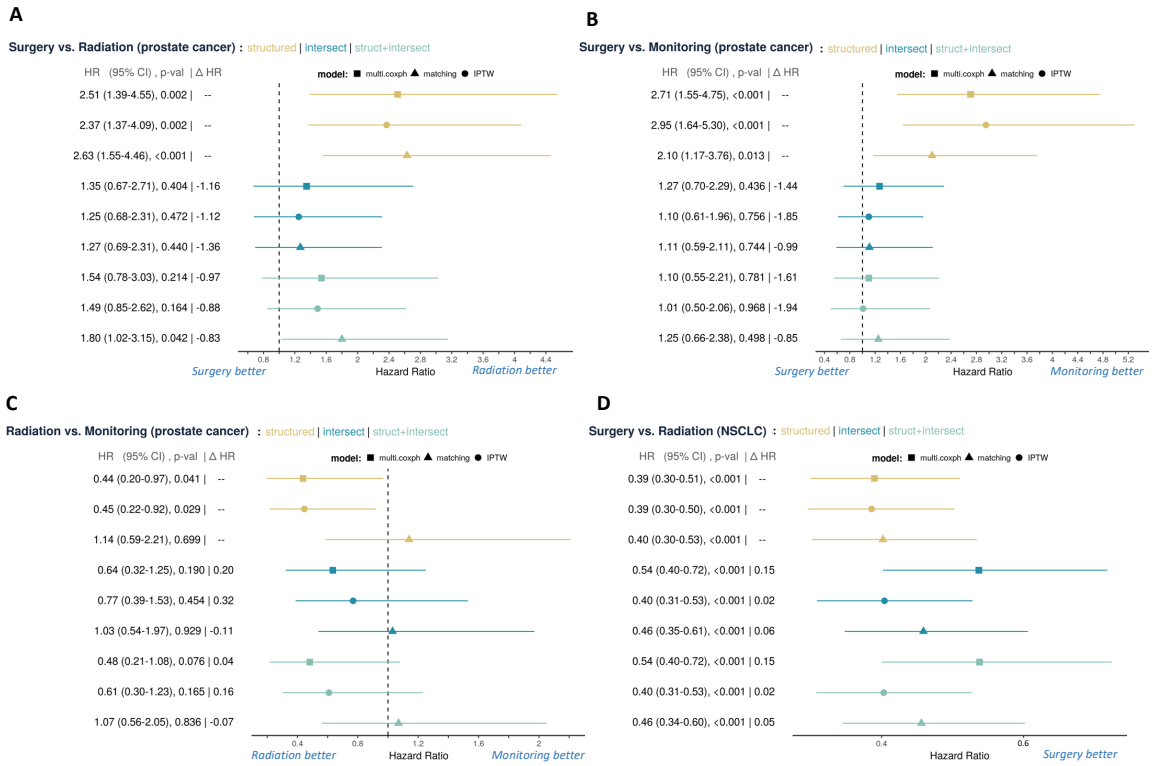


Figure 3: Forest plots of each of the comparison groups. The left-hand label is shown as: HR (95% CI), p-value | Δ HR. The Δ HR measure is the difference of the current HR estimate and the baseline, structured, HR estimate. We see that the inclusion of our potential confounders shift the HR point estimate in the direction of the RCT and reduces the selection bias. The blue labels below each graph indicate which treatment is better in terms of HR comparison. For D, *radiation better* is not displayed because the HR value is not shifted beyond 1.0, the direction of radiation better.

2.3 Potential Confounder Interpretation

We show that the potential confounders we have uncovered are interpretable through clinical expertise. We examine the effect on survival for each selected covariate in term of univariate and multivariate survival analysis with a Cox-PH model. In univariate analysis, a single covariate is regressed on the survival outcome and describes the survival with respect to a single covariate. In multivariate analysis, all the selected covariates are regressed on the survival outcome and describe each covariate's effect on survival while adjusting for the impact of all selected covariates. For a particular variable, an HR below 1 indicates that the covariate is a positive predictor of survival, an HR above 1 indicates a negative predictor of survival, and an HR equal to 1 means that the variable does not seem to effect survival.

2.3.1 Prostate Cancer

For *surgery* vs. *radiation* and *surgery* vs. *active monitoring*, `pat_age`, `bladder`, and `urothelial` are chosen as intersection covariates. Moreover, they are also shown to be significant through both univariate and multivariate covariate analysis in Tables 3 and 4.

Patient age is a known confounder in treatment decision and survival outcomes. Older patients are more likely to receive radiation due to surgery risk. However, older patients also have higher mortality.

Patients with bladder cancer or bladder issues are more likely to get surgery than radiation. Radiation does not work well for bladder cancer. Patients with bladder problems may prefer surgery because radiation can irritate the bladder and cause urinary problems. However, these are also patients with higher mortality and more health issues.

Moreover, for this particular dataset, we note that the confounding appear to be observable. The bias of *surgery* being worse than *radiation* and *monitoring* is due to a group

of patients who are diagnosed with prostate cancer through a resection for bladder cancer or other bladder issues. When a patient with bladder cancer has a cystoprostatectomy in which the bladder and prostate are both removed, a pathologist can sometimes find a prostate tumor in the pathology specimen. Bladder cancer patients tend to be older, have more medical issues, and a higher mortality rate. However, these patients often have low clinical stage for prostate cancer. The terms `bladder` and `urothelial` describe this group of patients. Our method is able to capture some characteristics of this group and use this to reduce selection bias.

For *radiation* vs. *active monitoring*, we do not observe confounders that are significant in Table 5. `bladder` is picked up but does not seem to have a significant effect. We do see some interesting terms that could indicate poor health, such as `abnormal`, `dysfunction`, or `wound`. We also observe selection of medical conditions that could confound treatment, such as `lymphoma` or `oropharynx`. It can be that the confounding here is not as easily observable or our method is unable to identify it.

2.3.2 Lung Cancer

We examine Table 6 for the intersection covariates through univariate and multivariate analysis. We observe that some of the significant terms are `pat_age`, `male`, `race_api`, `diagnosis_year`, `alk`, `left.low`, and `severe`.

We note that age, gender, race, and diagnosis year are known confounders for treatment decision and outcome.

The covariate `alk` points to the ALK mutation for NSCLC. About 5% of NSCLCs have a rearrangement in a gene called ALK; the ALK gene rearrangement produces an abnormal ALK protein that causes the cells to grow and spread. This change is often seen in non-smokers (or light smokers) who are younger and who have the adenocarcinoma subtype of

NSCLC [39]. We hypothesize that `alk` is significant because the ALK mutation is mutually exclusive from the EGFR mutation [40]. The EGFR mutation is often present in asian patients and EGFR patients typically have better survival. Hence, the significance of `alk` can be due to picking up on the absence of the EGFR mutation.

The covariate `left.low` can point to NSCLC on the lower left node of the lung. Studies have observed that lung cancer on the lower lobe or lower left lobe has worse survival [41, 42]. The covariate `severe` could be pointing to a severe conditions related to lung and other problems that indicate poor overall health. Similarly, we also observe other terms that could describe the severity of lung cancer - such as `squamous.cell`, `pulmonary.nodule`, or `silhouette` - or overall health levels - `alert`, `attention`, `cyst`, or `discomfort`.

Overall, we are able to uncover some potential confounders that are easy to interpret and capture useful clinical insights.

3 Discussion

We have demonstrated how causal inference methods can be used to draw more reliable conclusions from population-based studies. Our paper shows that 1) clinical notes, or unstructured data, can be an important source for uncovering confounders, and 2) current clinical tools can be augmented with machine learning methods to provide better decision support. Furthermore, our experimental framework can be easily adapted to use textual data to reduce selection bias in retrospective studies more generally.

Our method can be used to improve clinical practice. Due to the simplicity of the machine learning tools employed, our method can be easily implemented as an additional step in the design of observational CER studies. Our results also show that our method is generalizable to different types of cancer and for various types of study cohort comparisons. With the continued digitization of clinical notes and the increasing access to EMRs, we

Table 3: Univariate and multivariate covariate-specific HR for *surgery* vs. *radiation* for prostate cancer. The $*$ denotes intersection terms identified by our method. The lower block of covariates represent terms extracted from clinical notes.

