Proximal Causal Inference With Text Data

Jacob M. Chen
Williams College
jmc8@williams.edu

Rohit Bhattacharya
Williams College
rb17@williams.edu

Katherine A. Keith
Williams College
kak5@williams.edu

Abstract

Recent text-based causal methods attempt to mitigate confounding bias by including unstructured text data as proxies of confounding variables that are partially or imperfectly measured. These approaches assume analysts have supervised labels of the confounders given text for a subset of instances, a constraint that is not always feasible due to data privacy or cost. Here, we address settings in which an important confounding variable is completely unobserved. We propose a new causal inference method that splits pre-treatment text data, infers two proxies from two zero-shot models on the separate splits, and applies these proxies in the proximal g-formula. We prove that our text-based proxy method satisfies identification conditions required by the proximal g-formula while other seemingly reasonable proposals do not. We evaluate our method in synthetic and semi-synthetic settings and find that it produces estimates with low bias. This combination of proximal causal inference and zero-shot classifiers is novel (to our knowledge) and expands the set of text-specific causal methods available to practitioners.

1 Introduction

Data-driven decision making relies on estimating the effect of interventions, i.e. *causal effect estimation*. For example, a doctor must decide which medicine she will give her patient, ideally the one with the greatest effect on positive outcomes. Many causal effects are estimated via randomized controlled trials—considered the gold standard in causal inference; however, if an experiment is unfeasible or unethical, one must use observational data. In observational settings, a primary obstacle to unbiased causal effect estimation is confounding variables, variables that affect both the treatment (e.g., which medicine) and the outcome.

Recently, some studies have attempted to mitigate confounding by incorporating (pre-treatment) unstructured text data as proxies for confounding variables or specifying linguistic properties as the confounding variables themselves, e.g., topic (Veitch et al., 2020; Roberts et al., 2020), tone (Sridhar and Getoor, 2019), or use of specific word types (Olteanu et al., 2017). A wide range of fields have used text in casual estimates, including medicine (Zeng et al., 2022), the behavioral social sciences (Kiciman et al., 2018), and science-of-science (Zhang et al., 2023). See Keith et al. (2020); Feder et al. (2022); Egami et al. (2022) for general overviews of text-based causal estimation.

If all confounders are directly observed, then causal estimation is relatively[1] straightforward with *backdoor adjustment* (Pearl, 2009). However, some applications use supervised classifiers to predict the confounding variables from text data. Because text classifiers rarely achieve perfect accuracy, analysts must account for *measurement error*. To address this, another line of work has developed post-hoc corrections of causal estimates in the presence of imperfect classifiers (Wood-Doughty et al., 2018; Fong and Tyler, 2021; Egami et al., 2023; Mozer et al., 2023). Yet, these approaches require ground-truth labels of the confounding variables for a subset of instances, a constraint that is not always feasible due to privacy restrictions, high costs, or lack of expert labor for labeling.

Our work fills this gap. We address the causal estimation setting for which a practitioner has specified a confounding variable that is truly unmeasured (we have no observations of the variable), but unstructured text data could be used to infer proxies. For this setting, our method combines *proximal causal inference* with zero-shot classifiers.

Proximal causal inference (Miao et al., 2018; Tchetgen Tchetgen et al., 2020) can identify the true causal effect given *two* proxies for the unmeasured confounder that satisfy certain causal identifi-

[1] Setting aside challenges of high-dimensional covariate selection for causal estimation, e.g., see Tamarchenko (2023).

cation conditions. A major criticism of this method is that it can be difficult to find two suitable proxies among the structured variables; however, we conjecture that unstructured text data (if available) could be a rich source of potential proxies.

In our proposed method, we estimate two proxies from text data via zero-shot classifiers, i.e. classifiers that perform an unseen task with no supervised examples. In subsequent sections, we expand upon the necessary conditions required for our method and empirically validate our method on synthetic and semi-synthetic data with real-world clinical notes. Since large pre-trained language models (LLMs) have promising performance on zero-shot classification benchmarks (Yin et al., 2019; Brown et al., 2020; Wei et al., 2021; Sanh et al., 2021, *inter alia*), we use LLMs for one (or both) of the proxies in our experimental pipeline. Our combination of proximal causal inference and zero-shot classifiers is not only novel, but also expands the set of text-specific causal designs available to practitioners.[2]

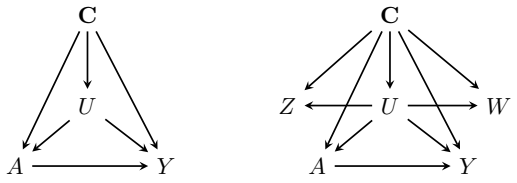
In summary, our contributions are

- We propose a new causal inference method that splits pre-treatment text data, infers two proxies from two different zero-shot models on the separate splits, and applies these in the proximal g-formula (Tchetgen Tchetgen et al., 2020).
- We provide theoretical proofs that our method satisfies the identification conditions of *proximal causal inference* and prove that other seemingly reasonable alternative methods do not.
- We propose a falsification heuristic that uses the odds ratio of the proxies conditional on observed covariates as an approximation of the (untestable) proximal causal inference conditions.
- In synthetic and semi-synthetic experiments using MIMIC-III's real-world clinical notes (Johnson et al., 2016), our odds ratio heuristic correctly flags when identification conditions are violated. When the heuristic passes, causal estimates from our method have low bias and confidence intervals that cover the true parameter; others do not.

2 Problem Setup And Motivation

To motivate our approach, let us imagine we are an applied practitioner who is tasked with determining the effectiveness of thrombolytic (clot busting) medications relative to blood thinning medications to treat clots arising from an ischemic stroke. Such

[2]Supporting code is available at https://github.com/jacobmchen/proximal_w_text.



(a) (b)

Figure 1: Causal DAGs (a) depicting unmeasured confounding and (b) compatible with the canonical assumptions used for *proximal causal inference*.

medications are usually administered within three hours of the stroke to improve chances of patient recovery (Zaheer et al., 2011). Given the urgency and the short treatment window, running a randomized experiment to compare these drugs is infeasible. This leaves us with observational data, so we examine electronic health records (EHRs) from a database like MIMIC-III (Johnson et al., 2016).

We formalize our causal estimand as follows: let A denote a binary treatment variable corresponding to clot busting ($A = 1$) or blood thinning medication ($A = 0$), and let Y denote measurements of the D-dimer protein in the patient's blood which directly measures how much of the clotting has dissolved. In do-calculus notation (Pearl, 2009), the target causal estimand is the average causal effect, $\text{ACE} := \mathbb{E}[Y|\,\text{do}(A = 1)] - \mathbb{E}[Y|\,\text{do}(A = 0)]$.

