New Directions for Causal Inference With Complex Data in Health Care, Social Science, and Beyond

New directions for causal inference with complex data in health care, social science, and beyond

A DISSERTATION PRESENTED
BY
REAGAN MOZER
TO
THE DEPARTMENT OF STATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
STATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 2019

Dissertation advisor: Professor Luke Miratrix Reagan Mozer

New directions for causal inference with complex data in health care, social science, and beyond

Abstract

With the rise of the online marketplace and the digitization of data from administrative databases and electronic health records, we now have access to more comprehensive and more complex data than ever before. And while the digital revolution has created a new frontier for empirical research that promises exciting insights about human health and behavior, it has also given rise to a number of theoretical and computational complexities that call for statistical innovation. This dissertation focuses on research directions that apply classical statistical frameworks to large, messy modern datasets with a focus on settings where inferences are complicated due to the presence of unstructured text or electronic health data. Both types of data have the potential to help advance scientific understanding - for example, about the impacts of a non-randomized intervention on the health or sentiments of a population - but these data are commonly overlooked because of the theoretical or computational obstacles they impose. In particular, text is inherently high-dimensional and difficult to model, and electronic health data are often inconsistent or incomplete due to large measurement errors and non-ignorable missing values. In the chapters that follow, we first provide some background on the potential outcomes framework, describe how the literature has evolved over time, and characterize a new frontier for methodological research in causal inference with complex, modern data. We then present new statistical methods to address challenges that arise when making causal inferences with text data and in longitudinal observational studies. Throughout, we illustrate the proposed methods with case studies examining real data in health care and social science.

Contents

To Reese.

Acknowledgments

First and foremost, I want to thank my advisors: Luke Miratrix, Mark Glickman, and Fabrizia Mealli, for being a constant source of guidance, support, and inspiration. Each of them has greatly influenced not only the way I think about statistics, but also the way I approach research and problem-solving in general. I am also incredibly grateful for my peers in the department and in the greater community at Harvard for their friendship and camaraderie over the years. Among them, I want to specially acknowledge Zach Branson and Luis Campos, without whom I could not have completed this degree.

I would also like to express my gratitude to the many other advisors, faculty members, and collaborators within the Harvard community, including Michael Parzen, Neil Shephard, Jose Zubizarreta, and the members of the Miratrix CARES Lab, who have all contributed to my research and professional development in countless ways during my years as a graduate student.

Finally, I want to thank my family, including my father, Kevin Rose, my sister, Heather Boehme, and my brothers Grant and Nicholas Rose. Thank you for always encouraging me to pursue my dreams and for everything you have done to make those dreams possible. In addition to my Texas

family, I am also infinitely grateful for the love and support of my parents-in-law, Peter and Deborah Mozer, who have repeatedly gone above and beyond what was expected to help me accomplish my goals, both professionally and personally. I also cannot fail to recognize my husband, Reese, who been at my side for every step of my graduate career, helping me to overcome every trial and to celebrate every success. Last but not least, I want to thank my mentor, role-model, and dearest friend Brenda Osuna, whose enthusiasm for statistics inspired me to pursue this degree. Brenda, I owe so much of my success to you.

Twenty years ago most data was still collected manually. The cost of collecting it was proportional to the amount collected. This made the cost of collecting large amounts prohibitively expensive. The goal was to carefully design experiments so that maximal information could be obtained with the fewest possible measurements. This of course has changed.

Jerome H. Friedman

0 Introduction

Over the past decade, the digital revolution has transformed the landscape for statistical inference with advancements in technology that make it possible for researchers to collect, store, and process larger and more complex sources of data than ever before. For instance, the introduction of electronic medical records has given rise to large healthcare databases that contain detailed clinical information about treatment practices and patient profiles. Similarly, large amounts of text data are now being produced and documented online at a rate faster than social scientists can use them. But

these "big data" are only as useful as the methods we have to analyze them in meaningful ways.

This dissertation explores the changing landscape of statistics research in the digital age and describes various methodological advancements in this domain. More specifically, we consider how to conceptualize the causal effects of a particular action, behavior, or intervention on the physical world and investigate methods that allow us to precisely measure and quantify those impacts. We begin in Chapter 1 with a brief review of the foundations of causal inference, starting from the earliest approaches for experimental evaluation of causal effects. We discuss the evolution of methodology for estimating causal effects over the past several decades and describe some common challenges faced by modern-day practitioners, including topics of ongoing work. We then outline several emerging frontiers of causal inference research related to these challenges, highlighting areas for future development. In each of the subsequent chapters, we offer a thorough investigation of specific challenges for estimating causal effects through case studies using real data in health care and social science. In Chapter 2, we consider how to make precise and principled quantitative comparisons between collections of text documents in a manner that aligns with human judgment of written language. Chapter 3 then presents a novel statistical approach for estimating the effects of a medical intervention using observational data collected from electronic health records.

This thesis was originally composed as three separate, self-contained articles and includes materials presented in Mozer & Mealli (2019); Mozer et al. (2019) and Mozer & Glickman (2019). To maintain integrity of the original content, a separate introduction and discussion section is provided in each chapter. As a result, some notation and definitions may be repeated throughout the thesis.

1

Frontiers for causal inference in the big data era

1.1 Introduction

"Big data" have the potential to help answer important questions of causality in the medical and social sciences that previously would have been impossible (Grimmer, 2015). For example, genome sequencing technology now allows us to study microscopic patterns in gene expressions that may determine an individual's predisposition to certain diseases and ailments. But along with the potential to facilitate scientific discovery, these types of rich new data sources also bring new philosophical and methodological challenges for causal inference.

Recently, much methodological work has focused on developing methodology for causal inference in these new and more complicated settings; however, this work is largely spread across a variety of disciplines, including statistics, machine learning, and economics, and there have been few attempts to unify the diverse literature. Researchers interested in causal inference with complex data

are left without a clear place to turn to learn about existing research, new methods, and current standards of practice for addressing these complexities. In this paper, we describe a number of common challenges that arise in modern applications of causal inference with complex data, such as non-randomized treatment assignment, interference between units, and high-dimensional covariates, and provide a structure for thinking about these problems. We summarize the multidisciplinary literature as it relates to these issues, highlighting certain areas where methodology is dense and approaches to inference are widely applicable and identifying gaps in the literature where more work is still needed.

This chapter proceeds as follows. Section 1.2 provides a brief review of some foundational concepts in causal inference and introduces notation that will be used throughout the chapter. In Sections 1.3 and 1.4, we consider two settings where complexities in data induce methodological challenges for causal inference and where the literature is relatively well-established. Specifically, in Section 1.3, we discuss approaches to inference in randomized experiments that require adjustment for post-treatment variables, and in Section 1.4, we describe various identification strategies for observational studies where the assignment mechanism is unknown. Next, in Section 1.5, we describe an area of more recent research, namely, in settings with high-dimensional covariates. In Section 1.6, we discuss the use of sensitivity analysis as a general tool for problems in causal inference. Finally, in Section 1.7, we offer more guidance for practitioners and provide a discussion of our views on the direction of modern-day causal inference research.

1.2 Notation and background

Throughout this paper, we consider approaches to causal inference based on the potential outcomes framework, now commonly referred to as the Rubin Causal Model (RCM) (Holland, 1986). The RCM consists of two fundamental parts, and a third optional one. The first part is the use of po-

tential outcomes to define causal effects in all situations, whether from a randomized experiment or observational study. The second part is the specification of an *assignment mechanism*, a probabilistic model for how some units received treatment and other units received control, which allows us to learn from the observed data. Inferences about causal effects can be made using these two parts alone through randomization-based modes of inference. Alternatively, the third and optional part of the RCM allows Bayesian or model-based analysis of observed data to draw inferences for causal effects.

Consider a study with N units, indexed by $i = 1, \ldots, N$, where each unit receives treatment assignment Z_i, which equals 1 for units receiving the active treatment and 0 for units receiving the control version. For each subject, suppose we also observe a vector of p pre-treatment covariates $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$, which are variables that are unaffected by treatment assignment. The objective of causal inference is to learn about the effect of administering the active treatment compared to the control version on an outcome variable Y for a given unit, subset of units, or population of units. Specifically, the causal effect of the active treatment for unit i is the comparison of unit i's outcome under assignment to treatment, $Y_i(1)$, and their outcome under assignment to control, $Y_i(0)$. A causal estimand involves a comparison of $Y_i(0)$ and $Y_i(1)$ across all N units, or on a common subset of units defined by their X_i. Implicit within this notation is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980a), which requires that there is no interference between units and that there are no hidden versions of treatments. Notably, the first component of SUTVA that asserts no interference between units can be viewed as special case of the latter assumption of no hidden forms of treatment. See Appendix A for a detailed discussion of this proposition.

Under SUTVA, any causal estimand, τ, can be constructed as a function of all potential outcomes, treatment assignments, and covariates for all units: $\tau = \tau(Y(0), Y(1), X, Z)$. In general, these estimands can take different forms, but we are often interested in unit-level treatment effects defined by simple differences:

Definition 1 (Unit-level Causal Effect).

$$\tau_i = Y_i(1) - Y_i(0).$$

Also of interest are various summary measures of these unit-level effects, which may be calculated with respect to a particular finite sample of units or for an entire population. Estimands that condition on the finite set of units for which we observe covariates, treatment assignments, and outcomes are referred to as *finite sample* causal effects:

Definition 2 (Finite-sample Average Treatment Effect).

$$\tau_{fs}^{ATE} = \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - Y_i(0).$$

When the set of units for which we observe values is believed to have been randomly sampled from a larger, potentially infinite population, we may also consider causal estimands that average over this *super-population*:

Definition 3 (Super-population Average Treatment Effect).

$$\tau_{sp}^{ATE} = \mathbb{E}\left[Y_i(1) - Y_i(0)\right],$$

where the expectation is over all units in the super-population. For both finite sample and super-population based inferences, we may also be interested in estimating causal effects that directly condition on observed covariates. In general, we define the conditional average treatment effect (CATE) in both settings as follows:

Definition 4 (Finite-sample Conditional Average Treatment Effect).

$$\tau_{fs}^{CATE}(x) = \frac{1}{N(x)} \sum_{i:X_i=x} Y_i(1) - Y_i(0),$$

Definition 5 (Super-population Conditional Average Treatment Effect).

$$\tau_{sp}^{CATE}(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i = x\right].$$

These estimands allow for straightforward inference in settings with discrete covariates, where the finite sample or super-population can be partitioned into subsamples of units with equal covariate values. When such a partition exists, estimates of the CATE within each subsample will be unbiased for the effects of interest. However, when there are covariate values that are unique to either the treatment or control group it is generally impossible to construct estimators that are exactly unbiased. As a result, super-population conditional average treatment effects are not always well-defined for continuous x.

The "fundamental problem facing causal inference" (Rubin, 1978a) is that we cannot observe both potential outcomes for unit i, but rather the observed outcome $Y_i^{obs} = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. The unobserved potential outcomes can therefore be viewed as missing data (Rubin, 1974), which must be estimated in order to perform inference. To do so requires an assignment mechanism, which specifies the conditional probability of each vector of treatment assignments given all observed covariates and potential outcomes: $P(Z|X, Y(0), Y(1))$. The assignment mechanism allows us to conceptualize the underlying experiment, whether real or hypothetical, that lead to the data that was actually observed. When an assignment mechanism is "unconfounded" meaning that treatment assignment is independent of the potential outcomes given all observed covariates (i.e., $P(Z|X, Y(0), Y(1)) = P(Z|X)$), the observed outcomes can be used in a straightforward man-

ner to estimate the missing potential outcomes. Often implicit within the assumption of unconfoundedness is the assumption that all units have a positive probability of receiving treatment (i.e., $0 < P(Z = 1|X) < 1$ for all X). In this regard, an assignment mechanism that is both unconfounded and probabilistic may also referred to as "strongly ignorable", a term that is closely linked with the concept of missing at random in the missing data literature (Little & Rubin, 2014; Frumento et al., 2012).

Strongly ignorable assignment mechanisms are true by design in randomized experiments and typically must be assumed in observational studies. There is a large and growing literature on techniques for pre-processing observational data in order to make the assumption of strong ignorability more plausible, and a separate body of work focused on evaluating the sensitivity of inferential results to violations of this assumption. These methods are described in more detail later on.

1.2.1 Modes of inference

For statistical analysis of data from a strongly ignorable assignment mechanism, there are two distinct modes of randomization-based inference, one due to Neyman (1923) and the other due to Fisher (1925b). Neyman's form of inference focuses on evaluating the expectations of statistics over the distribution induced by the assignment mechanism. This approach is based on constructing an unbiased estimator of the average effect of treatment, then using the properties of statistical procedures under repeated sampling to construct a confidence interval. Fisher's approach focuses instead on evaluating the sharp null hypothesis of no unit-level causal effect (i.e., $H_0 : Y_i(0) = Y_i(1)$). Under the sharp null hypothesis, both potential outcomes are known for all units, which allows for inference through what is essentially a "proof by contradiction" (Imbens & Rubin, 2015), whereby the observed value of a test statistic is compared against its randomization distribution under the null. The same procedure can be applied to construct interval estimates, called Fisher intervals. Fisher's approach is the simplest conceptually of the two and is nonparametric, making it an attractive strat-

egy that is straightforward to implement in a wide variety of settings. However, the approach is limited by the sharp null hypothesis in that it does not immediately apply to settings where the estimand of interest is defined as an average treatment effect. Further, Fisher's approach is limited to inference on the finite sample and cannot generalize beyond the particular set of units being analyzed. Neyman's approach offers a framework that is free from these limitations, but, adversely, relies on parametric assumptions (asymptotically) that make it more difficult to implement in practice. Both approaches also suffer from a "lack of prescription" (Little & Rubin, 2000), in the sense that they are best suited for evaluating proposed procedures or hypothesized treatment effects, but typically are not useful for developing procedures that can be used for inference in complicated settings.

Alternatively, Bayesian, or model-based, inference builds a probabilistic model for the underlying data, $P(X, Y(0), Y(1))$, which can be combined with the model for the assignment mechanism to derive the posterior distribution of the desired causal effect. In particular, this approach explicitly addresses estimation of the missing potential outcomes, Y^{mis}, by taking samples of Y^{mis} from the posterior predictive distribution $P(Y^{mis}|X, Y^{obs}, Z)$. Each drawn value of Y^{mis} can be used to construct a complete data set (Y^{obs}, Y^{mis}), from which the causal effect of interest can be calculated directly. Thus, the basis of the Bayesian approach is that missing potential outcomes are multiply imputed in order to generate a posterior distribution for the causal effects. This approach is direct, intuitive, and can be used to conduct inference for a wide variety of causal estimands, including those defined for the super-population, that randomization-based approaches cannot accommodate. However, this form of inference requires careful specification of a model for the data as well as specification of prior distributions governing the unknown parameters. In practice, specifying a model and constructing efficient priors may be difficult without some scientific information or prior knowledge about the underlying data. Further, inferences made using the Bayesian approach, especially in non-randomized settings, may be sensitive to minor changes in the procedures and specifications used.

One important issue to note is how the choice of inferential approach is linked to the estimand of interest, which is generally guided by the population of units to which we want to generalize. In particular, randomization-based approaches are inherently geared toward inferences about the finite sample, since these approaches condition explicitly on the observed sample of N units. This differs from super-population inference in which some aspects of the observed data are assumed to be independent and identically distributed (iid) draws from some distribution (Ding et al., 2016; Pashley & Miratrix, 2017).

In the subsections below, we consider a number of complications that commonly arise in modern applications of causal inference. For each complication, we consider which types of estimands, defined on either the finite population or super-population, may be appropriate to define the causal effects of interest in that setting. We then summarize a number of recent and popular approaches to addressing each complication, highlighting how the choice of estimand may affect the utility of different procedures and identifying potential areas for future work. Throughout this paper, we place a particular emphasis on recent work from the Bayesian perspective, because many complications, modern and otherwise, that arise in real world studies of causal effects can often be addressed more flexibly using the model-based approach than with randomization-based methods.

1.3 Adjusting for intermediate variables

Many classical approaches to causal inference have focused on addressing issues of confounding through the *prospective* process of experimental design (Fisher, 1925a, 1937; Kempthorne, 1952; Cochran & Cox, 1957; Cox, 1958). The theme of these approaches is that randomized experiments can be designed and conducted in a manner that ensures objective and valid results. By explicitly addressing potential sources of bias through randomization, carefully designed experiments should, in theory, facilitate straightforward analysis. However, not all issues can be addressed by design. In

fact, there are many settings where a potential confounder becomes apparent only after randomization has occurred. Such a potential confounder is commonly referred to as an *intermediate variable*, since it arises after treatment and must be addressed before estimating the treatment effect. An active area of research therefore considers the problem of causal inference in the presence of post-treatment complications (Mealli & Pacini, 2013a; Feller et al., 2017; Forastiere et al., 2018).

In this section, we discuss strategies for addressing intermediate variables in applications of causal inference. This topic encompasses a range of settings such as treatment noncompliance, outcomes that are censored due to death or dropout, surrogate endpoints, and causal mediation analysis.

1.3.1 The principal stratification framework

As a running example, we first consider randomized experiments with noncompliance, where for some units, the treatment *assigned*, denoted by Z, may be different from the treatment actually *received*, denoted by W. For example, some of the units assigned to take the active treatment may take the control treatment instead, and some assigned to take the control may manage to take the active treatment. Because W_i is observed post-treatment, each unit i has two potential versions: $W_i(0)$ and $W_i(1)$, from which we observe only one, denoted $W_i^{obs} = W_i(Z_i)$.

The standard inferential procedure for randomized studies with noncompliance is called Intention to Treat (ITT) analysis. This method ignores observed compliance information and compares those assigned to treatment to those assigned to control. This procedure gives a valid estimate of the effect of treatment assignment on outcome, at least from the randomization- based perspective."As-treated" and "per-protocol" are two other ways that data of this type could be analyzed. An as-treated analysis compares those who received treatment with those who received control, ignoring treatment assignment. Per-protocol analysis compares people who were assigned to and received treatment with those who were assigned to and received control, i.e., compares those who appeared to comply with the protocol.

A *basic* principal stratification with respect to W is a partition of units based on their joint potential outcomes principal strata: $G_i = (W_i(0), W_i(1))$. More generally, a principal stratification with respect to an intermediate variable is a partition of units, whose sets are unions of sets in the basic principal stratification. In the simplest case of all-or-none compliance, W is a binary variable that equals 1 if the active treatment is received and equals 0 if the control treatment is received. In this case, there are at most four principal strata: compliers, defiers, never-takers, and always-takers (Angrist et al., 1996). Never-takers, denoted $G_i \equiv n = (0, 0)$, are units who would not take the treatment regardless of the assignment; compliers, denoted $G_i \equiv c = (0, 1)$, are units who would take the treatment if assigned and would not if not assigned; defiers are units who would take the opposite of the assigned treatment, denoted $G_i \equiv d = (1, 0)$; and always-takers, denoted $G_i \equiv a = (1, 1)$, are units who would not take the treatment regardless of the assignment. The key property of principal strata is that they are, by definition, not affected by the treatment assignment, and thus can be regarded as a pre-treatment variable. Therefore, comparisons of $Y(1)$ and $Y(0)$ within a principal stratum are well-defined causal effects in the sense of Rubin (1978a) since they compare quantities defined on the same set of units. These effects are called principal causal effects (PCEs), defined generally as

$$\text{PCE}(w_0, w_1) \equiv E[Y_i(1) - Y_i(0) \mid G_i = (w_0, w_1)]. \tag{1.3.1}$$

One example of an interesting PCE is the complier average causal effect (CACE), also commonly referred to as the local average treatment effect (LATE) (Imbens & Angrist, 1994). The CACE is also a special case of the causal estimand that is targeted by the Instrumental Variables (IV) approach used in econometrics (Imbens & Angrist, 1994). This estimand conditions explicitly on W and, under a set of conditions known as "exclusion restrictions", is interpreted as the causal effect of the treatment actually received. As such, PCEs are always *local* effects, in the sense that they quantify the

average effects of treatment for a latent subset of units within a given sample.

To make inferences on PCEs some structural assumptions are usually invoked. Specifically, the standard unconfoundedness assumption is modified to incorporate the intermediate variable as

Assumption 1 (Unconfoundedness). $Z \perp Y(1), Y(0), W(1), W(0)|X$.

This assumption has important implications. First, it implies conditional independence between the treatment assignment and stratum membership given the covariates (i.e., $Z_i \perp G_i|X_i$). Stated differently, unconfoundedness implies that the principal strata have the same distribution in both treatment arms, within cells defined by the observed covariates. Second, 1 asserts that the potential outcomes are independent of treatment assignment within principal strata (i.e., $Z_i \perp \{Y_i(1), Y_i(0)\}|\{W_i(1), W_i(0), X_i\}$). This condition is known as *latent unconfoundedness*, and it is the condition that is exploited, for example, in IV estimation. While in general, despite unconfoundedness, we cannot contrast potential outcomes conditional on the observed value W_i^{obs}, latent unconfoundedness allows us to contrast potential outcomes conditional on a principal stratum.

However, individual principal stratum memberships, G_i, are often missing or only partially observed (e.g., units with $(Z_i = 1, W_i^{obs} = 1)$ can be either compliers or always-takers, and units with $(Z_i = 0, W_i^{obs} = 0)$ can be either compliers or never-takers). As a result, PCEs are generally not identifiable without additional structural assumptions. Alternatively, one can derive nonparametric bounds for these effects under only unconfoundedness (e.g. Grilli & Mealli, 2008; Zhang & Rubin, 2003a; Mealli & Pacini, 2013b). Under the former approach, two assumptions commonly enforced for inference on PCEs are as follows:

Assumption 2 (Monotonicity). $W_i(1) \geq W_i(0)$, *for all i.*

Assumption 3 (Stochastic Exclusion Restriction for non-compliers). *For* $g = a, n \Pr(Y_i(1) \mid G_i = g) = \Pr(Y_i(0) \mid G_i = g)$.

In general, the monotonicity assumption rules out defiers, and the exclusion restriction (ER) rules out the effect of treatment assignment for never-takers and always-takers. Note that Assumption 3 is a strong and substantive assumption asserting that, for certain units, treatment assignment does not affect the outcome except for through the treatment actually received. While this may be plausible in certain randomized settings, it is not implied by design. Under assumptions 1-3, PCEs for compliers, always-takers and never-takers are nonparametrically identifiable in the sense that moment-based estimators exist and are consistent (Angrist et al., 1996; Zhang & Rubin, 2003b).

We advocate the use of Bayesian model-based inference for PCEs due to its flexibility. In particular, from a Bayesian perspective, PCEs are always identified and identifying assumptions such as 1-3 are therefore not strictly necessary (Imbens & Rubin, 1997b). This is because proper prior distributions on the model parameters will always yield proper posterior distributions of the causal estimands. Thus, any identification issues will be reflected through regions of flatness in the posterior distribution (Imbens & Rubin, 1997b; Feller et al., 2016). In settings where the estimands of interest are weakly identified, inference for causal inferences can be sharpened by additional assumptions such as 2 and 3. The Bayesian framework is also useful in this setting since it provides a natural structure for sensitivity analysis, which allows us to assess how the posterior distribution for causal estimands changes when we strengthen or relax certain assumptions.

However, it is important to note that there may be cases when identifying assumptions are implausible. For example, consider a single-blinded randomized trial of a medical intervention where, given the knowledge that a patient has been assigned to control, a clinician may provide other forms of care to reduce the likelihood of a bad outcome (e.g., death). This clearly violates the exclusion restriction. In this case, inferences for PCEs can be sharpened by secondary outcomes and covariates - for instance, the exclusion restriction may hold between the treatment assignment and a secondary outcome (e.g., side effects; Mealli & Pacini, 2013b; Mercatanti et al., 2014. Similarly, when the assumption of monotonicity is infeasible, we may be unable to uniquely identify strata proportions.

The consequence of this is not necessarily that we end up with biased estimates, but rather that the parameters we care about might only be partially identified, resulting in increased variability in the posterior distribution of the estimated causal effects.

To conduct Bayesian inference with intermediate variables given a global parameter θ, one needs to specify two models for the data: one for the conditional distribution of potential outcomes given principal strata (and covariates), $\Pr(Y_i(0), Y_i(1) \mid G_i, X_i, \theta)$ and another for the distribution of principal strata conditional on the covariates $\Pr(G_i \mid X_i, \theta)$. Bayesian inference also requires specification of the prior distribution $p(\theta)$. Here, the posterior distribution of θ is generally not tractable. However, posterior inference can proceed by straightforward application of Markov Chain Monte Carlo (MCMC) techniques, such as the Gibbs sampler (Geman & Geman, 1984; Gelman et al., 2014) combined with a data augmentation step where the missing potential intermediate variables, and therefore principal stratum membership, are drawn from their posterior predictive distribution (Imbens & Rubin, 1997b; Tanner & Wong, 1987) These random draws provide posterior inference for the causal estimands. Here, as before, population PCEs do not depend on the association between $Y_i(0)$ and $Y_i(1)$ as long as the association parameters are *a priori* independent of the other parameters, but finite sample PCEs generally do depend on the association (Jin & Rubin, 2008a).

1.3.2 Extensions of principal stratification

More complex examples of noncompliance exist, even when active treatment is not available to those not assigned to take it. For example, compliance can be "partial" in the sense that only a fraction of an assigned dose of pills is taken. We might also encounter "extended noncompliance," where even those assigned "control" may not take their assigned dose (Jin & Rubin, 2008b). These issues often co-occur; for instance, in placebo-controlled randomized clinical trials, units assigned to control may experience placebo effects that vary in magnitude depending on underlying psychological characteristics (Mozer et al., 2017). Below, we discuss several extensions of principal stratification

that attempt to address complications such as these.

Non-ignorable missing data and censoring due to "death". In both experimental and observational studies, outcomes may be missing for a subset of units because of non-response or loss to follow-up. Alternatively, outcomes may be censored or undefined for some units due to death. The principal stratification framework can be applied in each of these scenarios, where principal strata are defined by the non-response behavior under all possible treatment levels Mattei et al. (2014). For instance, Rubin (2006a) characterized a class of intervention studies where patients in the experiment may die after treatment, but before the primary outcome is observed; for these patients, the outcome is "censored" due to death. In this setting, the censoring mechanism can be represented as an intermediate variable defined by an indicator for survival at the end of the study period. Clearly, the causal effect of treatment is well-defined only for "always-survivors" — the principal stratum of units who would survive irrespective of the treatment ($G_i = (1, 1)$), and the target causal estimand is the always-survivor average causal effect (SACE). Examples of censoring due to "death" include evaluating the causal effects of a Breast Self-Examination (BSE) teaching course on quality of BSE execution, where W is the execution (Mealli et al., 2004; Mattei & Mealli, 2007), or evaluating the effects of a HIV vaccine on the viral load where W is the HIV infection status (Gilbert et al., 2003). Principal stratification can also be applied to address censoring in other, non-clinical domains, for example, to evaluate the causal effects of job training programs on wages where W is employment status (Zhang et al., 2008, 2009; Frumento et al., 2012), or to evaluate the effects of an educational intervention on final test scores where W is the graduation status (Zhang & Rubin, 2003a).

Multiple intermediate variables. In many applications, there can be more than one intermediate variable. For example, in evaluating the effects of Jobs Corp—a large randomized job training program—on wages, the data are characterized by non-compliance, nonignorable missing outcomes and censoring due to unemployment (Frumento et al., 2012). The number of possible principal strata increases exponentially with the number of intermediate variables. The estimation strategy is

usually through imposing more structural assumptions such as exclusion restrictions and versions of monotonicity to limit the number of principal strata (Mattei & Mealli, 2007). Forastiere et al. (2016) tackle the two problems of noncompliance and interference (e.g., spillover effects) in clustered encouragement designs, designed with the purpose of increasing the uptake of the treatment of interest when the treatment cannot be enforced because of ethical or practical constrains.

Non-binary intermediate variables. The majority of research on principal stratification focuses on binary intermediate variables, but non-binary intermediate variables are common. For example, partial compliance often arises in randomized trials (e.g. Efron & Feldman, 1991), and continuous mediators are often present in mediation studies. Conceptually it is straightforward to extend principal stratification to continuous intermediate variables: one can define a principal stratum for all possible values of (w_0, w_1) and the causal estimands $\text{PCE}(w_0, w_1)$ become a *surface*. However, this leads to an infinite number of possible principal strata, each of which is, in theory, an empty set, introducing substantial complications to inference and interpretation. The standard method of dichotomizing the continuous intermediate variable is subject to information loss and arbitrary choice of cutoff points. In the context of partial compliance in randomized trials, fully parametric Jin & Rubin (2008a); Zigler et al. (2012) or semi-parametric Bartolucci & Grilli (2011); Schwartz et al. (2011) frequentist and Bayesian models have been proposed. For instance, using the Dirichlet Process mixture model (DPMM) for the principal strata, one can accommodate complex data features such as outliers, skewness and multi-modality. More importantly, the clustering structure of the Dirichlet Process leads to data-driven, *a-posterori* coarsening of the principal strata, which offers more natural interpretation of the results. Similarly, recent work by Comment et al. (2019) develops a Bayesian model-based approach for inference in longitudinal settings where principal strata are defined over a potentially arbitrary continuous time interval.