	Univariate analysis			**Multivariate analysis**		
Covariates	*HR*	*95% CI*	*p-value*	*HR*	*95% CI*	*p-value*
W.surgery	1.27	[0.77, 2.1]	0.352	1.09	[0.59, 2]	0.777
pat_age*	594.88	[87, 4.1e+03]	<0.001	35.96	[3.5, 3.7e+02]	0.003
race_white	0.92	[0.44, 1.9]	0.822	0.65	[0.22, 1.9]	0.439
race_api	0.63	[0.18, 2.2]	0.467	0.67	[0.14, 3.3]	0.622
race_black	1.63	[0.33, 8.1]	0.551	4.04	[0.64, 25]	0.137
hispanic	0.85	[0.2, 3.6]	0.831	1.52	[0.33, 7]	0.593
clinical_stage	0.30	[0.042, 2.2]	0.237	1.02	[0.14, 7.4]	0.987
tumor_grade	0.05	[0.0013, 2]	0.111	0.10	[0.00038, 24]	0.406
grade_unknown	0.55	[0.028, 11]	0.698	0.99	[0.0029, 3.4e+02]	0.996
diagnosis_year	0.12	[0.024, 0.57]	0.008	0.17	[0.025, 1.2]	0.075
bladder*	207.51	[79, 5.4e+02]	<0.001	35.95	[9.3, 1.4e+02]	<0.001
urothelial*	1919.54	[4.2e+02, 8.7e+03]	<0.001	44.07	[4.4, 4.4e+02]	0.001

HR = Hazard Ratio; CI = confidence interval; *: intersection terms

Table 4: Univariate and multivariate covariate-specific HR for *surgery* vs. *active monitoring* for prostate cancer. The $*$ denotes intersection terms identified by our method. The lower block of covariates represent terms extracted from clinical notes.

	Univariate analysis			**Multivariate analysis**		
Covariates	*HR*	*95% CI*	*p-value*	*HR*	*95% CI*	*p-value*
W.surgery	1.67	[0.99, 2.8]	0.057	1.02	[0.55, 1.9]	0.957
pat_age*	3669.74	[5.3e+02, 2.5e+04]	<0.001	143.94	[11, 1.9e+03]	<0.001
race_white	0.87	[0.41, 1.8]	0.709	0.68	[0.23, 2]	0.478
race_api	0.59	[0.15, 2.3]	0.443	0.81	[0.16, 4.1]	0.799
race_black	2.04	[0.41, 10]	0.384	5.16	[0.82, 32]	0.080
hispanic	1.09	[0.29, 4.1]	0.898	1.68	[0.41, 6.8]	0.471
clinical_stage	2.58	[0.31, 21]	0.378	3.75	[0.35, 40]	0.275
tumor_grade	0.22	[0.014, 3.7]	0.296	2.37	[0.0084, 6.6e+02]	0.764
grade_unknown	0.06	[0.00094, 4.2]	0.198	0.02	[2.5e-05, 14]	0.235
diagnosis_year	0.12	[0.027, 0.55]	0.006	0.50	[0.073, 3.4]	0.483
bladder*	160.34	[65, 3.9e+02]	<0.001	24.17	[6.6, 89]	<0.001
urothelial*	2178.75	[5e+02, 9.6e+03]	<0.001	68.15	[7.4, 6.3e+02]	<0.001

HR = Hazard Ratio; CI = confidence interval; *: intersection terms

Table 5: Univariate and multivariate covariate-specific HR for *radiation* vs. *active monitoring* for prostate cancer. The $*$ denotes intersection terms identified by our method. The lower block of covariates represent terms extracted from clinical notes.

	Univariate analysis			**Multivariate analysis**		
Covariates	*HR*	*95% CI*	*p-value*	*HR*	*95% CI*	*p-value*
W.radiation	1.22	[0.63, 2.4]	0.551	0.62	[0.24, 1.6]	0.316
pat_age*	265.19	[9.1, 7.8e+03]	0.001	275.03	[3.7, 2.1e+04]	0.011
race_white	0.43	[0.15, 1.3]	0.129	0.39	[0.099, 1.5]	0.170
race_api	1.38	[0.29, 6.7]	0.687	0.62	[0.09, 4.3]	0.626
race_black	1.69	[0.18, 16]	0.646	1.90	[0.19, 19]	0.582
hispanic	0.38	[0.014, 11]	0.572	0.39	[0.016, 9.7]	0.567
clinical_stage	2.42	[0.2, 30]	0.491	1.12	[0.056, 22]	0.939
tumor_grade	0.89	[0.034, 23]	0.942	533.75	[0.015, 1.9e+07]	0.240
grade_unknown	0.12	[0.0017, 8.9]	0.337	0.00	[3e-07, 32]	0.220
diagnosis_year	0.69	[0.053, 9.1]	0.781	4.96	[0.1, 2.4e+02]	0.417
carotid*	44.60	[4.1, 4.9e+02]	0.002	9.63	[2.1, 43]	0.003
resident*	185839.25	[80, 4.3e+08]	0.002	1288062.91	[1e+03, 1.6e+09]	<0.001

HR = Hazard Ratio; CI = confidence interval; *: intersection terms

Table 6: Univariate and multivariate covariate-specific HR for *surgery* vs. *radiation* for stage I NSCLC. The $*$ denotes intersection terms identified by our method. The lower block of covariates represent terms extracted from clinical notes.

	Univariate analysis			**Multivariate analysis**		
Covariates	*HR*	*95% CI*	*p-value*	*HR*	*95% CI*	*p-value*
W.surgery	0.309	[0.24, 0.4]	<0.001	0.594	[0.45, 0.79]	<0.001
pat_age*	52.9	[17, 1.6e+02]	<0.001	18.5	[5.9, 58]	<0.001
male*	3.07	[1.8, 5.1]	<0.001	1.88	[1.1, 3.3]	0.026
race_white	0.842	[0.53, 1.4]	0.476	0.471	[0.27, 0.84]	0.010
race_api*	0.0557	[0.019, 0.17]	<0.001	0.0572	[0.018, 0.19]	<0.001
race_black	2.03	[0.62, 6.7]	0.245	2.48	[0.78, 7.9]	0.122
hispanic	0.664	[0.19, 2.3]	0.514	0.271	[0.067, 1.1]	0.067
diagnosis_year*	0.0166	[0.007, 0.039]	<0.001	0.0488	[0.015, 0.15]	<0.001
alert*	9.46e-12	[3.6e-16, 2.5e-07]	<0.001	0.00727	[1.2e-07, 4.5e+02]	0.381
alk*	1.17e+04	[38, 3.7e+06]	0.001	3.54e+04	[7.5e+02, 1.7e+06]	<0.001
appearance*	5.46e-09	[2.6e-12, 1.1e-05]	<0.001	0.141	[0.00027, 74]	0.539
attention*	1.03e-12	[4.6e-20, 2.3e-05]	0.001	0.00124	[2.4e-09, 6.4e+02]	0.319
bedtime*	1.87e-10	[3.6e-16, 9.7e-05]	<0.001	7e-05	[1.7e-09, 2.9]	0.078
cyst*	0.00719	[9.7e-05, 0.54]	0.025	0.0341	[0.00025, 4.7]	0.180
discomfort*	1.06e+04	[12, 9.6e+06]	0.008	17.7	[0.012, 2.6e+04]	0.441
feel*	6.64e-10	[5.7e-16, 0.00077]	0.003	8.13e-05	[1.7e-09, 3.9]	0.087
left.low*	29.1	[5.2, 1.6e+02]	<0.001	9.4	[1.4, 61]	0.019
nipple*	2.57e-09	[1.6e-15, 0.0041]	0.007	7.2e-05	[6.4e-09, 0.81]	0.045
preppe*	0.00475	[1.5e-07, 1.5e+02]	0.312	0.074	[1.4e-05, 3.9e+02]	0.552
pulmonary.nodule*	0.00138	[1.9e-05, 0.099]	0.003	0.173	[0.0074, 4.1]	0.276
rib*	688	[40, 1.2e+04]	<0.001	23	[1.2, 4.3e+02]	0.036
severe*	601	[56, 6.5e+03]	<0.001	62.2	[9.3, 4.2e+02]	<0.001
silhouette*	2.14e-05	[6.9e-10, 0.66]	0.042	9.51e-06	[6.4e-11, 1.4]	0.057
sponge*	2.04e-07	[3.3e-11, 0.0013]	<0.001	0.00961	[2e-05, 4.7]	0.142
squamous.cell*	94.8	[16, 5.6e+02]	<0.001	4.46	[0.6, 33]	0.144
venous*	0.0464	[0.00089, 2.4]	0.128	0.0665	[0.0017, 2.6]	0.145

HR = Hazard Ratio; CI = confidence interval; *: intersection terms

recommend this as an essential step for any researcher seeking to draw clinical insights from observational data. The terms uncovered with our method can not only be used to improve observational CERs, but also be used to generate interpretable insights about current clinical practice. The uncovering of relevant information and subsequent insights can then be used to inform high-stakes medical decisions.