Examining the EHRs, we find structured variables—variables with defined values in tabular form—that are potential confounders, including biological factors, such as age, sex, and blood pressure, as well as socio-economic factors, such as income. We denote the observed confounders as the set $\mathbf{C}$. However, we are worried about biased causal effects because atrial fibrillation (irregular heart rhythms) is an important confounder corresponding to a pre-existing heart condition, and it is not recorded in the structured data. We denote this unmeasured confounder as U. Figure 1(a) depicts this problem setup in the form of a causal directed acyclic graph (causal DAG) (Spirtes et al., 2000; Pearl, 2009). In this case, it is well known that adjusting for just the observed confounders via the backdoor formula $\sum_{\mathbf{C}}(\mathbb{E}[Y|A = 1, \mathbf{C}] - \mathbb{E}[Y|A = 0, \mathbf{C}]) \times p(\mathbf{C})$ will give a biased estimate of the ACE (Pearl, 1995).

In response to this issue, we consider work that uses proxy variables of the unmeasured confounder. However, we are subject to the following **restriction**:

(R1) We do not have access to any rows of data where U is measured.

This kind of restriction is common in healthcare or social science settings when data privacy is crucial and hand labeling of unstructured data is impossible or infeasible due to high costs or lack of expert labor; we elaborate in Appendix B.

Under (R1), the classic *effect restoration* method developed in Pearl (2010) cannot be applied because this method requires us to estimate a distribution $P(W|U)$, where W is the proposed proxy of U, a task impossible without access to U. Building from Kuroki and Pearl (2014), a more recent line of work called *proximal causal inference* (Miao et al., 2018; Tchetgen Tchetgen et al., 2020) is able to identify the true causal effect as long as the analyst proposes two proxies W and Z satisfying the following independence conditions

(P1) Independence of proxies: $W \perp\!\!\!\perp Z \mid U, \mathbf{C}$

(P2) One of the proxies, say W, does not depend on values of the treatment: $W \perp\!\!\!\perp A \mid U, \mathbf{C}$

(P3) The other proxy, Z, does not depend on values of the outcome: $Z \perp\!\!\!\perp Y \mid A, U, \mathbf{C}$

A canonical example of proxies that satisfy these conditions is shown in Figure 1(b)[3]. In addition to these independence relations that impose the absence of certain edges in the causal DAG, e.g., no edge can be present between Z and W to satisfy (P1), there is an additional completeness condition that imposes the existence of $U \to Z$ and $U \to W$. This condition is akin to the relevance condition in the instrumental variables literature (Angrist et al., 1996) and ensures that the proxies W and Z exhibit sufficient variability relative to the variability of U.

(P4) Completeness: for any square integrable function $v(\cdot)$ and for all values $w, a, \mathbf{c}$, we have

$$\mathbb{E}[v(U) \mid w, a, \mathbf{c}] = 0 \iff v(U) = 0, \text{ and}$$
$$\mathbb{E}[v(Z) \mid w, a, \mathbf{c}] = 0 \iff v(Z) = 0.$$

Intuitively, these completeness conditions do not hold unless Z and W truly hold some predictive value for the unmeasured confounder U.[4] Under (P1-P4), each piece of the ACE, $\mathbb{E}[Y|\operatorname{do}(A=a)]$, is identified via the *proximal g-formula*,

$$\mathbb{E}[Y|\operatorname{do}(a)] = \sum_{w,\mathbf{c}} h(a, w, \mathbf{c}) \times p(w, \mathbf{c}), \quad (1)$$

where $h(a, w, \mathbf{c})$ is referred to as an outcome confounding bridge function that is a solution to the equation $\mathbb{E}[Y|a, z, \mathbf{c}] = \sum_w h(a, w, \mathbf{c}) \times p(w|a, z, \mathbf{c})$ (Miao et al., 2018). The existence of a solution is guaranteed under (P1-P4), but solving it can still be difficult. However, there exist simple two-stage regression estimators for the proximal g-formula (Tchetgen Tchetgen et al., 2020; Mastouri et al., 2021) which we make use of in Section 5 once we have identified a valid pair of proxies. This brings us to the primary issue at hand.

Primary question. In practice, how do we find two proxies W and Z among the structured variables such that they happen to satisfy all of (P1-P4)?

Our answer. We cannot, at least not without a high degree of domain knowledge which is unavailable in many real-world settings. Additionally, since these conditions are untestable in the observed data, empirical verification of (P1-P4) is also impossible. Instead, in this work, we propose relying on raw unstructured text data (e.g., clinical notes) in an attempt to infer proxies that satisfy (P1-P4) ***by design***.

Returning to our motivating problem, our approach applies zero-shot models (due to (R1)) to two separate splits of the pre-treatment clinical notes of each patient and then obtains two different predictions, W and Z, for atrial fibrillation (our U).

In order to simplify our method and make it easier to implement, we make two relatively weak assumptions.

(S1) The unmeasured confounder U between A and Y can be specified as a binary variable.

(S2) The text only causes W and Z (and no other variables).

We make (S1) since text classification typically performs better empirically than text regression (Wang et al., 2022). Assumption (S2) asserts that the text data considered serves as a record of events rather than actionable data.[5]

[3] One could also add the edges $W \to Y$ and $Z \to A$ to Figure 1(b), but these relations will not show up in our text-based setting. Shpitser et al. (2023) also propose a general proximal identification algorithm that is compatible with other causal DAGs, but we focus on the canonical proximal learning assumptions stated in Tchetgen Tchetgen et al. (2020).

[4] See Miao et al. (2018) for more details on completeness.

[5] We leave to future work more complicated scenarios, e.g., if the reader's perception of text differs from the writer's intent (Pryzant et al., 2021).

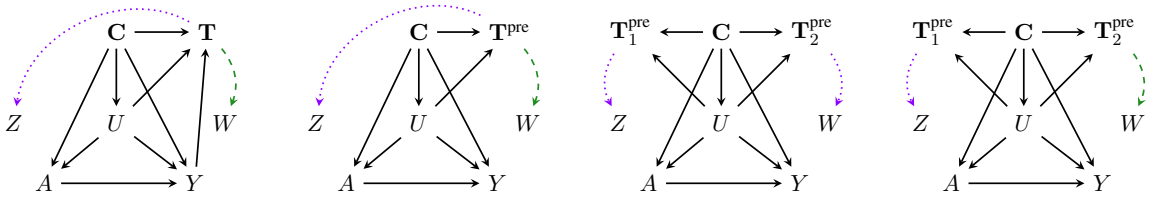


Figure 2: Causal DAGs depicting depicting several different scenarios for inferring text-based proxies. Edges with different colors and patterns, e.g., $\mathbf{T} \dashrightarrow Z$ and $\mathbf{T} \dashrightarrow W$, indicate that different zero-shot models were used.