1.3.3 Causal mediation analyses

Conceptualizing the mediating role of an intermediate variable in the causal pathways between treatment and outcome is a difficult task that has received increasing attention in recent years. The role of principal stratification and formal mediation analysis when dealing with issues concerning causal mechanisms has led to heated debates among the causal community. Mediation analysis and principal stratification analysis generally focus on different causal estimands, answer different questions and involve different sets of assumptions; ultimately, these approaches utilize the information provided by the data in a substantially different ways. VanderWeele (2008) shows the relationships between mediation analysis and principal stratification from a theoretical point of view; Mealli & Mattei (2012) further investigate the relationship between these approaches.

If one is seeking information on causal mechanisms, principal stratification analysis starts by looking at the effects of treatment on outcome that are associative and dissociative with the effects of treatment on the mediating variable (Gallop et al., 2009; Elliott et al., 2010). Associative PCEs are causal effects within principal strata where the mediating variable is affected by treatment ($w_0 \neq w_1$), while dissociative PCEs are causal effects within principal strata where the mediating variable is unaffected by treatment ($w_0 = w_1 = w$). Dissociative PCEs naturally provide information on the existence of a unchanneled causal effect of the treatment on the primary outcome for the sub-population of units for whom treatment does not affect the intermediate variable (Rubin, 2004). If dissociative PCEs are all zero, then there is no evidence on the unchanneled (direct) effect of the treatment after controlling for the mediator (Mattei & Mealli, 2011). In addition if associative effects are large in magnitude relative to dissociative effects and unchanneled effects are homogeneous across principal strata, then it is reasonable to believe that the mediator channels a part of the treatment effect on the outcome. On another hand, associative effects that are similar in magnitude relative to the dissociative effects suggest that the treatment affects the outcome mainly through

other causal pathways rather than through the mediator of interest.

Mediation analysis focuses on disentangling direct and indirect effects, which are generally defined at the individual level and averaged over the whole population. Causal estimands in mediation analysis are typically a function of potential outcome outcomes that are indexed on both the treatment and the intermediate (mediating) variable $Y_i(Z_i, W_i)$. Some of these potential outcome are therefore regarded as *a priori* counterfactuals, primitives that are explicitly avoided in the principal stratification framework. For example, a natural direct effect is defined as $NDE(0) = E[Y_i(Z_i = 1, W_i = W_i(0)) - Y_i(Z_i = 0, W_i = W_i(0))]$, that is, it is the effect of treatment if, possibly contrary to fact, the mediator under treatment were set to the value it would take on under control. Obviously, the quantity $Y_i(Z_i = 1, W_i = W_i(0))$ is *a priori* counterfactual for subjects such that $W_i(1) \neq W_i(0)$. Baccini et al. (2017) use the Bayesian approach for inference, and estimate both associative and dissociative principal strata effects arising in principal stratification, as well as natural effects from mediation analysis. Other applications of this approach include evaluating to what extent the causal effect of birth control on thrombosis in women is mediated by the effect of being on the pill on pregnancy (Pearl, 2001), and assessing to what extent the causal effect of installing scrubbers at coal-fired power plants on the ambient concentration of a certain pollutant is mediated through the reduced emission of other pollutants (Kim et al., 2019).

Surrogate endpoints A special case of mediation is surrogate endpoint analysis. Often in randomized experiments, especially clinical trials, the primary outcome may be rare, late-occurring or costly to obtain. Instead, researchers may rely on easier-to-measure variables known to have a strong association with the true endpoint to reliably extract information about the effect of treatment. Such variables are called "surrogate" or "biomarker" endpoints. For example, in a randomized trial to evaluate the efficacy of a HIV vaccine, the primary outcome is HIV infection, which may take place after a considerable period of time. Count of CD_4 cells in blood has long been used as a surrogate endpoint for HIV infection (Hernan et al., 2000; Gran et al., 2016; Cain et al., 2011). There have

been different definitions of surrogate endpoints (e.g. Prentice, 1989); Frangakis & Rubin (2002) proposed the "principal surrogate" definition with a causal interpretation, using the concepts of associative and dissociative principal causal effects. Under this formulation, W is a principal surrogate if the associative effect is zero for all w. The quality of W as a surrogate is determined by its associative effects relative to its dissociative effects, referred to as "causal effect predictiveness" by Gilbert & Hudgens (2008). Intuitively, difference in the potential values of a good surrogate under different treatment conditions would strongly predict difference in the potential outcomes of the final endpoint. Therefore, a surrogate with close-to-zero associative effects and high dissociative effects would be regarded as a good surrogate. Gilbert & Hudgens (2008) extended the framework to continuous-valued surrogates and advocated using baseline covariates to improve estimation. Bayesian estimation and modeling strategies have been developed in Li et al. (2011); Zigler & Belin (2012); Conlon et al. (2014). See also Mealli & Mattei (2012).

1.3.4 Summary and open questions

In general, there is no standard approach for causal inference in broken randomized experiments. There are, however, some interesting cases where features of the assignment mechanism allow us to draw credible causal inferences on certain *local* effects. In contrast, inferences on population-level causal effects are typically only possible under strong assumptions. It is difficult to discuss these identification issues in general, since assumptions that are plausible in some settings may be infeasible in other contexts. However, as we will see in later sections, the analysis of randomized experiments with post-treatment complications also serves as the gold standard for the analysis of observational studies by suggesting the assumptions required to identify and estimate causal effects.

As we have shown, the literature on causal inference with complications due to intermediate variables, and on principal stratification in particular, is rapidly expanding, and Bayesian inference appears the natural mode of inference for this setting. However, many questions remain in this area.

Model specification and model diagnostics play a crucial role in principal stratification analysis; so far posterior predictive p-values have been used (e.g. Mattei & Mealli, 2007; Mattei et al., 2013), as well as Bayesian goodness-of-fit methods such as Bayes factors and marginal likelihood (e.g. Chib, 1995), but further research is required. Another open issue is how to tackle the curse of dimensionality in the presence of multiple intermediate variables; specifically, whether a data-driven coarsening is to be preferred to a subject-matter *a priori* coarsening through structural assumptions (see also, Daniel et al. (2015)).

1.4 Inference in observational studies

For credible and precise causal inference, it is desirable to compare treated and control units that are as similar as possible on background characteristics such that any observed differences can be reasonably attributed to the effects of treatment. Randomized experiments are advantageous in this regard because the randomization guarantees that treatment and control groups will be balanced, in expectation, on all background characteristics that affect the outcome of interest. In observational studies, on the other hand, the researcher has no control over the assignment of treatment to units. This lack of control makes such studies inherently more controversial than randomized experiments, since units may select their own treatments, or their environments may impose treatments upon them, in a manner that leads to systematic differences between treatment and control groups. For example, in an observational study evaluating the effect of smoking on longevity, individuals who choose to smoke may be more likely than non-smokers to engage in other risky behaviors that decrease their life expectancy, so comparisons of outcomes between these groups may reflect these differences rather than effects of smoking itself. To obtain unbiased estimates of the causal effects of interest, it is therefore important to control for these naturally occurring differences between treatment and control groups.

Methods for drawing valid inferences about causal effects in the face of non-randomized treatment assignment, through both the design and analysis phases of an observational study, can generally be categorized according to one of two approaches. The first class of approaches rely on the assumption that the unknown assignment mechanism is unconfounded given observed covariates:

Assumption 4 (Selection on observables).

$$(Y(0), Y(1)) \perp\!\!\!\perp Z|X$$

That is, it is assumed that all covariates that affect both the treatment assignment and the potential outcomes are observed in the study, such that it is plausible that the unknown assignment mechanism is unconfounded given these covariates. Another important assumption typically required here is that of positivity, also commonly referred to as covariate overlap, which asserts that all units have a non-zero probability of assignment to each treatment condition:

Assumption 5 (Positivity).

$$0 < P(Z = 1|X) < 1$$

Under assumptions 4-5, the assignment mechanism for an observational study can be interpreted as if, within populations of units with the same value for the covariates, a randomized experiment was conducted, although the assignment probabilities for the units are unknown. Thus, after adjusting for observed covariates, causal estimands such as the ATE can be estimated using methods for analyzing randomized data.

For example, one might use covariate data to create matched samples of pairs of treated and control units, then analyze the outcome data for these samples as if they had arisen from a randomized experiment with a matched pairs design. Alternatively, one might identify subgroups of units with similar covariate distributions, then compare outcomes between treatment and control groups

within each subgroup and average across subgroups to estimate the average treatment effect. There are a number of strategies for assessing the degree of balance in the covariate distributions between treatment and control groups (see Imbens & Rubin, 2015, chapter 14), and identifying balanced samples of treated and control units is the subject of a large and rapidly expanding body of work on methods for matching and subclassification (e.g., Rosenbaum, 2012; Zubizarreta et al., 2014a).

Approaches based on the unconfoundedness assumption are attractive in the sense that estimation and inference are relatively straightforward after controlling for observed confounders; however, these assumptions may not be plausible in all settings, especially in studies with few observed covariates. The second class of approaches offer alternatives to these strong identifying assumptions and still allow for unbiased estimation of some causal estimands. To allow for violations of unconfoundedness, we may rely on the presence of additional information in the study and consider alternative assumptions about the data-generating process. For instance, the popular IV approach relies on the presence of additional information made available through an instrument, which is used to uncover the causal effect of treatment. Other methods for causal inference with observational data that make different assumptions have been the subject of much recent work (e.g., regression discontinuity designs; Lee & Lemieuxa, 2010; Li et al., 2015; Mattei & Mealli, 2016).

1.4.1 Inference under the assumption of unconfoundedness

When the assignment mechanism of an observational study can be assumed to be unconfounded after conditioning on observed covariates, valid causal inferences can be obtained by adjusting estimated treatment effects for the covariates. Recent methods that directly and flexibly balance covariate distributions include Diamond & Sekhon (2013); Zubizarreta (2012, 2015). These last methods are implemented in the packages "designmatch" and "sbw" in R (Zubizarreta & Kilcioglu, 2016). Other recent work has focused on extending methods from machine learning to this setting, either to flexibly estimate the outcome surface or propensity scores, or to combine them using a "doubly-

robust estimator" (Robins et al., 1995; Hirano & Imbens, 2001). Such estimators attempt to adjust directly for both the association between the covariates and the treatment assignment and the association between the covariates and the potential outcomes (Athey et al., 2017). See Van der Laan & Rose (2011); Athey & Imbens (2015) for an introduction to these methods and discussion of their utility for causal inference in both observational studies and randomized experiments.

Other methods adjust for an estimate of the propensity score, $e(X_i) = p(W_i = 1|X_i)$. The propensity score is typically estimated using a logistic regression model with the treatment indicator Z regressed on covariates X. However, recent advancements in the machine learning literature have given rise to a number of new approaches for estimating the propensity score using techniques such as random forests, generalized boosting methods, and variable selection tools Austin et al., 2007; Setoguchi et al., 2008; Lee et al., 2010. Adjustment methods based on the propensity score include adding the estimated propensity score as a covariate in a regression model, matching or subclassification of treated units and controls based on the estimated propensity score, or weighting cases by estimates of the propensity score. For example, using matching, treated-control pairs might be formed with similar values of the propensity score (and perhaps the covariates). Alternatively, through subclassification, subjects might be cross-classified by treatment group and by quintiles of the estimated propensity score; separate estimates of treatment effects are then made within each propensity score subclass and then combined across the five subclasses.

1.4.2 Inference with weaker identification strategies

Obviously, the assumption of an unconfounded treatment assignment given the observed covariates that is required for direct inference about causal effects in observational studies is a strong one and may not be plausible in some settings. Other approaches for causal inference from observational data offer alternatives to the standard unconfoundedness assumption and still allow for unbiased estimation of some causal estimands. For example, the Instrumental Variables (IV) approach relies

on the presence of additional information in the data made available through an "instrument", which is a variable known a priori to have a causal effect on the treatment assignment, but no effect on the outcome of interest. This instrument is used to uncover the causal effect of treatment. For example, continuing the earlier example on studying the effect of smoking on longevity, one might use the tax rate on cigarettes as an instrument, since this variable is likely to be related to longevity only through its influence on the development of smoking habits. These methods have historically been widely applied in practice (Angrist et al., 1996; Imbens & Rubin, 1997a; Frumento et al., 2012; Mattei et al., 2013) and are described in detail in Angrist & Krueger (2001).

There has been much work recently on developing other alternatives to the unconfoundedness assumption, such as approaches based on regression discontinuity (e.g., Lee & Lemieux, 2010; Li et al., 2015; Mattei & Mealli, 2016. However, weaker identification strategies often come with an added cost. For example, in studies that use an RDD to estimate the local average treatment effect of an intervention, there are often too few units close to the discontinuity, so we must rely on data from units that are further away. This type of extrapolation can make inferences imprecise and may lead to poor generalizability of results.

1.5 Causal inference with high-dimensional covariates

While statisticians historically have grappled with how to make precise inferences in low dimensions with small sample sizes, today a more common challenge concerns how to perform variable selection with high-dimensional data. A practical problem for researchers trying to make causal inferences with "big data" in both randomized and non-randomized settings is the decision of what variables should be included. This problem manifests in both the design and analysis phases of causal inference. For example, how should one perform randomization in a single-cell RNA-sequencing experiment in order to balance treatment and control groups across millions of gene expressions (Bacher

& Kendziorski, 2016)?

Intuitively, a large number of covariates in a randomized experiment or observational study should translate to more precise estimates of the causal effects of interest, because, in principle, more covariates should contain more information relevant for imputing the missing potential outcomes. Thus, by conditioning on a large number of observed covariates in a randomized experiment, one might expect to increase efficiency of estimation, regardless of the inferential approach. Similarly, in observational studies, the unconfoundedness assumption is often more plausible if a large number of pre-treatment variables are included in the analysis. Despite the intuitive appeal, in both randomized experiments and observational studies the task of adjusting for high-dimensional covariates often creates computational challenges that greatly decrease the efficiency of standard approaches to treatment effect estimation. In practice, the number of potential complications that arise for drawing valid causal inferences typically grows with the number of covariates.

1.5.1 Adjustment with high-dimensional covariates

Randomized experiments. Covariates are commonly used to increase accuracy of treatment effect estimates in randomized experiments. However, when covariates are high-dimensional, many of the available features may be irrelevant for predicting the outcome. This setting is the focus of a class of "approximately sparse" regression models. The problem of how to select a parsimonious model from many possible predictors is further complicated when there are more observed covariates than observations (i.e., "$N < p$ problems"). Regularized regression models such as the LASSO (Tibshirani, 1996) can be used to select the optimal number of covariates and produce valid causal estimates in such a setting. For instance, Belloni et al. (2014a,b) propose a double-selection procedure that first uses the LASSO to select covariates that are correlated with the outcome, and then again to select covariates that are correlated with the treatment. The outcome can then be regressed on the union of the two sets of covariates, greatly improving the properties of the estimators for the average

treatment effect (Belloni et al., 2014a). Another approach estimates a regression model where the treatment indicator is interacted with covariates, and uses LASSO as a variable selection algorithm for determining which covariates are most important (Imai et al., 2013). Covariate selection via the LASSO can also be implemented within the Bayesian framework to facilitate causal inference (e.g., Ratkovic & Tingley, 2017; Shortreed & Ertefaie, 2017). The methods are widely used in practice, although best practices for the choice of appropriate regularized estimators and methods for data-driven selection of regularization parameters are topics of ongoing research (e.g., Abadie & Kasy, 2017).

Observational studies. Another important type of complication that arises in high-dimensional studies regards how to define covariate balance over high-dimensional covariates. D'Amour et al. (2017) discusses how evaluating, and even conceptualizing covariate balance in high-dimensions is difficult, since "everything is far away in high dimensions". This may pose a challenge in studies even when $N > p$. There is a large and growing literature on the specific techniques that can be used to balance covariate distributions between treatment and control groups, but most methods first attempt to define a sufficient reduction, or lower-dimensional metric, that contains all of the information within the covariates that is relevant for determining treatment assignment. This is usually the propensity score but may take other forms. Another approach estimates weights that directly balance covariates or functions of the covariates between treatment and control groups, so that once the data has been re-weighted, it mimics more closely a randomized experiment (e.g., Zubizarreta, 2015; Imai & Ratkovic, 2014). See also Wang et al. (2015) and Athey et al. (2017).

Alternatively, Bayesian methods are intuitive and flexible for model building with many predictors. Spertus & Normand (2018) propose an approach for estimating Bayesian propensity scores, which allow model building while propagating the necessary uncertainty in the treatment model to allow for valid causal inference. Specifically, the authors advocate using a regularized Bayesian logistic model or BART to model the treatment assignment (Spertus & Normand, 2018). This allows the

use of horseshoe priors, which shrink noisy coefficients toward zero while avoiding over-shrinking of true confounders. In general, this approach can be applied using any proper prior distribution to control model uncertainty in the context of covariate selection. Wang et al. (2015) proposes a related approach called Bayesian Adjustment for Confounding (BAC), which allows model-building without a-priori certainty about which among a large set of covariates are important. This is done by specifying a functional form for the association between potential confounders with both the exposure (treatment assignment) and the outcome.

1.5.2 Heterogeneous treatment effects

While many classical approaches to causal inference have focused on treatment effects that average across a population, in practice it is clear that different units may have unique responses to treatment. Heterogeneous treatment effects (HTE) refer to the phenomenon where the treatment effect for an individual unit, defined as a comparison of potential outcomes under treatment and control for that unit, may differ systematically from the average treatment effect. Suppose that for each unit i with potential outcomes $Y_i(0)$ and $Y_i(1)$, we observe some pre-treatment covariates X_i that influence the magnitude of the effect of treatment. For example, when evaluating the effects of a medical intervention on health care utilization post-treatment, it may be of interest to estimate heterogeneous treatment effects as a function of prior utilization. The concern in this domain is that searching over many covariates and subsets of the covariate space may lead to "false discoveries," that is, spurious findings of treatment effect differences.

There is a growing literature on methodology that attempts to identify systematic variation in response to treatment that is not due to measurement error or simple random variation. One approach to estimating heterogeneous treatment effects is to use regression trees (Imai & Strauss, 2011) or Bayesian Additive Regression Trees (BART, Chipman et al., 2010; Green & Kern, 2012). In these methods, the sample is repeatedly split into subgroups in order to minimize within-group variation

in outcomes. Aggregates of outcomes within these groups can then be used to estimate the CATE described in Section 1.2. For instance, Athey & Imbens (2015) develop a method based on regression trees that allows researchers to partition the covariate space into subgroups based on treatment effect heterogeneity. The output of the method is a treatment effect and a confidence interval for each subgroup. This approach can also be implemented using random forests (Wager & Athey, 2018).

While these methods allow for flexible modeling with minimal assumptions, it is generally difficult to know if a method will perform well for a particular data set. Thus, rather than use only one method, Grimmer et al. (2017) advocate for *ensemble* methods, which use weighted averages of estimates from individual models, for estimating heterogeneous effects. This idea has been reiterated by Ding et al. (2016), who suggest that, depending on the extent to which covariates are believed to be predictive of treatment effect variation, hybrid methods may be the most powerful approach for testing heterogeneity.

A related but distinct area of methodological work considers how to estimate the causal effects of heterogeneous, or high-dimensional, treatments. Researchers are increasingly making use of experimental designs that include a large number of treatment conditions, in order to examine how (potentially subtle) differences in treatment content or treatment administration affect outcomes across a population of units (Hainmueller et al., 2014). For instance, Bavli & Mozer (2019) recently conducted an experiment to evaluate the causal effects of presenting mock jurors with prior-award information on subsequent award determinations using a factorial design with 22 treatment conditions. Similar factorial experiments have been implemented to evaluate...

1.6 Sensitivity analyses

The RCM provides a clear mathematical framework for defining causal estimands and specifies the precise set of assumptions that are necessary to make valid and unbiased causal inferences in both

randomized and non-randomized studies. However, causal conclusions are generally more defensible if robust to deviations from these assumptions. To evaluate the robustness of causal estimates to violations of these typically unverifiable assumptions, sensitivity analyses are often conducted that attempt to precisely bound the magnitude of the causal effects as a function of the degree to which the assumptions are violated (Ding & VanderWeele, 2016).

Sensitivity analysis is different from model testing because the identifying assumption is intrinsically untestable since the observed data are uninformative about the distribution of $Y(0)$ for treated units and $Y(1)$ for control units. In practice, we generally do not know that we have available a set of covariates that is adequate to support the claim of an unconfounded assignment mechanism. Thus, even if we have successfully adjusted for all covariates at hand, we cannot be sure that there is not some hidden unmeasured covariate that may bias the results. A sensitivity analysis posits the existence of such a covariate and how it relates to both treatment assignment and outcome, and it examines how the results change.

For instance, Rosenbaum & Rubin (1983a) propose a strategy for assessing the robustness of the estimated causal effects with respect to assumptions about an unobserved binary covariate that is associated with both the treatment and the outcome of interest. The central assumption of this approach is that the assignment to treatment is not unconfounded given the set of observable variables X (i.e., $P(Z|Y(0), Y(1), X) \neq P(Z|X)$), but unconfoundedness does hold given X and an unobserved binary covariate, denoted by U (i.e., $P(Z|Y(0), Y(1), X, U) = P(Z|X, U)$). Given these assumptions, Rosenbaum & Rubin (1983a) propose a parametric approach that requires specifying the distribution of U and identifying parameters that characterize association between U and Z, $Y(1)$, and $Y(0)$ given observed covariates X. The full likelihood can then be derived and maximized, holding the sensitivity parameters as fixed and known values. It is then possible to judge the sensitivity of inferential conclusions with respect to certain plausible variations of the association parameters. If conclusions are relatively insensitive over a range of plausible assumptions about U,

causal inference is more defensible.

Other approaches to sensitivity analysis include methods based on the randomization distribution (Rosenbaum, 2002b), and bounding methods (Manski, 2003; Horowitz & Manski, 2000). Nonparametric and semi-parametric versions have also been proposed in the literature (Rosenbaum, 1987, 2002a; Ichino et al., 2008; Imbens, 2003; Ding & VanderWeele, 2016).

1.7 Discussion

Throughout this paper, we have attempted to convey the power of the potential outcomes formulation of causal effects in a variety of settings and provide an overview of a number of methodologies that may be relevant for statistical researchers. We by no means provide a comprehensive review of the vast literature on the analysis of causal effects. Current research in causal inference is exceedingly lively, involving extensive interactions between psychologists, behavioral scientists, computer scientists, economists, epidemiologists, philosophers, statisticians, and others. Armed with simple but powerful ideas such as the potential outcomes definition of causal effects, we look forward to future methodological developments in this crucial and fascinating area of empirical research.

2

Methods for matching to facilitate causal comparisons with text as data

2.1 Introduction

Recently, Roberts et al. (2018) introduced an approach for matching text documents in order to address confounding in observational studies of substantive and policy-relevant quantities of interest. Matching is a statistical tool primarily used to facilitate causal inferences about the effects of a particular treatment, action, or intervention from non-randomized data in the presence of confounding covariates (Rubin, 1973b; Rosenbaum, 2002b; Rubin, 2006b; Stuart, 2010). The principles behind matching can also be used to create sharp, targeted comparisons of units in order to, for example, create more principled rankings of hospitals (Silber et al., 2014). The core idea of matching is to find sets of units from distinct populations that are in all ways similar, other than some specific aspects of interest; one can then compare these remaining aspects across the populations of interest to ascertain differences foundational to these populations. In short, matching provides a strategy for making

precise comparisons and performing principled investigations in observational studies.

Though widely used in practice, matching is typically used in settings where both the covariates and outcomes are well-defined, low-dimensional quantities. Text is not such a setting. With text, standard contrasts of outcomes between groups may be distorted estimates of the contrasts of interest due to confounding by high-dimensional and possibly latent features of the text such as topical content or overall sentiment. How to best capture and adjust for these features is the core concern of this work. In particular, we consider the problem of matching documents within a corpus made up of distinct groups (e.g., a treatment and control group), where interest is in finding a collection of matched documents that are fundamentally "the same" along key dimensions of interest (in our first application, for example, we find newspaper articles that are about the same events and stories). These matched documents can then be used to make unbiased comparisons between groups on external features such as rates of citation or online views, or on features of the text itself, such as sentiment. In the case where group membership can be thought of as the receipt of a particular intervention (e.g., documents that were censored vs. not, such as in Roberts et al. 2018), this allows us draw causal inferences about effects of interest.

This paper makes three contributions to guide researchers interested in this domain. Our first contribution is a deconstruction and discussion of the elements that constitute text matching. This formulation identifies a series of choices a researcher can make when performing text matching and presents an approach for conceptualizing how matching can be used in studies where the covariates, the outcome of interest, or both are defined by summary measures of text. Our second contribution is to investigate these choices using a systematic multi-factor human evaluation experiment to examine how different representations and distance metrics correspond to human judgment about document similarity. Our experiment explores the efficiency of each combination of choices for matching documents in order to identify the representations and distance metrics that dominate in our context in terms of producing the largest number of matches for a given dataset without sacri-

ficing match quality. We also present a general framework for designing and conducting systematic evaluations of text-matching methods that can be used to perform similar investigations in different contexts. Our third contribution is twofold.

First, we present a novel application of template matching (Silber et al., 2014) to compare news media organizations' biases, beyond choices of which stories to cover, in order to engage with a running debate on partisan bias in the news media. Through template matching on text, we identify similar samples of news articles from each news source that, taken together, allow for a more principled (though not necessarily causal) investigation of how different news sources may differ systematically in terms of partisan favorability. In our second application, we illustrate the utility of text matching in a more traditional causal inference setting, namely, in an observational study evaluating the causal effects of a binary treatment. Here we demonstrate how matching on text obtained from doctors' notes can be used to improve covariate balance between treatment and control groups in an observational study examining the effects of a medical intervention. We further discuss how researchers might leverage text data to strengthen the key assumptions required to make valid causal inferences in this non-randomized context.

Our work builds on Roberts et al. (2018), the seminal paper in this literature, which introduces text matching and operationalizes the text data by using topic modeling coupled with propensity scores to generate a lower-dimensional representation of text to match on. They also present several applications that motivate the use of text matching to address confounding and describe several of the methodological challenges for matching that arise in these settings. Specifically, Roberts et al. (2018) discuss the limitations of direct propensity score matching and coarsened exact matching (CEM) on the raw text for matching with high dimensional data and introduce Topical Inverse Regression Matching (TIRM), which uses structural topic modeling (STM) (Roberts et al., 2016a) to generate a low-dimensional representation of a corpus and then applies CEM to generate matched samples of documents from distinct groups within the corpus. Building upon this work, we de-

velop a general framework for constructing and evaluating text matching methods. This allows us to consider a number of alternative matching methods not considered in Roberts et al. (2018), each characterized by one representation of the corpus and one distance metric. Within this framework, we also present a systematic approach for comparing different matching methods through our evaluation experiment, which identifies methods that can produce more matches and/or matches of higher quality than those produced by TIRM. Overall, we clarify that there is a trade-off between match quality and the number of matches, although many methods do not optimize either choice.

2.2 Background

2.2.1 Notation and problem setup

Consider a collection of N text documents, indexed by $i = 1, \ldots, N$, where each document contains a sequence of terms. These documents could be any of a number of forms such as news articles posted online, blog posts, or entire books, and each document in the dataset need not be of the same form. Together, these N documents comprise a corpus, and the set of V unique terms used across the corpus define the vocabulary. Each term in the vocabulary is typically a unique, lowercase, alphanumeric token (i.e., a word, number, or punctuation mark), though the exact specification of terms may depend on design decisions by the analyst (e.g., one may choose to include as terms in the vocabulary all bigrams observed in the corpus in addition to all observed unigrams). Because the number and composition of features which may be extracted from text is not well defined, documents are generally regarded as "unstructured" data in the sense that their dimension is *ex ante* unknown.* To address this issue, we impose structure on the text through a representation, X, which maps each document to a finite, usually high-dimensional, quantitative space.

*In particular, the number and composition of features which may be extracted from a given corpus is not well-defined and may vary depending on researcher focus.