We believe that our work is the first to explore the potential of including unstructured clinical notes to reduce selection bias in oncology settings. We are also one of the first works to incorporate unstructured data into causal inference estimators and Cox-PH models. Although our method has been developed to address a specific problem in oncology and applied in the clinical setting, it can also be easily adapted for application in any observational study that seeks to incorporate unstructured text. There is much work to be done in using NLP and unstructured text for causal inference. Our work present a simple and flexible way to generate interpretable causal insights from text of any sort. Our study also has several limitations. First, we use simple NLP methods to process the clinical notes and extract the top 500 or 1000 features for variable selection. In the process, much information in the text nodes is discarded and the sequence of past medical events are not taken into account. We choose this setup due to the the small sample size of oncology study cohorts, which makes it difficult to train more complicated models for textual processing. In theory, the more work that is placed into the clinical notes preprocessing and the higher quality of the features generated from these notes, the more informative the uncovered potential confounders will be. For future work, we hope to explore how other NLP techniques, such as topic modeling or clustering, can be used to build even higher quality features from the unstructured text. There are also an increasing number of deep learning models that can be used to identify interpretable insights [18]. We are interested in how these deep learning methods can be applied to generate causal insights on another

study population with larger sample size.

Second, our method can only uncover potential confounders that can be observed in notes. There are many sources of confounding in observational data and even rich EMR data cannot capture everything. If the confounding is unknown and unobservable, no method to our knowledge will be able to adjust for it. Hence, it would be good practice to perform sensitivity analysis to evaluate the result's robustness to unknown confounding.

Third, our approach of selecting intersection covariates is an empirical approach designed for uncovering the most valuable potential confounders. While our approach seemed to work well empirically in this study, one could do simulation studies to help verify its validity in the future.

Finally, the validity of causal inference models cannot be determined without prospective experimental data. Therefore, the uncovered confounders and estimated HR can only be validated by clinicians. We are identifying potential candidates for the bias and then evaluating these candidates of bias against RCTs.

Many challenges still remain for employing unstructured data for causal inference analysis and medical settings. We hope this work interests both clinical practitioners augmenting existing clinical support tools and researchers using textual data to reduce confounding in observational data. We hope our workflow, problem framing, and experimental design can serve as such a sandbox for testing more complex algorithms or adapting to other application areas. Ultimately, we hope this research will find causal information in clinical notes and provide a transparent way for machine learning to inform medical decision making.

4 Methods

4.1 Dataset

With approval of the Stanford Institutional Review Board (IRB), we curate a dataset of non-metastatic prostate and lung cancer patients from the Stanford Cancer Institute Research Database (SCIRDB). The database includes patients seen in the Stanford Health Care (SHC) system from 2008 to 2019 for prostate cancer and 2000 to 2019 for lung cancer. SHC clinical sites include one academic hospital, one freestanding cancer center, and several outpatient clinics. From SCIRDB, we pull a total of 3,638 prostate cancer patients with 552,009 clinical notes and 3,274 non-small cell lung cancer (NSCLC) patients with 648,505 clinical notes. The clinical notes include progress notes, letters, discharge summaries, emergency department notes, history and physical notes, and treatment planning notes.

For each patient, we also pull the structured EMR and data from the inpatient billing system. From the California Cancer Registry (CCR), we pull the available initial treatment information, cancer staging, tumor description, date of diagnosis, date of death, and date of last follow-up for these selected patients. For NSCLC, we also pull the recorded Epic cancer staging information.

4.2 Study Cohort

We build our study cohorts from SCIRDB with reference to existing observational study principles and clinical expertise.

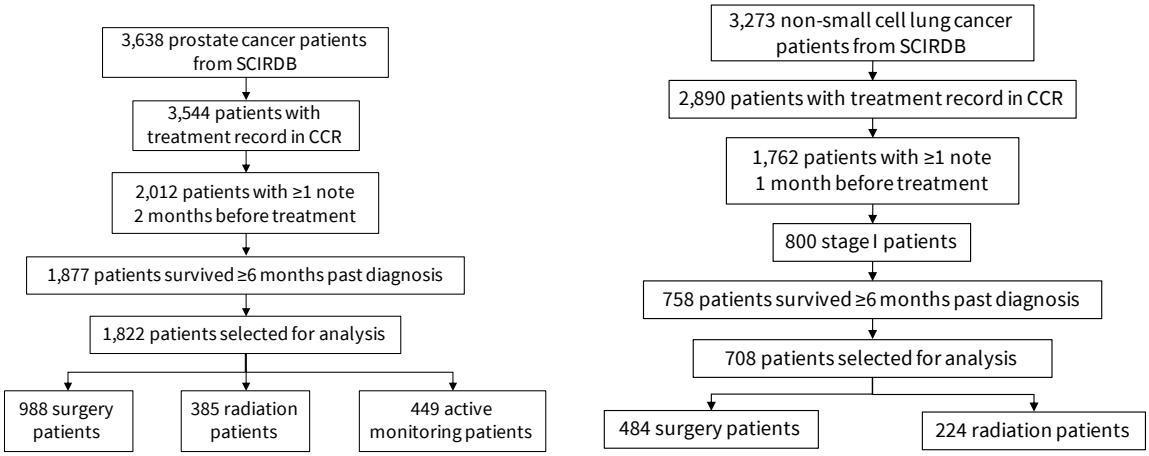
For each patient, we combine all treatments with the same Diagnosis ID in the CCR as the initial line of treatment. For patients with multiple diagnosis id, we keep the first record of treatment. For prostate cancer, patients without a recorded treatment are labeled as *active monitoring*. To avoid explicit revelation of the treatment choice, we only include

notes more than 2 months before treatment start date for prostate cancer and 1 month for NSCLC. We then select for patients with at least one note before the specified time. We also perform landmark analysis to avoid immortal-time bias [43] and select only patients who survived at least 6 months past their date of diagnosis. Landmark analysis refers to the practice of designating a time point during the follow-up period and analyzing only patients who have survived until the landmark time [43]. To ensure the proportional hazard condition, patients who are still living are censored at time of last follow-up [44]. The patient filtering and cohort selection process is shown in Figure 4.

For patients with unknown clinical stage but known pathological stage, we impute the clinical stage by training a clinical stage classification model using the pathological stage and other patient information. Pathological stage is usually a little higher than clinical stage due to the staging based on biopsy samples instead of imaging; hence, it is inaccurate to group them together. Clinical stage is more frequently used for similar observational studies [20, 36] and it is more rigorous to impute the missing clinical stage with a model trained on the pathological stage and other relevant covariates. We train the clinical stage imputation model with `pat_age`, `pathological_stage`, `diagnosis_year`, and `tumor_grade`. For NSCLC, `tumor_grade` is not included due to missing information. For both prostate and NSCLC, we train and validate a random forest model [45, 46] on patients with both clinical and pathological stage available. The imputed stages are used as the clinical stage for those patients. For patients with both clinical and pathological stage missing, we are able to fill in some through clinical chart reviews.

To form the treatment groups, we assign patients with only surgery records into the surgery group and patients with only radiation records into the radiation group. For patients with both radiation and surgery, patients who received surgery first are assigned to the *surgery* group and patients who received radiation first to the *radiation* group. For

prostate cancer, patients with no recorded treatment are assigned to the *active monitoring* group. For NSCLC, only patients with clinical stage I are included.