3 Designing Text-Based Proxies

In this section, we describe our method for designing text-based proxies. In doing so, we describe various "gotchas," pitfalls in attempting to use these text-based proxies in causal effect estimation. Later, in Section 5, we explore these pitfalls empirically and find, as expected, they result in biased causal effect estimates. Finally, we describe how exploring these gotchas leads us to our final recommended design, given by Figure 2(d).

Gotcha #1: Not handling predictions as proxies

Suppose we try to avoid the complications of proximal causal inference by using the predictions from one of our proxy zero-shot models, say W, as the confounding variable itself.

Proposition 1. *Unless the predicted W is exactly equal to U, using W as a direct replacement for U in the backdoor adjustment formula will give biased estimates of the ACE.*

Proof. Suppose for some subset of instances $W \neq U$. Then, the backdoor path through U is not blocked, and the ACE is biased (Pearl, 1995). □

In other words, we would need a zero-shot classifier with 100% accuracy in order to use W directly, a scenario that is extremely unlikely in the real world. Further, under (R1)[6], we cannot measure accuracy at inference time since we do not have any observations of U. As expected, in our semi-synthetic experiments in Section 5, we find using predictions of W directly in a backdoor adjustment formula results in biased estimates; see Figure 3.

Gotcha #2: Using post-treatment text

While it is well-known that adjusting for post-treatment covariates in the backdoor formula often leads to bias (Pearl, 2009), it is not obvious what might go wrong when using post-treatment text to infer proxies for the proximal g-formula.

Proposition 2. *If both W and Z are inferred from zero-shot models on text that contain post-treatment information, then the resulting proxies violate either (P2) or (P3), or both.*

Proof. Consider Figure 2(a), where the proxies are produced using text that is potentially post-outcome and thus also post-treatment. We show that this violates both (P2) and (P3). Clearly this violates (P3)—by a simple d-separation argument we see that $Z \not\perp\!\!\!\perp Y \mid A, U, \mathbf{C}$ due to the open path $Y \to \mathbf{T} \dashrightarrow Z$. Similarly, (P2) is violated from the open path $A \to Y \to \mathbf{T} \dashrightarrow W$. □

Thus, before performing zero-shot inference, it is important that the text for each individual is filtered in such a way that it contains only the text preceding treatment[7]. In our running clinical example, this is easily achieved since clinical notes have time stamps and information about when the patient was treated and discharged.

Gotcha #3: Predicting both proxies from the same passage of text

After filtering to only pre-treatment text, $\mathbf{T}^{\text{pre}}$, for each individual, the intuitive next step is to use $\mathbf{T}^{\text{pre}}$ to infer W and Z. Yet, we show this will result in biased estimates.

Proposition 3. *If W and Z are inferred via zero-shot models on the same passage of pre-treatment text, the resulting proxies violate (P1).*

Proof. Consider the causal DAG with proxies W and Z in Fig 2(b). By d-separation we have $W \not\perp\!\!\!\perp Z \mid U, \mathbf{C}$ due to the path $Z \dashleftarrow \mathbf{T}^{\text{pre}} \dashrightarrow W$. □

[6] In the absence of (R1), we direct readers to work that adjusts via measurement error estimates or assumptions that U is "missing at random" (Wood-Doughty et al., 2018).

[7] In Appendix A we provide an example where the proxies could be generated using a mix of pre and post-treatment text while satisfying (P1-P3). However, the pre-treatment rule is simpler so we recommend it for our method.

As expected, in our synthetic experiments in Section 5, we find inferring proxies from the same text results in biased estimates; see Table 1.

Gotcha #4: Using a single zero-shot model

To avoid Gotcha #3, we split[8] the pre-treatment text into two halves, $\mathbf{T}_1^{\text{pre}}$ and $\mathbf{T}_2^{\text{pre}}$. Now, should we apply the same zero-shot model to these two passages of text to infer the proxies W and Z, as in Figure 2(c), or should we apply two separate zero-shot models as in Figure 2(d)?

We find different answers in theory and practice. Proposition 4 shows that, in theory, both are valid as long as $\mathbf{T}_1^{\text{pre}} \perp\!\!\!\perp \mathbf{T}_2^{\text{pre}} \mid U, \mathbf{C}$.[9] However, we later describe how our semi-synthetic experiments demonstrate the need for two different models in practice in Section 5. We first establish the theoretical validity of both strategies.

Proposition 4. *If W and Z are inferred using zero-shot classification on two separate splits of pre-treatment text such that $\mathbf{T}_1^{pre} \perp\!\!\!\perp \mathbf{T}_2^{pre} \mid U, \mathbf{C}$, then these proxies satisfy (P1-P3). Additionally, if the proxies are predictive of U, i.e., $Z \not\perp\!\!\!\perp U \mid \mathbf{C}$ and $W \not\perp\!\!\!\perp U \mid \mathbf{C}$, then (P4) holds.*

Proof. Suppose we apply zero-shot classification models to two splits of pre-treatment text in a way that results in causal DAGs shown in Figure 2(c) or (d), depending on whether we use one or two models, respectively. Applying d-separation confirms that the conditions (P1-P3) hold in both cases.

Let $|\mathfrak{X}_V|$ denote the number of categories of a variable V. Kuroki and Pearl (2014); Tchetgen Tchetgen et al. (2020) state that when W and Z are predictive of U (as stated in the proposition), a sufficient condition for (P4) is $\min(|\mathfrak{X}_Z|, |\mathfrak{X}_W|) \geq |\mathfrak{X}_U|$. Since U is binary under (S1) and since W and Z are inferred from classifiers, they are discrete variables, and this condition is satisfied. Hence, (P4) is satisfied. □

Our Final Design Procedure

Based on the suggestions accumulated in this section, our design procedure is summarized by the causal DAG in Figure 2(d)—obtain pre-treatment text, split it into two halves, and apply two distinct zero-shot models to obtain W and Z.

In Proposition 4, we formalized how this procedure can be used to design proxies that satisfy the proximal identification conditions (P1-P4). However, this result relied on two important preconditions: (1) the conditional independence of the two splits of the text, and (2) W and Z being (at least weakly) predictive of U. Yet, both of these conditions are untestable. This motivates our *falsification heuristic* in the next section.