To make principled comparisons between groups of documents within the corpus, we borrow from the notation and principles of the Rubin Causal Model (RCM) (Holland, 1986). Under the RCM, each document has an indicator for treatment assignment (i.e., group membership), Z_i, which equals 1 for documents in the treatment group and 0 for documents in the control group. Interest focuses on estimating differences between these groups on an outcome variable, which, under a causal view, would take the value $Y_i(1)$ if document i is in the treatment group and $Y_i(0)$ if document i is in the control group. These outcomes may be separate from the text of the document (e.g., the number of times a document has been viewed online) or may be a feature of the text (e.g., the length of the document or level of positive sentiment within the document).[†] Credible and precise causal inference revolves around comparing treated and control documents that are as similar as possible. However, in observational studies, Z_i is typically not be randomly assigned, leading to systematic differences between treatment and control groups. Matching is a strategy that attempts to address this issue by identifying samples of treated and control documents that are comparable on covariates in order to approximate random assignment of Z_i (i.e., to satisfy $Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1))|X_i$) (Rosenbaum, 2002b; Rubin, 2006b). Under this key assumption of "selection on observables," which states that all covariates that affect both treatment assignment and potential outcomes are observed and captured within X, comparisons of outcomes between matched samples can be used to obtain unbiased estimates of the quantities of interest (Rosenbaum, 2002b). For example, in our second application examining the effects of a medical intervention, we argue that matching on both a set of numerical covariates and the text content of the patients chart allows us to identify two groups of patients, one treated and one not, that are similar enough on pre-treatment variables such that any systematic differences in their outcomes can be plausibly attributed to the impact of the intervention.

[†]In the latter case, care must be taken to ensure the features of the representation X used to define the covariates are suitably separated from features that define the potential outcomes. This issue is discussed further in Section 2.3 and in Appendix B.1.4.

These causal inference tools can be used more broadly, however, to produce clearly defined comparisons of groups of units even when a particular intervention is not well-defined. For example, Silber et al. (2014) introduces *template matching* as a tool for comparing multiple hospitals that potentially serve different mixes of patients (e.g., some hospitals have a higher share of high-risk patients). The core idea is to compare like with like: by comparing hospitals along an effective "score card" of patients, we can see which hospitals are more effective, on average, given a canonical population. In general, we focus on this general conception of matching, recognizing that often in text there is no treatment that could, even in concept, be randomized. For example, a comparison of style between men and women could not easily be construed as a causal impact. Nevertheless, the framing and targeting of a controlled comparison, a framing inherent in a causal inference approach, can still be useful in these contexts. This broader formulation of matching is used in our first application in Section 2.5 investigating different aspects of bias in newspaper media.

2.2.2 Promises and pitfalls of text matching

Matching methods generally consist of five steps: 1) identify a collection of potential confounders (covariates) that would compromise any causal claims if they were systematically different across the treatment groups, 2) define a measure of distance (or similarity) to determine whether one unit is a good match for another, 3) match units across groups according to the chosen distance metric, 4) evaluate the quality of the resulting matched samples in terms of their balance on observed covariates, possibly repeating the matching procedure until suitable balance is achieved, and 5) estimate treatment effects from these matched data (Stuart, 2010). Different choices at each step of this process produce an expansive range of possible configurations. For instance, there are distance metrics for scalar covariates (Rubin, 1973b), for multivariate covariates summarized through a univariate propensity score (Rosenbaum & Rubin, 1983b, 1985), and multivariate metrics such as the Mahalanobis distance metric (Rubin, 1978b; Gu & Rosenbaum, 1993).

Similarly, there is a large and diverse literature on matching procedures (Rosenbaum, 2002b; Rubin, 2006b), and the choice of procedure depends on both substantive and methodological concerns. Some procedures match each unit in the treatment group to its one "closest" control unit and discard all unused controls (e.g., one-to-one matching with replacement), while other procedures allow treated units to be matched to multiple controls (e.g., ratio matching; Smith, 1997) and/or matching without replacement (e.g., optimal matching Rosenbaum 1989). Match quality is often evaluated with a number of diagnostics that formalize the notion of covariate balance such as the standardized differences in means of each covariate (Rosenbaum & Rubin, 1985). Unfortunately, determinations of what constitutes "suitable" balance or match quality are often based on arbitrary criteria (Imai et al., 2008; Austin, 2009), and assessing whether a matching procedure has been successful can be quite difficult. That being said, once a suitable set of matches is obtained, one can then typically analyze the resulting matched data using classic methods appropriate for the type of data in hand. Stuart (2010) outlines a number of common analytical approaches.

The rich and high-dimensional nature of text data gives rise to a number of unique challenges for matching documents using the standard approach described above. From a causal inference perspective, in many text corpora there is going to be substantial lack of overlap, i.e., entire types of documents in one group that simply do not exist in the other groups. This lack of overlap is exacerbated by the high-dimensional aspect of text: the richer the representation of text, the harder it will be to find documents similar along all available dimensions to a target document (D'Amour et al., 2017). This makes the many design decisions required to operationalize text for matching such as defining a distance metric and implementing a matching procedure especially challenging. Distance metrics must be defined over sparse, high-dimensional representations of text in a manner that captures the subtleties of language. If these representations are overly flexible, standard matching procedures can fail to identify good (or any) matches in this setting due to the curse of dimensionality.

Lack of overlap can come from substantive lack of overlap (the documents are inherently differ-

ent), but also aspects of the text representation that are not substantive (this is akin to overfitting the representation model). Ideally a good representation and distance metric will preserve the former but not the latter. All of the matching procedures discussed in this work can be thought of as carving out as many high quality matches as they can find, implicitly setting parts of the corpus aside to have good comparisons across groups. This is in effect *isolating* (Zubizarreta et al., 2014b) a focused comparison within a larger context. In a causal context, this can shift the implied estimand of interest to only those units in the overlap region. For further discussion of the approaches commonly used to address overlap issues, see, for example, Fogarty et al. (2016); Dehejia & Wahba (2002); Stuart (2010).

In addition to these difficulties, the rich nature of text data also provides an opportunity in that it lends itself to more straightforward, intuitive assessments of match quality than are typically possible with quantitative data. Specifically, while it is difficult to interpret the quality of a matched pair of units using numerical diagnostics alone due to being high dimensional, the quality of a matched pair of text documents is generally intuitive to conceptualize. With text data, human readers can quickly synthesize the vast amount of information contained within the text and quantify match quality in a way that is directly interpretable. Thus, when performing matching with text data, final match quality can be established in a manner that aligns with human judgment about document similarity. This is a version of "thick description," discussed in Rosenbaum (2010, pg. 322). This also allows for comparing different matching methods to each other in order to find methods that, potentially by using more sparse representations of text or more structured distance measures, can simultaneously find more matched documents while maintaining a high degree of match quality.

2.2.3 Different types of text-based confounding

Text is quite multifaceted, but that does not necessarily mean that the researcher needs to attend to all aspects of the text in order to appropriately control for any confounding. The confounding

feature of the text may be superficial and reducible to keywords, for example whether a news story covers politics, or it may be latent and difficult to deterministically measure, like a news story's ideological content.

In the simplest case, for example, consider a study with a single confounding feature that affects both assignment to treatment and the outcome of interest. Suppose that feature is defined as the presence in the text of a single word or phrase that is known *ex ante*. Since this can be measured deterministically using the available text data, then one can easily construct a statistic to capture that confounding (e.g., a binary variable indicating whether or not each document contains the word or phrase of interest). In this setting, the "best" text matching method will be the one that produces the best balance on that single critical word or phrase, calculated directly as the difference in means between prevalence of that word or phrase in treatment corpus and its prevalence in the control corpus.

In more complex settings, it may be necessary to control for some *latent* feature of the text, which might manifest in the text data as a set of related words. For instance, in the medical study described in Section 2.5.2, a patient's degree of frailty (i.e., healthiness or lack thereof) is a potentially confounding factor that is not measured numerically. This latent construct may manifest in the text data as a number of different key terms or phrases (e.g., "wheelchair bound"). If all such text-based indicators for the underlying construct of interest can be identified *ex ante* based on subject matter expertise and/or substantive theory, then it may be possible to directly quantify the latent variable by applying some hand-coded decision rules to the text. (In Section 2.5.2, we invert this procedure as a validation study of our more involved matching methods: if it is possible to avoid confounding by controlling for a set of pre-specified terms, then the most successful general text matching method will be the one that produces the best aggregate balance on those key words.) Again, in this circumstance, we may simply calculate these features for our documents and use classic matching methods from there.

The still more difficult scenario, the scenario that is the focus of this paper, is one in which the latent confounding feature of interest is challenging to measure directly, e.g., is not reducible to key words or phrases; these are the cases where we advocate for our more involved matching process that deals with general representation and distance metrics. In particular, many studies may have important confounding features that are inherently subjective (e.g., a hospital patient's level of optimism or a news story's partisan content). For example, in Section 2.5.1, we control for a subjective and latent feature of news articles: the story being covered. Since there are many different stories covered across all news articles, this confounding feature is a categorical variable in high dimension. As such, while there may be keywords which perfectly identify any one story in particular, for example the flight numbers of plane crashes or the names of important figures, compiling a complete list of all such keywords would be impossible. It is contexts such as these that we hope matching on more general representations of text without generating a set of hand-coded and targeted covariates will still allow for principled comparisons between groups of documents. But these automated methods may not work in a given context, and thus we also recommend in such contexts relying on human evaluation to verify that the matching process is controlling for those aspects of the text considered most critical to obtain ones "selection on observables" assumption.

2.3 A framework for matching with text data

When performing matching, different choices at each step of the process will typically interact in ways that affect both the quantity and quality of matches obtained. This can lead to different substantive inferences about the causal effects of interest. Therefore, it is important to consider the combination of choices as a whole in any application of matching. Although some guidelines and conventional wisdom have been developed to help researchers navigate these decisions, no best practices have yet been identified in general, let alone in settings with text data, where, in addition to the

usual choices for matching, researchers must also consider how to operationalize the data. We extend the classic matching framework to accommodate text documents by first identifying an appropriate quantitative representation of the corpus that ideally focuses attention on those aspects we are attempting to control for, then applying the usual steps for matching using this representation. Our framework applies in settings where summary measures of text are used to define the confounding covariates, the outcomes, or both.

The general procedure to match documents based on aspects of text that we propose is the following:

1. Choose a representation of the text and define explicitly the features that will be considered covariates and those, if any, that will be considered outcomes, based on this representation.[‡]

2. Define a distance metric to measure the similarity of two documents based on their generated covariate values that ideally focuses attention on the aspects of text considered the most important to account for (i.e., biggest potential confounders).

3. Implement a matching procedure to generate a matched sample of documents.

4. Evaluate match quality across the matched documents, and potentially repeat Steps 1-3 until consistently high quality matches are achieved.

5. Estimate the effects of interest using the final set of matched documents.

In the subsections below, we briefly introduce a number of different choices available in steps 1-3 of the above procedure and discuss the benefits and limitations of each. These options are summarized in Table 2.1. We then, in Section 2.4, present an approach for step 4 based on a human evaluation experiment. Finally, we illustrate step 5 through two different applications in Section 2.5.

[‡]There are additional considerations and steps required when both the covariates and outcome are characterized by text; see Appendix B.1.4.

Table 2.1: Common choices at each of the first three stages of the text matching procedure and examples of additional specifications required by each choice.

Step	Description	Common Choices	Specifications Required
1	Text representation	Term-Document Matrix	Dimension of vocabulary, weighting scheme, sparsity reduction
		Statistical Topic Model	Vocabulary size, number of topics, prior distributions, weighting scheme
		Document Embedding	Embedding dimension, training data, neural network architecture
2	Distance metric	Exact	None
		Coarsened Exact	Coarsening rules
		Continuous	Functional form (e.g., Euclidean, Cosine, Mahalanobis)
3	Matching procedure	Nearest neighbor	Replacement protocol, caliper, number of matches
		Optimal	Caliper, trimming, objective function
		Cardinality	Caliper, trimming

These steps and choices required to perform matching should be familiar to those with experience in standard matching, as many of the choices are directly parallel to a standard matching procedure. Because text is such a rich source of data, however, to determine the most suitable specification for matching requires careful consideration about what features of the text are most important to adjust for. For a more thorough discussion and description of the various choices within these steps, see Appendix B.1.

2.3.1 Text representations

The *representation* of a text document transforms an ordered list of words and punctuation into a vector of covariates, and is the most novel necessary component of matching with text. To choose a representation, the researcher must first formulate a definition for textual similarity that is appropriate for the study at hand. In some cases, all of the information about potential confounders

captured within the text data may be either directly estimable (e.g., frequency of a particular keyword) or may be plausible to estimate using a single numerical summary (e.g., the primary topic of a document estimated using a topic model). In other cases, such a direct approach may not be possible.

The most common general representation of text is as a "bag-of-words," containing unigrams and often bigrams, collated into a term-document matrix (TDM); the TDM may also be rescaled according to Term Frequency-Inverse Document Frequency (TF-IDF) weighting. Without additional processing, however, these vectors are typically very long; more parsimonious representations involve calculating a document's factor loadings from unsupervised learning methods like factor analysis or Structural Topic Models (STM) (Roberts et al., 2016a), or calculating a scalar propensity score for each document using the bag-of-words representation (Taddy, 2013). Finally, we also consider a Word2Vec representation (Mikolov et al., 2013), in which a neural network embeds words in a lower-dimensional space and a document's value is the weighted average of its words.

Each of these methods involves a number of tuning parameters. When using the bag-of-words representation, researchers often remove very common and very rare words at arbitrary thresholds, as these add little predictive power, or choose to weight terms by their inverse document frequency; these pre-processing decisions can be very important (Denny & Spirling, 2018). Topic models such as the STM are similarly sensitive to these pre-processing decisions (Fan et al., 2017) and also require specification of the number of topics and selecting covariates, which are often unstable. Word2Vec values depend on the dimensionality of the word vectors as well as the training data and the architecture of the neural network.

Overall, when choosing a representation, researchers need to consider what aspects of the text are confounding the outcome. For example, in our evaluation study that used matched pairs of news articles from Fox News and CNN, we were interested in identifying pairs of stories that were about the same general topic (e.g., plane crashes versus public policy) and that also utilized the same set

of keywords (e.g., "AirAsia" or "Obama"); this may suggest (as we found) that representations that preserve the details of different keywords was important for obtaining good matches. Generally, when the objective is to identify exact or nearly exact matches, we recommend using text representations that retain as much information in the text as possible. In particular, documents that are matched using the entire term-vector will typically be similar with regards to both topical content and usage of keywords, while documents matched using topic proportions may only be topically similar.

When the aspects of text are more targeted or specific, simply directly computing the relevant covariates constructed by hand-coded rules may be the best option. That being said, one might imagine that generally matching on the content of the text—as represented by the specific words and phrases used—will frequently capture much of what different researchers in different contexts may view as the necessary component for their selection on observables assumption. Clearly this is an area for future work; as we see more matching with text in the social sciences, we will also see a clear picture as to what structural aspects of text are connected to the substantive aspects of text that researchers find important.

2.3.2 Distance metrics

Having converted the corpus into covariate representations, the second challenge is in *comparing* any two documents under the chosen representation to produce a measure of distance. The two main categories of distance metrics are exact (or coarsened exact) distances, and continuous distances. Exact distances consider whether or not the documents are identical in their representation. If so, the documents are a match. Coarsened exact distance bins each variable in the representation, then identifies pairs of documents which share the same bins. If the representation in question is based on a TDM, these methods are likely to find only a small number of high quality matches, given the large number of covariates that all need to agree either exactly or within a bin. The alterna-

tive to exact distance metrics is continuous distance metrics such as Euclidean distance, Mahalanobis distance, and cosine distance. Counter to exact and coarsened exact metrics, which identify matches directly, these metrics produce scalar values capturing the similarity between two documents.

2.3.3 Matching procedures

After choosing a representation and a distance metric, the choice of matching procedure often follows naturally, as is the case in standard matching analyses. Exact and coarsened exact distance metrics provide their own matching procedure, while continuous distance metrics require both a distance formula and a *caliper* for specifying the maximum allowable distance at which two documents may be said to still match. The calipers may be at odds with the desired number of matches, as some treated units may have no control units within the chosen caliper, and may subsequently be "pruned" by many common matching procedures. Alternatively, researchers may allow any one treated unit to match multiple controls, or may choose a greedy matching algorithm.

2.4 Experimental evaluation of text matching methods

In the previous section, we presented different forms of representations for text data and described a number of different metrics for defining distance using each type of representation. Any combination of these options could be used to perform matching. However, the quantity and quality of matches obtained depend heavily on the chosen representation and distance metric. For example, using a small caliper might lead to only a small number of nearly-exact matches, while a larger caliper might identify more matches at the expense of overall match quality. Alternatively, if CEM on a STM-based representation produces a large number of low-quality matches, applying the same procedure on a TDM-based representation may produce a smaller number of matches with more apparent similarities.

We investigate how this quantity versus quality trade-off manifests across different combinations of methods through an evaluation experiment performed with human subjects. Applying several variants of the matching procedure described in Section 2.3 to a common corpus, we explore how the quantity of matched pairs produced varies with different specifications of the representation and distance metric. Then, to evaluate how these choices affect the quality of matched pairs, we rely on evaluations of human coders.

In this study, we consider five distance metrics (Euclidean distance, Mahalanobis distance, cosine distance, distance in estimated propensity score, and coarsened exact distance), as well as 26 unique representations,[§] including nine different TDM-based representations, 12 different STM-based representations, and five Word2Vec embedding-based representations. Crossing these two factors produces 130 combinations, where each combination corresponds to a unique specification of the matching procedure described in Section 2.3. Among these combinations, 5 specifications are variants of the TIRM procedure developed in Roberts et al. (2018). Specifications of each of the procedures are provided in Appendix B.2.

To compare the different choices of representation and distance metric considered here, we apply each combination to a common corpus to produce a set of matched pairs for each. We use a corpus of $N = 3,361$ news articles published from January 20, 2014 to May 9, 2015, representing the daily front matter content for each of two online news sources: Fox News ($N = 1,796$) and CNN ($N = 1,565$). The news source labels were used as the treatment indicator, with $Z = 1$ for articles published by Fox News and $Z = 0$ for articles published by CNN. These data are posted to the Dataverse of Political Analysis (Mozer, 2019).

[§]Because estimation and distance calculations with high-dimensional text representations can be computationally intensive, we restrict our analyses to this set of 26 possible representations, which we believe provide an adequate representation of the spectrum of possible text-representations that could be used for applications of text-matching. However, we emphasize that the methods presented in this paper, including the procedure for text-matching and the framework for performing systematic evaluations of text-matching methods, can be extended to include any number of additional variants to the representations considered here.

To match, we first calculate the distances between all possible pairs of treated and control units based on the specified representation and distance metric. Each treated unit is then matched to a set of control units with whom its distance was within the specified caliper.[¶] Using this procedure, 13 of the original 130 specifications considered did not identify any matched pairs. Each of the remaining 117 procedures identified between 23 and 1635 matched pairs of articles (with an average of 568 matched pairs per procedure). The union of matched pairs across all specifications resulted in 33,907 unique pairs of articles, where each unique pair was identified, on average, by 1.91 of the 117 different procedures. We view the frequency of each unique pair within the sample of 66,505 pairs identified as a rough proxy for match quality because, ideally when performing matching, the final sample of matched pairs identified will be robust to different choices of the distance metric or representation. Thus, we expect that matched pairs that are identified by multiple procedures will have higher subjective match quality than singleton pairs.

2.4.1 Measuring match quality

In standard applications of matching, if two units that are matched do not appear substantively similar, then any observed differences in outcomes may be due to poor match quality rather than the effect of treatment. Usual best practice is to calculate overall balance between the treatment and control groups, which is typically measured by the difference-in-means for all covariates of interest. If differences on all matched covariates are small in magnitude, then the samples are considered balanced, and thus, typically, well-matched.

As previously discussed, to calculate balance in settings where the covariates are text data, these standard balance measures typically fail to capture meaningful differences in the text. Further, due to the curse of dimensionality in these settings, it is likely that at least some (and probably many)

[¶] The caliper was calculated as the 0.1th quantile of the distribution of distances for all $1796 \times 1565 = 2{,}810{,}740$ possible pairs of articles under each specification.

covariates will be unbalanced between treatment and control groups. Thus, to measure match quality we rely on a useful property of text: its ease of interpretability. A researcher evaluating two units that have been matched on demographic covariates, for example, may be unable to verify the quality of a matched pair. However, depending on what aspects of text the researcher is substantively attempting to match on, human coders who are tasked with reading two matched text documents are often amply capable of quantifying their subjective similarity if given instructions as to what to attend to. We leverage this property to measure match quality using an online survey of human respondents, where match quality is defined on a scale of 0 (lowest quality) to 10 (highest quality).

To obtain match quality ratings, we conducted a survey experiment using Amazon's Mechanical Turk (MTurk) and the Digital Laboratory for the Social Sciences (DLABSS) (Enos et al., 2016). Online crowd-sourcing platforms such as these have been shown to be effective for similarity evaluations in a number of settings (Mason & Suri, 2012). For instance, a study by Snow et al. (2008) that tasked non-expert human workers on MTurk with five natural language evaluations reported a high degree of agreement between the crowd-sourced results and gold-standard results provided by experts. In the present study, respondents were first informed about the nature of the task and then given training on how to evaluate the similarity of two documents. After completing training, participants were then presented with a series of 11 paired newspaper articles, including an attention check and an anchoring question, and asked to assign a similarity rating. For each question, participants were instructed to read both articles in the pair and rate the articles' similarity from zero to ten, where zero indicates that the articles are entirely unrelated and ten indicates that the articles are covering the exact same event. Snapshots of the survey are presented in Appendix B.3.

We might be concerned that an online convenience sample may not be an ideal population for conducting this analysis, and that their perceptions of article similarity might differ from the overall population, or from trained experts. To assess the reliability of this survey as an instrument for measuring document similarity, we leverage the fact that we performed two identical pilot surveys

prior to the experiment using respondents from two distinct populations and found a high correlation ($\rho = 0.85$) between the average match quality scores obtained from each sample. Additional details about this assessment are provided in Appendix D in the Supplement. We take note that these populations, MTurkers and DLABSS respondents, are both regularly used as coders to build training data sets for certain tasks in machine learning; the hallmark of these tasks is that they are easily and accurately performed by untrained human respondents. We argue that this task of identifying whether two articles discuss related stories falls squarely in this category, and our inter-coder reliability test (described in Appendix D in the Supplement) supports this argument.[‖]

In an ideal setting, for each unique matched pair identified using the procedure described above, we would obtain a sample of similarity ratings from multiple human coders. Aggregating these ratings across all pairs in a particular matched data set would then allow us to estimate the average match quality corresponding to each of the 130 procedures considered, with the quality scores for the 13 procedures that identified no matches set to zero. Though this is possible in principle, to generate a single rating for each unique matched pair requires that a human coder read both documents and evaluate the overall similarity of the two articles. This can be an expensive and time-consuming task. Thus, in this study, it was not possible to obtain a sample of ratings for each of the 33,907 unique pairs.

Instead, we took a stratified, weighted sample of pairs such that the resulting sample would be representative of the population of all 33,907 unique matched pairs as well as the population of 2,776,833 pairs of documents that were not identified by any of the matching procedures. Specifically, the sample was chosen such that each of the 130 matching procedures that identified a non-zero number of matches would be represented by at least four pairs in the experiment. For each stratum, the sampling weights for each pair were calculated proportional to the estimated match quality of

[‖]For researchers interested in conducting their own text matching evaluation studies, we note that MTurk and DLABSS populations may not always be applicable, especially in contexts where domain expertise is required.

that pair, calculated using a predictive model trained on human-coded data from a pilot experiment. We also sampled an additional 50 unique pairs from the pool of 2,776,833 pairs not identified by any matching procedures.

Ratings obtained from these pairs can be used to obtain a reference point for interpreting match quality scores. The resulting sample consisted of 505 unique pairs ranging the full spectrum of predicted match quality scores. Each respondent's set of nine randomly selected questions were drawn independently such that each pair would be evaluated by multiple respondents. Using this scheme, each of the 505 sampled pairs was evaluated by between six and eleven different participants (average of 9). Question order was randomized, but the anchor was always the first question, and the attention check was always the fifth question.

We surveyed a total of 505 respondents. After removing responses from 52 participants who failed the attention check,** all remaining ratings were used to calculate the average match quality for each of the 505 sampled pairs evaluated. These scores were then used to evaluate each of the 130 combinations of methods considered in the evaluation, where the contribution of each sampled pair to the overall measure of quality for a particular combination of methods was weighted according to its sampling weight. This inferential procedure is described more formally in Appendix B.5.

2.4.2 Results

Which automated measures are most predictive of human judgment about match quality?

Our primary research question concerns how unique combinations of text representation and distance metric contribute to the quantity and quality of obtained matches in the interest of identi-

**The attention check consisted of two articles with very similar headlines but completely different article text. The text of one article stated that this question was an attention check, and that the respondent should choose a score of zero. Participants who did not assign a score of zero on this question are regarded as having failed the attention check.

fying an optimal combination of these choices in a given setting. We can estimate the quality of the 130 matching methods considered in the evaluation experiment using weighted averages of the scores across the 505 pairs evaluated by human coders. However, it is also of general interest to be able to evaluate new matching procedures without requiring additional human experimentation. We also want to maximize the precision of our quality estimates for the 130 methods considered in this study. To these ends, we examine if we can predict human judgment about match quality based on the distance scores generated by each different combination of one representation and one distance metric. If the relationship between the calculated match distance and validated match quality is strong, then we may be confident that closely-matched documents, as rated under that metric, would pass a human-subjects validation study.

To evaluate the influence of each distance score on match quality, we take the pairwise distances between documents for each of the 505 matched pairs used in the evaluation experiment under different combinations of the representations and distance metrics described in Section 2.3. After excluding all CEM-based matching procedures, under which all pairwise distances are equal to zero or infinity by construction, all distances were combined into a data set containing 104 distance values for each of the 505 matched pairs. Figure 2.1 gives six examples of how these distances correlate with observed match quality based on human ratings of similarity, along with the fitted regression line obtained from quadratic regressions of average match quality on distance. Here, the strong correlations suggest that automated measures of match quality could be useful for predicting human judgment. The particularly strong relationship between the cosine distance metric calculated over a TDM-based representation provides additional evidence in favor of matching using this particular combination of methods. These findings also suggest that the increased efficiency achieved with TDM cosine matching is not attributable to the cosine distance metric alone, since the predictive power achieved using cosine distance on a Word2Vec (W2V) representation or a STM-based representation is considerably lower than that based on a TDM-based representation.

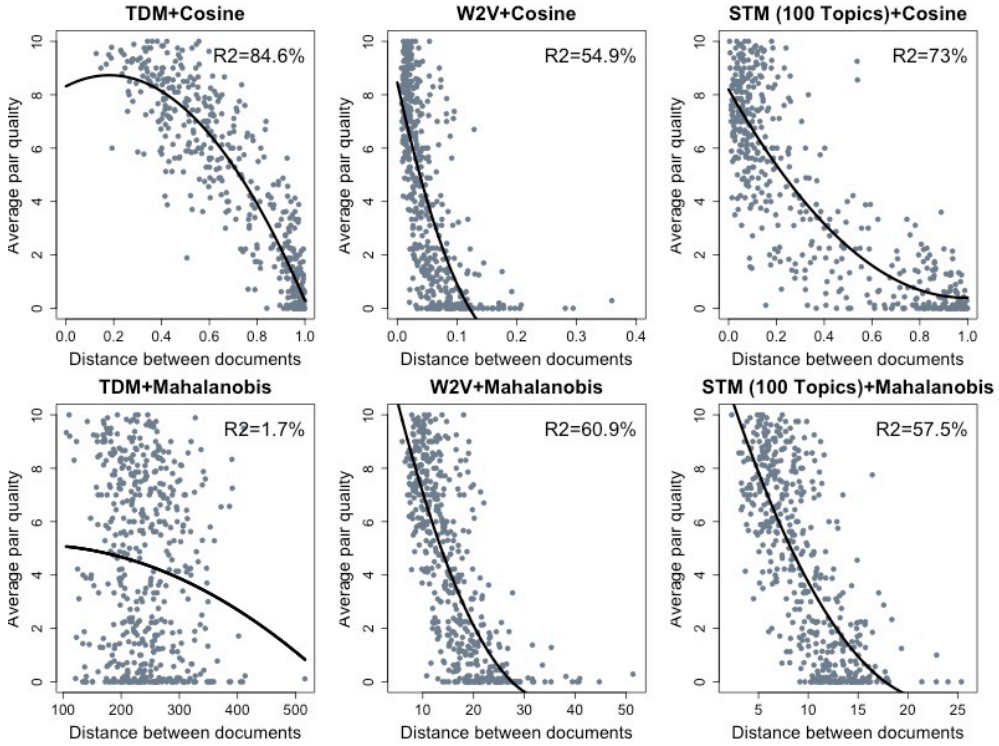


Figure 2.1: Distance between documents and match quality based on the cosine distance measured over a TDM-based representation (top left) exhibit a stronger relationship than cosine distance measured over both a W2V-based representation (top center) and a STM-based representation (top right), and a much stronger relationship than the Mahalanobis distance measured over a TDM-based representation (bottom left), a W2V-based representation (bottom center) or a STM-based representation (bottom right).