(a) Prostate Cancer (b) Lung Cancer

Figure 4: Patient cohort selection process for prostate and lung cancer patients.

4.3 Data Processing and Representation

We build the covariates used for uncovering confounders through the process shown in Figure 1A. We compile the data from SCIRDB, CCR, and Epic for each patient.

We include age, race, ethnicity, clinical stage, and diagnosis year as part of the structured data. For prostate cancer, we also include tumor grade. For NSCLC, we also include gender. Based on age range categories used in Li et al. [47], we form the categorical variable `pat_age` by splitting age into ranges of ≤ 49 years-old, 5-year buckets from $50 - 84$ years-old, and ≥ 85 years-old. Race and ethnicity are encoded as one-hot vectors, with each feature indicating one race or ethnicity. Race is combined based on what is done in Li et al. [47]. We select these structured covariates because they are commonly accepted by clinicians as potential confounders and often included in CER studies [5]. For race, `race_unknown` is not included as a covariate. For ethnicity, only `hispanic` is included as a

covariate. For tumor grade, patients with unknown grade are imputed with the median grade value. The indicator variable `grade_unknown` is added to indicate which patients have been imputed. The covariates `tumor_grade` and `grade_unknown` are not included for NSCLC due to missing information of tumor grade and clinical judgement. In the end, we have 9 structured covariates for prostate cancer and 7 structured covariates for NSCLC.

We build word frequency representations of the clinical notes for the unstructured covariates. For each patient, we compile notes within the specified time (i.e. 2 months prior to treatment start date for prostate cancer and 1 month prior for NSCLC). We only use notes from before treatment so that we are not predicting survival outcome with information unavailable at the time of treatment decision. The different time windows for the two diseases was selected as NSCLC treatment generally starts more quickly than prostate cancer treatment due to the more rapidly progressing nature of the cancer. The notes are segmented based on clinical field labels (e.g. "IMPRESSION:", "HISTORY:"), tab spaces, NLTK sentence tokenization [48]. To remove noise, we remove clinical field labels and two sentences from the beginning and end of each document. We also remove sentences with common locations (e.g. "Stanford Medical Center", "Palo Alto") and medical doctor names (e.g. "xx xx, M.D.") as these are often prefix or suffix to note documents. To avoid including conditions patients do not have, we remove sentences if they contain a negation term (i.e. "no", "denies", "does not", "none") and contain less than 15 words.

We then identify biomedical entities from the preprocessed clinical notes with scispaCy [49]. scispaCy is a spaCy [50] based model for processing biomedical, scientific, and clinical text. The scispaCy models identifies a list of all the entities in the text that exist in a biomedical dictionary, such as the Unified Medical Language System [51]. We then lemmatize and combine all biomedical entities identified from the sentences for each patient into a single document. To further remove noise, we remove stop words using a combination

of the NLTK stopwords [48] and data-specific stopwords such as medical units (e.g., "lb", "oz", "mmhg"), time terms (e.g., "months", "days"), and medical or Stanford specific terms (e.g., "stanford", "patient", "doctor") that are very common but irrelevant to the task at hand. We also create a dictionary of synonyms in the dataset and use the dictionary to combine these words. The dictionary includes lexical variations that are not reduced to the same root during lemmatization (e.g. "abnormality" → "abnormal", "consult" → "consultation"), abbreviations (e.g. "hx" → "history", "fu" → "followup"), and common synonyms (e.g. "assistance → "service", "action" → "movement").

Finally, we remove punctuation and generate term frequency representations of the text using bag-of-words (BOW) with Term frequency–inverse document frequency (TF-IDF) weighting [52]. Bag-of-words (BOW) model is a simplifying representation in natural language processing. It represents text (such as sentence or document) as a vector of word occurrence count. TF-IDF, is a score that reweighs the BOW matrix to reflect how important a word is to a document in a collection or corpus. We implement this with scikit-learn [46]. For prostate cancer, we use $max_features = 500$ with $ngrams = (1, 1)$. Although there are more prostate cancer patients, the lower number of death events makes it difficult to include a large number of covariates. For NSCLC, we use $max_features = 1000$, $ngrams = (1, 2)$ (including single words and two word phrases), and $max_df = 0.7$ (filters out dataset-specific stopwords). Hence, we have 500 unstructured covariates for prostate cancer and 1000 unstructured covariates for NSCLC.

We scale and normalize both the structured and unstructured covariates before concatenating them. In total, we build 509 covariates for prostate cancer and 1007 covariates for NSCLC. These covariates are then used to uncover potential dataset-specific confounders.

4.4 Outcomes

We define our survival outcome as (Y_i, E_i), where $Y_i \in \mathbb{Z}^+$ is the number of survival days since the diagnosis and $E_i \in \{0, 1\}$ is an indicator for whether a death event has been observed during follow-up. The treatment, $W_i \in \{0, 1\}$, is an indicator for either surgery, radiation, or monitoring, depending on the treatment group. The covariates, X_i, includes the structured dataset pulled from the EMR data and the bag-of-words matrix representation generated from EMR notes.

4.5 Uncover and Evaluate Confounders

We uncover interpretable potential confounders from the covariates and evaluate the confounders we've identified with survival analysis. The approach is shown in Figure 1B.

We find the potential confounders by identifying covariates that are predictive of both treatment and survival outcome. We train prediction models for treatment ($W_i = 1$) and the survival outcome (Y_i, E_i) with Lasso [10] using **glmnet** [53]. Lasso is a $L1-$penalized linear that can produce coefficients for covariates that are exactly zero, and is, hence, often used for creating sparse models [54] or variable selection [9]. We select the intersection of covariates with non-zero coefficients from both the treatment and survival outcome models as potential confounders. For *surgery* vs. *radiation* and *surgery* vs. *active monitoring* for prostate cancer, we use *lambda.1se* to select the intersection covariates. With *radiation* vs. *monitoring* for prostate cancer and *surgery* vs. *radiation* for stage I NSCLC, we use *lambda.min* to select the intersection covariates. The intersection terms selected are more stable with *lambda.1se*. However, we choose *lambda.min* for the latter two treatment groups because *lambda.1se* did not select any covariates from the text.

We then evaluate each of the covariate combinations with propensity score adjusted survival analysis. Propensity scores for patient i is the probability of receiving the treat-

ment of interest, $W_i = 1$, given the covariates X_i [55]. Conditional on the propensity score, the distribution of observed covariates is expected to be the same in both branches of the treatment group. It is often used to reduce the effect of confounding in observational studies [55, 56]. In survival analysis, the hazard rate $h(t|X)$ is the probability the patient will die within time t given covariates X. The HR is the ratio of the hazard rate of the two treatments. In survival outcomes analysis, the HR is treated as the treatment effect of choosing the treatment of interest, $W_i = 1$.

We use the Cox-proportional hazard (Cox-PH) model to perform survival regression [57]. We assume the proportional hazards condition [58], which states that covariates are multiplicatively related to the hazard, e.g., a covariate may halve a subject's hazard at any given time t while the baseline hazard may vary. Hence, the effect of covariates estimated by any proportional hazards model can be reported as the HR of the covariate.

In a Cox-PH model, the hazard rate of an individual is a linear function of their static covariates and a population-level baseline hazard that changes over time. We adjust for covariates (e.g. `pat_age`, `race_white`, etc.) against duration of survival and a binary variable indicating whether the outcome event has occurred. We estimate

$$h(t|X) = h_0(t) \exp\Big(b_w W + \sum_{j=1}^{p} b_j X_j\Big),$$

where p is the number of covariates, $b_0(t)$ the baseline hazard, b_W the effect size of the treatment, and b_j the effect size of the jth covariate. The HR for a covariate is equal to e^{b_i}. We define the HR of the treatment as e^{b_w}.

We use 3 methods to estimate the HR:

- Nearest-Neighbor Matching on Propensity Score (**matching**) [14]: We perform nearest-neighbor propensity score matching (NNM) on selected covariates and estimate the

HR on the matched population using a univariate Cox-PH model regressed on the treatment.