4 Falsification: Odds Ratio Heuristic

In practice, a major challenge for our procedure—and indeed all causal methods—are its assumptions. Sometimes, causal models imply testable restrictions on the observed data that can be used in *falsification* or *confirmation tests* of model assumptions, see Wang et al. (2017); Bhattacharya and Nabi (2022); Chen et al. (2023) for tests of some popular causal models. In our case, the proximal model implies no testable restrictions (Tchetgen Tchetgen et al., 2020), so the best we can do is provide a *falsification heuristic* that allows analysts to detect serious violations of (P1-P4) when using the inferred proxies. We design our heuristic based on the odds ratio function described below.

Given arbitrary reference values w_0 and z_0, the conditional odds ratio function for W and Z given covariates $\mathbf{X}$ is defined as (Chen, 2007),

$$\text{OR}(w, z \mid \mathbf{x}) = \frac{p(w \mid z, \mathbf{x})}{p(w_0 \mid z, \mathbf{x})} \frac{p(w_0 \mid z_0, \mathbf{x})}{p(w \mid z_0, \mathbf{x})}.$$

This function is important because $W \perp\!\!\!\perp Z \mid \mathbf{X}$ if and only if $\text{OR}(w, z \mid \mathbf{x}) = 1$ for all values $w, z, \mathbf{x}$. We summarize this odds ratio as a single free parameter, $\gamma_{WZ.\mathbf{X}}$, and, for the simplicity of our pipeline, we estimate it under a parametric model for $p(W|Z, \mathbf{X})$.[10]

Now, we describe our proximal conditions in terms of odds ratio parameters. If (P1-P3) are satisfied, then $W \perp\!\!\!\perp Z \mid U, \mathbf{C}$ and $\gamma_{WZ.U\mathbf{C}} = 1$. Further, if the zero-shot models are truly predictive of U, then (P4) is satisfied and $W \not\perp\!\!\!\perp Z \mid \mathbf{C}$ which means $\gamma_{WZ.\mathbf{C}} \neq 1$. Ideally, we would want to estimate both of these odds ratio parameters to confirm (P1-P4) empirically; however, $\gamma_{WZ.U\mathbf{C}}$ cannot be computed from observed data alone due to (R1).

[8] While there exist many strategies to split the text, we find simply splitting in half works well in practice. For concerns regarding settings with short texts, we direct the reader to our odds ratio heuristic in Section 4.

[9] Note this independence condition does not imply the two pieces of text are completely uncorrelated. Since the text is written based on observations of the same individual we certainly expect $\mathbf{T}_1^{\text{pre}} \not\perp\!\!\!\perp \mathbf{T}_2^{\text{pre}}$; we simply require that the two pieces are correlated only due to $\mathbf{C}$ and U.

[10] More generally, the odds ratio can also be treated as a finite p-dimensional parameter vector $\boldsymbol{\gamma}$ and estimated under parametric or semi-parametric restrictions on $p(W|Z, \mathbf{X})$ and $p(Z|W, \mathbf{X})$ (Chen, 2007; Tchetgen Tchetgen et al., 2010).

Using a parameter we can estimate from observed data, $\gamma_{WZ.\mathbf{C}}$, we propose an **odds ratio falsification heuristic** in lines 6-10 of Algorithm 1. Now, we explain why, if this heuristic holds, an analyst can be reasonably confident in using their inferred text-based proxies for estimation.

First, we examine a lower bound on $\gamma_{WZ.\mathbf{C}}$. Based on our previous discussion, if $\gamma_{WZ.\mathbf{C}}$ is close to 1, then we should suspect that one or both of our zero-shot models failed to return informative predictions for U. Next, let us treat $\gamma_{WZ.\mathbf{C}}$ as an imperfect approximation of $\gamma_{WZ.U\mathbf{C}}$. Let W, Z, U be binary with reference values $w_0 = z_0 = u_0 = 0$. VanderWeele (2008) proposed the following three conditions under which an odds ratio $\gamma_{WZ.\mathbf{C}}$ that fails to adjust for an unmeasured confounder U is an *overestimate* of the true odds ratio $\gamma_{WZ.U\mathbf{C}}$: (i) $\{U\} \cup \mathbf{C}$ satisfies the backdoor criterion with respect to W and Z; (ii) U is univariate or consists of independent components conditional on $\mathbf{C}$; (iii) $\mathbb{E}[W|z, u, \mathbf{c}]$ is non-decreasing in U for all z and $\mathbf{c}$ and $\mathbb{E}[Z|u, \mathbf{c}]$ is non-decreasing in U for all $\mathbf{c}$. Condition (i) is satisfied from Graph 2(d), and condition (ii) is satisfied by assumption (S1). Finally, condition (iii) is satisfied when our zero-shot models are reasonable predictors of the unmeasured confounder U by the following argument. Notice that $\mathbb{E}[W|z, u, \mathbf{c}] = \mathbb{E}[W|u, \mathbf{c}] = p(W = 1|u, \mathbf{c})$, where the first equality follows from d-separation in Graph 2(d) and the second follows from definition of expectation for binary W. Then we should expect, if the zero-shot models are reasonably accurate, that $p(W = 1|U = 1, \mathbf{c}) > p(W = 1|U = 0, \mathbf{c})$. Therefore, the first part of condition (iii) is satisfied. Similar logic holds for $\mathbb{E}[Z|u, \mathbf{x}]$.

Hence, we have shown that under ideal conditions that satisfy (P1-P4), we should expect $\gamma_{WZ.\mathbf{C}} > \gamma_{WZ.U\mathbf{C}} = 1$, and we should reject proxies W and Z if we get an odds ratio $\gamma_{WZ.\mathbf{C}} \leq 1$. Next, we examine the upper bound on $\gamma_{WZ.\mathbf{C}}$.

Consider the extreme case where $\gamma_{WZ.\mathbf{C}} = \infty$. This corresponds to a situation where $W = Z$, so (P1) is clearly not satisfied. In general, if $\gamma_{WZ.\mathbf{C}}$ is higher than some threshold γ_{high}, corresponding to the maximum association that one could reasonably explain by a single open path through U, we should suspect that perhaps the proxies W and Z are associated with each other due to additional paths through other unmeasured variables that make it so that the two halves of text are not independent of each other, i.e., $\mathbf{T}_1^{\text{pre}} \not\perp\!\!\!\perp \mathbf{T}_2^{\text{pre}} \mid U, \mathbf{C}$.