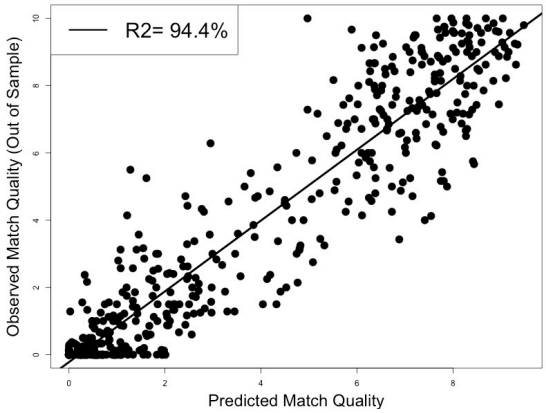


Figure 2.2: Predictive model for match quality trained on human evaluations has a correlation of 0.944 with observed quality scores obtained in a separate human evaluation experiment on a different set of pairs, indicating high out-of-sample predictive accuracy.

To leverage the aggregate relationship of the various machine measures of similarity on match quality, we developed a model for predicting the quality of a matched pair of documents based on the 104 distance scores, which we then trained on the 505 pairs evaluated in our survey experiment. For estimation, we use the LASSO (Tibshirani, 1996), implemented with ten-fold cross validation (Kohavi et al., 1995). Here, for each of the 505 pairs, the outcome was defined as the average of the ratings received for that pair across the human coders, and the covariates were the 104 distance measures. We also included quadratic terms in the model, resulting in a total of p=208 terms. Of these, the final model obtained from cross-validation selected 23 terms with non-zero coefficients and achieved 88% out-of-sample predictive accuracy. However, our results suggest that the majority of the predictive power of this model can be attributed to two terms: cosine distance over the full, unweighted term-document matrix and cosine distance over an STM with 100 topics; see Appendix B.6 for additional details.

The high predictive accuracy of our fitted model suggests that automated measures of similarity could be effectively used to evaluate new matched samples or entirely new matching procedures

without requiring any additional human evaluation.[††] We can also use it to enhance the precision of our estimates of match quality for the 130 matching methods considered in the evaluation experiment using model-assisted survey sampling techniques.

Which methods make the best matching procedures?

To compare the performance of the final set of 130 matching procedures considered in our study, we, for each method, estimate the average quality of all pairs selected by that method. We increase precision of these estimates using model-assisted survey sampling. In particular, we first use the predictive model described above to predict the quality of all matched pairs of a method. This average quality estimate is then adjusted by a weighted average of the residual differences between predicted and actual measured quality for those pairs directly evaluated in the human experiment. (The average quality scores for the 13 procedures that identified no matches are all set equal to zero.) This two-step process does not depend on the model validity and is unbiased.[‡‡] We assess uncertainty with a variant of the parametric bootstrap. See Appendix B.5 for further details of the estimation approach and associated uncertainty quantification. Figure 2.3 shows the performance of each of the 130 procedures in terms of average predicted match quality vs. number of pairs identified, with uncertainty intervals estimated using a parametric bootstrap; see Appendix D in the Supplement for a tabular summary of these results. We group the procedures by the large-scale choices of representation and distance metric used. Within each tile of the larger plot are different procedures corresponding to different design decisions within a general approach such as tuning parameters such as number of

[††]Since this model was trained on human evaluations of matched newspaper articles, extrapolating predictions may only be appropriate in settings with similar types of documents. However, our experimental framework for measuring match quality could be implemented using text data to build a similar predictive model in other contexts.

[‡‡]Nearly unbiased that is. There is a small bias term due using a Hajek-style approach rather than Horvitz-Thompson. This comes from the sample having a random total weight due to using the weighted sampling method.

topics used in a topic model. As sensitivity check, see Appendix B.6 in the Supplement for results using the simple weighted means of the sampled pairs of each method; results are broadly similar.

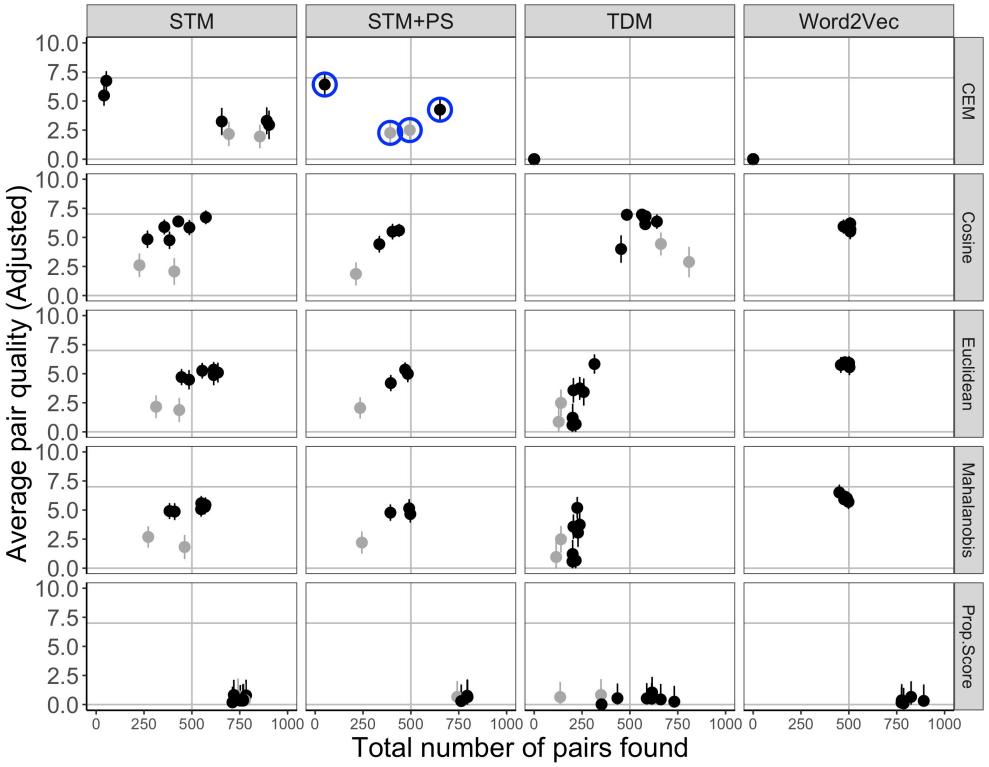


Figure 2.3: Number of matches found versus average model-assisted match quality scores for each combination of matching methods. Grey points indicate procedures with extreme reduction in information (e.g., procedures that match on only stop words). Blue circles highlight procedures that use existing state-of-the-art methods for text matching. One procedure with many low quality pairs at coordinates (1605,1.39) is excluded from this plot.

The methods which generally produce the highest quality matches for our study are those based on cosine distance calculated over a TDM-based representation. The method that produces the most matches out of all 130 procedures considered uses STM on ten topics with sufficient reduction and CEM in 2 bins and identifies over 1600 matched pairs. However, this method is among the lowest scoring methods in terms of quality, with a sample-adjusted average match quality of

1.14. Conversely, a procedure that uses STM on 30 topics with sufficient reduction and CEM in 3 bins, appears to produce considerably higher quality matches, with an average match quality of 6.39, but identifies relatively few matched pairs. In comparison, a method that combines a bounded TDM with TF-IDF weighting with the cosine distance metric identified 579 matches with an average match quality of 7.11. This illustrates an important weakness of CEM: too few bins produce many low quality matches, while too many bins produce too few matches, even though they are high quality. While in many applications there may be a number of bins which produce a reasonable number of good quality matches, that is not the case in our setting. Here, two bins produce poor matches while three bins produce far too few. This trade-off does not appear to be present for matching procedures using cosine distance with a TDM-based representation, which dominate in both number of matches found and overall quality of those matched pairs. In addition, the matching procedures based on this combination appear to be more robust to various the pre-processing decisions made when constructing the representation than procedures that use an alternative distance metric or representation, as illustrated by the tight clustering of the variants of this general approach on the plot.

Overall, our results indicate that, in our context, matching on the full TDM produces both more and higher quality matches than matching on a vector of STM loadings when considering the content similarity of pairs of news articles. Moreover, TDM-based representations with cosine matching appear relatively robust to tuning parameters including the degree of bounding applied and the choice of weighting scheme. STM-based representations, on the other hand, appear to be somewhat sensitive to tuning parameters, with representations that include a large number of topics achieving higher average match quality than those constructed over a smaller number of topics. This result provides further support for the findings in Roberts et al. (2018). In that paper, the authors found that matching on more topics generally led to better results in terms of recovering pairs of nearly identical documents.

2.4.3 Evaluating text matching methods

In our applied examples, we find that text representations that use the TDM or Word2Vec embeddings paired with cosine distance achieve the best results in terms of maximizing predicted match quality and the number of matches identified. But these results may well not be general. We therefore emphasize that applied researchers conducting their own text matching analyses need to conduct their own systematic evaluations to determine which representations and distance metrics work best in their domains. Here we offer some thoughts on how to think about and design convenient and flexible systematic evaluations.

First, until we have more general research knowledge in the field, we recommend implementing a suite of text matching procedures that include a diverse set of representations and distance metrics, and then comparing the matches identified by the different methods. If there is substantial overlap across all methods, it may be that no evaluation is necessary. However, in most cases, the sets of identified matches will largely diverge. In this case, we recommend formally evaluating which methods best capture the confounding previously identified.

In our discussion of different types of confounding, above, we noted that if the aspects of text that are most important are directly measurable, then these more general text matching approaches are not needed, strictly speaking. In this case we recommend directly assessing balance on specific covariates built from the text. But if general text matching methods are used in such contexts, we believe checking these core covariates to still be of use as signals as to which methods are at least achieving balance on some core summary statistics. This is akin to viewing mean balance as a proxy for covariate balance in classic matching.

If the potential confounding truly hinges on the more complex and latent aspects of text, however, then one could ideally leverage human judgment to hand evaluate the full set of possible matched pairs of text documents. In our case, for example, we could, given unlimited resources,

ask human coders to read through the entire corpus of news articles and put them into bins according to which stories they cover. Even untrained human coders could be reliably good at this task. This, of course, is generally not possible, but we hope the methods described above serve a similar function.

As we have seen, we can evaluate the success of such an attempt by inverting the full human-coding procedure to generate a test: we identify a set of possible matches using automated text matching methods and then and present a subset of them to trained human coders. These human coders can then evaluate sample pairs of matched documents to determine which matches are systematically "best" according to their own judgment. Using this information we can then see which methods appear to best match on the targeted aspects of text. This human coding task is of utmost importance, requiring both careful pretesting and substantial guidance to ensure the humans attend to the aspects of text deemed most important as potential confounders. In particular, the primary concern is instructing the human coders to evaluate similarity *along the latent dimension of interest*, which in our media case is whether any two articles truly cover the same events or issues.

One final circumstance bears discussion: it may be the case that the identified latent dimension of interest is challenging or impossible for human evaluators to reliably code. For example, even two experienced medical doctors may systematically disagree in their readings of patient data such as X-rays (Steiner et al., 2018). In such cases, human evaluations may not serve as a reliable ground truth to which automated text match quality may be compared. It is still possible that automated text matching methods would work well in these cases, but researchers cannot validate those results in this framework.

This and other open questions, including identifying what contexts would have topic model- or propensity score-based representations outperform TDM-based or Word2Vec embedding-based representations, we leave to future research.

2.5 Applications

2.5.1 Decomposing media bias

While American pundits and political figures continue to accuse major media organizations of "liberal bias," scholars, after nearly two decades of research on the issue, have yet to come to a consensus about how to *measure* bias, let alone determine its direction. A fundamental challenge in this domain is how to disentangle the component of bias relating to *how* a story is covered, often referred to as "presentation bias" (Groseclose & Milyo, 2005; Gentzkow & Shapiro, 2006; Ho et al., 2008; Gentzkow & Shapiro, 2010; Groeling, 2013), from the component relating to *what* is covered, also known as "selection bias" (Groeling, 2013) or "topic selection." In particular, systematic comparisons of *how* stories are covered by different news sources (e.g., comparing the level of positive sentiment expressed in the article) may be biased by differences in the content being compared. We present a new approach for addressing this issue by using text matching to control for selection bias.

We analyze a corpus consisting of $N = 9,905$ articles published during 2013 by each of 13 popular online news outlets. This data was collected and analyzed in Budak et al. (2016). The news sources analyzed here consist of Breitbart, CNN, Daily Kos, Fox News, Huffington Post, The Los Angeles Times, NBC News, The New York Times, Reuters, USA Today, The Wall Street Journal, The Washington Post, and Yahoo. In addition to the text of each article, the data include labels indicating each articles' primary and secondary topics, where these topics were chosen from a set of 15 possible topics by human coders in a separate evaluation experiment performed by Budak et al. (2016). The data also include two human-coded outcomes that measure the ideological position of each article on a 5-point Likert scale. Specifically, human workers tasked with reading and evaluating the articles were asked "on a scale of 1-5, how much does this article favor the Republican party?", and similarly, "on a scale of 1-5, how much does this article favor the Democratic party?"

To perform matching on this data, we use the optimal procedure for identifying articles cover-

ing the same underlying story identified by our prior evaluation experiment: cosine matching on a bounded TDM. Note that because the outcomes of interest in this analysis are human-coded measures of favorability toward democrats and republicans, we limit the vocabulary of the TDM to include only nouns and verbs to avoid matching on aspects of language that may be highly correlated with these outcomes. Because in this example we have a multi-valued treatment with 13 levels, each representing a different news source, we follow the procedure for template matching described in Silber et al. (2014) to obtain matched samples of 150 articles across all treatment groups. In brief, the template matching procedure first finds a representative set of stories across the entire corpus, and uses that template to find a sample of similar articles within each source that collectively cover this canonical set of topics. This allows us to identify a set of articles sampled from each source that are all similar to the same template and therefore similar to each other.

Before matching, our estimates of a news source's average favorability are a measure of overall bias, which includes biases imposed through differential selection of content to publish as well as biases imposed through the language and specific terms used when covering the same content. The matching controls selection biases due to some sources selecting different stories that may be more or less favorable to a given party than other stories. Differences in estimated favorability on the matched articles can be attributed to presentation bias. The difference between estimates of average favorability before matching (overall bias) and estimates after matching (presentation bias) therefore represent the magnitude of selection biases imposed by the sources. Large differences between pre- and post-matched estimates indicate a stronger influence of selection bias relative to presentation bias.

Figure 2.4 shows the average favorability toward Democrats (blue) and Republicans (red) for each news source overall, and the average favorability among the template matched documents. Arrows begin at the average score before matching, and terminate at the average score after matching. The length of the arrows is the estimated magnitude of the bias of each source that is attributable to

differences in selection.

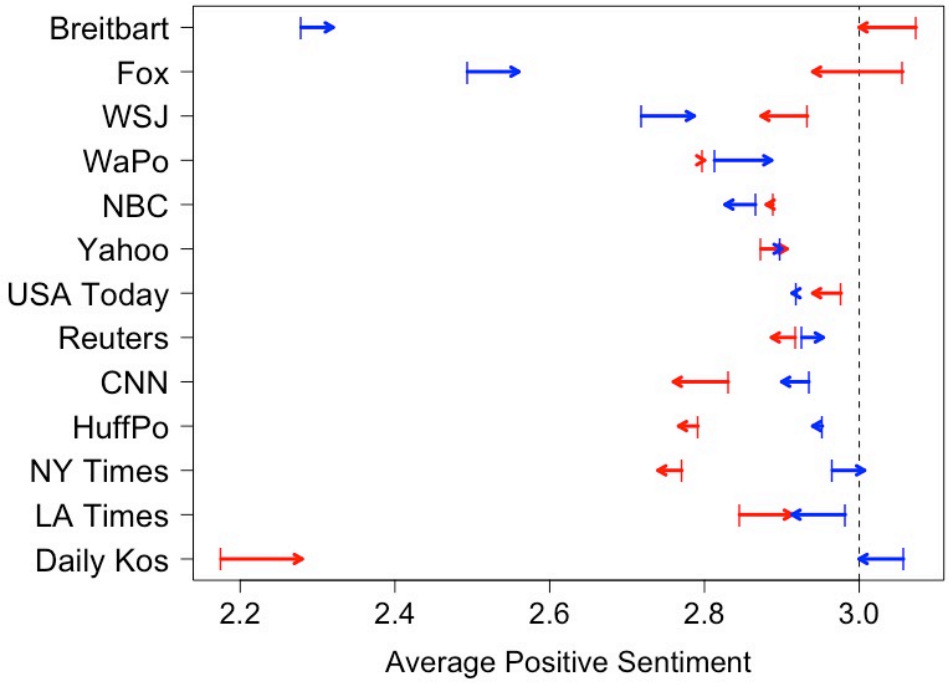


Figure 2.4: Estimates of average favorability toward Democrats (blue) and Republicans (red) for each source both before and after matching.

Before discussing the pattern of shifts, we first look at overall trends of favorability across sources. First, overall sentiment towards Republicans generally hovers around 2.8 to 3.1, slightly less, on average, than the partisan neutrality of $x = 3$, which corresponds to a response of "neither favorable nor unfavorable." The one exception is the Daily Kos, which is unfavorable. Other sources (CNN, the Huffington Post, the NY Times, and the LA Times) are at the low end of this range, indicating some negative sentiment. For the Democrats, there is somewhat more variation, however, with Brietbart being the least favorable, followed by Fox and WSJ, and the Daily Kos being the most.

Furthermore, it is primarily the more extreme sources that show selection effects. Breitbart, Fox

and WSJ, for example, all become more positive towards Democrats and less positive towards Republicans when we adjust for story. This suggests they tend to select stories that are biased more towards Republicans and away from Democrats, a selection bias effect. Similarly, the LA Times and Daily Kos show the opposite trends, again showing selection bias effects in the opposite direction. The remaining sources do not appear significantly be impacted by controlling for selection.

We performed a series of sensitivity checks to assess the stability of our results to different specifications of the matching procedure and/or different choices of template sample. We also examine the variability due to randomly matching documents to assess how much estimation uncertainty is present in our analysis. Details of these analyses are provided in Appendix B.7. Generally, we see that estimating the selection effect of an individual source is difficult, and that the magnitude of the selection effects tends to be small, indicating that the choice of what stories to cover is not driving the overall favorability ratings. In other words, most differences in favorability appear to be driven by presentation bias.

2.5.2 Improving covariate balance in observational studies

In our second application, we demonstrate how text matching can be used to strengthen inferences in observational studies with text data. Specifically, we show that text matching can be used to control for confounders measured by features of the text that would otherwise be missed using traditional matching schemes.

We use a subset of the data first presented in Feng et al. (2018), which conducted an observational study designed to investigate the causal impact of bedside transthoracic echocardiography (TTE), a tool used to create pictures of the heart, on the outcomes of adult patients in critical care who are diagnosed with sepsis. The data were obtained from the Medical Information Mart for Intensive Care (MIMIC) database (Johnson et al., 2016) on 2,401 patients diagnosed with sepsis in the medical and surgical intensive care units at a Massachusetts Institute of Technology university hospital

located in Boston, Massachusetts. Within this sample, the treatment group consists of 1,228 patients who received a TTE during their stay in the ICU (defined by time stamps corresponding to times of admission and discharge) and the control group is comprised of 1,173 patients who did not receive a TTE during this time. For each patient we observe a vector of pre-treatment covariates including demographic data, lab measurements, and other clinical variables. In addition to these numerical data, each patient is also associated with a text document containing intake notes written by nursing staff at the time of ICU admission. The primary outcome in this study was 28-day mortality from the time of ICU admission.

Because the treatment in this study was not randomly assigned to patients, it is possible that patients in the treatment and control groups may differ systematically in ways that affect both their assignment to treatment versus control and their 28-day mortality. For instance, patients who are in critical condition when admitted into the ICU may die before treatment with a TTE has been considered. Similarly, patients whose health conditions quickly improve after admission may be just as quickly discharged. Therefore, in order to obtain unbiased estimates of the effects of TTE on patient mortality, it is important to identify and appropriately adjust for any potentially confounding variables such as degree of health at the time of admission.

We apply two different matching approaches to this data: one that matches patients only on numerical data and ignores the text data, and one that matches patients using both the numerical and text data. In the first procedure, following Feng et al. (2018), we match treated and control units using optimal one-to-one matching (Hansen & Klopfer, 2006) on estimated propensity scores (calculated by fitting a logistic regression of the indicator for treatment assignment on the observed numerical covariates). We enforce a propensity score caliper equal to 0.1 standard deviations of the estimated distribution, which discards any treated units for whom the nearest control unit is not within a suitable distance. In the second approach, we perform optimal one-to-one text matching within propensity score calipers. Intuitively, this procedure works by first, via the calipers, reducing

the space of possible treated-control pairings in a way that ensures adequate balance on numerical covariates. By then performing text matching within this space to select a specific match given a set of candidate matches all within the calipers, we obtain matched samples that are similar with respect to all observed covariates, including the original observed covariates and any variables that were not recorded during the study but can be estimated by summary measures of the text.

Identifying the optimal text-matching method here requires careful consideration of how text similarity should be defined and evaluated in this medical context. Here, the ideal text-matching method is one that matches documents on key medical concepts and prognostic factors that could both impact choice of using TTE as well as the outcome (i.e., potential confounders) that are captured within the text data. Unlike in the previous application, these features cannot be reliably evaluated by non-expert human coders due to the domain expertise and familiarity with medical jargon necessary to make comparisons between medical documents. Thus, to perform a systematic evaluation of text matching methods in this study, we adopt an information retrieval approach for comparing medical texts that has been widely applied in the biomedical literature (Aronson, 2001; Zeng et al., 2007).

In particular, by consulting with medical professionals, we first obtained a mapping of the texts to a set of clinically meaningful concepts that could be used to characterize ICU patients. Following the approach of MacLean & Heer (2013), we then calculated the Jaccard similarity over this mapping between matched pairs of documents as an omnibus measure of match quality. We treat these scores as a working gold standard for this particular application; these scores are based on careful consideration from domain experts who have the medical background required to extract potentially confounding information from this type of nuanced text. When such a mapping is available, the Jaccard similarity metric offers a practical alternative to human evaluation for obtaining estimates of match quality that can be used to compare the relative performance of different matching procedures. However, this metric may not be appropriate for evaluating new texts or for measuring

text similarity in other contexts. Figure 2.5 shows the average pairwise Jaccard similarity achieved after matching (within propensity score calipers based on the numerical covariates) using each of the 130 text matching specifications described in Section 2.3. The best-performing procedure matches each treated unit to its nearest control based on the cosine distance calculated over a bounded TDM, where treated units whose nearest control is outside the specified caliper are discarded.

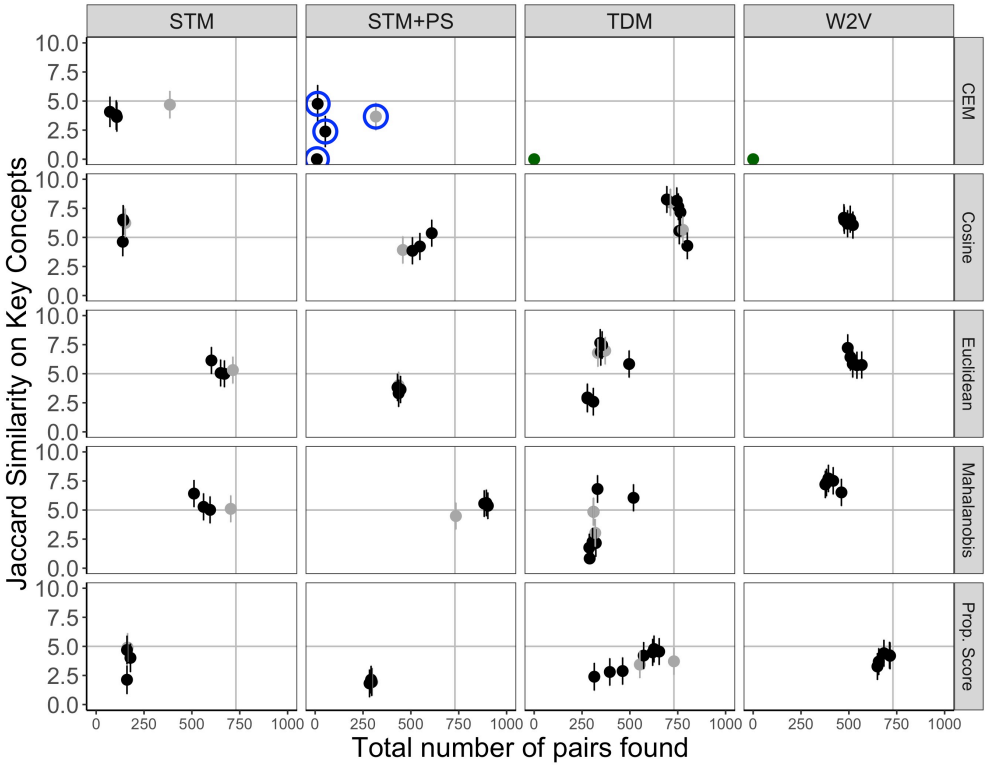


Figure 2.5: Number of matches found versus average pairwise Jaccard similarity for each combination of matching methods. Grey points indicate procedures with extreme reduction in information (e.g., procedures that match on only stop words). Blue circles highlight procedures that use existing state-of-the-art methods for text matching.

Figure 2.6 shows the covariate balance between treatment and control groups on both quantitative and text-based covariates before matching, after propensity score matching (PSM) on numeric

covariates alone, and after text matching using our preferred method (using cosine distance on a bounded TDM) within propensity score calipers. Here, each of the five text-based covariates are calculated using summary measures based on word-counts from the patient-level text documents.

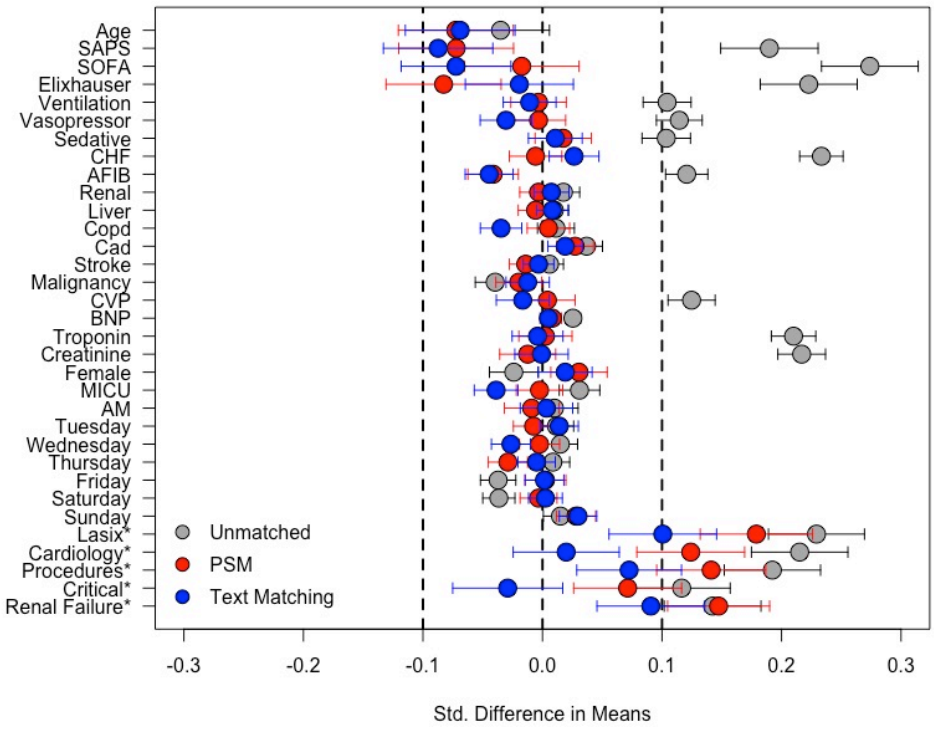


Figure 2.6: Standardized differences in means with 95% confidence intervals between treatment and control groups on 26 numerical covariates and 5 text-based covariates (denoted by *) before matching (gray), after propensity score matching (red), and after text matching (blue).