- Inverse Propensity of Treatment Weighting (**IPTW**) [59, 14]: We estimate the HR using a univariate Cox-PH model regressed on the treatment with inverse propensity score weighting with stabilization [59]. The weights are defined as

$$w_i = W_i + (1 - W_i)\Big[\frac{e(X_i)}{1 - e(X_i)}\Big]$$

- Multivariate Cox Proportional Hazard (**multi.coxph**) [57, 60, 5]: We estimate the HR using a multivariate regression model on the treatment and selected covariates to see how covariates interact with each other. The multivariate model is also weighted with the inverse propensity scores above to form a doubly-robust model.

All Cox-PH models are trained using the **survival** R package [61] with robust variance. NNM is performed using the **Matching** R package [62].

We estimate the propensity scores using logistic regression [63] with **glmnet** [53], stochastic gradient boosting [64] with **gbm** [65], and generalized random forests with **grf** [24]. We select the propensity score estimation method with the best overlap and covariate balance post propensity score adjustment.

We then compare the 3 methods for estimating HR using forest plots.

For each covariate in struct+intersect, we also show the univariate and multivariate Cox-PH model HR, 95% HR confidence interval, and p-value, using the analysis presented in Bradburn et al. [60]. Note that for the multivariate Cox-PH covariate analysis, we do not weight the model with the inverse propensity scores.

5 Data availability

The datasets analyzed for the study are not publicly available. The EHR data cannot be redistributed to researchers other than those approved through the Stanford Institutional Review Board. We have therefore given detailed description of our data selection and processing pipeline in the Methods section.

6 Code availability

The training and statistical evaluation code can be made available upon request to the corresponding author.

7 Author Contributions

J.Z. contributed to data acquisition, data processing, study design, methodology, implementation, interpretation of results, and drafted and revised the paper. M.F.G. contributed to the acquisition of data, study design, interpretation of results, and revised the paper draft. D.L.R. contributed to the acquisition of data, provision of computational resources, methodology, and revised the paper draft. S.A. contributed to study design, methodology, interpretation of results, and revised the paper draft. R.D.S. contributed to the acquisition of data, study design, methodology, interpretation of results, and revised the paper draft. All authors contributed to the study conception, provided feedback during the work development, and gave approval for the submission.

8 Acknowledgements

For data acquisition, we also thank A. Solomon Henry and Douglas Wood. The research is supported by funding from the Stanford Human-Centered Artificial Intelligence Institute and Department of Management Science and Engineering.

9 Competing Interests

The authors declare that there are no competing interests.

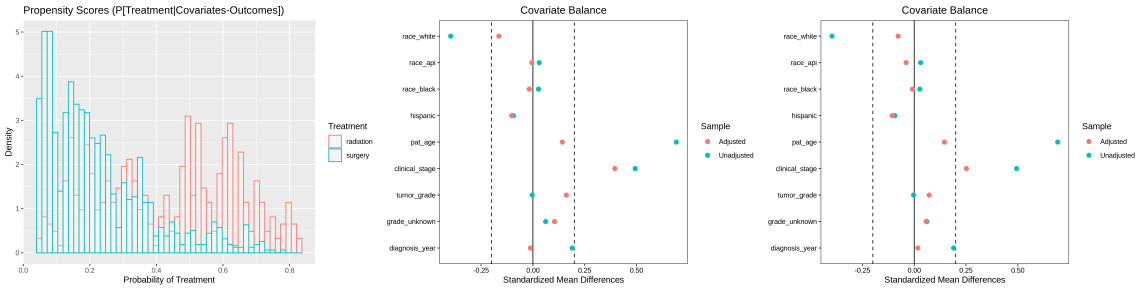
A Supplement

A.1 Propensity Score and Covariate Balance Plots

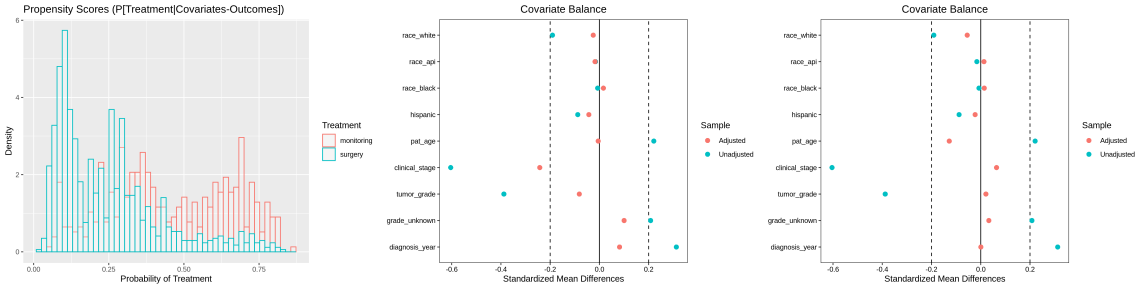
We show the propensity scores and covariate balance plots for each of the results plotted in Figure 3. In Figure 5, we show the plots for *structured*. In Figure 6, we show the plots for *intersect*. In Figure 7, we show the plots for *struct+intersect*.

A.2 Covariate Correlation

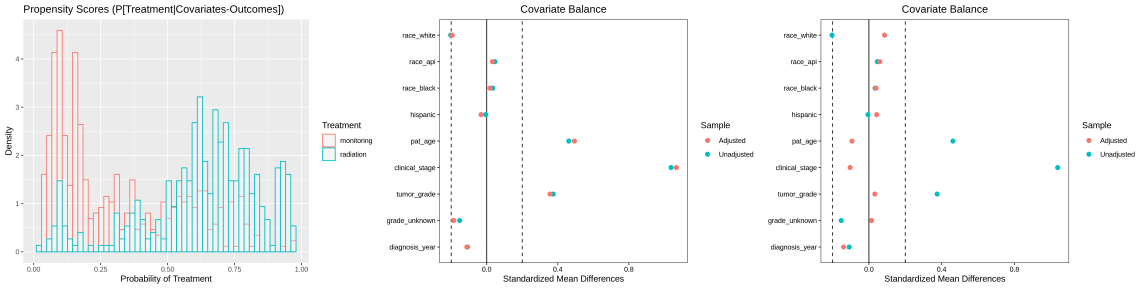
We show the R^2 values for all combinations of the selected covariates for each of the treatment groups. The R^2 is the square of the correlation from linear regression. It measures the proportion of variation in the dependent variable that can be attributed to the independent variable. In Table 7, we show the R^2 values for *surgery vs. radiation* for prostate cancer. In Table 8, we show the R^2 values for *surgery vs. monitoring* for prostate cancer. In Table 9, we show the R^2 values for *radiation vs. monitoring* for prostate cancer. In Table 10, we show the R^2 values for *surgery vs. radiation* for NSCLC.



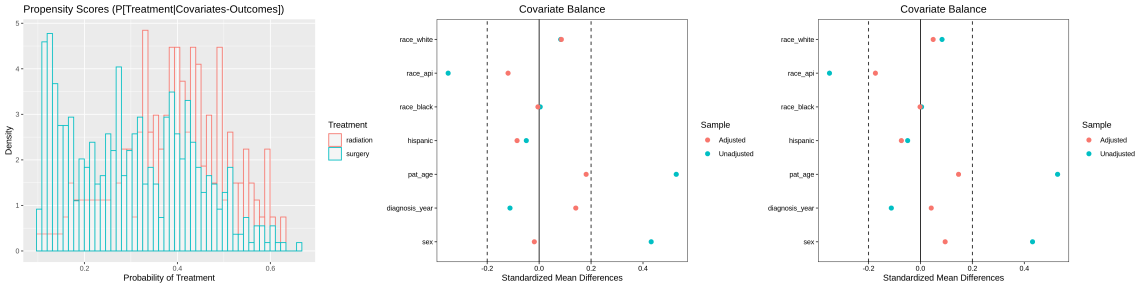
(a) Surgery vs Radiation (prostate, **grf**)



(b) Surgery vs Monitoring (prostate, **glmnet**)