Algorithm 1 for inferring two text-based proxies

1: **Inputs**: Observed confounders $\mathbf{C}$; Text $\mathbf{T}$; Zero-shot models $\mathcal{M}_1$, $\mathcal{M}_2$; Specified γ_{high}
2: Extract pre-treatment text $\mathbf{T}^{\text{pre}}$ from $\mathbf{T}$
3: Split $\mathbf{T}^{\text{pre}}$ into two halves $\mathbf{T}_1^{\text{pre}}$ and $\mathbf{T}_2^{\text{pre}}$
4: $Z \leftarrow \mathcal{M}_1(\mathbf{T}_1^{\text{pre}})$ and $W \leftarrow \mathcal{M}_2(\mathbf{T}_2^{\text{pre}})$
5: // Odds Ratio Falsification Heuristic
6: **if** $1 < \gamma_{WZ.\mathbf{C}} < \gamma_{\text{high}}$ **then**
7: return W and Z
8: **else**
9: return "stop"
10: **end if**

Following standard practice in *sensitivity analysis*, e.g., Liu et al. (2013); Leppälä (2023), we leave it to the analyst to specify the upper bound γ_{high} based on domain knowledge. In our experiments in Section 5, we found that, when the proximal conditions are not satisfied, $\gamma_{WZ.\mathbf{C}}$ far exceeds any reasonable setting of γ_{high}. Hence, our heuristic works quite well in practice even with a generous suggestion for an upper bound. We describe our full design procedure with the diagnostic in Algorithm 1. We now evaluate its effectiveness for downstream causal inference.

5 Empirical Experiments and Results

In this section, we explore the following empirical research questions:

RQ: How does Algorithm 1 compare to other alternatives in terms of bias and confidence interval coverage of the estimated causal effects? Does our odds ratio heuristic effectively flag when to stop or proceed?

In causal inference, empirical evaluation is difficult because it requires ground-truth labels for counterfactual outcomes of an individual under multiple versions of the treatment, data that is generally impossible to obtain (Holland, 1986); see Gentzel et al. (2019); Keith et al. (2024). Thus, we turn to synthetic data and semi-synthetic data so we have access to the true ACE and U to evaluate methods. We describe the experimental set-ups, the causal estimation procedure used by all experiments, and finally, the results to our RQ.

5.1 Fully Synthetic Experiments

We create our fully synthetic DGP based on the DAG in Figure 2(d); see Appendix C for full details. To summarize, A and U are binary, and Y and C are continuous. We generate (very simple)

Estimation Pipeline	$\gamma_{WZ.\mathbf{C}}$	Est. ACE	Bias	Conf. Interval (CI)	CI Cov.
P1M	$2.799^{\checkmark}$	1.332	**0.032**	$(1.285, 1.379)$	**Yes**
P1M with Gotcha #3	4.795×10^{25}	1.509	0.209	$(1.475, 1.545)$	No
P2M	$3.761^{\checkmark}$	1.315	**0.015**	$(1.268, 1.358)$	**Yes**
P2M with Gotcha #3	1.047×10^{2}	1.464	0.164	$(1.430, 1.498)$	No

Table 1: **Fully synthetic results** with the true ACE equal to 1.3. Here, ✓ distinguishes settings that passed the odds ratio heuristic from those that failed it, with $\gamma_{\text{high}} = 5$.

synthetic text data with four continuous variables, X_1, X_2, X_3, X_4, as functions of U and C. For training, we generate two realizations of these variables, which we call $\mathbf{X}_1^{\text{train}}$ and $\mathbf{X}_2^{\text{train}}$, and likewise two realizations for inference time, $\mathbf{X}_1^{\text{inf}}$ and $\mathbf{X}_2^{\text{inf}}$.

At inference time, we explore four different strategies for inferring proxies. First, we explore using one or two distinct models, which we refer to as *Proximal 1-Model* (P1M) and *Proximal 2-Model* (P2M), respectively. For one model, we train a logistic regression classifier to predict the true U from an aggregated variable $\widetilde{\mathbf{X}}^{\text{train}} = (\mathbf{X}_1^{\text{train}} + \mathbf{X}_2^{\text{train}})/2$ as $P_\theta(U = 1|\widetilde{\mathbf{X}}^{\text{train}})$.[11] For the other model, we use the following heuristic: predict 1 if $X_1 > 1.1$ else 0. P1M uses the logistic regression model and P2M uses both the logistic regression and heuristic models. Next, we vary whether Gotcha #3 holds at inference time for both P1M and P2M, i.e. whether the models infer Z and W from only $\mathbf{X}_1^{\text{inf}}$ or both $\mathbf{X}_1^{\text{inf}}$ and $\mathbf{X}_2^{\text{inf}}$.

5.2 Semi-Synthetic Experiments

For our semi-synthetic data, we use real-world clinical notes and structured variables from MIMIC-III, an anonymized dataset of patients admitted to critical care units at a large tertiary care hospital (Johnson et al., 2016). We provide full details of pre-processing steps and the DGP in Appendix D. To summarize, we use the following real data from MIMIC-III: we choose `age` and `gender` of patients as $\mathbf{C}$; we create three different settings for U: patients' binary diagnostic status of `atrial fibrillation` (Afib), `congestive heart failure` (Heart), or `acute kidney failure` (Kidney). As $\mathbf{T}^{\text{pre}}$, we use clinician notes and (to avoid Gotcha #2) explicitly exclude discharge summaries which are post-treatment. We split at the middle character index to create $\mathbf{T}_1^{\text{pre}}$ and $\mathbf{T}_2^{\text{pre}}$. Then, consistent with the DAG in Figure 2(d), we synthetically generate binary A and continuous Y. In total, we use 29,451 patient records.

P1M and P2M. For one of our zero-shot models, we use Flan-T5 XXL, an instruction-tuned large language model, (Chung et al., 2022). Following Ziems et al. (2023), we use the prompt template *Context: {$\mathbf{T}_1^{pre}$} Is it likely the patient has {U}? Constraint: Even if you are uncertain, you must pick either 'Yes' or 'No' without using any other words.* We assign 1 when Flan-T5 outputs *"Yes"* and 0 otherwise. For the other zero-shot model, we use keyword matching. The model outputs 1 if the text contains one of the following keywords: "atrial" for U=Afib; "family" for U=Heart; and "liver" for U=Kidney;[12] otherwise it outputs 0. P1M uses Flan-T5 for both Z and W and P2M uses Flan-T5 for W and keyword matching for Z.

Baselines. We compare P1M and P2M to two baselines. First, we randomly generate W and Z from the distribution $p(W = 1) = p(Z = 1) = 0.5$. Second, we use the inferred W from Flan-T5 on $\mathbf{T}_1^{\text{pre}}$ directly in a backdoor adjustment formula, corresponding to Gotcha #1.