These variables, according to medical experts consulted on this project, all could indicate potential confounds that could bias estimates of impact if not controlled. Our general text matching methods do not directly balance these covariates; the improved balance is a consequence of matching on the more general overall distance metric and representation used.

In general, common wisdom (e.g., (Imbens & Rubin, 2015)) is to condition on *all* available data

that could indicate potential confounding influences when making inferences using observational data. While PSM is able to adequately balance the numerical covariates and some of the text-based covariates most correlated with these numerical measures, it fails to sufficiently adjust for differences between treatment and control groups on a number of potential confounders captured only by the text. For instance, both the unmatched data and the matched sample generated using PSM have large imbalances between treatment and control groups on references to Lasix, a medication commonly used to treat congestive heart failure. In the unmatched sample, only 10% of treated units have documents containing references to this medication compared to 28% of control units who are associated with the medication. Matching on the estimated propensity scores reduces this imbalance only slightly, while cosine matching within propensity score calipers shows a considerable improvement in the balance achieved between treatment groups on this variable. Incorporating the text data into the matching procedure leads to similar improvements in balance for the other five text-based variables while also maintaining suitable overall balance on the numerical covariates.

Table 2.2: Survival rates for treatment and control groups and estimated treatment effects before and after propensity score matching (PSM) and text matching within propensity score calipers.

Procedure	Effective Sample Size	Survival Rate		Difference (Std. Error)
		Treatment	Control	
Before Matching	1186	72.5%	71.2%	1.3% (1.8%)
PSM	807	72.5%	67.7%	4.8% (2.2%)
Text Matching	894	72.5%	67.5%	5.0% (2.1%)

Table 2.2 summarizes the survival rates in the treatment and control groups within each matched sample along with the effective sample sizes (i.e., the equivalent number of matched pairs) in the final matched samples. Generally, there appears to be some confounding, with the adjusted impacts being larger than the naïve differences. The matched sample identified using text matching is slightly

larger in terms of effective sample size than simple PSM, although they are not significantly different. This increase in effective sample size highlights the efficiency of text matching; when evaluating multiple control units that are eligible matches for a single treated unit in terms of quantitative covariates, the text-based distance offers a more refined measure of pairwise similarity than distances based on the propensity score. Further, when text matching within propensity score calipers, small differences in estimated propensity scores across control units will be offset by any large differences in text. In the present application, this allows for more precise and efficient optimization of the matched sample.

Of course, conducting a matched analysis is rooted in thoughtful design. In particular, the researcher must decide which variables are important potential confounders, and which are not. This is especially important when balancing the trade-offs between achieving better balance on some variables at the expense of others. The purpose of highlighting text matching in this context is to demonstrate how information from the text can also be included in these decisions of what to attend to. If the text is deemed not informative, then of course it should not be an important consideration with matching. But, as in this case, if the text is considered to indicate significant aspects of patient condition that should be attended to, the general matching procedures we have discussed can provide a way forward. And if it is uncertain what is important, then sensitivity checks that focus balance on different groups of variables can further strengthen causal claims in these contexts.

2.6 Discussion

In this paper we have made three primary contributions. First, we have provided guidance for constructing different text matching methods and evaluating the match quality of pairs of documents identified using such methods. Second, we empirically evaluated a series of candidate text matching procedures constructed using this framework along with the methods developed in Roberts et al.

(2018). Third, we have applied our methods to a data set of news media in order to engage with a long-standing theoretical debate in political science about the composition of bias in news, and to an observational study evaluating the effectiveness of a medical intervention.

Text matching is widely applicable in the social sciences. Roberts et al. (2018) show how text matching can produce causal estimates in applications such as international religious conflict, government-backed internet censorship, and gender bias in academic publishing. We believe the framework presented in this paper will help expand the scope and usability of text matching even further and will facilitate investigation of text data across a wide variety of disciplines. For instance, the methods described here could enhance state-of-the-art techniques for plagiarism detection and text reuse, techniques that are widely used in political science. By identifying bills that are textually similar to an original legislative proposal, our approach could be used to improve upon work tracking the spread of policy through state legislatures (Kroeger, 2016); and by comparing social media posts to a matched source article, our methods could detect the dispersion of false news topics through a social network. Secondly, our framework could be used to construct *networks* of lexical similarity, for instance of news sources, politicians, or national constitutions. As well, the metrics we consider for measuring text similarity could themselves resolve measurement problems in cases where lexical *divergence* is the quantity of interest, for example in cases of studying ideological polarization using text data (Peterson & Spirling, 2018).

We urge, however, that researchers consider how similar their use cases are to ours when extrapolating from results based on our evaluation experiments. In particular, while cosine distance and TDM-based representations produced high quality results in both of our applied examples, this finding should not be taken as conclusive evidence that these choices are the best in any application of text matching. Further, we emphasize to researchers that the results of our human evaluation experiment depend on the crucial assumption that humans are able to distinguish between textual differences that represent potential confounders, which may bias inferential results if not appropri-

ately controlled for and extraneous differences that are not relevant for the purposes of inference. This assumption may not be plausible in all settings, and we therefore encourage future researchers to conduct their own evaluation studies, especially when using text matching to control for linguistic features other than content similarity, for example stylistic, topic, tone, or semantic similarity. We hope such future evaluations, in connection with this one, will advance our collective understanding of best practices in this important domain.

3

Methods to estimate the effects of medical interventions in longitudinal studies with treatment by indication

3.1 Introduction

In certain observational studies, particularly in a medical context, interest centers on estimating the causal effects of an action, treatment, or intervention on a time-to-event outcome (e.g., the effects of surgery on post-operative time to death). Insights about the effects of these non-randomized treatments have the potential to help answer important questions both in population health and in individualized medicine. For instance, a physician considering whether or not to prescribe a medication to a patient despite the known side effects would benefit from information about the expected effects of treatment on longevity among patients with similar characteristics (Sox et al., 2009). In these settings, measuring time-to-event outcomes in a treated group is straightforward; time is measured from treatment initiation. However, a key obstacle for inference is the lack of observed initiation

times in the control group, without which a time-to-event outcome among the controls is not well-defined.

Prior attempts at using observational data to establish clinical guidelines include studies evaluating the effects of different medical interventions on coronary heart disease (Hernán et al., 2008; Danaei et al., 2013) and studies evaluating the effectiveness of antiretroviral therapy for reducing mortality in HIV-infected populations (Cain et al., 2011). To make valid causal inferences about clinical strategies in these settings requires researchers to formulate a research question that is relevant for decision makers, define treatment and control groups within an appropriate patient population that can be used to study this question, and adjust for observed differences between these groups that may confound the treatment effects of interest. In studies that rely on data from electronic medical records or administrative databases, these tasks are especially difficult (Byar, 1980; Levine & Julian, 2008; Freidlin & Korn, 2012). For instance, while the initiation of an intervention such as a pharmaceutical therapy is typically recorded in administrative databases for patients receiving treatment, defining the initialization time of "control" presents a philosophical challenge for causal inference. This issue is of particular concern in studies with a time-to-event outcome, which is defined relative to the time of treatment assignment, where the unknown times of assignment to control must be inferred in order to measure outcome values.

Another common concern among researchers in this domain is how to address confounding due to "treatment by indication" (Poses et al., 1995). This is based on the idea that a good clinician will prescribe a new medication or recommend a medical procedure to a patient only when there is evidence that the treatment in question is necessary (Poses et al., 1995). In these settings, treatment is typically initiated once the patient's health reaches a state such that the potential benefits to the patient's overall health offered by the candidate treatment outweigh the anticipated risks or side effects associated with treatment. As a result, patients who receive treatment over a given time period may differ systematically from patients who are untreated during that period in terms of their need for

medical care. Further, among the set of untreated patients who do not receive treatment during the observation period, it is unclear whether treatment was consciously withheld from those patients at some point during the study or if treatment was simply never considered due to lack of indications.

In this paper, we consider how to draw causal inferences about a binary treatment, which can either be initiated for a patient or withheld from the same patient at a single point in time. This time point, which we refer to as the "indication time", is the time at which a patient presents with a particular set of symptoms or pre-specified indications that necessitate clinical intervention. The indication time for any individual patient might be a function of the patient's behavior (e.g., when a patient requests a medication therapy), or might be solely determined by clinical factors. For instance, in our applied example, we consider a population of patients in the VA health system who are diagnosed with pulmonary hypertension (PH) types 2 and 3, where patients may receive medication therapy in the form of phosphodiesterase-5-inhibitors (PDE5Is) for treatment of PH-related symptoms only when it is determined that the medication may be beneficial for the patient given their current state of health. As a result, indication times may vary greatly across patients in a sample, which often induces a complicated time structure in the observed data. To accommodate this structure when attempting to draw causal inferences, researchers often focus on estimands that are defined relative to a time-varying treatment (e.g., the effect of initiating treatment now versus later). We propose an approach for treatment effect estimation that views indication times as a fixed pre-treatment covariate (i.e., a variable that is unaffected by assignment to treatment or control) representing a physiological characteristic related to the patient's health. We then condition on these indication times to construct an estimator of the causal effect of treatment versus control that is independent from the time-structure of the underlying data.

The paper proceeds as follows. In Section 3.2, we first review existing methods in this domain and describe a strategy for making causal inferences using longitudinal observational data by approximating a sequence of randomized experiments. We then present an alternative conceptualization

of the hypothetical randomized experiment that can be used to facilitate causal inferences, which is based on separating the process that governs the time of indication from the mechanism that determines the receipt of treatment versus control upon indication. In Section 3.3, we present a framework for designing comparative effectiveness studies to approximate this underlying randomized experiment and describe the formal assumptions required for inference. The core of this framework is a joint state-space model for predicting indication times as a function of longitudinal covariates, described in Section 3.4, which we use to infer the missing indication times for untreated patients. In this Section, we also propose an approach for estimating treatment effects that directly incorporates uncertainty about the inferred indication times. We then illustrate the proposed framework in Section 3.5 to study the effects of prescribing contraindicated PDE5Is for treatment of PH using administrative data from the VA health care system.

3.2 Designing comparative effectiveness studies to approximate randomized experiments

3.2.1 Background

When attempting to draw causal inferences with non-randomized data, a common first step is to characterize the hypothetical randomized experiment that could have led to the observed data (Rubin, 2010). This requires us to identify the patient population of interest and describe the indications that trigger clinical intervention within this population. Given this information, we can then attempt to precisely formulate the medical actions, treatments, or decisions that are to be compared. Also important to consider is the timing of data collection, including the times when the treatments under study were initiated or withheld from patients and the times when the outcome data were recorded. Next, in the design phase of the study, we attempt to recover a subset of the observed data that approximates this hypothetical experiment using techniques such as matching. The core idea

of matching is to find sets of treated and untreated units that are in all ways similar, except in regard to their treatment assignment (Rubin, 1973b, 2006b). If a sub-sample of units can be identified such that the covariate balance between treatment and control groups approximates the balance that one would expect under randomization, the analysis phase of the study can then proceed by applying standard inferential procedures to this sub-sample as if the data had resulted from a randomized experiment (Rosenbaum, 2002b).

While this design-based approach to covariate adjustment in observational studies has been widely applied to estimate causal effects using cross-sectional data, little work has focused on extending these methods to longitudinal settings. In fact, how to properly design and analyze longitudinal observational studies remains a point of controversy (Rubin, 2010). One challenge that commonly arises when designing longitudinal observational studies that approximate randomized experiments for the purposes of causal inference is related to the lack of a well-defined control intervention (Huitfeldt et al., 2015). For instance, in a study evaluating the efficacy of a particular medication, drug A, for treating depression, initiation of control might be marked by the receipt of an alternative medication, drug B. Alternatively, if the control intervention is defined as the decision to withhold drug A in favor of an non-pharmacological approach (e.g., self-care or psychotherapy), then receipt of assignment to control may not be available in the observed data. In these settings, unknown values of treatment assignment may be missing due to measurement errors (e.g., if receipt of the control intervention is not recorded) or may be censored due to follow-up (e.g., for patients whose treatment assignment occurs after the study has ended). Care must therefore be taken to define the control intervention and identify the corresponding control group, since simply defining the control group to be the set of all untreated patients over the course of the study may introduce an explicit source of confounding.

How to address challenges such as these is the focus of many studies in comparative effectiveness research (CER), which aims to use observational data from real clinical scenarios to inform deci-

sions about what treatments are likely to work well in practice (Marko & Weil, 2010; Armstrong, 2012). Within the CER literature, a number of methods have been proposed that attempt avoid the issue of control group identification in longitudinal settings by framing the data-generating process in terms of a sequence of randomized experiments occurring over time, where at each time treatment can either be initiated or withheld (Hernán et al., 2008; Kennedy et al., 2010; Danaei et al., 2013; Li et al., 2001). This conceptualization represents a single medical intervention of interest as a time-varying treatment. The implicit assumption within this approach is that, unlike in a classical randomized experiment, all patients will receive treatment eventually and it is rather their *times* of treatment that are a result of randomization. Accordingly, unit-level causal effects are defined by contrasting the potential outcomes that would be observed for each patient if treatment were initiated at different time points (e.g., the effects of initiating treatment now versus later on a patient's survival time). Since it is possible to observe at most one of these potential outcomes for each patient, causal inferences are therefore based on comparisons between groups of patients who received treatment at each time point with groups of patients who were not-yet-treated at that point.

Methods based on sequential randomization may be useful to practitioners interested in estimating the causal effects of delaying treatment with a single medical intervention. Inferences based on this conceptual model may also help answer questions about the optimal time to initiate some treatment in order to maximize its expected benefit to a particular patient. However, these inferences may not be appropriate in settings with treatment by indication, where the time of initiation is not under the control of the clinician. In these settings, when a patient presents with symptoms that indicate the need for medical intervention, their clinician is faced with a decision about *which* among available treatments to initiate at that time. Here, it is not sensible to hypothesize about how a patient's outcomes would have changed if their indications had presented at a different time (Angrist et al., 1996). Rather, the time of indication for treatment should be viewed as a fixed pre-treatment covariate and causal effects should be framed as contrasts of outcomes under different treatment

strategies given the observed indication times.

3.2.2 A novel conceptualization of the underlying experiment

To motivate our proposed framework, consider the following hypothetical randomized experiment with a binary treatment, where interest is in estimating the causal effects of treatment on a time-to-event outcome for patients with a particular medical condition. Here, treatment is defined as the decision to apply the intervention and the control treatment is defined as the decision to withhold the intervention. Assume that patients become eligible for inclusion in the experiment only if and when their health reaches a point where clinical intervention may be beneficial to the patient despite any associated risks or side effects (the so-called "indication time"). Upon enrollment, suppose that patients are then randomized to receive either treatment or control with probability that depends on their indication time. After randomization, suppose each patient is observed at regular time intervals until the earliest occurrence of a pre-specified event (e.g., renal failure) or the end of follow-up, whichever comes first. The outcome is an event time calculated with respect to indication time (e.g., time to renal failure, time to hospital discharge). In this idealized experiment, contrasts of outcomes between treated and control units with similar indication times will be unbiased estimates of the causal effects of interest.

This conceptualization of the underlying randomized experiment explicitly defines the control group as the set of untreated patients for whom treatment was indicated during the study period, but treatment was consciously withheld at that time ("true controls"). The remaining untreated patients are those for whom no indications for treatment were present during the study ("ineligible controls"). Because these units are not assigned to treatment or control during the study period, they are not relevant units for the purposes of causal inference. When indication times are known, as in the ideal randomized experiment described above, these ineligible controls can easily be identified and discarded. However, in practice when analyzing non-randomized data, indication times may

be censored due to follow-up or death and are often only partially observed, typically for patients receiving treatment. In addition, the probabilities of assignment to treatment over time are generally unknown and may be difficult to infer without expert domain knowledge.

3.3 Framework for conditioning on time of treatment indication

3.3.1 General notation

Consider a study with N units (patients) diagnosed with a medical condition, indexed by $i = 1, \ldots, N$. For each patient, we observe a vector of p covariate measurements collected at K discrete time periods over a fixed study period $X_{iK} = (X_{i0}, \ldots, X_{iK})$, where X_{i0} is a p-vector of baseline covariates observed at the time of diagnosis. These covariates capture characteristics of each unit or the unit's environment (such as age, gender, physiological factors, diet, medical treatments, and environmental exposures) over the course of the study. Let T_i denote the indication time of patient i, which may occur at a discrete time within the study or may be censored at the end of follow-up (i.e., $T_i \in [0, K] \cup \{c\}$). By construction, we assume that patients become eligible for treatment and are assigned to treatment or control only upon indication for treatment. Thus, treatment assignment (and the causal effect of assignment to treatment) is only well-defined for the subset of patients whose indication times fall within the specified study period. We therefore let $S_i = 1\{T_i \in [0, K]\}$ represent eligibility for treatment assignment, where $S_i = 1$ for patients whose indication times occur at some point within the study period and $S_i = 0$ for patients whose indication times are censored. Similarly, we let M_i be an indicator for the missingness of T_i with $M_i = 1$ for patients whose indication times are unobserved. Finally, let Z_i be an indicator for assignment to treatment upon indication, which equals 1 for patients assigned to treatment upon indication and equals for patients assigned to control upon indication. Unlike the classical setting of a randomized experiment, we suppose that indication times T over the study period are observed only for the treatment group

(i.e., $M_i = 0$ if and only if $Z_i = 1$ and $T_i \in [0, K]$).

3.3.2 Causal estimands and assumptions for identification

Interest is in estimating the causal effects of assignment to treatment versus control on a time-to-event outcome defined as the time from treatment indication to the time of an event of interest (e.g., death, hospital discharge, symptom remission). We denote this outcome by Y_T, where the subscript signifies that the event times are measured relative to the times of indication. Under the Rubin Causal Model (RCM; Holland, 1986), each participant has two potential outcomes, $Y_{iT_i}(0)$ and $Y_{iT_i}(1)$, which represent the outcomes that would be observed for unit i under assignment to control treatment, respectively, when assigned at the indication time T_i. Let $Y_T = (Y_T(0), Y_T(1))$ denote the complete set of potential outcomes for all units. We make the Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1980b), which asserts that there is no interference between units and no hidden forms of treatment. Under this assumption, the average treatment effect (ATE) at a single indication time is defined as:

$$\tau_T = E[Y_T(1) - Y_T(0)],$$

where $E[\cdot]$ is the expectation over all units. In longitudinal studies where indication may occur over a fixed study period $[0, K]$, we can construct an aggregate measure of these time-specific effects as

$$\tau = \frac{1}{K} \sum_{t=1}^{K} \tau_t.$$

In settings where the outcome of interest Y is defined relative to a time of death or failure, τ captures the average change in survival time under treatment compared to control for units who present with indications for treatment over the study period. Alternatively, as we will see in Section 3.5,

causal estimands can also be specified to quantify the average difference in survival rates at specified points during a follow-up period. To construct an unbiased estimator of this quantity using non-randomized data, we assume that treatment assignment is conditionally independent of the potential outcomes given all pre-treatment covariates including the indication times (i.e., $Y_{iT} = (Y_{iT_i}(0), Y_{iT_i}(1)) \perp\!\!\!\perp Z_i | X_{iT_i}, T_i$). In the clinical context we consider, this asserts that assignment to treatment versus control is unconfounded given indication times, T, and observed pre-treatment covariates, X. Under this assumption, we can obtain an unbiased estimate of the treatment effect as:

$$\hat{\tau} = \frac{1}{N_1} \sum_{Z=1} Y_T(1) - \frac{1}{N_0} \sum_{Z=0} Y_T(0), \tag{3.3.1}$$

where $N_1 = \sum_{i=1}^{N} S_i Z_i$ and $N_0 = \sum_{i=1}^{N} S_i(1 - Z_i)$ are the numbers of treated and control units who are eligible for analyses, respectively. Here, in contrast to the classical setting, the outcomes that are actually observed for each unit depend not only on their treatment assignment but also on their indication time. For units whose indication times are not observed, these missing times must be inferred in order to calculate observed values of Y_T needed in 3.3.1. In the section below, we develop a strategy to infer the missing times of indication that can be used to facilitate inference in this setting.

3.3.3 Overview of inferential approach

Our conceptual framework implies by design that the missingness of the times of treatment assignment, T^{mis}, and the indicators for assignment to control, $Z = 0$, are completely dependent on the values of those missing measurements. That is, we assume that the data (T_i, Z_i) for a unit i will be missing if $T_i > K$ (regardless of the value of Z_i) or if $Z_i = 0$ (regardless of the value of T_i). This means that the missing measurements indicated by $M_i = 1$ are "missing not at random" (MNAR; Imbens & Rubin, 2015). By inferring the missing times of assignment for the untreated units, we can

minimize the information loss that arises from the missing data mechanism in order to make more precise inferences from the observed data.

The first goal is therefore to build a model for predicting the observed indication times T^{obs} based on baseline and time-varying covariates X, which we will use to infer the missing indication times. This model will also induce a probability distribution on the potential outcomes in the control group, since the observed outcomes must be calculated relative to the indication times. By applying the assumptions described above in Section 3.3.2, we can separate the joint distribution of the complete data (including all observed potential outcomes Y_T^{obs} as well as both the observed and unobserved indication times, $T = (T^{obs}, T^{mis})$ given some global parameter θ with partition $\theta = (\theta_1, \theta_2, \theta_3)$ as:

$$p(Y_T^{obs}, T, Z|X, \theta) = p(Y_T^{obs}|T, Z, X, \theta_1)p(Z|T, X, \theta_2)p(T|X, \theta_3). \tag{3.3.2}$$

For Bayesian inference with prior distribution $p(\theta) = p(\theta_1, \theta_2, \theta_3)$ on θ, the posterior distribution of θ given the complete data with is:

$$p(\theta|Y_T^{obs}, T, Z) \propto p(\theta_1, \theta_2, \theta_3)p(Y_T^{obs}|T, Z, X, \theta_1)p(Z|T, X, \theta_2)p(T|X, \theta_3). \tag{3.3.3}$$

Posterior inference on θ can then proceed by straightforward application of Markov Chain Monte Carlo (MCMC) techniques, such as the Gibbs sampler (Geman & Geman, 1984; Gelman et al., 2014). For example, in each iteration of the Gibbs sampler, we draw the missing indication times T^{mis} from the conditional posterior predictive distribution of T^{obs} given covariates X and the current draw of the parameter θ.

The complete indication times can then be used to classify untreated patients into distinct groups of true controls and ineligible controls based on eligibility S, where the true control group consists of patients with $M = 1$ and $S = 1$. For the true controls, we can then calculate values for the po-

tential outcomes $Y_T(0)$ given the generated values of T. These values are then regarded as observed potential outcomes, denoted Y_T^{obs}, which are equal to the calculated $Y_T(0)$ for units classified as true controls and equal to $Y_T(1)$ for treated units. Given the observed potential outcomes Y_T^{obs}, the complete times of indication $T = (T^{obs}, T^{mis})$ and the corresponding assignment vector Z, we can then update the parameters θ_1, θ_2 and θ_3 by drawing from the joint conditional posterior:

$$p(\theta_1, \theta_2, \theta_3 | Y_T^{obs}, T, X, Z) \propto p(Y_T^{obs} | T, Z, X, \theta_1) p(Z | T, X, \theta_2) p(T | X, \theta_3) p(\theta_1, \theta_2, \theta_3). \quad (3.3.4)$$

For posterior inference on the causal effects of interest, we can continue this sampling procedure after approximate convergence in distribution. Thus, in each iteration, we construct a dataset consisting of all observed indication times, the simulated indication times, and all observed potential outcomes, and then use this completed data to calculate an estimate of the treatment effect as in Equation 3.3.1. Alternatively, we could specify a joint distribution for the potential outcomes $Y_T = (Y_T(0), Y_T(1))$ that we could then use to impute the missing potential outcomes Y_T^{mis} for each patient in each iteration by drawing from the conditional distribution with density function $p(Y_T^{mis} | Y_T^{obs}, T, X, \theta)$. Repeating this process over many such simulated datasets produces the approximate posterior distribution for all causal effects of interest. In the same way, posterior samples of θ can provide posterior estimates of the parameters that characterize the data-generating process; this is described in greater detail in Section 3.4.

3.4 State-space model for time of treatment indication

3.4.1 Model formulation

Based on the conceptual framework presented in Section 3.3, we propose a specific and pragmatic model for predicting the time of indication for treatment - the earliest time at which a patient presents with indications for clinical intervention - as a function of both fixed and time-varying covariates. In particular, we hypothesize that observed covariate measurements that reflect worsening health and diminished functional capacity will be predictive of indication times. Similarly, we assume that covariates capturing provider characteristics or temporal features (e.g., month or year when measurements are recorded) are independent of the indication times but may influence the probability of assignment to treatment upon indication. For instance, in the application presented in Section 3.5, we expect the probability of treatment to decline over time as clinicians become more informed about populations where the medication of interest may be contraindicated. We capture these separate dependencies through separate components of the hierarchical model; the first component characterizes the longitudinal process that governs patients' indication times, and the second component describes the conditional probability of assignment to treatment given those indication times. Specifically, we model a patient-level health process generated by time-varying covariates together with random fluctuations over time and adopt a so-called "threshold approach" (Albert & Chib, 1993), which views the indication time as the first hitting time (FHT) of this latent process. Similarly, we assume that the conditional probability of assignment to treatment versus control given the indication time for each patient varies based on institutional preferences, which may systematically change over time with widespread changes in the established guidelines for treatment.

Suppose that the time series of covariate measurements for each unit, X_{iK}, is independent from the measurements of other units, and let $\theta_{i,1:K} = (\theta_{i1}, \ldots, \theta_{iK})$ be a state variable representing fluctuations in unit i's overall health over the course of the study that are not explained by the covari-

ates. We assume a standard normal distribution for the daily health fluctuations for unit i between days t and $t-1$, such that

$$\theta_{it} = \rho\theta_{i,t-1} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathcal{N}(0,1), \tag{3.4.1}$$

where $\mathcal{N}$ denotes the normal distribution. We assume $|\rho| < 1$ (i.e., the latent state process $\theta_{1:T}$ follows a stationary autoregressive model of order one). To initialize this process, we assume a standard normal prior distribution for θ_0. Here, the stochastic component ε_{it} captures the unexplained variation of unit i's health over time. Conditional on the latent states $\theta_{1:K}$, we then define the observation process $\Psi_{1:K}$, which relates the overall health trajectory of unit i to their indication time according to a probit regression model as

$$P(\Psi_{it} = 1|\theta_{it}, X_{it}) = \Phi(\theta_{it} + X_{it}\beta), \tag{3.4.2}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal random variable and $\beta \in \mathbb{R}^p$ is a vector of regression coefficients. Under the latent variable representation of 3.4.1 and 3.4.2 (also referred to as a state-space mixed model; Czado & Song, 2008), the indication time for each unit i can be expressed as

$$T_i = \inf\{t \in [0,K] : \Psi_{it} = 1\}. \tag{3.4.3}$$

Note that although T is a continuous random variable by nature, the observed times of assignment to treatment $T^{obs} \in \{0,1,\ldots,K\}$ are discrete and interval-censored due to measurement times. Thus, our proposed model considers a discretized form of the assignment times. The model for indication times T induces a conditional distribution for the time of indication given a vector of longitudinal, pre-treatment covariates. One of the main advantages of this latent state represen-

tation is that it can flexibly accommodate both fixed and time-varying covariates. The proposed modeling approach can also be easily extended to settings with non-linear covariate effects by re-expressing equation 3.4.2 as a generalized linear model with link function $g(\cdot)$.

We consider a separate model for the assignment mechanism, which determines assignment to treatment versus control upon indication. Here, our model formulation is based on the assumption that, in many settings, variation in treatment practices may be due to clinician and/or institutional preferences rather than differences in patient characteristics (Slaughter et al., 2017). In particular, we assume that each unit is assigned to treatment with a probability that depends on their indication time. Conditional on T, the assignment mechanism can be expressed as

$$Z_i|T_i \sim Bernoulli(\pi_{iT}), \quad \text{logit}(\pi_{iT}) = \delta_0 + \delta_1 f(T_i), \tag{3.4.4}$$

where $f(T)$ is a deterministic transformation of the indication time (e.g., year or month). This model can be easily extended to accommodate nonlinear effects or other artifacts of the study that are believed to influence the probability of treatment assignment. For example, in Section 3.5, we require that δ_1 is strictly positive such that the probability of receiving the treatment is monotonically decreasing over time.