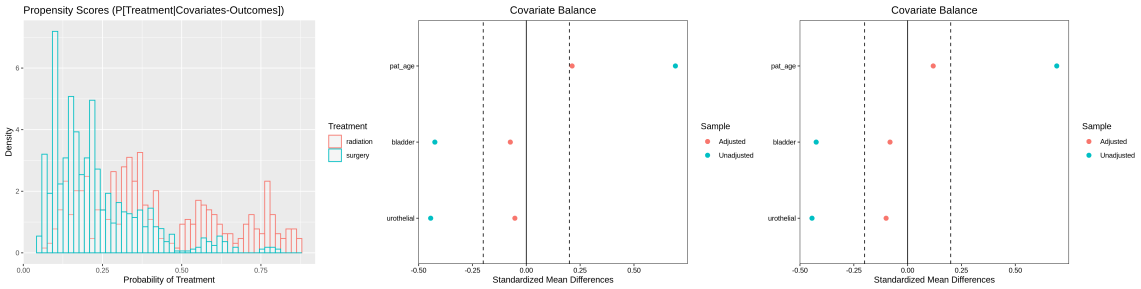


(c) Monitoring vs Radiation (prostate, **glmnet**)

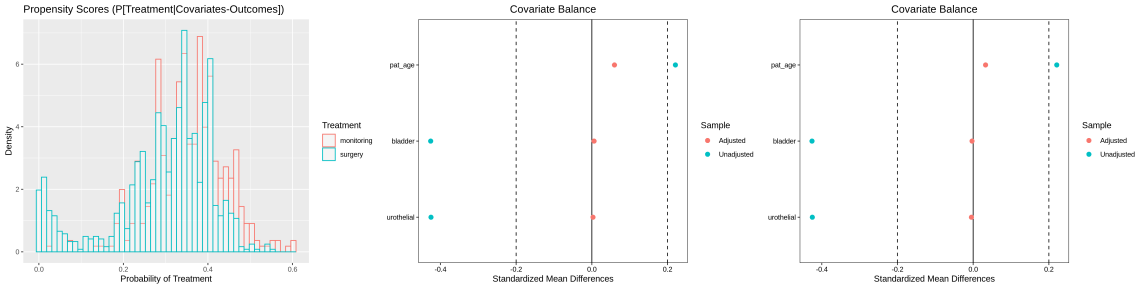


(d) Surgery vs Radiation (lung, **grf**)

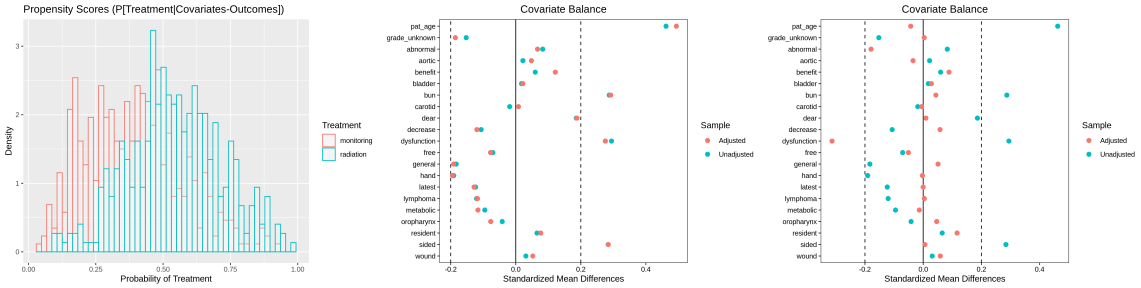
Figure 5: Supplementary plots to the *structured* results presented in Figure 3. (left) Propensity score plot with *structured*. (middle) Covariate balance plot for matching. (right) Covariate balance plots for IPTW.



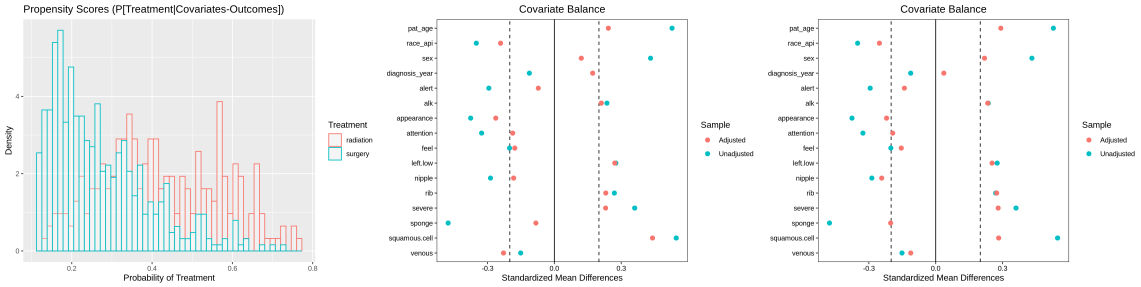
(a) Surgery vs Radiation (prostate, **grf**)



(b) Surgery vs Monitoring (prostate, **glmnet**)

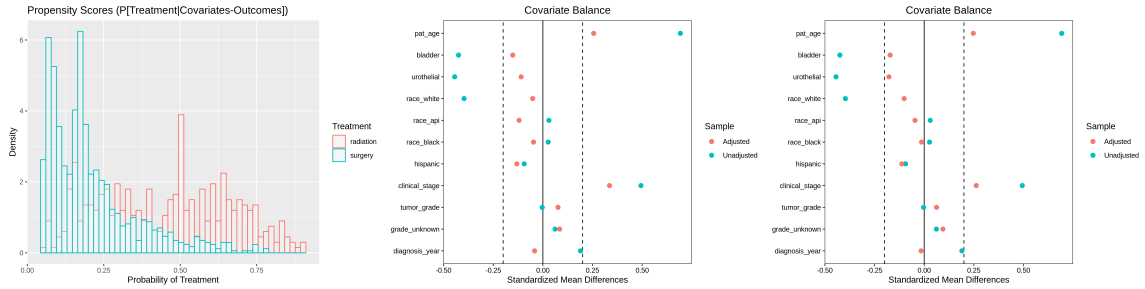


(c) Monitoring vs Radiation (prostate, **glmnet**)

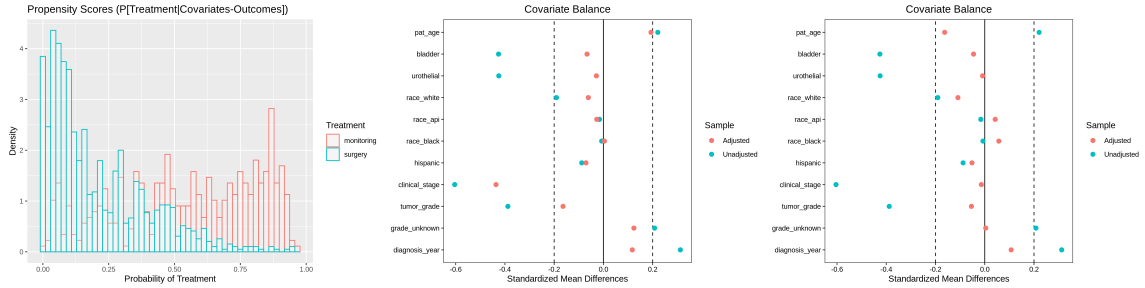


(d) Surgery vs Radiation (lung, **grf**)

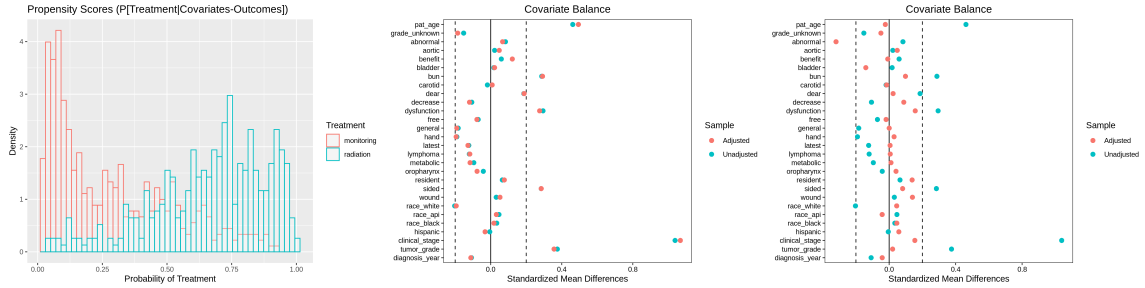
Figure 6: Supplementary plots to the *intersect* results presented in Figure 3. (left) Propensity score plot with *intersect*. (middle) Covariate balance plot for matching. (right) Covariate balance plots for IPTW.