5.3 Estimation of Proximal g-formula

For all experiments, we estimate the ACE by using the inferred W and Z in a two-stage linear regression estimator for the proximal g-formula provided by Tchetgen Tchetgen et al.. Although the linearity assumption is restrictive, it allows us to focus on evaluating the efficacy of our proposed method for inferring text-based proxies as opposed to complications with non-linear proximal estimation (Mastouri et al., 2021). Briefly, we first fit a linear regression $\mathbb{E}[W|A, Z, \mathbf{X}]$. Next, we infer $\widehat{W}$, continuous predictions for W, using the fitted model. For the second stage, we fit a linear model for $\mathbb{E}[Y|A, \widehat{W}, \mathbf{X}]$. The coefficient for A in this

[11] Of course, having access to the true U may not qualify as "zero-shot" in the strict sense of the term, but we complement this idealized scenario with the difficult true zero-shot scenario in semi-synthetic experiments.

[12] Justifications for these keywords are in Appendix D.

U	$\gamma_{WZ.U\mathbf{C}}$ P1M	$\gamma_{WZ.U\mathbf{C}}$ P2M	$\gamma_{WZ.\mathbf{C}}$ P1M	$\gamma_{WZ.\mathbf{C}}$ P2M
Afib	4.655	1.071	7.753	1.719✓
Heart	28.327	1.513	37.852	2.836✓
Kidney	55.907	3.31×10^7	56.366	3.44×10^7

Table 2: **Semi-synthetic odds ratio heuristic** calculated with $\gamma_{WZ.\mathbf{C}}$ as well as the oracle $\gamma_{WZ.U\mathbf{C}}$. Here, ✓ distinguishes settings that passed the heuristic from those that failed it, with $\gamma_{\text{high}} = 5$.

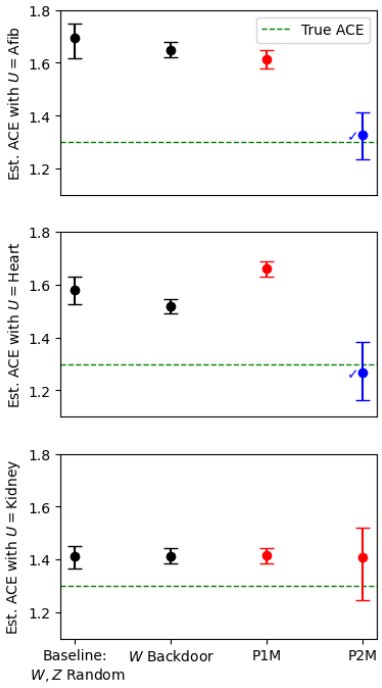


Figure 3: **Semi-synthetic results** for ACE point estimates (dots) and confidence intervals (bars). Here, ✓ distinguishes settings that passed the odds ratio heuristic from those that failed it, with $\gamma_{\text{high}} = 5$.

second linear model is the estimated ACE. We calculate 95% confidence intervals for the ACE via the bootstrap percentile method (Wasserman, 2004).

5.4 Results

Results for synthetic experiments are in Table 1 and for semi-synthetic experiments are in Tables 2, 3, 4 and Figure 3. In short, our empirical results corroborate preference for Algorithm 1 over alternatives.

Gotchas. First, we discuss the "gotcha" methods we showed to be theoretically incorrect in Section 3. Regarding Gotcha #1, Figure 3 shows that, across all settings of U, using the inferred W directly in the backdoor adjustment formula results in estimates with large bias. Regarding Gotcha #3, using the same realization $\mathbf{X}_1^{\text{inf}}$ results in high bias—0.209 and 0.164 for P1M and P2M, respectively, in Table 1—whose CIs do not cover the true ACE. Regarding Gotcha #4, as in Proposition 4, using a single zero-shot model (P1M) results in low bias in the idealized setting of the synthetic DGP (Table 1; first row). However, using real clinical notes and Flan-T5, the third column of Figure 3 shows that P1M has estimates with high bias across all three settings of U. We hypothesize this is due Flan-T5 predicting Z and W with raw agreement rates higher than 0.9 for all three Us (Table 3).

Odds ratio heuristic. For all experiments we set $\gamma_{\text{high}} = 5$.[13] Even with this liberal setting, we find low bias and good CI coverage for scenarios that pass our heuristic (P1M and P2M for synthetic and P2M for U equal to Afib and Heart for semi-synthetic) and poor estimates in all other cases. Examining Table 2, we see that $\gamma_{WZ.\mathbf{C}} > \gamma_{WZ.U\mathbf{C}}$ for all settings, which corroborates the discussion in Section 4. This gives us confidence that Algorithm 1 appropriately flags when to stop or proceed.

6 Conclusion and Future Work

In this work we proposed a novel causal inference method for estimating causal effects in observational studies when a confounding variable is completely unobserved but unstructured text data is available to infer potential proxies. Our method splits the pre-treatment text data, infers two proxies using two zero-shot models on the separate splits, and applies these proxies in the proximal g-formula. We have shown why one should prefer our method to alternatives, both with theory and the results of synthetic and semi-synthetic experiments.

Although we use a clinical setting as our running example, our method is applicable to many other domains where it is infeasible to obtain ground-truth U labels due to privacy constraints, e.g., social media or education studies with private messaging data or sensitive student coursework. Possible directions for future work include incorporating non-linear proximal estimation strategies, expanding beyond text to learned proxies from other modalities (see Knox et al. (2022)), more guidance for setting γ_{high} in our odds ratio heuristic, extending our method to incorporate categorical U, W, Z, and evaluating the efficacy of using soft probabilistic outputs from the zero-shot classifiers.

[13] This means, conditional on $\mathbf{C}$, we would expect the odds of W to be able to increase 5-fold for every increase in Z, an extremely liberal constraint given W and Z are binary.

References

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Rohit Bhattacharya and Razieh Nabi. 2022. On testability of the front-door model via Verma constraints. In *Proceedings of the 38th Conference On Uncertainty in Artificial Intelligence*, pages 202–212. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Hua Yun Chen. 2007. A semiparametric odds ratio model for measuring association. *Biometrics*, 63:413–421.

Jacob M Chen, Daniel Malinsky, and Rohit Bhattacharya. 2023. Causal inference with outcome-dependent missingness and self-censoring. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, pages 358–368. PMLR.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652.

Naoki Egami, Musashi Hinck, Brandon M Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Christian Fong and Matthew Tyler. 2021. Machine learning predictions as regression covariates. *Political Analysis*, 29(4):467–484.