3.4.2 Inference procedure

Our estimation procedure uses the Kalman Filter to marginalize out the latent state parameters $\theta_{1:T}$ for more efficient estimation. The full likelihood of the parameters $\Omega = (\Psi_{1:K}, \rho, \beta, \delta_0, \delta_1)$ can be written as

$$\mathcal{L}(\Omega) = \prod_{i=1}^{N} \Big(p(T_i^{obs} = t|\Omega) p(Z_i = 1|T_i^{obs} = t, \Omega) \Big)^{1-M_i} (p(M_i = 0|\Omega))^{M_i} \tag{3.4.5}$$

where

$$p(M = 0|\Omega) = 1 - \sum_{t=0}^{K} p(T_i^{mis} = t|\Omega)p(Z_i = 1|T_i^{mis} = t, \Omega).$$

The associated posterior density is therefore

$$p(\Omega|T^{obs}, Z, X, M) \propto p(T^{obs}|X, \Omega)p(Z|T^{obs}, \Omega)p(M = 0|\Omega)p(\Omega).$$

We assume the prior distribution as

$$\begin{aligned} p(\Omega) &= p(\rho, \beta, \delta_1, \delta_0, \Psi_{1:T}) \\ &= p(\rho)p(\beta)p(\delta_1)p(\delta_0)p(\Psi_{1:T}|\rho, \beta) \end{aligned}$$

with

$$\begin{aligned} p(\rho) &\sim \mathcal{N}_{[-1,1]}(\rho_0, \sigma_\rho) \\ p(\beta) &\sim \mathcal{N}_p(\beta_0, \Sigma_0) \\ p(\delta_0) &\sim N(a, b) \\ p(\delta_1) &\sim Gamma(c, d) \\ p(\Psi_{1:T}|\rho, \beta) &= p(\Psi_1) \prod_{j=2}^{T} p(\Psi_j|\Psi_{j-1}, \rho, \beta) \end{aligned}$$

Here, the last equation is derived by standard application of the Kalman filter (Carlin & Polson, 1992), which allows for more efficient evaluation of the marginal log-posterior of $p(\rho, \beta|\Psi_{1:K})$. Incorporating the Kalman filter also allows us to draw samples $\Psi_{1:K}^{\star}$ from the marginal posterior

distribution of $p(\Psi_{1:K}|T^{obs}, Z = 1, M = 0)$. Both distributions are straightforward to evaluate using standard software for implementing MCMC methods such as JAGS (Plummer et al., 2003). As previously described, these samples allow us to measure the missing times of treatment assignment and assignment to control according to 3.4.3, where $T^{mis} = c$ for units with $\Psi_{1:T} = 0$. Given the inferred indication times, our framework implies that $Z_i = 0$ for units with $T_i^{mis} \in [0, K]$. Recall that treatment assignment is undefined for those whose time of treatment assignment was censored by the end of the study (i.e. units with $T_i^{mis} = c$). Finally, we can calculate the observed outcomes for units inferred to belong to the group of true controls by computing the difference between the observed event times and the inferred indication times for those units.

3.4.3 Bayesian analysis with inferred indication times

One of the benefits of the proposed approach for estimating treatment effects in this setting is that it allows us to make inferences about the missing indication times that are free from the outcome analysis. Our approach allows for flexible specification of the causal estimands of interest and also allows researchers to choose any mode of inference for analysis of the outcomes that they see appropriate.

For example, one might use the posterior mode of inferred times of indication for each unexposed unit calculated over a large number of MCMC samples as the point estimate for that unit's indication time. This can then be viewed as a single imputation of the missing values, and conditional on these estimates one could estimate the treatment effects by simple comparisons of means of the time-to-event outcomes (e.g., using a Neymanian or Fisherian mode of inference). Alternatively, our framework also allows for more sophisticated analysis of outcomes. For instance, we could first obtain a large number of posterior samples for the missing times of indication across for all untreated units and use these samples to form unit-level empirical posterior distributions of the missing indication times. Then, one might iteratively impute values from these distributions and calculate the corresponding observed potential outcomes in each iteration. Taking the difference in

means between treatment and control groups across all iterations would then produce a distribution of treatment effect estimates (given the potential outcomes) that incorporates uncertainty about the missing indication times.

Important to note is that this type of approach is only partially Bayesian in that we are using Bayesian methods to impute missing values that we need in order to measure the observed outcomes. To make this a fully Bayesian approach, as previously mentioned, one could also specify imputation model for the missing potential outcomes (i.e., the missing time-to-event outcomes under alternative assignment) conditional on the partially observed outcomes and the assignment vector.

3.5 Application

To illustrate our proposed methods, we analyzed data from a recent study evaluating the impacts of inappropriate prescribing practices for the treatment of pulmonary hypertension using electronic medical records from the VA health system (Freiman et al., 2015). Pulmonary hypertension is a condition of high blood pressure that affects arteries in the lungs and heart. One common treatment for PH is a class of medications called phosphodiesterase-5-inhibitors (PDE5Is), which act on enzymes causing blood vessels to relax in order to lower blood pressure. While PDE5Is have been shown to be effective for treating some rare forms of PH, Freiman et al. (2015) identified the use of these drugs as wasteful, ineffective, and potentially harmful for treating patients with more common types of PH caused by left heart disease (Group 2) or hypoxemic lung disease (Group 3). Despite its contraindication for patients with these types of PH, a study of veterans diagnosed with these types of PH over the years of 2005 to 2012 identified over 2,000 prescriptions for PDE5Is that were inappropriately administered to patients in the VA health system. To understand the impact of these inappropriate treatment practices on patient outcomes, it is of interest to measure the causal effects of prescribing PDE5Is to patients with PH Groups 2 and 3 on the time lag between the application of the inter-

vention and the occurrence of a clinical event of interest (e.g., time from treatment to acute renal failure).

3.5.1 Data

The data contains demographic and laboratory measurements as well as records of the utilization of medications, inpatient and outpatient services for over 350,000 veterans who were diagnosed with PH types 2 and 3 and received prescription medications from the VA from 2005 to 2016. For all patients, health-related measurements were collected at the time of PH diagnosis as well as at intermittent observation times corresponding to patient-provider interactions within either the VA health system or Medicare (e.g., an inpatient or outpatient visit). The exact number of measurements recorded and the time elapsed between subsequent measurements varied by patient. Observations with implausible values for lab measurements or demographic variables were excluded prior to analysis (e.g., variables measured as proportions were constrained to the simplex).

For the present analysis, we considered only male patients who at the time of PH diagnosis were between 65 to 95 years of age, were eligible for Medicare benefits, and who had not received prescriptions for a PDE5I medication prior to PH diagnosis. For data integrity purposes, we further excluded any patients who were receiving Medicare part C at the time of diagnosis, since these patients may have received PH related care from private providers. We also excluded patients who did not receive any prescriptions within one year prior to their diagnosis. Finally, we excluded any patients who received a prescription for nitrates after PH diagnosis and prior to their first PDE5I prescription, since nitrates are contraindicated for treatment with PDE5Is. Outcomes for these patients therefore reflect off-label use of PDE5Is and cannot be used to make causal inferences about the effects of treatment when administered appropriately.

Among patients who met all initial eligibility criteria, we defined the "treatment group" as the set of patients who filled at least one prescription for a PDE5I medication (PHP5I with 15+ pills) during

the study period. Because we observed the prescription dispense dates rather than the dates when the medications were first prescribed, we excluded from our analyses any treated patient whose first dispense date occurred more than 60 days after their preceding hospital visit. After applying all exclusion restrictions, the remaining sample was comprised of 534 patients who received treatment for PH with a PDE5I medication within a one-year period following their diagnosis date and 167,701 potential controls who did not receive a PDE5I during that period.

Our outcome of interest is survival time in the period following indication for treatment, which we observed for treated units and must be inferred for units in the control group. We consider the date of PH diagnosis as the time of "earliest eligibility" for indication. For each potential control patient, we base our inferences on clinical data observed at intermittent intervals from the time of earliest eligibility until death or the end of follow-up in December 2016, whichever occurred first.

3.5.2 Constructing comparison groups and time-varying covariates

To make the assumptions of our proposed framework described in Section 3.3 more plausible in this setting, our first task was to select a set of potential control units who appeared similar to the treatment group at baseline based on health-related measurements collected in the 1 year prior to PH diagnosis (excluding the date of diagnosis). Here, we considered a set of 17 baseline covariates identified as predictive of the time of assignment to treatment within the treatment group. See Appendix C.2.1 for details of this variable selection procedure. Potential controls were selected using pairwise Mahalanobis distance matching with replacement (Rubin, 1978b; Gu & Rosenbaum, 1993), whereby each treated unit is matched to its closest control unit based on the Mahalanobis distance calculated over baseline covariates. This produced a final sample of 534 treated units and 531 matched control units who were similar to the treatment group at their times of PH diagnosis but did not receive a PDE5I at any point during the observation period. Diagnostics performed after matching confirmed that the covariate distributions were adequately balanced between the treatment group

and the matched potential control group. Table 3.1 shows descriptive statistics on baseline variable for the final matched samples. After matching, we proceeded under the assumption that overall health status is unconfounded given baseline covariates such that patients with similar covariates at baseline can be expected to have similar health trajectories and therefore, similar indication times over the course of study.

Table 3.1: Baseline covariate data for potential controls and treatment group

	Potential Control	Treatment
PH Group 2	88%	88%
Recently Hospitalized	13%	12%
Recent Procedure	65%	75%
Recent ER Visit	52%	53%
Age	74.8	74.5
Weight	203.3	204.5
Height	69.5	69.5
Resting Heart Rate	74.2	75.1
Systolic Blood Pressure	130.1	130.5
Diabolic Blood Pressure	71.2	71.2
Inpatient Days	1.19	1.22
Outpatient Days	27.3	27.9
Number of Comorbidities	0.52	0.54
Cardiac Events	1.60	1.49
Pulmonary Events	0.61	0.69
Organ Failure Events	1.09	1.16
Number of Medications	11.4	11.8

In addition to baseline covariates, we also included in our analyses a number of time-varying covariates collected at intermittent observation times throughout the study. Specifically, we considered six time-varying covariates that indicate changes in binary variables over time, including an indicator for whether the patient was most recently observed in an outpatient versus inpatient setting, an indicator for the presence of new comorbidities, and separate indicators for recent hospitalization, organ failure events, cardiac events or pulmonary procedures recorded during the 30 days prior

to each visit. We also include as a single time-varying covariate the Mahalanobis distance calculated between each patient's laboratory measurements at baseline (e.g., heart rate and blood pressure) and values of the same laboratory variables collected at each follow-up visit. This allowed us to greatly reduce the dimension of the covariate space and operationalize the laboratory variables in terms of gain scores.

We observed a strong positive correlation between the empirical probabilities of survival and indication times for patients in the treatment group. Among 77 patients who receive treatment within 7 days following their PH diagnosis (i.e., patients with $T_i \leq 7$), only 53 patients (68.8%) were alive one year post-diagnosis. Averaged over all 534 patients whose indication times occur within one year of diagnosis (i.e., the treatment group), this survival rate increases to 80.6%. In contrast, the survival rate at one-year after PH diagnosis for the 531 matched potential controls is approximately 81.7%; however, because we do not observe the indication times for these patients, this is a naive estimate that is likely negatively biased. Also note that these survival probabilities are defined relative to *diagnosis times*, which we regard as a fixed pre-treatment covariate. As a result, these measurements are purely descriptive and should not be regarded as treatment effects.

3.5.3 Results

Our primary objective in this application is to estimate the causal effects of inappropriate prescribing practices on survival outcomes among veterans receiving treatment for PH. To accomplish this, our analysis is based on the crucial assumption that patients present with indications for treatment at times that vary as a function of their overall health. We also assume that as soon as a patient is indicated for treatment, they will consult with their health-care provider to decide to either initiate the treatment of interest or withhold it in favor of an alternative therapy. Using the approach described in Section 3.3, we fit the model defined by equations (3.4.1)–(3.4.4) using MCMC posterior simulation. We constrained the parameter δ_1, which captures the temporal component of the probability

of assignment to treatment versus control, to be negative for this application since the use of PDE5Is for treatment of PH was believed to be monotonically decreasing over the course of the study. This is because PDE5Is were the standard, first-line therapy for treatment of PH at the beginning of the study with use steadily decreasing as knowledge of its contraindication for some PH patients spread throughout the medical community.

To estimate the treatment effects of interest, we fit eight distinct models, each corresponding to indication times occurring within 1-14 days, 15-30 days, 31-60 days, 61-90 days, 91-120 days, 121-180 days, 181-270, and 271-365 days after PH diagnosis. For each interval, we estimated the causal effects of treatment versus control on survival at follow-up one year after indication (e.g., the conditional average treatment effect for patients whose indication times occur within 1-14 days following PH diagnosis). To estimate these effects, we first identified the subset of units whose indication times (either observed or inferred) were within the specified study period based on the current values sampled using MCMC. Within this subset, we then calculated survival rates at one year post-indication for treated units as well as for units classified as true controls. Finally, we measured the treatment effect as the difference between these two proportions. For each of the specified study periods, we ran the MCMC sampler with four parallel chains each run for 20,000 iterations, where the first 5000 draws of each chain were discarded as a burn-in period. With the resulting 60,000 samples, we calculated the posterior means and 95% credibility intervals (CI) for all model parameters. In all cases, the MCMC simulated model parameters and quantities of interest passed the convergence test of Geweke (1992) and Gelman et al. (1992). As an additional sensitivity check, we evaluated the performance of the proposed model in each setting under different choices of hyper parameters using the deviance information criterion (DIC; Spiegelhalter et al., 2002) and found the results to be generally robust.

Posterior means and 95% credible intervals for the treatment effects of interest are presented in Table 3.2. Here, the causal estimand at each time point is defined as the average effect of treatment

on one-year survival for all patients with indication times occurring at or before that time (i.e., $\tau_K = E[Y_i(1) - Y_i(0)|T_i \leq K]$ for $k = 14, 30, 60, 90, 120, 180, 270, 365$). Table 3.2 also summarizes the cumulative number of potential control units identified as "true controls" at each time point based on the posterior median of the 60,000 MCMC samples (i.e., $\hat{N}_{0K} = \sum_{Z_i=0} 1\{\hat{T}_i \leq K\}$).

Table 3.2: Median number of true control units inferred from matched sample of potential controls at each time point and estimated effects of treatment compared to control with 95% posterior intervals.

Study period	Number treated	Inferred number of "true controls"		One-year survival after indication		Estimated impact of PDE5I	
(K)	(N_1)	$\hat{N}_{0K}$	95% CI	Treatment	Control	$\hat{\tau}_K$	95% CI
14 Days	109	48	(34, 62)	71.6%	82.0%	-10.4%	(-18.6%, -1.6%)
30 Days	164	88	(72, 103)	73.2%	82.8%	-9.6%	(-14.5%, -3.8%)
60 Days	233	113	(96, 127)	75.1%	83.3%	-8.2%	(-12.8%, -5.0%)
90 Days	280	127	(111, 141)	72.5%	83.3%	-10.8%	(-14.7%, -6.8%)
120 Days	318	134	(119, 147)	73.0%	85.1%	-10.1%	(-14.1%, -6.4%)
180 Days	380	150	(135, 162)	73.4%	84.3%	-10.9%	(-14.1%, -7.5%)
270 Days	456	165	(151, 180)	73.9%	84.5%	-10.6%	(-13.9%, -7.5%)
365 Days	534	183	(170, 195)	72.3%	84.5%	-12.2%	(-14.6%, -9.8%)

These results suggest that the majority of the matched potential controls were ineligible to receive treatment during the one year period following their PH diagnosis. For these patients, lack of treatment can therefore be interpreted as a lack of *indication* for treatment. On the other hand, potential controls with inferred indication times occurring at or before each specified time point are regarded as "true controls" for whom, upon indication for treatment, PDE5Is were actively withheld, possibly in favor of alternative medication or treatment strategy. Given the inferred times of indication, these patients provide a credible comparison group with which we can make causal inferences. In particular, our findings indicate that among patients with PH Types 2 and 3 who are indicated for treatment at any point in the one year following their diagnosis, treatment with PDE5Is has a large, negative effect on one-year survival probability. Figure 3.1 shows the effects of treatment over a one

year period following PH diagnosis calculated using a smoothing spline (Marsh & Cormier, 2001), which we specified with eight knots located at each posterior mean treatment effect estimated using our model.

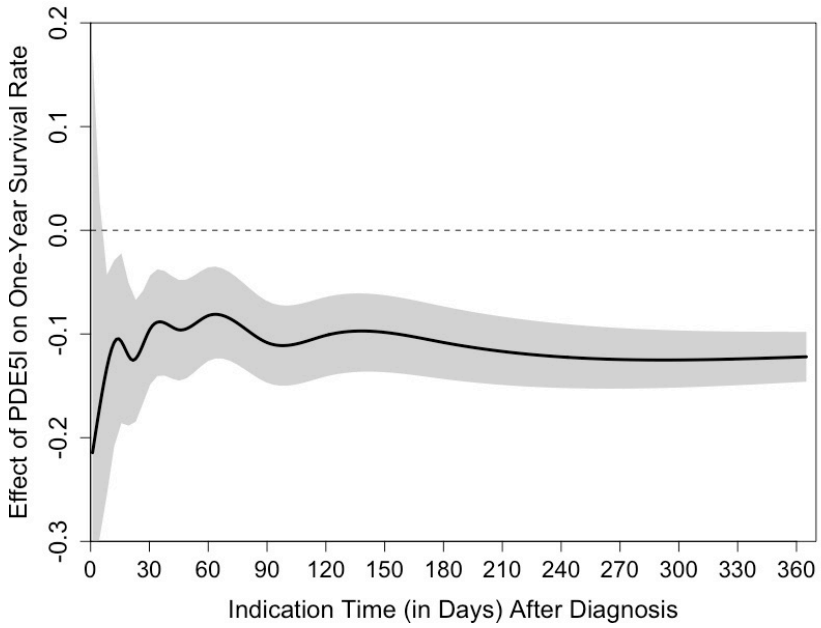


Figure 3.1: Spline curve for estimated effects of PDE5I compared to control on one-year survival rate based on time from diagnosis to indication. Shaded area shows 95% posterior credibility interval.

Table 3.3 shows posterior estimates for the correlation between patients' latent health states, ρ, and parameters governing the probability of assignment to treatment upon indication, δ_0 and δ_1. Here again, our results are largely consistent across the settings with credible intervals shrinking as study period widens. In general we find a small negative correlation between latent health measurements, which decreases in magnitude over time. This suggests there may be some variation in patients' overall health that is not explained by the observed covariates in the time period immediately following PH diagnosis, but this systematic variation dissipates over time. Posterior estimates

for parameters of the treatment assignment mechanism are also relatively stable across the different study periods. Baseline probabilities of treatment upon indication range from $\pi_1 = 0.79$ for the study period of 14 days to $\pi_1 = 0.73$ for the study period of 365 days, with probability of treatment slowly decreasing over time.

Table 3.3: Posterior medians with 95% credibility intervals for parameters ρ δ_0, and δ_1 in each study period.

	ρ	δ_0	δ_1
14 Days	-0.34 (-0.67, -0.06)	1.10 (0.47, 1.82)	-0.05 (-0.15, 0.00)
30 Days	-0.21 (-0.34, -0.21)	0.76 (1.1, 0.76)	-0.02 (-0.08, 0.00)
60 Days	-0.15 (-0.31, 0.02)	0.87 (0.54, 1.29)	-0.02 (-0.08, 0.00)
90 Days	-0.10 (-0.22, 0.04)	0.93 (0.62, 1.27)	-0.02 (-0.07, 0.00)
120 Days	-0.05 (-0.17, 0.10)	0.99 (0.69, 1.35)	-0.02 (-0.07, 0.00)
180 Days	-0.05 (-0.17, 0.05)	1.00 (0.71, 1.33)	-0.02 (-0.07, 0.00)
270 Days	-0.07 (-0.14, -0.03)	1.10 (0.83, 1.32)	-0.01 (-0.05, 0.00)
365 Days	-0.05 (-0.13, 0.01)	1.00 (0.75, 1.29)	-0.02 (-0.05, 0.00)

A key feature of our model is that it allows us to directly evaluate which of the baseline and time-varying covariates carry more or less information about patients' times of indication. In the present study, posterior inferences for covariate effects within each of the study periods were similar. In general, our findings suggest that patients with PH Type 2 were more likely to have indications for treatment shortly after PH diagnosis than patients with PH Type 3. Results also indicate that receipt of one or more incidental medical procedures (e.g., cardiac surgery or pulmonary function testing) within 30 days prior to PH diagnosis is strongly associated with earlier indication times. Further, we

found that patients who regularly receive care in an inpatient setting generally have earlier indication times than patients whose follow-up visit typically occur in an outpatient setting. Among the other baseline covariates included in our analysis, receipt of one or more pulmonary disease events (e.g., pneumonia) within 30 days prior to PH diagnosis and number of comorbidities present at baseline were also positively associated with indication for treatment during the study period. These results may offer insights for clinicians about best practices for health management of PH patients, and may also be used to guide modeling decisions in other applications.

3.6 Discussion

In this paper, we propose a novel conceptualization of longitudinal observational studies with treatment by indication that can be used to design comparative effectiveness studies that approximate a single randomized experiment with a binary treatment. In studies with time-varying exposures, which are typically analyzed using marginal structural models or risk-set matching, the conceptual framework we present offers an alternative formulation of the underlying causal problem that is both intuitive and relatively straightforward to implement in our setting. This conceptualization reformulates the problem of evaluating a time-varying treatment in the presence of time-varying confounders as one of confounding due to sequential enrollment. Our hope is that this simplified representation of a traditionally complex data structure will allow for more straightforward analyses of health data in the digital age (e.g., data from electronic medical records).

The merit of the proposed model is that it allows for model-based assessments of the times of treatment indication. Our approach allows for systematic evaluation of the underlying health factors that may be most influential in determining a patient's need for treatment. Inferences based on this modeling strategy may be useful to address a number of important questions in health services research - for example, what preventative health strategies might be most effective for delaying the

onset of indications for treatment? The model also allows us to obtain conditional probabilities for a subject having indications for treatment at different points over time, which could be used to inform preventative treatment strategies. We note that the proposed model assumes that subjects' overall health fluctuates under natural conditions.

As described in Section 3.4, the proposed model for time of indication for treatment can accommodate both fixed and time-varying covariates, which can be useful for explaining differences in the aspects of health associated with subject-specific characteristics and/or conditions that vary between hospital visit within a subject. Alternatively, covariate data could be excluded entirely from the model to make inferences about indication times that are viewed as fully stochastic. Another possibility for modeling variability in the parameters involves the inclusion of facility-based, provider-level, or geographic-specific random effects. However, this complicates the calculation of the log-likelihood considerably, rendering model fitting more challenging.

Finally, we note that our application and inferential results should be regarded as an illustrative example of the proposed methods rather than an attempt to provide definitive answers about the causal effects of inappropriate prescribing practices on health-related outcomes for PH patients.

A

Supplemental Materials for Chapter 1

A.1 A special case of equivalence for the two components of SUTVA

Consider a randomized experiment with a binary treatment. For an experiment with N units with a specified assignment mechanism (e.g., a Bernoulli experiment), denote the set of all possible randomizations across N units as $\mathcal{Z}$. For a particular random assignment $Z \in \mathcal{Z}$, suppose the assumption of no interference is violated such that the potential outcomes for a given unit depend

on the treatment assignments of the other units in the experiment (i.e., $Y_i(Z) = Y_i(Z_i, Z_{-i}) \neq Y_i(Z_i)$ for unit i). First suppose that there is only interference between unit i and unit j such that $Y_i(Z_i, Z_{-i}) = Y_i(Z_i, Z_j)$ and that this interference is one-way (i.e. $Y_j(Z) = Y_j(Z_j)$. Let $F = f(Z)$ be a binary vector indicating the presence of interference under the given assignment Z and define $\mathcal{F} = \{f(Z) : Z \in \mathcal{Z}\}$ to be the set of all possible interference vectors defined by the assignment mechanism. Thus, for any particular assignment vector Z, by randomly assigning each unit to either treatment or control, we are also implicitly assigning to units to a particular level of interference F. We can define the treatment levels accordingly by crossing the receipt of treatment versus control (2 levels) with receipt of interference versus no interference (2 levels). Crossing these factors produces four general levels of treatment: 1) control + no interference, 2) control + interference, 3) treatment + no interference, 4) treatment + interference. Let $Z^\star$ denote one such assignment of N units to a combination of treatment and interference. Then in setting described above, $Y_i(Z^\star) = Y_i(Z_i^\star)$. Under this formulation, the assumption of no interference is satisfied. However, because we cannot observe receipt of interference in a randomized experiment, two of the four treatment levels will be hidden.

For example, suppose interference is only present under assignment to control such that $Y_i(0, Z_j = 0) \neq Y_i(0, Z_j = 1)$.Let 0^C denote assignment to control given that $Z_j = 0$ and 0^T denote assignment to control given that $Z_j = 1$. Note that in the absence of interference, assignment to 0^C is equivalent to assignment to 0^T. Using this notation, we can re-express the treatment assignment for all units as $Z^\star \in \{0, 0^C, 0^T, 1\}$. Here, $Y_i(Z^\star) = Y_i(Z_i^\star) \in \{Y_i(0^C), Y_i(0^T), Y_i(1)\}$ for unit i and $Y_k(Z^\star) = Y_k(Z_k^\star) \in \{Y_k(0), Y_k(1)\}$ for all $N - 1$ other units. Thus, the assumption of no interference between units is satisfied under $Z^\star$, but the assumption of no hidden forms of treatment is now violated because it is only possible to observe $Z = 0$ or $Z = 1$.

This formulation can be easily extended to capture different degrees of interference across units (e.g., defining multiple levels of interference corresponding to no interference, mild interference,

high interference). In particular, we could define $f(Z)$ continuously, where F_i represents the impact of interference on the potential outcomes of unit i under the random assignment of all units defined by Z.

B

Supplemental Materials for Chapter 2

B.1 Text Representations and Distance Metrics

In Section 2.3 we describe a framework for text matching involving choosing both a text representation and a distance metric; we then briefly outline the options for each. Here we expand that discussion.

B.1.1 Choosing a representation

To operationalize documents for text matching, we must first represent the corpus in a structured, quantitative form. There are two important properties to consider when constructing a representation for text with the goal of matching. First, the chosen representation should be sufficiently low-dimensional such that it is practical to define and calculate distances between documents. If a representation contains thousands of covariates, calculating even a simple measure of distance may be computationally challenging or may suffer from the curse of dimensionality. Second, the chosen representation should be meaningful; that is, it should capture sufficient information about the corpus so that matches obtained based on this representation will be similar in some clear and interpretable way. As discussed in Section 2.2, text matching is only a useful tool for comparing groups of text documents when the representation defines covariates that contain useful information about systematic differences between the groups.

In this paper, we explore three common types of representations: the term-document matrix (TDM), which favors retaining more information about the text at the cost of dimensionality, statistical topic models, which favor dimension reduction at the potential cost of information, and neural network embeddings, which fall somewhere in between. There are a number of alternative text representations that could also be used to perform matching within our framework, including other representations based on neural networks (Bengio et al., 2003) or those constructed using document embeddings (Le & Mikolov, 2014; Dai et al., 2015), but these are left as a topic for future research.

Representations based on the term-document matrix

Perhaps the simplest way to represent a text corpus is as a TDM. Under the common "bag-of-words" assumption, the TDM considers two documents identical if they use the same terms with the same frequency, regardless of the ordering of the terms (Salton & McGill, 1986). When match-

ing documents, it is intuitive that documents that use the same set of terms at similar rates should be considered similar, so the TDM provides a natural construction for representing text with the goal of matching. However, the dimensionality of a standard TDM may give rise to computational challenges when calculating pairwise distances between documents in some corpora. There are many dimension-reduction strategies that can be applied to help mitigate this issue including techniques based on matrix rescaling using a scheme such as TF-IDF scoring (Salton, 1991), and techniques for bounding the vocabulary to eliminate extremely rare and/or extremely common terms. However, it should be noted that in large corpora, a bounded and rescaled TDM may still have a dimension in the tens of thousands, setting known to be difficult for matching (Roberts et al., 2018).

Representations based on statistical topic models

An alternative representation for text, popular in the text analysis literature, is based on statistical topic models (Blei, 2012), e.g., LDA (Blei et al., 2003) and STM (Roberts et al., 2016a). The main argument for matching using a topic-model-based representation of text is that document similarity can adequately be determined by comparing targeted aspects of the text rather than by comparing the use of specific terms. That is, topic-model-based representations imply that two documents are similar if they cover a fixed number of topics at the same rates. Topic models provide an efficient strategy for considerably reducing the dimension of the covariates while retaining all information that is relevant for matching. In contrast to the tens of thousands of covariates typically defined using a representation based on the TDM, representations built using topic models typically contain no more than a few hundred covariates at most. However, consistent estimation of topic proportions is notoriously difficult due to issues with multimodality of these models, which gives rise to a number of issues for applications of matching in practice (Roberts et al., 2016b).