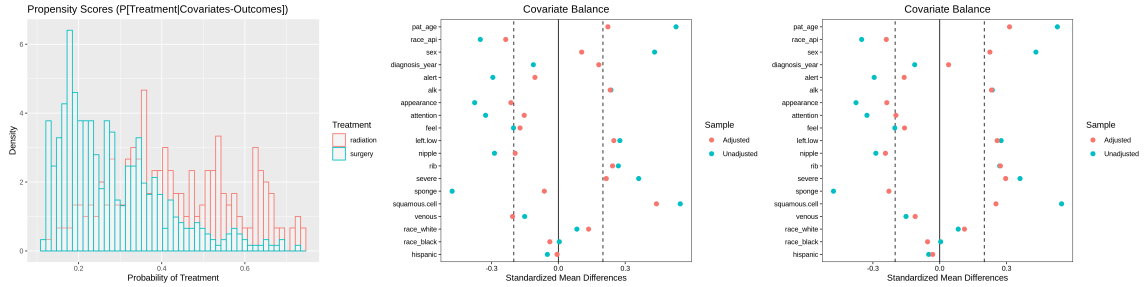
(a) Surgery vs Radiation (prostate, **grf**)



(b) Surgery vs Monitoring (prostate, **glmnet**)



(c) Monitoring vs Radiation (prostate, **glmnet**)



(d) Surgery vs Radiation (lung, **grf**)

Figure 7: Supplementary plots to the *struct+intersect* results presented in Figure 3. (left) Propensity score plot with *struct+intersect*. (middle) Covariate balance plot for matching. (right) Covariate balance plots for IPTW.

Table 7: R^2 correlation of the **struct+intersect** covariates of *surgery vs. radiation* for prostate cancer.

	pat_age	bladder	urothelial	race_white	race_api	race_black	hispanic	clinical_stage	tumor_grade	grade_unknown	diagnosis_year
pat_age		0.04	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.02
bladder	0.04		0.47	0.01	0.00	0.00	0.00	0.04	0.03	0.00	0.00
urothelial	0.01	0.47		0.01	0.00	0.00	0.00	0.04	0.03	0.00	0.00
race_white	0.00	0.01	0.01		0.22	0.07	0.05	0.00	0.00	0.00	0.03
race_api	0.00	0.00	0.00	0.22		0.00	0.01	0.00	0.00	0.00	0.00
race_black	0.00	0.00	0.00	0.07	0.00		0.00	0.00	0.00	0.00	0.00
hispanic	0.01	0.00	0.00	0.05	0.01	0.00		0.00	0.00	0.00	0.00
clinical_stage	0.00	0.04	0.04	0.00	0.00	0.00	0.00		0.04	0.01	0.03
tumor_grade	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.04		0.51	0.21
grade_unknown	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.51		0.02
diagnosis_year	0.02	0.00	0.00	0.03	0.00	0.00	0.00	0.03	0.21	0.02	

Table 8: R^2 correlation of the **struct+intersect** covariates of *surgery vs. monitoring* for prostate cancer.

	pat_age	carotid	resident	race_white	race_api	race_black	hispanic	clinical_stage	tumor_grade	grade_unknown	diagnosis_year
pat_age		0.02	0.00	0.00	0.01	0.01	0.00	0.04	0.01	0.01	0.07
carotid	0.02		0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
resident	0.00	0.00		0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
race_white	0.00	0.00	0.00		0.17	0.06	0.02	0.00	0.00	0.00	0.00
race_api	0.01	0.00	0.00	0.17		0.00	0.00	0.00	0.00	0.00	0.01
race_black	0.01	0.00	0.01	0.06	0.00		0.00	0.00	0.00	0.00	0.00
hispanic	0.00	0.00	0.00	0.02	0.00	0.00		0.00	0.00	0.00	0.00
clinical_stage	0.04	0.00	0.00	0.00	0.00	0.00	0.00		0.05	0.01	0.06
tumor_grade	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.05		0.66	0.22
grade_unknown	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.66		0.04
diagnosis_year	0.07	0.00	0.00	0.00	0.01	0.00	0.00	0.06	0.22	0.04	

Table 9: R^2 correlation of the **struct+intersect** covariates of *radiation vs. monitoring* for prostate cancer.

	pat_age	carotid	resident	race_white	race_api	race_black	hispanic	clinical_stage	tumor_grade	grade_unknown	diagnosis_year
pat_age		0.02	0.00	0.00	0.01	0.01	0.00	0.04	0.01	0.01	0.07
carotid	0.02		0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
resident	0.00	0.00		0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
race_white	0.00	0.00	0.00		0.17	0.06	0.02	0.00	0.00	0.00	0.00
race_api	0.01	0.00	0.00	0.17		0.00	0.00	0.00	0.00	0.00	0.01
race_black	0.01	0.00	0.01	0.06	0.00		0.00	0.00	0.00	0.00	0.00
hispanic	0.00	0.00	0.00	0.02	0.00	0.00		0.00	0.00	0.00	0.00
clinical_stage	0.04	0.00	0.00	0.00	0.00	0.00	0.00		0.05	0.01	0.06
tumor_grade	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.05		0.66	0.22
grade_unknown	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.66		0.04
diagnosis_year	0.07	0.00	0.00	0.00	0.01	0.00	0.00	0.06	0.22	0.04	

Table 10: R^2 correlation of the **struct+intersect** covariates of *surgery vs. radiation* for NSCLC.

	pat_age	race_api	sex	diagnosis_year	alert	alk	allergy	appearance	attention	cyst	discomfort	eye	fever	inguinal	nipple	rib	severe	silhouette	sponge	squamous	race_white	race_black	hispanic
pat_age		0.00	0.01	0.04	0.01	0.00	0.00	0.02	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00
race_api	0.00		0.00	0.03	0.02	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.30	0.00	0.01
sex	0.01	0.00		0.04	0.01	0.01	0.01	0.01	0.01	0.02	0.00	0.01	0.00	0.02	0.02	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.00
diagnosis_year	0.04	0.03	0.04		0.11	0.01	0.03	0.02	0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.00
alert	0.01	0.02	0.01	0.11		0.00	0.02	0.04	0.01	0.00	0.01	0.01	0.05	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00
alk	0.00	0.00	0.01	0.01	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
allergy	0.00	0.00	0.01	0.03	0.02	0.00		0.02	0.00	0.00	0.00	0.01	0.04	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
appearance	0.02	0.03	0.01	0.02	0.04	0.00	0.02		0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
attention	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
cyst	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
discomfort	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
eye	0.00	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fever	0.00	0.00	0.00	0.01	0.05	0.00	0.04	0.03	0.00	0.00	0.00	0.00		0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
inguinal	0.00	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01		0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
nipple	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
rib	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00
severe	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00		0.00	0.01	0.00	0.00	0.00	0.01
silhouette	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00
sponge	0.01	0.00	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00		0.00	0.01	0.00	0.00
squamous	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00
race_white	0.00	0.30	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00		0.04	0.04
race_black	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04		0.00
hispanic	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.04	0.00	

References

[1] Vinay Prasad, Andrae Vandross, Caitlin Toomey, Michael Cheung, Jason Rho, Steven Quinn, Satish Jacob Chacko, Durga Borkar, Victor Gall, Senthil Selvaraj, et al. A decade of reversal: an analysis of 146 contradicted medical practices. In *Mayo Clinic Proceedings*, volume 88, pages 790–798. Elsevier, 2013.

[2] Payal D Soni, Holly E Hartman, Robert T Dess, Ahmed Abugharib, Steven G Allen, Felix Y Feng, Anthony L Zietman, Reshma Jagsi, Matthew J Schipper, and Daniel E Spratt. Comparison of population-based observational studies with randomized trials in oncology. *Journal of Clinical Oncology*, pages JCO–18, 2019.

[3] Peter M Rothwell. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93, 2005.