Amanda M Gentzel, Dan Garant, and David Jensen. 2019. The case for evaluating causal models using interventional measures and empirical data. *Advances in Neural Information Processing Systems*, 32.

Paul W Holland. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344.

Katherine A Keith, Sergey Feldman, David Jurgens, Jonathan Bragg, and Rohit Bhattacharya. 2024. RCT rejection sampling for causal estimation evaluation. *(Forthcoming) Transactions of Machine Learning Research*.

Emre Kiciman, Scott Counts, and Melissa Gasser. 2018. Using longitudinal social media analysis to understand the effects of early college alcohol use. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Dean Knox, Christopher Lucas, and Wendy K Tam Cho. 2022. Testing causal theories with learned proxies. *Annual Review of Political Science*, 25:419–441.

Manabu Kuroki and Judea Pearl. 2014. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Kalle Leppälä. 2023. Sensitivity analysis on odds ratios. *American Journal of Epidemiology*, page kwad137.

Weiwei Liu, S Janet Kuramoto, and Elizabeth A Stuart. 2013. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science*, 14:570–580.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. 2021. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International conference on machine learning*, pages 7512–7523. PMLR.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. 2018. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

Reagan Mozer, Aaron R Kaufman, Leo A Celi, and Luke Miratrix. 2023. Leveraging text data for causal inference using electronic health records. *arXiv preprint arXiv:2307.03687*.

Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 370–386.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Judea Pearl. 2010. On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 425–432.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 2021. Causal effects of linguistic properties. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4095–4109.

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

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2021. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Ilya Shpitser, Zach Wood-Doughty, and Eric J Tchetgen Tchetgen Tchetgen. 2023. The proximal ID algorithm. *Journal of Machine Learning Research*, 24(188):1–46.

Peter L Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. MIT press.

Dhanya Sridhar and Lise Getoor. 2019. Estimating causal effects of tone in online debates. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1872–1878.

Elijah Tamarchenko. 2023. Combining optimal adjustment set selection and post selection inference in unknown causal graphs. Undergraduate thesis, Williams College.

Eric J Tchetgen Tchetgen, James M Robins, and Andrea Rotnitzky. 2010. On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, 97(1):171–180.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. 2020. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*.

Tyler J VanderWeele. 2008. The sign of the bias of unmeasured confounding. *Biometrics*, 64(3):702–706.

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

Linbo Wang, James M Robins, and Thomas S Richardson. 2017. On falsification of the binary instrumental variable model. *Biometrika*, 104(1):229–236.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696.

Larry Wasserman. 2004. *All of statistics: a concise course in statistical inference*, volume 26. Springer.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4598.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923.

Zahid Zaheer, Thompson Robinson, and Amit K Mistri. 2011. Thrombolysis in acute ischaemic stroke: An update. *Therapeutic Advances in Chronic Disease*, 2(2):119–131.

Jiaming Zeng, Michael F Gensheimer, Daniel L Rubin, Susan Athey, and Ross D Shachter. 2022. Uncovering interpretable potential confounders in electronic medical records. *Nature Communications*, 13(1):1014.

Raymond Zhang, Neha Nayak Kennard, Daniel Smith, Dan McFarland, Andrew McCallum, and Katherine Keith. 2023. Causal matching with text embeddings: A case study in estimating the causal effects of peer review policies. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1284–1297.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

A Using Post-Treatment Text

Here, we describe a scenario for which we may infer valid proxies using both post-treatment and pre-treatment text. In Figure 4, $\mathbf{T}^{\text{post}}$ is post-treatment text whereas $\mathbf{T}^{\text{pre}}$ is pre-treatment text. Through simple d-separation arguments, we can see that each of (P1-P3) are fulfilled. Hence, it is still possible to use post-treatment text to generate a valid proxy as long as we use pre-treatment text to generate the other proxy. However, the pre-treatment rule is simpler to implement and validate, so we recommend it in our final method.

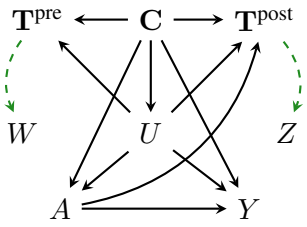


Figure 4: Using both pre-treatment and post-treatment text to generate valid proxies.

B Elaboration on Problem Restriction

Restriction (R1) is clearly present in settings for which we need to adjust for a confouding variable that is impossible or difficult to measure—for example, atrial fibrillation can go undiagnosed for years. A logical first attempt to mitigate this constraint is to train a supervised classifier on a subset of the data to create proxies for the rest of the dataset, as explored in Wood-Doughty et al. (2018). This, however, requires humans to hand-label large amounts of text data. If labeling takes place on a crowd-sourcing platform, e.g., Amazon's Mechanical Turk, crowd-sourcing costs can quickly sky-rocket and often exceed tens of thousands of dollars, even for small datasets. Furthermore, many datasets—particularly those in clinical settings—require domain expertise (e.g., trained medical doctors) which will likely increase costs significantly and limit the availability of labelers.

Cost and expertise of labelers aside, the possibility of supervised learning is further restricted by patient privacy legislation. We typically cannot transport sensitive data regarding patients' personal data to platforms such as Amazon's Mechanical Turk for labeling due to the Health Insurance Portability and Accountability Act (HIPPA)[14] in the United States. In addition, legal acts such as the General Data Protection Regulation (GDPR)[15] in the European Union and the California Consumer Privacy Act (CCPA)[16] restrict the movement and repurposing of user data across platforms. Our proposed method overcomes restriction (R1) by using zero-shot learners that do not require labeled examples.

C Fully Synthetic Data-Generating Process

The DAG representing the fully synthetic data-generating process is shown in Figure 5. We simulate U, a binary variable, as follows

$$U \sim \text{Bernoulli}(0.48)$$

Next, we simulate baseline confounders and synthetic "text" X_1, X_2, X_3, X_4 as follows:

$$\begin{aligned}
C &\sim \mathcal{N}(0,1) \\
X_1 &\sim \mathcal{N}(0,1) + 2 * U + 1.5 * C \\
X_2 &\sim \mathcal{N}(0,1) + \exp(X_1) + U + 1.5 * C \\
X_3 &\sim \mathcal{N}(0,1) + 1.3 * U + 1.5 * C \\
X_4 &\sim \mathcal{N}(0,1) + {X_3}^2 + 0.5 * {X_3}^3 \\
&\quad + U + 1.5 * C
\end{aligned}$$

Finally, the treatment and outcome variables are generated via

$$\begin{aligned}
p(A=1) &= \text{expit}(0.8 * U + C - 0.3) \\
A &\sim \text{Bernoulli}(p(A=1)) \\
Y &\sim \mathcal{N}(0,1) + 1.3 * A + 1.4 * U + 1.2 * C
\end{aligned}$$

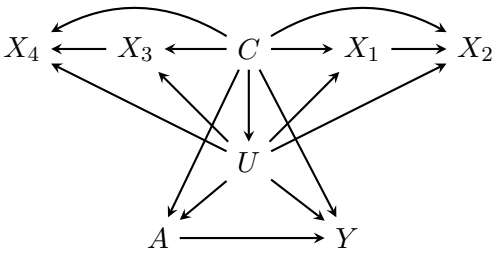


Figure 5: DAG showing the fully synthetic data-generating process.