Representations based on neural network embeddings

Mikolov et al. (Mikolov et al., 2013) introduce a neural network architecture to embed words in an $n-$dimensional space based on its usage and the words which commonly surround it. This architecture has proven remarkably powerful with many intriguing properties. For example, it performs very well in a series of "linguistic algebra" tasks, successfully solving questions like "Japan" − "sushi" + "Germany" = "bratwurst."

Propensity scores

When matching in settings with multiple covariates, a common technique is to first perform dimension reduction to project the multivariate covariates into a univariate space. A popular tool used for this purpose is the propensity score, defined as the probability of receiving treatment given the observed covariates (Rosenbaum & Rubin, 1983b). Propensity scores summarize all of the covariates into one scalar, and matching is then performed by identifying groups of units with similar values of this score. In practice, propensity scores are generally not known to the researcher and must be estimated using the observed data. When applied to text, propensity scores can be used to further condense the information within a chosen higher-dimensional representation into a summary of only the information that is relevant for determining treatment assignment. Propensity scores representations can be constructed using a quantitative text representation. For example, using STM-based representations or Word2Vec-based representations where dimension of the covariate space is less than the number of documents, standard techniques such as simple logistic regression can be used to estimate propensity scores. To construct propensity score representations over larger a covariate space, such as those typically spanned by a TDM, we use Multinomial Inverse Regression (MNIR; Taddy, 2013), which provides a novel estimation technique for performing logistic regression of phrase counts from the TDM onto the treatment indicator. After estimating this model,

we can calculate a sufficient reduction score that, in principle, will contain all the information from the TDM that is relevant for predicting treatment assignment. Performing a forward regression of the treatment indicator on this sufficient reduction score produces the desired propensity score estimates.

B.1.2 Design choices for representations

Here we discuss a number of design choices that are required for the different representations considered in our study.

TDM-based representations. Each of the TDM-based representations is characterized by a bounding scheme, which determines the subset of the vocabulary that will be included in X, and a weighting scheme, which determines the numerical rule for how the values of X are measured. We consider standard term-frequency (TF) weighting, TF-IDF weighting, and L2-rescaled TF-IDF weighting. We also consider a number of different screening schemes, including no screening, schemes that eliminate high and low frequency terms, and schemes that consider only high and low frequency terms.

STM-based representations. Each STM-based representation is characterized by a fixed number of topics (K=10, 30, 50, or 100) and takes one of three distinct forms: 1) the vector of K estimated topic proportions ("S1"), 2) the vector of K estimated topic proportions and the SR score ("S2"), or 3) a coarsened version of the vector of K estimated topic proportions ("S3"). This coarsened representation is constructed using the following procedure. For each document, we first identify the three topics with the largest estimated topic proportions. We retain and standardize these three values and set all remaining $K - 3$ topic proportions equal to 0, so that the resulting vector of coarsened topic proportions, $\hat{\theta}_i^{\star}$, contains only three non-zero elements. We then calculate the

"focus" of each document, denoted by F_i, a metric we define as the proportion of topical content that is explained by the three most prominent topics. Focus scores close to one indicate content that is highly concentrated on a small number of topics (e.g., a news article covering health care reform may have nearly 100% of its content focused on the topics of *health* and *policy*); conversely, focus scores close to zero indicate more general content covering a wide range of topics (e.g., a news article entitled "The ten events that shaped 2017" may have content spread evenly across ten or more distinct topics). To estimate this score for each document, we take the sum of the raw values of the three non-zero topic proportions identified as above (i.e., $\hat{F}_i = \hat{\theta}_{i[1]} + \hat{\theta}_{i[2]} + \hat{\theta}_{i[3]}$ where $\hat{\theta}_{i[j]}$ is the jth order statistic of the vector $\hat{\theta}$). Appending this estimated focus score to the coarsened topic proportion vector produces the final $(K+1)$-dimensional representation.

TIRM representations. The TIRM procedure of Roberts et al. (2018) uses an STM-based representation with an additional representation based on document-level propensity scores estimated using the STM framework. These separate representations are then combined within the TIRM procedure using a CEM distance. Each variant of the TIRM procedure considered in this paper is characterized by a fixed number of topics and a set coarsening level (2 bins, 3 bins, or 4 bins).

Word Embedding representations. Google and Stanford University have produced a variety of pre-trained word embedding models. Google's GoogleNews model, where each word vector is length 300 using a corpus of 100 billion words, draws from the entire corpus of Google News; this corpus is therefore extremely well-suited to our analysis. As well, we consider several of Stanford's GloVe embeddings (Pennington et al., 2014). In particular, we employ their models with word vectors of length 50, 100, 200, and 300. For each of these five embeddings, we produce document-level vectors by taking the weighted average of all word vectors in a document (Kusner et al., 2015).

B.1.3 Defining a distance metric

After a representation is chosen, applying this representation to the corpus generates a finite set of numerical covariate values associated with each document (i.e., X_i denotes the covariates observed for document i for all $i = 1, \ldots, N$). The next step in the matching procedure concerns how to use these covariate values to quantify the similarity between two documents. There are two main classes of distance metrics. Exact and coarsened exact distances regard distances as binary: the distance between two units is either zero or infinity, and two units are eligible to be matched only if the distance between them is equal to zero. Alternatively, continuous distance metrics define distance on a continuum, and matching typically proceeds by identifying pairs of units for whom the calculated distance is within some allowable threshold ("caliper").

Exact and coarsened exact distances

The exact distance is defined as $D_{ij} = 0$ if $X_i = X_j$ and is equal to infinity otherwise. Matching over this metric (exact matching) generates pairs of documents between treatment and control groups that match exactly on every covariate. Although this is the ideal, exact matching is typically not possible in practice with more than a few covariates. A more flexible metric can be defined by first coarsening the covariate values into "substantively indistinguishable" bins, then using exact distance within these bins (Iacus et al., 2012). For example, using a topic-model-based representation, one might define a coarsening rule such that documents will be matched if they share the same primary topic (i.e., if the topic with the maximum estimated topic proportion among the K topics is the same for both documents). Roberts et al. (2018) advocates using CEM for matching documents based on a representation built using an STM, but, in principle, this technique can also be used with TDM-based representations. For example, one might coarsen the term counts of a TDM into binary values indicating whether each term in the vocabulary is used within each document. Though

it is possible in principle, coarsening does not scale well with the dimension of the covariates and so may not be practical for matching with TDM-based representations. This type of distance specification may also create sensitivities in the matching procedure, since even minor changes in the coarsening rules can dramatically impact the resulting matched samples.

Continuous distances

Various continuous distance metrics can be used for matching, including linear distances based on the (estimated) propensity score or best linear discriminant (Rosenbaum & Rubin, 1983b), multivariate metrics such as the Mahalanobis metric (Rubin, 1973a), or combined metrics, such as methods that match on the Mahalanobis metric within propensity score calipers (Rosenbaum & Rubin, 1985). When matching on covariates defined by text data, care must be taken to define a metric that appropriately captures the complexities of text. For instance, linear distance metrics such as Euclidean distance may often fail to capture information about the relative importance of different covariates. To make this more clear, consider two pairs of documents containing the texts: "obama spoke", "obama wrote" and 'he spoke", "he wrote". Under a TDM-based representation, the Euclidean distances between units in each of these pairs are equal; however, the first pair of documents is intuitively more similar than the second, since the term "obama" contains more information about the content of the documents than the term "he". Similarly, the Euclidean distance between the pair documents "obama spoke", "obama obama" is equivalent to the distance between the pair "obama spoke", "he wrote", since by this metric distance increases linearly with differences in term frequencies. These issues also arise when using linear distance metrics with topic-model-based representations.

A metric that is less vulnerable to these complications is Mahalanobis distance, which defines the between documents i and j as $D_{ij} = (X_i - X_i)^T \Sigma^{-1} (X_i - X_j)$, where Σ is the variance-covariance matrix of the covariates X. This is essentially a normalized Euclidean distance, which weights covari-

ates according to their relative influence on the total variation across all documents in the corpus. Calculating Mahalanobis distance is practical for lower-dimensional representations, but because the matrix inversion does not scale well with the dimension of X, it may not be computationally feasible for matching using larger, TDM-based representations.

An alternative metric, which can be efficiently computed using representations defined over thousands of covariates, is cosine distance. Cosine distance measures the cosine of the angle between two documents in a vector space:

$$D_{ij} = 1 - \frac{\sum X_i X_j}{\sqrt{\sum X_i^2}\sqrt{\sum X_j^2}}.$$

Cosine distance is commonly used for determining text similarity in fields such as informational retrieval and is an appealing choice for matching because, irrespective of the dimension of the representation, it captures interpretable overall differences in covariate values (e.g., a cosine distance of one corresponds to a 90 degree angle between documents, suggesting no similarity and no shared vocabulary). In general, the utility of a particular continuous distance metric will largely depend on the distribution that is induced on the covariates through the representation.

Calipers and combinations of metrics

When pruning treated units is acceptable, exact and coarsened exact matching methods have the desirable property that the balance that will be achieved between matched samples is established a-priori. Treated units for whom there is at least one exact or coarsened exact match in the control group are matched, and all other treated units are dropped. On the other hand, matching with a continuous distance metric requires tuning after distances have been calculated in order to bound the balance between matched samples. After the distances between all possible pairings of treated and control documents have been calculated, one then chooses a caliper, D_{max}, such that any pair

of units i and j with distance $D_{ij} > D_{max}$ cannot be matched. Here, when pruning treated units is acceptable, any treated units without at least one potential match are dropped. Calipers are typically specified according to a "rule of thumb" that asserts that D_{max} be set equal to the value of 0.25 or 0.5 times the standard deviation of the distribution of distance values over all possible pairs of treated and control units, but in some special cases, the caliper can be chosen to reflect a more interpretable restriction. For example, using the cosine distance metric, one might choose a caliper to bound the maximum allowable angle between matched documents.

B.1.4 Text as covariates and outcomes

The procedure described in Section 2.3 is relatively straightforward to apply in studies where text enters the problem only through the covariates. However, in more complicated settings where both the covariates and one or more outcomes are defined by features of text, additional steps may be necessary to ensure these components are adequately separated.

In practice it is generally recommended that outcome data be removed from the dataset before beginning the matching process to preclude even the appearance of "fishing," whereby a researcher selects a matching procedure or a particular matched sample that leads to a desirable result (Rubin, 2007). However, this may not be possible when evaluating a text corpus, since both the covariates and outcome may often be latent features of the text (Egami et al., 2017). For instance, suppose we are interested in comparing the level of positive sentiment within articles based on the gender of the authors. One can imagine that news articles that report incidences of crime will typically reflect lower levels of positive sentiment than articles reporting on holiday activities, regardless of the gender of the reporter. Thus, we might like to match articles between male and female reporters based on their topical content and then compare the sentiment expressed within these matched samples. Here, we must extract both the set of covariates that will be used for matching (i.e., topical content) and the outcome (level of positive sentiment) from the same observed text. Because these different

components may often be related, measuring both using the same data poses two important challenges for causal inference: first, it requires that the researcher use the observed data to posit a model on the "post-treatment" outcome, and, second, measurement of the covariates creates potential for fishing. In particular, suppose that positive sentiment is defined for each document as the number of times terms such as "happy" are used within that document (standardized by each document's length). Suppose also that we use the entire vocabulary to measure covariate values for each document (e.g., using a statistical topic model). In this scenario, matching on topical content is likely to produce matches that have similar rates of usage of the term "happy" (in addition to having similar rates of usage of other terms), which may actually diminish our ability to detect differences in sentiment.

To address this issue, we recommend that researchers interested in inference in these settings define the covariates and outcome over a particular representation, or set of distinct representations, such that measurement of the outcome can be performed independently of the measurement of covariates. For example, one might measure the covariates using a representation of text defined over only nouns, and separately, measure outcome values using a representation defined over only adjectives. Or, continuing the previous example, one might divide the vocabulary into distinct subsets of terms, where one subset is used to measure topical content and the other is used to measure positive sentiment. In settings where the chosen representation of the text must be inferred from the observed data (e.g., topic-model-based representations), cross-validation techniques can also be employed, as described in Egami et al. (2017). For instance, one might randomly divide the corpus into training set and test set, where the training set is used to build a model for the representation, and this model is then applied to the test set to obtain covariate values that will be used in the matching procedure.

B.2 Index of representations evaluated

Table B.1: Representations considered in human evaluation experiment

Type	Name	Description	Dimension
TDM	T1	TF Bounded from 4-1000	10726
	T2	TF-IDF Bounded from 4-1000	10726
	T3	TF-IDF Bounded from 4-100	9413
	T4	TF-IDF Bounded from 4-10	4879
	T5	TF-IDF Bounded from 10-500	6000
	T6	TF-IDF Bounded from 500-1000	154
	T7	L2 Rescaled TF-IDF Bounded from 4-1000	10726
	T8	TF on unbounded TDM	34397
	T9	TF-IDF on unbounded TDM	34397
STM	S1-10	STM on 10 Topics	10
	S2-10	10 Topics + estimated sufficient reduction	11
	S3-10	10 Topics, top 3 topics + focus	11
	S1-30	30 Topics	30
	S2-30	30 Topics + estimated sufficient reduction	31
	S3-30	30 Topics, top 3 topics + focus	31
	S1-50	50 Topics	50
	S2-50	50 Topics + estimated sufficient reduction	51
	S3-50	50 Topics, top 3 topics + focus	51
	S1-100	100 Topics	100
	S2-100	100 Topics + estimated sufficient reduction	101
	S3-100	100 Topics, top 3 topics + focus	101
Word2Vec	W1	Word embedding of dimension 50 (Google)	50
	W2	Word embedding of dimension 100 (Google)	100
	W3	Word embedding of dimension 200 (Google)	200
	W4	Word embedding of dimension 300 (Google)	300
	W5	Word embedding of dimension 300	300

B.3 Survey used in human evaluation experiment

The figures below show snapshots of different components of the survey as they were presented to participants in each of our human evaluation experiments. In particular, Figure B.1 shows the survey landing page, where participants were informed about the nature of the task. Participants were then presented with the scoring rubric shown in Figure B.2 and completed a series of training tasks as depicted in Figure B.3.

Figure B.1: The survey landing page informed participants about the nature of the task.

Thank you for beginning our survey! Please answer every question. If you skip questions, you may not receive payment. **Read each question carefully. Some of them are attention checks.**

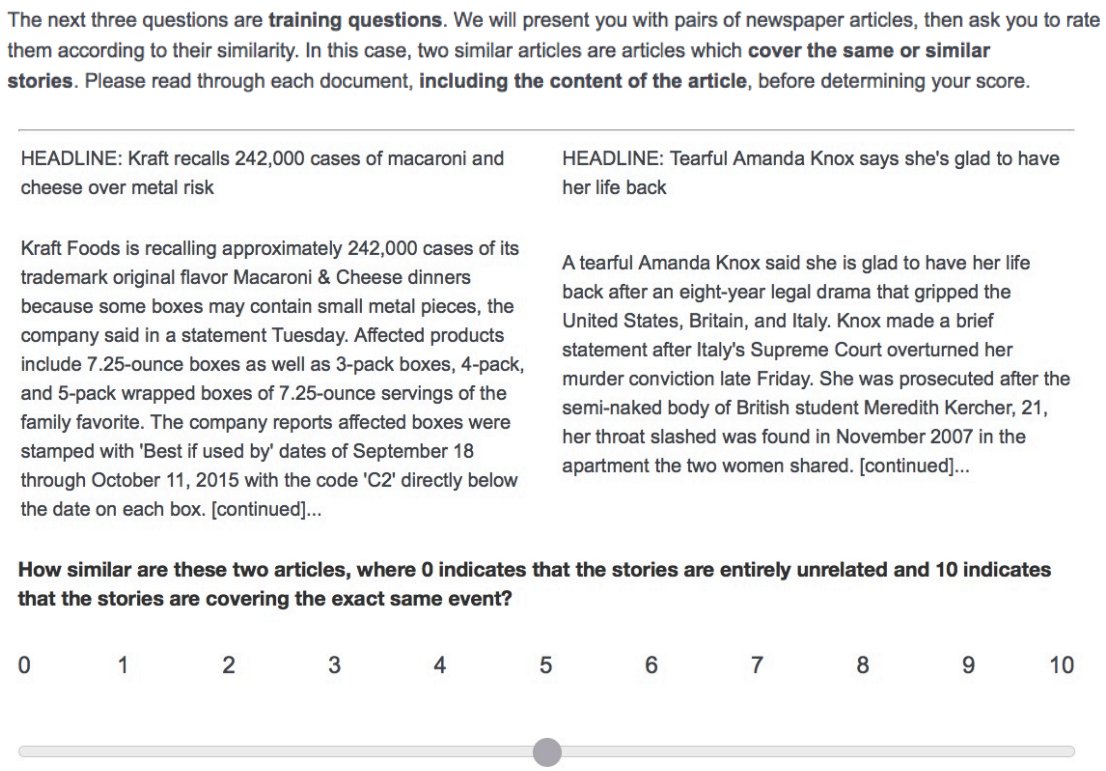
We are going to show you a series of pairs of newspaper articles and ask you to rate the similarity of the documents in each pair. We are interested in how similar they are in terms of the stories they are covering. For example, some of the pairs might be about the exact same event, some pairs might be two stories covering similar but distinct events, and some pairs might be about entirely unrelated things.

Figure B.2: After enrolling in the experiment, participants were presented with a scoring rubric to use as a guide for determining the similarity of a pair of documents.

You will rate similarity on a scale from 0 (no similarity) to 10 (extremely similar). The scoring rubric below provides a guide to help you determine the similarity of a pair of articles.

Score	Description
0	The articles are completely different and cover entirely unrelated events.
3	The articles cover different events that are somewhat related.
5	The articles cover different events, but the events are similar kinds of events.
7	The articles cover the same event, but the specific details presented about that event may be different.
10	The articles are nearly identical in terms of the event they are covering and the details being presented about that event.

Figure B.3: In the first training task of the survey, participants were ask to read and score a pair of articles and were then informed that the anticipated score for this pair was zero. Specifically, they were told "We think these articles' similarity is 0 out of 10. The first article is related to macaroni and cheese, while the second article is about a murder trial."



B.4 Sensitivity of match quality scores to the population of respondents

To determine the generalizability of the match quality ratings obtained from our survey experiment, we compare two identical pilot surveys using respondents from two distinct populations. The first pilot survey was administered through Mechanical Turk, and the second pilot was administered through the Digital Laboratory for the Social Sciences (Enos et al., 2016). For each survey, respondents were asked to read and evaluate ten paired articles, including one attention check and one anchoring question. Each respondent was randomly assigned to evaluate eight matched pairs from a sample of 200, where this pilot sample was generated using the same weighted sampling scheme described above. Figure B.4 shows the average match quality scores for each of the 200 matched pairs evaluated based on sample of 337 respondents from Mechanical Turk and 226 respondents from DLABSS. The large correlation between average matched quality scores across samples (ρ=0.88) suggests that our survey is a useful instrument for generating consistent average ratings of match quality across diverse populations of respondents. In particular, even though individual conceptions of match quality may differ across respondents, the average of these conceptions both appears to meaningfully separate the pairs of documents and to be stable across at least two different populations.

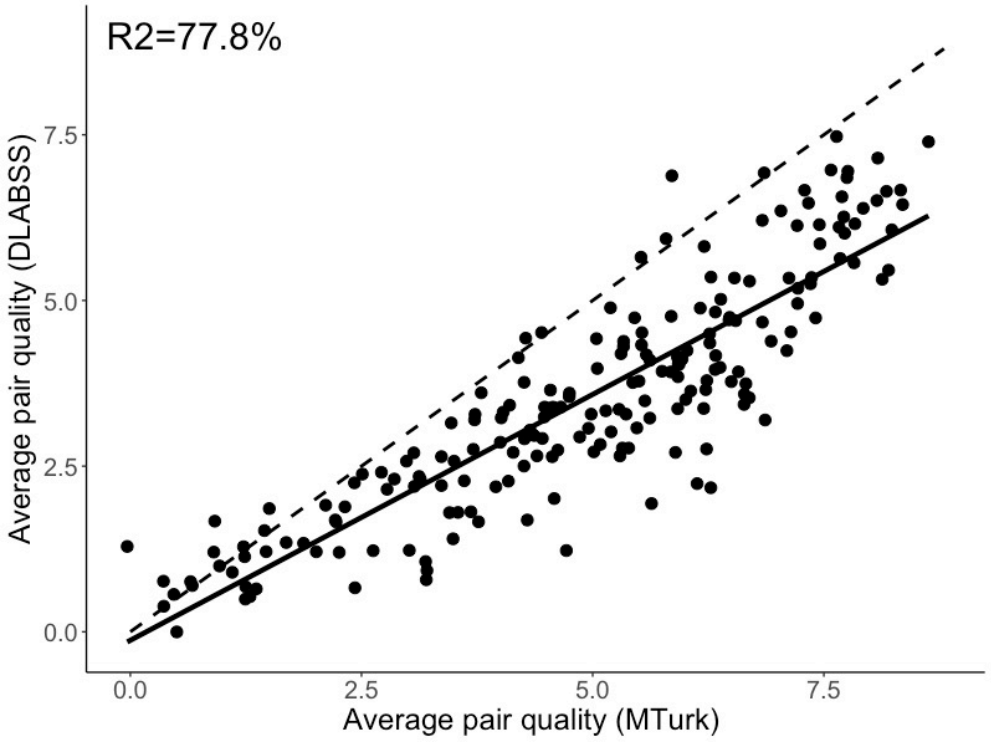


Figure B.4: The strong linear relationship between the average match quality scores for 200 pairs of articles evaluated in two separate pilot studies (solid line) compared to a perfect fit (dotted line) suggests that the survey produces consistent results across samples, when averaged across multiple respondents.

B.5 Technical details of the evaluation of match quality of pairs of news articles

In this section we more fully describe the design and analysis of the human evaluation experiment for the newspaper matching example. We start by discussing how we generated our sampling strategy and weights, and then discuss how we used model-assisted survey sampling to estimate average match quality for the different methods along with associated uncertainty.

B.5.1 Details of the sampling design

The study presented in this paper is in fact a replication study as our initial study did not directly assess all procedures considered (in particular, we did not initially evaluate the Word2Vec procedures). We therefore designed our second study to both directly extend our findings, verify the prior results, and further investigate the predictive accuracy of our models to out-of-sample pairs. In order to achieve this, we designed a sampling scheme that has three components: (1) we sampled 4 pairs from each procedure considered, (2) we directly sampled pairs that were previously evaluated to assess the stability of the evaluation process, and (3) we sampled pairs not selected by any method to examine differences between selected and non-selected pairs. The first stage sampled pairs with weights based on the predicted quality of the pairs in order to sample predicted high-quality pairs more heavily. We used the prior study's fit predictive model to generate these predictions. The second and third stage sampled a fixed number of pairs within each tier of quality (from 0 to 8+) to see the full range of pair qualities in our sample (simple random sampling would not work since the vast majority of pairs are scored as quality 1 or lower). This overall process resulted in a sample of 505 pairs that fully represents all possible pairs (selected and not). For each pair we have an initial predicted quality score, a sampling probability π_i, and an associated sampling weight $w_i \propto 1/\pi_i$.

Because many of the procedures generally select the same high-quality pairs, the sequential sam-

pling of 4 pairs for each procedure tends to give many of the same pairs back. This is by design, and means that our sample primarily consists of pairs shared by multiple procedures which gives greater precision in estimating these procedures' average quality. We simply take the unique set of pairs sampled as our final evaluation sample.

We calculate the actual sampling weights of each pair for this scheme using simulation. In particular, we conduct our sampling scheme 100,000 times and calculate how often each pair is selected into the sample. These provide (up to monte carlo error) the true selection probabilities π_i; inverting them provides the true sampling weights w_i. For the out-of-matched pairs sampling stage (3), we averaged these final weights across groups of pairs that all have the same probability of selection to increase precision.

The stage (1) sampling scheme intentionally induces selection bias into the sample by discouraging rare pairs, especially singleton pairs, which are expected to be low quality with little variability, in favor of pairs that are identified by multiple matching procedures. Regardless, because the sampling probabilities are fixed a-priori, weighted averages of the pairs' match quality gives good estimates of the average quality of the pairs selected by each procedures; this approach is simply classic survey sampling as described in, e.g.,Sarndal et al. (2003). All this complexity in the sampling design is to ensure that the sample evaluated is targeted to give information on as many procedures as possible, a difficult task when evaluating 130 procedures with a sample size of about 500.

B.5.2 Estimating pair and procedure quality.

Let $u_{t,c}$ denote a potential pairing of treatment and control documents, where t is the index of the treated unit and c is the index of the control unit. In our evaluation study, $t = 1, \ldots, 1565$ and $c = 1, \ldots, 1796$. For matching procedure j, let $\mathcal{R}_j$ denote the set of n_j matched pairs of articles identified using procedure j. The set of all unique pairs selected by any of the J procedures consid-

ered in the evaluation experiment, denoted $\mathcal{R}$, is defined by the union of these subsets:

$$\mathcal{R} = \cup_{j=1}^{J} \mathcal{R}_j.$$

We index the pairs with $i = 1, \ldots, N$.

The frequency of how often each pair u_i in $\mathcal{R}$ was selected by a procedure is:

$$F_i = \sum_{j=1}^{J} 1\{u_i \in \mathcal{R}_j\},$$

where $1\{i \in \mathcal{R}_j\}$ is an indicator variable taking value 1 if pair u_i is identified using matching procedure j and 0 otherwise.

From the human evaluation, we, for each element i of $\mathcal{S}$, where $\mathcal{S}$ is the set of all sampled pairs, observe m_i similarity ratings, $q_{i,1}^{obs}, \ldots, q_{i,m_i}^{obs}$ where $q_{i,\cdot}^{obs} \in [0, 10]$. We estimate the match quality for each evaluated pair i using the average of observed ratings for that pair, $\bar{q}_i^{obs}$.*

We wish to estimate, for each procedure, the finite-population quantities of the average true quality of the pairs selected. In particular, if we let q_i be the average quality score score we would see if we had an arbitrarily large number of human respondents evaluate that pair, our targets of inference are, for each procedure j,

$$Q_j = \frac{1}{N_j} \sum_{u_i \in \mathcal{R}_j} q_i.$$

The Q_j are population quantities of how the matching procedure did in the specific context considered. This estimand does not necessarily take into account how the methods would perform on other corpora, even ones similar to this one.

To estimate Q_j for any matching procedure j in our evaluation we use a weighted average of the

*We also explored modeling these ratings to account for rater effects and variable number of ratings per question, but as the results were essentially unchanged, elected to use the simple averages.

match quality estimates across the pairs contained in $\mathcal{R}_j \cap \mathcal{S}$, where weights for each pair are equal to the inverse probability of being sampled:

$$\hat{Q}_{samp,j} = \frac{1}{Z_j} \sum_{u_i \in \mathcal{R}_j} \frac{1}{\pi_i} S_i \bar{q}_i^{obs} \text{ with } Z_j = \sum_{u_i \in \mathcal{R}_j} \frac{1}{\pi_i} S_i. \tag{B.5.1}$$

with S_i an indicator of whether pair i was sampled for evaluation, with sampling probability π_i, and Z_j a normalizing constant. This is a simple Haj́ek estimator and is known to have good properties.

Unfortunately, despite the sampling scheme, some of our methods only had a small number of pairs sampled for evaluation. Estimating the average match quality for such procedures could therefore be fairly imprecise. We address this by using our model for predicting the match quality of a pair of documents based on different machine measures of similarity to construct model-assisted survey sampling estimators that use the predicted qualities to adjust these estimated average quality scores. We describe this analysis approach next.

B.5.3 Improving the estimates of procedure quality.

To enhance our predictions of match quality for our procedures, we use a model trained on the pairs in $\mathcal{S}_{pre}$, the sample collected in our initial study, to calculate the predicted match quality, $\hat{q}_i$ for all pairs $i = 1, \ldots, N$. These $\hat{q}_i$ are fixed, and do not depend on the analyzed (i.e., second) random sample. We can use these predictions to adjust our estimates of the average quality of all pairs for each procedure using survey sampling methods.