[4] Sharon H Giordano, Yong-Fang Kuo, Zhigang Duan, Gabriel N Hortobagyi, Jean Freeman, and James S Goodwin. Limits of observational data in determining outcomes from cancer therapy. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 112(11):2456–2466, 2008.

[5] Christopher JD Wallis, Refik Saskin, Richard Choo, Sender Herschorn, Ronald T Kodama, Raj Satkunasivam, Prakesh S Shah, Cyril Danjoux, and Robert K Nam. Surgery versus radiotherapy for clinically-localized prostate cancer: a systematic review and meta-analysis. *European urology*, 70(1):21–30, 2016.

[6] DH Yeh, S Tam, K Fung, SD MacNeil, J Yoo, E Winquist, DA Palma, and AC Nichols. Transoral robotic surgery vs. radiotherapy for management of oropharyngeal squamous cell carcinoma–a systematic review of the literature. *European Journal of Surgical Oncology (EJSO)*, 41(12):1603–1614, 2015.

[7] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019.

[8] Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, pages 615–625, 2004.

[9] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014.

[10] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[11] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

[12] Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

[13] Huzhang Mao, Liang Li, Wei Yang, and Yu Shen. On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. *Statistics in medicine*, 37(26):3745–3763, 2018.

[14] Peter C Austin. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in medicine*, 33(7):1242–1258, 2014.

[15] Imon Banerjee, Selen Bozkurt, Jennifer Lee Caswell-Jin, Allison W Kurian, and

Daniel L Rubin. Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. *JCO clinical cancer informatics*, 3:1–12, 2019.

[16] Albee Y Ling, Allison W Kurian, Jennifer L Caswell-Jin, George W Sledge, Nigam H Shah, and Suzanne R Tamang. Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. *JAMIA Open*, 2019.

[17] Imon Banerjee, Michael Francis Gensheimer, Douglas J Wood, Solomon Henry, Sonya Aggarwal, Daniel T Chang, and Daniel L Rubin. Probabilistic prognostic estimates of survival in metastatic cancer patients (ppes-met) utilizing free-text clinical narratives. *Scientific reports*, 8(1):10037, 2018.

[18] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

[19] Anthony C Nichols, Julie Theurer, Eitan Prisman, Nancy Read, Eric Berthelet, Eric Tran, Kevin Fung, John R de Almeida, Andrew Bayley, David P Goldstein, et al. Radiotherapy versus transoral robotic surgery and neck dissection for oropharyngeal squamous cell carcinoma (orator): an open-label, phase 2, randomised trial. *The Lancet Oncology*, 20(10):1349–1359, 2019.

[20] Freddie C Hamdy, Jenny L Donovan, J Athene Lane, Malcolm Mason, Chris Metcalfe, Peter Holding, Michael Davis, Tim J Peters, Emma L Turner, Richard M Martin, et al. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *New England Journal of Medicine*, 375(15):1415–1424, 2016.

[21] Stephen B Williams, Jinhai Huo, Karim Chamie, Marc C Smaldone, Christopher D

Kosarek, Justin E Fang, Leslie A Ynalvez, Simon P Kim, Karen E Hoffman, Sharon H Giordano, et al. Discerning the survival advantage among patients with prostate cancer who undergo radical prostatectomy or radiotherapy: the limitations of cancer registry data. *Cancer*, 123(9):1617–1624, 2017.

[22] Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, 2011.

[23] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[24] Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[25] Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

[26] Susan Athey, Guido Imbens, Thai Pham, and Stefan Wager. Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, 107(5):278–81, 2017.

[27] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[28] Tianyu Wang, Marco Morucci, M Awan, Yameng Liu, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Flame: A fast large-scale almost matching exactly approach to causal inference. *arXiv preprint arXiv:1707.06315*, 2017.

[29] Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*, 2018.

[30] Reagan Mozer, Luke Miratrix, Aaron Russell Kaufman, and L Jason Anastasopoulos. Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis*, 28(4):445–468, 2020.

[31] Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903, 2020.

[32] Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR, 2020.

[33] Michael F Gensheimer, A Solomon Henry, Douglas J Wood, Trevor J Hastie, Sonya Aggarwal, Sara A Dudley, Pooja Pradhan, Imon Banerjee, Eunpi Cho, Kavitha Ramchandran, et al. Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. *JNCI: Journal of the National Cancer Institute*, 111(6):568–574, 2019.

[34] Xi Sheryl Zhang, Fengyi Tang, Hiroko H Dodge, Jiayu Zhou, and Fei Wang. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2487–2495, 2019.

[35] Joseph A Miccio, Wesley J Talcott, Vikram Jairam, Henry S Park, B Yu James, Michael S Leapman, Skyler B Johnson, Martin T King, Paul L Nguyen, and Ben-

jamin H Kann. Quantifying treatment selection bias effect on survival in comparative effectiveness research: findings from low-risk prostate cancer patients. *Prostate Cancer and Prostatic Diseases*, pages 1–9, 2020.

[36] Joe Y Chang, Suresh Senan, Marinus A Paul, Reza J Mehran, Alexander V Louie, Peter Balter, Harry JM Groen, Stephen E McRae, Joachim Widder, Lei Feng, et al. Stereotactic ablative radiotherapy versus lobectomy for operable stage i non-small-cell lung cancer: a pooled analysis of two randomised trials. *The Lancet Oncology*, 16(6): 630–637, 2015.

[37] Frank E Harrell Jr, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2):143–152, 1984.

[38] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

[39] Benjamin Solomon and C Lovly. Anaplastic lymphoma kinase (alk) fusion oncogene positive non-small cell lung cancer. *Uptodate. Waltham, MA: UpToDate. Accessed July*, 5, 2020.

[40] Justin F Gainor, Anna M Varghese, Sai-Hong Ignatius Ou, Sheheryar Kabraji, Mark M Awad, Ryohei Katayama, Amanda Pawlak, Mari Mino-Kenudson, Beow Y Yeap, Gregory J Riely, et al. Alk rearrangements are mutually exclusive with mutations in egfr or kras: an analysis of 1,683 patients with non–small cell lung cancer. *Clinical cancer research*, 19(15):4273–4281, 2013.

[41] Hyun Woo Lee, Young Sik Park, Sangshin Park, and Chang-Hoon Lee. Poor prognosis

of nsclc located in lower lobe is partly mediated by lower frequency of egfr mutations. *Scientific reports*, 10(1):1–8, 2020.

[42] Yujin Kudo, Hisashi Saji, Yoshihisa Shimada, Masaharu Nomura, Jitsuo Usuda, Naohiro Kajiwara, Tatsuo Ohira, and Norihiko Ikeda. Do tumours located in the left lower lobe have worse outcomes in lymph node-positive non-small cell lung cancer than tumours in other lobes? *European journal of cardio-thoracic surgery*, 42(3):414–419, 2012.

[43] Charity J Morgan. Landmark analysis: A primer, 2019.

[44] Patrick Royston and Mahesh KB Parmar. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC medical research methodology*, 13(1):152, 2013.

[45] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[47] Jun Li, David A Siegel, and Jessica B King. Stage-specific incidence rates and trends of prostate cancer by age, race, and ethnicity, united states, 2004–2014. *Annals of epidemiology*, 28(5):328–330, 2018.

[48] Edward Loper and Steven Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[49] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: Fast

and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*, 2019.

[50] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1), 2017.

[51] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[52] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

[53] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 (1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

[54] Tim Hesterberg, Nam Hee Choi, Lukas Meier, Chris Fraley, et al. Least angle and 1 penalized regression: A review. *Statistics Surveys*, 2:61–93, 2008.

[55] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[56] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[57] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[58] Norman E Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57, 1975.

[59] Stephen R Cole and Miguel A Hernán. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine*, 75(1):45–49, 2004.

[60] Mike J Bradburn, Taane G Clark, Sharon B Love, and Douglas G Altman. Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3):431–436, 2003.

[61] Terry Therneau. A package for survival analysis in s. r package version 2.37-7, 2014.

[62] Jasjeet S Sekhon. Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software, Forthcoming*, 2008.

[63] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[64] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[65] Greg Ridgeway. Generalized boosted models: A guide to the gbm package. *Update*, 1 (1):2007, 2007.