[14] https://www.hhs.gov/hipaa/index.html

[15] https://tinyurl.com/europegdpr

[16] https://oag.ca.gov/privacy/ccpa

D Creating Semi-Synthetic Data

MIMIC-III pre-processing. In the MIMIC-III dataset (Johnson et al., 2016), data is organized into multiple tables where each patient is assigned an anonymized identifier and each hospital admission for each patient is also given a unique identifier. Our first pre-processing step was to find a clinician's note for each patient that was not a discharge summary, and we dropped patients if they did not have any clinician's notes. Next, we selected the date of the hospital admission corresponding to each clinician's note for each patient. We also selected each patient's gender and date of birth. We inferred a patient's age by subtracting the date of their hospital admission by their date of birth. Note, the precise date of birth and hospital admission date for each patient are anonymized in MIMIC-III by shifting both dates by some amount of time unknown to us. We drop patients younger than 18 and those who have invalid age values, such as negative integers.

In MIMIC-III, each diagnosis corresponds to an ICD-9 Code: a unique numerical identifier. We select the diagnoses of the following conditions from MIMIC-III: (i) `Atrial fibrillation`; ICD-9 Code: 42731, (ii) `Congestive heart failure, unspecified`; ICD-9 Code: 4280, and (iii) `Acute kidney failure, unspecified`; ICD-9 Code: 5849. If the patient has the code in MIMIC-III, we set $U = 1$, otherwise we set $U = 0$.

Semi-synthetic DGP. The DAG representing the semi-synthetic data-generating process is shown in Figure 6. The variable U represents `atrial fibrillation`, `congestive heart failure`, or `acute kidney failure`, depending on the setting. We simulate draws of the binary variable A via

$$p(A = 1) = \text{expit}(0.8 * U + 0.8 * \text{Gender} + 0.8 * (\text{Age} - 67))$$
$$A \sim \text{Bernoulli}(p(A = 1))$$

Next, we simulate draws of the continuous variable Y from:

$$Y \sim \mathcal{N}(0, 1) + 1.3 * A + 1.4 * U + 0.8 * \text{Gender} + 0.5 * \text{Age}$$

When estimating the ACE using the two-stage linear regression described in Section 5, we condition on the baseline covariates `age` and `gender`.

Keywords Matching. Here, we provide brief justifications for the keywords we chose for the matching algorithm used by P2M. As a review, we used the following keywords: "atrial" for U=Afib; "family" for U=Heart; and "liver" for U=Kidney. Empirically, we found that the appearance of the keyword "atrial" in a clinician's note is a strong indication that the patient indeed suffered from atrial fibrillation. Similarly, we found empirically that the appearance of the keyword "family" was a strong predictor for congestive heart failure. After reading some clinicians' notes from which keyword matching correctly predicted patients' congestive heart failure statuses, we suspect that this is because congestive heart failures are often serious and warrant family visits for the patient. Finally, we also found that the keyword "liver" was empirically a strong predictor of acute kidney failure. We suspect that this is because acute kidney failure often occurs in patients that are already hospitalized for another condition, such as liver failure.

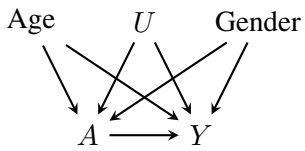


Figure 6: Causal DAG showing the semi-synthetic data-generating process.

E Additional Results from Semi-Synthetic Simulations

In Tables 3 and 4, we provide additional results from the semi-synthetic experiments.

Oracle?	Proxies	Metric	U=Afib	U=Heart	U=Kidney
Yes	–	1 - $p(U=1)$	0.721	0.736	0.796
Yes	W from Flan-T5 on $\mathbf{T}_1^{\text{pre}}$	$\gamma_{WU.\mathbf{C}}$	17.609	7.521	2.022
		Accuracy	0.766	0.782	0.795
		$p(W=1)$	0.062	0.169	0.002
		Precision	0.861	0.635	0.318
		Recall	0.191	0.408	0.002
Yes	Z from Flan-T5 on $\mathbf{T}_2^{\text{pre}}$	$\gamma_{ZU.\mathbf{C}}$	3.965	5.930	1.054
		Accuracy	0.731	0.768	0.794
		$p(Z=1)$	0.050	0.170	0.003
		Precision	0.595	0.595	0.203
		Recall	0.108	0.383	0.002
No	Z, W	Raw Agreement Rate; $p(W=Z)$	0.918	0.901	0.996

Table 3: **Additional metrics for P1M on semi-synthetic experiments**. The first column indicates whether the metric requires access to the oracle variable U variable (which is not available at true inference time). Precision, Recall, and Accuracy are calculated in reference to U.

Oracle?	Proxies	Metric	U=Afib	U=Heart	U=Kidney
Yes	–	$1-p(U=1)$	0.721	0.736	0.796
Yes	W from Flan-T5 on $\mathbf{T}_1^{\text{pre}}$	$\gamma_{WU.\mathbf{C}}$	17.609	7.521	2.022
		Accuracy	0.766	0.782	0.795
		$p(W=1)$	0.062	0.169	0.002
		Precision	0.861	0.635	0.318
		Recall	0.191	0.408	0.002
Yes	Z from Keyword Matching on $\mathbf{T}_2^{\text{pre}}$	$\gamma_{ZU.\mathbf{C}}$	2.569	12.326	1.287
		Accuracy	0.683	0.283	0.205
		$p(Z=1)$	0.261	0.981	0.999
		Precision	0.427	0.269	0.204
		Recall	0.399	0.998	0.999
No	Z and W	Raw Agreement Rate; $p(W=Z)$	0.723	0.186	0.003

Table 4: **Additional metrics for P2M on semi-synthetic experiments**. The first column indicates whether the metric requires access to the oracle variable U variable (which is not available at true inference time). Precision, Recall, and Accuracy are calculated in reference to U.