In particular, our model adjusted quality for procedure j is

$$\hat{Q}_{adj,j} = \frac{1}{n_j} \sum_{u_i \in \mathcal{R}_j} \hat{q}_i + \frac{1}{Z_j} \sum_{u_i \in \mathcal{R}_j} S_i \frac{1}{\pi_i} \left(\bar{q}_i^{obs} - \hat{q}_i \right)$$

Here $\hat{q}_i$ is the predicted quality based on the initial sample. Note the first term in the above is a fixed

constant, not dependent on the sample. The second term is random, depending on the sample, and, ignoring the small bias induced by Z_j being random, This is a *model-adjusted estimate*; the first summation gives the predicted average quality of the method. The second summation adds an adjustment based on the residuals for the actually sampled and evaluated pairs; this adjustment makes the overall estimate effectively unbiased[†] regardless of whether the predictive model is useful, predictive, or even correct. The more the predictive model aligns with the actual measured values, however, the more precise our estimates will be (as the residuals and adjustment part will get smaller and smaller as predictive accuracy grows).

B.5.4 Uncertainty estimation

Classic survey sampling results allowed us to estimate each procedure's average quality with the estimated qualities of our sampled pairs. We can also increase the precision of these estimates using model adjustment, using the predicted quality scores to adjust the same by population averaged characteristics. In both cases, the next step is to obtain appropriate uncertainty estimates (standard errors) for these point estimates. Unfortunately, the task of appropriately calculating uncertainty in this context for both the raw estimates and the model-adjusted estimates is a surprisingly difficult and subtle problem. In particular, while there are classic survey sampling formula that can be used to calculate uncertainty, they are asymptotic and are sensitive to extreme weights (which we have). This creates some perverse results (i.e. near zero standard errors) for some of the procedures that only had a few pairs sampled. To avoid this we, by instituting a homoscedastic assumption for the error terms, did a parametric simulation to calculate uncertainty in order to work around this problem. This procedure captures the variability induced by the varying sample weights and the measurement error due to the human evaluation. We describe this next.

[†]The bias is purely from using a Hajek rather than Horvitz-Thompson estimator, and comes from the normalizing Z_j being a random quantity. It is *not* a function of model misfit or misspecification.

Uncertainty estimates for the raw quality estimates. For the unadjusted quality measures, we estimate uncertainty using the principles of a case-wise bootstrap with some modifications. In particular, especially for those methods with very few (e.g. 4) sampled pairs, estimating the variability of quality of the pairs via case-wise bootstrap is unreliable unless we pool or partially pool estimates of variability across the different methods.

To see this consider a hypothetical method with 4 of its pairs sampled, 1 with very high weight due to being a rare pair and 3 with a low weight due to being selected by most methods. Any bootstrap sample that includes the high weight unit will essentially give an average quality score close to that of the high weight unit. Even bootstrap samples with multiple draws of the high weight unit will still get nearly that same average quality score since the values of these large elements will all be the same. Across bootstrap samples, this will give low variability, i.e., seemingly high precision. It does not take into account the variability of scores we might have actually seen across other units of similar weight. We address this with the a parametric approach that we describe next.

We first assess the typical variability of the quality scores of pairs within the procedures. For the unadjusted quality scores of the individual pairs we first calculated an estimate of the standard deviation of scores within a given match method (we did this by calculating the weighted standard deviation of scores). We then took the median of these values as our measure of within-method variation of pair quality. We use the median to avoid the impact of the extreme standard deviations due to the methods with small samples of pairs.[‡]

To calculate standard errors for our methods, we then simulated the pair sampling step followed by the scoring of sampled pairs step by first selecting pairs using the original sampling strategy, and then generating pseudo-quality scores with the same variance as we generally saw for pairs selected by a method. We then calculated the overall pseudo-quality for each of our methods based on these

[‡]We actually calculated this (pooled) standard deviation a variety of ways and took the largest to be maximally conservative.

scores and associated sample weights. Our standard errors are then the standard deviation of these generated overall pseduo-quality scores.

To compare, we also conducted a simple case-wise bootstrap. Here we sampled the evaluated pairs with replacement and calculated each methods' quality score using the bootstrap sample, finally obtaining standard errors using the standard deviation of the resulting values. This approach works well for those methods with 10 or more sampled pairs. Overall, our parametric approach generally produced larger standard errors, which is a mixture of the overall conservatism of our approach and of the aforementioned issue of the naíve approach giving small standard errors those methods with few pairs and a few high-weight pairs that dominate the overall quality measure. We thus report our parametric simulation-based standard errors.

Uncertainty for the model-adjusted approach. For the model-adjusted case, we again worried about those methods with few samples having less variability due to small numbers of high weight units giving nearly the same model adjustment with each step. We therefore follow the above process, but instead of generating synthetic outcomes we generated synthetic residuals by generating normally distributed noise with variance equal to the variance of the original residuals from our predictive model. These simulated residual-based standard errors were again conservative when compared to the naïve case-wise approach for those procedures with enough selected pairs to make this comparison.

Remarks. All our uncertainty estimation methods capture the uncertainty in the pair quality evaluation process as the variability of the pairs' quality scores captures both the measurement error and the structural variation of the pairs themselves. In our plots, we report the simulation-based standard errors for the model adjusted estimates. As noted in the text, the model-adjusted quality scores themselves were generally similar to unadjusted (for the directly evaluated methods where we

had both scores), and the differences between the two had no impact on our overall findings.

For methods that we did not initially identify for our human evaluation, we could calculate a predicted quality based on our model of

$$\hat{Q}_{pred,j} = \frac{1}{n_j} \sum_{u_i \in \mathcal{R}_j} \hat{q}_i.$$

This is extrapolation, however. If the new procedure was selecting pairs that systematically were better than predicted, for example, this extrapolation would be biased. Even if such a new method happened to use some pairs randomly selected for evaluation, we cannot use the survey adjusted $\hat{Q}_{adj,j}$ or raw estimate $\hat{Q}_{samp,j}$ since the pairs *not* in the sampling frame had no chance of selection. One could create a hybrid estimator by splitting the sample into potentially sampled, but we do not explore that further here.

B.5.5 Prior evaluation study details

As mentioned above, we performed an initial full study on an initial subset of the matching procedures considered (in particular, we did not initially evaluate the Word2Vec procedures). Overall, this study produced the same results as our final study.

We sampled pairs differently for our initial study. In particular, we did not have baseline predicted quality scores to calculate sampling weights from. We therefore, to produce a representative sample of matched articles for evaluation, did not take a sample from each procedure's pair list but instead took a weighted random sample of 500 pairs from the union of these lists, $\mathcal{R}$, with sampling weights roughly proportional to F_i, where F_i is the number of times pair i was selected by a procedure. Because singleton pairs comprised over 75% of the pairs in $\mathcal{R}$, we further downweighted pairs

with $F_i = 1$ by a factor of 5. Our overall sampling probabilities for pair i were then

$$w_i \propto \begin{cases} 0.20 \text{ if } F_i = 1, \\ F_i \text{ otherwise.} \end{cases}$$

We then calculated true sampling probabilities and weights via simulation as described above (due to high weights for some pairs and the sampling without replacement these initially weights are not truly proportional to inverse probability of selection).

B.6 Notes on the sample and unadjusted human experiment results

The final evaluation sample consisted of 33 pairs that were originally evaluated in the initial evaluation, 50 pairs that were not identified by any matching method considered, and 422 pairs that were used by at least one matching method evaluated. The sampling weights for those pairs that were selected by at least one method ranged from 0.02 to 10.7, with a median of 0.23. This corresponded to selection probabilities ranging from 1 in 1000 to 77%. 25% of the pairs had less than a 1% chance of being selected. The very rare pairs tend to come from the propensity score methods that had a large number of low-quality matches. Across procedures, some had only 4 pairs sampled, and some had up to 100. The average was 28 pairs.

The standard deviation of quality scores did depend on the sampling weight, with a standard deviation of around 2.5 for low pi_i and 1 for the highest pi_i. On the other hand, the standard deviation of scores for very low and very high predicted qualities was less than 0.5, rising to around 1.6 for pairs predicted to have a quality of 5. Within a given procedure, scores tended to have a standard deviation of around 2.37, for those procedures with 10 or more pairs sampled. If we look across all procedures the median decreases markedly due to poor estimates for small sample sizes. We used 2.37 in our simulation.

For the residual scores, residuals had a lower standard deviation near the endpoints (due to truncation) and peaked at around 1.6 for the middle scores. We therefore use a residual standard deviation of 1.6 in our simulations to calculate our standard errors, which will be generally conservative. Even with this conservative approximation, we are explaining 55% of our variation with our predictive score.

Figure B.5 shows the simple weighted average match quality of the directly evaluated pairs sampled for each of the 130 procedures considered in the evaluation experiment. The nominal 95% confidence intervals are from standard errors calculated from the parametric bootstrap described above.

The standard errors seem small, but some mild calculations suggest they are reasonable. In particular, with 28 pairs, if the pairs have a standard deviation of about 2, we would expect, roughly a standard error of $2/\sqrt{28} = 0.38$, which is what we tended to see. We also point out that we are considering the population of pairs selected by a method as fixed: this is a finite sample inference problem.

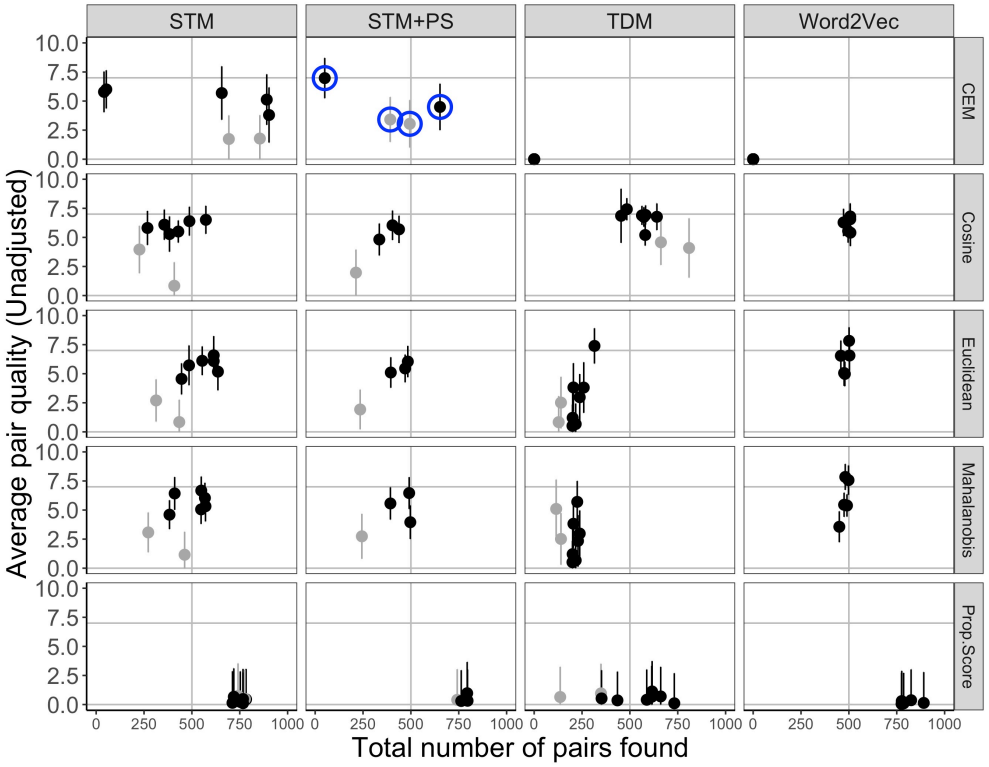


Figure B.5: Number of matches found versus estimated (unadjusted) average match quality scores for each combination of matching methods. Grey points indicate procedures with extreme reduction in information (e.g., procedures that match on only stop words). Blue circles highlight procedures that use existing state-of-the-art methods for text matching.

B.7 Template matching and sensitivity analyses for media bias application

To evaluate the robustness of our findings, we performed a series of sensitivity checks to assess how our results and subsequent conclusions change when using different specifications of the matching procedure. Figure B.6 shows the results produced by three alternative text matching methods. These robustness checks highlight the importance of the specification of the matching procedure: weaker methods (i.e., methods that produce low quality matches) typically lead to weaker inferences. For example, the results produced from template matching using the Mahalanobis distance metric on a vector of 100 topic proportions show generally smaller changes in average favorability within each source before and after matching than the results shown in Figure 2.4. The null results in this case provide further evidence in support of the claim that text matching is an effective strategy for reducing differences in the observed biases across news sources that are due to topic selection.

As a final robustness check of the results based on our template-matched sample, we performed the following consistency test. First, we randomly generated 10,000 pairs of documents containing 150 randomly selected articles from each news source. In each iteration of random sampling and for each news source, we then calculated the average favorability scores towards Democrats and Republicans within the matched sample. Figure B.7 shows the distributions of these favorability scores for each news source after 100 iterations of random matching. Finally, we calculated the total *change* in favorability observed after matching in each iteration, averaged across all 13 sources. More formally, for each iteration $i = 1, \ldots, 10000$ we calculated the test statistic:

$$T_i = \frac{1}{13} \sum_{j=1}^{13} \left(|\hat{Y}_j^{dem} - \hat{Y}_{j,M_i}^{dem}| + |\hat{Y}_j^{rep} - \hat{Y}_{j,M_i}^{rep}| \right),$$

where $\hat{Y}_j^{dem}$ and $\hat{Y}_j^{rep}$ denote the average favorability scores toward democrats and republicans, respectively, for all articles corresponding to source j in the original, unmatched sample. Quantities

$\hat{Y}^{dem}_{j,M_i}$ and $\hat{Y}^{rep}_{j,M_i}$ denote the partisan favorability scores averaged across the set of 150 articles from source j that were selected by random matching in iteration i. The sampling distribution of this test statistic provides a reference for values of the test statistic that may occur when comparing randomly selected sets of 150 articles across these 13 sources. Therefore, by comparing the value of our observed test statistic based on the results of our template-matching procedure described in Section 2.5 to the randomization distribution defined by T, we can estimate the probability that our template-matched results are due to random chance. Our results from this randomization test indicated that template matching on text removes a significant amount of the bias observed across sources that remains after adjusting for differences in topic selection (p=0.004).

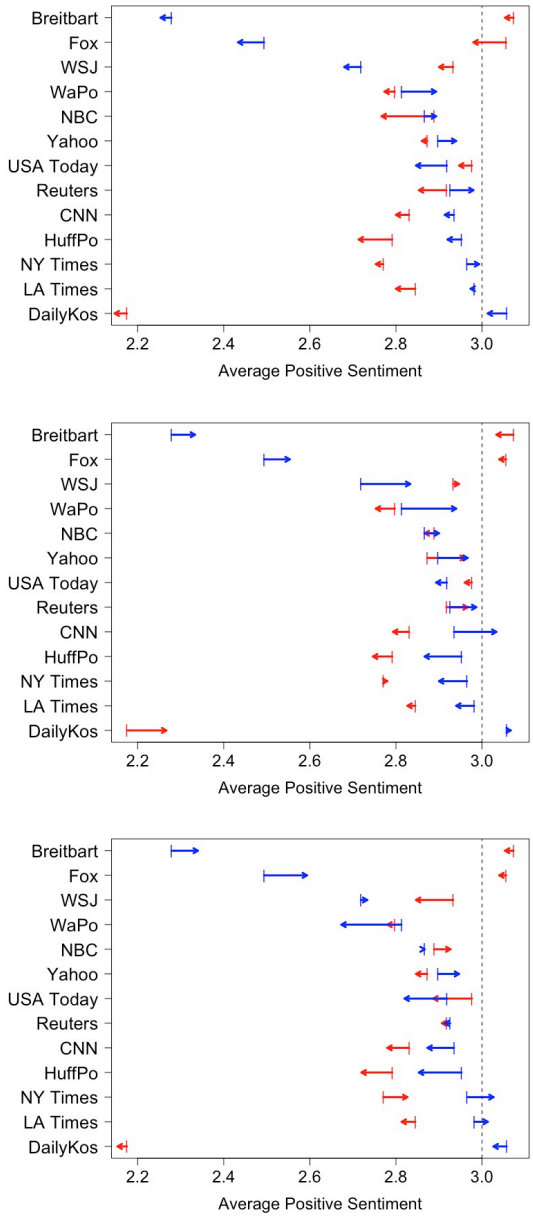


Figure B.6: Estimates of average favorability toward Democrats (blue) and Republicans (red) for each source both before and after matching using Mahalanobis matching on an STM with 100 topics (top), propensity score matching on an STM with 100 topics (center) and propensity score matching on a TDM (bottom).

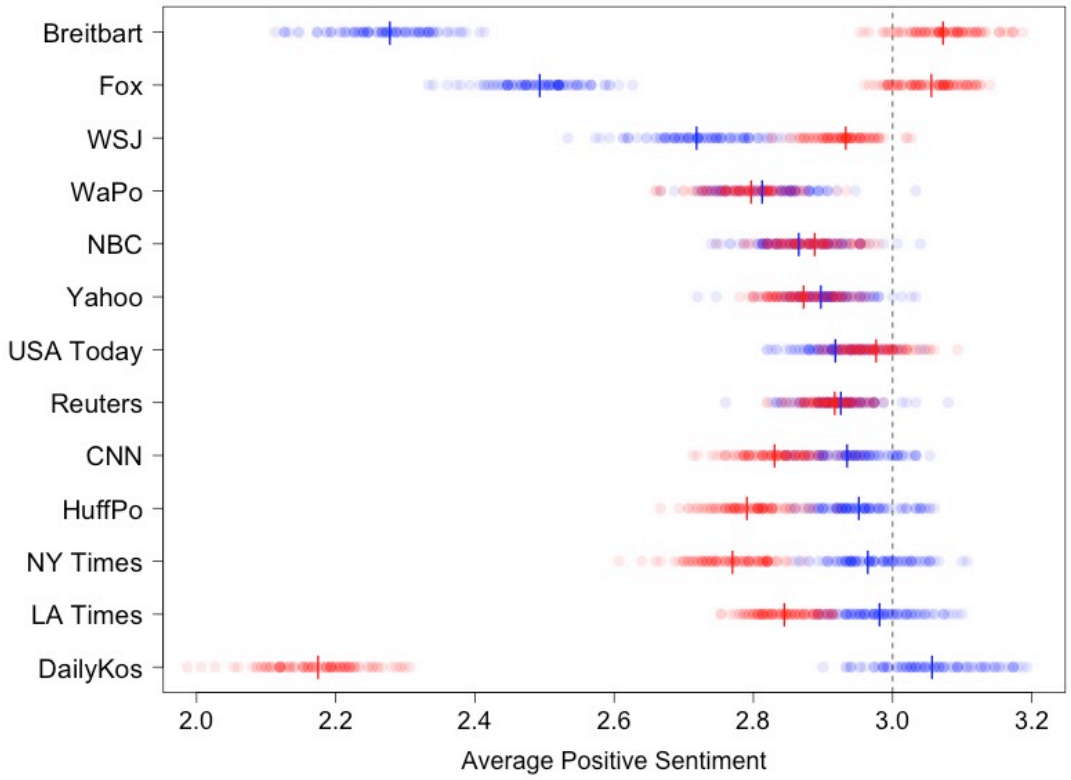


Figure B.7: Estimates of average favorability toward Democrats (blue) and Republicans (red) for each source for 100 iterations of random matching. Blue and red lines represent the average favorability scores before matching for Democrats and Republicans, respectively.

C

Supplemental Materials for Chapter 3

C.1 MCMC procedure

Posterior inference for this model can also be achieved using the following Gibbs sampling procedure (using the Kalman Filter in Steps 1-3):

1. Sample $\Psi_{1:T}$ from $p(\Psi_{1:T}|T_{obs}^{treat}, Z_{obs}, M, \rho, \beta, \delta_0, \delta_1)$ using the Kalman Filter.

2. Sample β from $p(\beta|\Psi_{1:T}, \rho, \delta_0, \delta_1)$

3. Sample ρ from $p(\rho|\Psi_{1:T}, \beta, \delta_0, \delta_1)$

4. Sample δ_0 and δ_1 from $p(\delta_0, \delta_1|\Psi_{1:T}, Z_{obs}, M) = p(\delta_0, \delta_1|T^{treat}, Z)$

5. Repeat steps (1-4) until convergence

C.2 Additional details of VA application

C.2.1 Covariate selection procedure

To determine which of the available covariates should be included in our model for predicting the time of assignment to treatment, we first performed a random forest analysis to identify a subset of the available time-varying covariates that are most predictive of receipt of assignment to treatment.

Covariate measurements were collected for each patient during each visit that occurred in the study period, and each of the 534 treated patients in our analysis was observed for between one to 611 distinct visits during the study period. For variable selection, we first constructed a new dataset containing observations for each of the 534 treated patients at each of two time periods: 1) the visit associated with assignment to treatment, and 2) the preceding visit. Patients who received assignment to treatment at their first visit were included in the dataset only once. Each observation vector in the resulting dataset therefore corresponds to the covariate values of a patient who has not yet been assigned or to a patient at their time of assignment. In principle, contrasts of covariates between these two groups should capture information about how the latent health process - and the corresponding indication for assignment to either treatment or control - varies with these longitudinal measurements.

For variable selection, we performed random forest analysis (Breiman, 2001) implemented using the "randomForest" package in R (Breiman & Cutler, 2003) to evaluate the relative importance of each of the time-varying covariates for predicting the time of indication for treatment. Using this

procedure, we identified the 20 most influential time-varying covariates from over 150 available variables. Among these 20 covariates were a number of covariates related to the current physical characteristics of the patient (e.g., current weight and blood pressure) as well as variables that captured changes in these physical measurements (e.g., change in weight). Other important covariates include the number of comorbidities that the patient has at the time of the visit and changes in health care utilization such as recent organ failure events (e.g., right heart failure), recent inpatient visits, or receipt of incidental procedures (e.g., echocardiograph).

C.2.2 Matching procedure for selection of potential controls

References

Abadie, A. & Kasy, M. (2017). The risk of machine learning. *arXiv preprint arXiv:1703.10935*.

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422), 669–679.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455.

Angrist, J. D. & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4), 69–85.

Armstrong, K. (2012). Methods in comparative effectiveness research. *Journal of Clinical Oncology*, 30(34), 4208.

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium* (pp.17).: American Medical Informatics Association.

Athey, S., Imbens, G., Pham, T., & Wager, S. (2017). Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, 107(5), 278–81.

Athey, S. & Imbens, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5).

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25), 3083–3107.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in medicine*, 26(4), 734–753.

Baccini, M., Mattei, A., & Mealli, F. (2017). Bayesian inference for causal mechanisms with application to a randomized study for postoperative pain control. *Biostatistics*, 18(4), 605–617.

Bacher, R. & Kendziorski, C. (2016). Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology*, 17(1), 63.

Bartolucci, F. & Grilli, L. (2011). Modeling partial compliance through copulas in the principal stratification framework. *Journal of the American Statistical Association*, 106, 469–479.

Bavli, H. & Mozer, R. (2019). The effects of comparable-case guidance on awards for pain and suffering and punitive damages: Evidence from a randomized controlled trial. *Yale Law & Policy Review*, 37.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2), 29–50.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608–650.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137–1155.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

Breiman, L. & Cutler, A. (2003). Manual for setting up. *Using, and Understanding Random Forest*, 4.

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80, 250–271.

Byar, D. P. (1980). Why data bases should not replace randomized clinical trials. *Biometrics*, (pp. 337–342).

Cain, L., Logan, R., Robins, J., Sterne, J., Sabin, C., Bansi, L., Justice, A., Goulet, J., Bucher, H., Esteve, A., et al. (2011). When to initiate combined antiretroviral therapy to reduce mortality and aids-defining illness in hiv-infected persons in developed countries: an observational study. *Annals of internal medicine*, 154(8), 509–515.

Carlin, B. P. & Polson, N. G. (1992). Monte carlo bayesian methods for discrete regression models and categorical time series. *Bayesian statistics*, 4, 577–586.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90, 1313–1321.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266–298.

Cochran, W. G. & Cox, G. M. (1957). Experimental designs.

Comment, L., Mealli, F., Haneuse, S., & Zigler, C. (2019). Survivor average causal effects for continuous time: a principal stratification approach to causal inference with semicompeting risks. *arXiv preprint arXiv:1902.09304*.

Conlon, A., Taylor, J., & Elliott, M. (2014). Surrogacy assessment using principal stratification with multivariate normal and gaussian copula models. *Statistcal Methods in Medical Research*.

Cox, D. R. (1958). Planning of experiments.

Czado, C. & Song, P. X.-K. (2008). State space mixed models for longitudinal observations with binary and binomial responses. *Statistical Papers*, 49(4), 691–714.

Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2017). Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*.

Danaei, G., Rodríguez, L. A. G., Cantero, O. F., Logan, R., & Hernán, M. A. (2013). Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease. *Statistical methods in medical research*, 22(1), 70–96.

Daniel, R., De Stavola, B., Cousens, S., & Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, 71(1), 1–14.

Dehejia, R. H. & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1), 151–161.

Denny, M. J. & Spirling, A. (2018). Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis*, (pp. 1–22).

Diamond, A. & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3), 932–945.

Ding, P., Feller, A., & Miratrix, L. (2016). Decomposing treatment effect variation. *arXiv preprint arXiv:1605.06566*.

Ding, P. & VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3), 368.

Efron, B. & Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86, 9–17.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2017). How to make causal inferences using texts. *arXiv preprint*.

Elliott, M., Raghunathan, T., & Li, Y. (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics*, 11, 353–372.

Enos, R. D., Hill, M., & Strange, A. M. (2016). Voluntary digital laboratories for experimental social science: The harvard digital lab for the social sciences. *Working Paper*.

Fan, A., Doshi-Velez, F., & Miratrix, L. (2017). Promoting domain-specific terms in topic models with informative priors. *arXiv preprint arXiv:1701.03227*.

Feller, A., Grindal, T., Miratrix, L., Page, L. C., et al. (2016). Compared to what? variation in the impacts of early childhood education by alternative care type. *The Annals of Applied Statistics*, 10(3), 1245–1285.

Feller, A., Mealli, F., & Miratrix, L. (2017). Principal score methods: Assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics*, 42(6), 726–758.

Feng, M., McSparron, J., Kien, D. T., Stone, D., Roberts, D., Schwartzstein, R., Vieillard-Baron, A., & Celi, L. A. (2018). When more is not less: A robust framework to evaluate the value of a diagnostic test in critical care. *Submitted*.

Fisher, R. (1925a). *Statistical Methods for Research Workers. First edition*. Edimburgh: Oliver and Boyd.

Fisher, R. A. (1925b). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22 (pp. 700–725).: Cambridge University Press.

Fisher, R. A. (1937). *The design of experiments*. Oliver And Boyd; Edinburgh; London.

Fogarty, C. B., Mikkelsen, M. E., Gaieski, D. F., & Small, D. S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association*, 111(514), 447–458.

Forastiere, L., Mealli, F., Miratrix, L., et al. (2018). Posterior predictive p-values with fisher randomization tests in noncompliance settings: Test statistics vs discrepancy measures. *Bayesian Analysis*, 13(3), 681–701.

Forastiere, L., Mealli, F., & VanderWeele, T. (2016). Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using bayesian principal stratification. *Journal of the American Statistical Association*, (pp. forthcoming).

Frangakis, C. & Rubin, D. (2002). Principal stratification in causal inference. *Biometrics*, 58(1), 21–29.

Freidlin, B. & Korn, E. (2012). Assessing causal relationships between treatments and clinical outcomes: always read the fine print. *Bone marrow transplantation*, 47(5), 626.

Freiman, M. R., Rose, A. J., Powell, R. W., Miller, D. R., & Wiener, R. S. (2015). Patterns of potentially inappropriate prescribing of phosphodiesterase inhibitors in pulmonary hypertension in the va. In *C13. ACCOUNTING FOR COSTS AND RESOURCE UTILIZATION IN RESPIRATORY HEALTH* (pp. A3889–A3889). American Thoracic Society.

Frumento, P., Mealli, F., Pacini, B., & Rubin, D. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, 107, 450–466.

Gallop, R., Small, D., Lin, J., Elliot, M., Joffe, M., & Have, T. T. (2009). Mediation analysis with principal stratification. *Statistics in Medicine*, 28(7), 1108–1130.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC, Boca Raton, FL.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4), 457–472.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6, 721–741.

Gentzkow, M. & Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2), 280–316.

Gentzkow, M. & Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1), 35–71.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian statistics*, 4, 641–649.

Gilbert, P., Bosch, R., & Hudgens, M. (2003). *Biometrics*, 59, 531–541.

Gilbert, P. & Hudgens, M. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics*, 64(4), 1146–1154.

Gran, J. M., Hoff, R., Røysland, K., Ledergerber, B., Young, J., & Aalen, O. O. (2016). Estimating the treatment effect of the treated under time-dependent confounding applied to simulated data and to data from the swiss hiv cohort study. *JR stat. Soc. C*.

Green, D. P. & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3), 491–511.

Grilli, L. & Mealli, F. (2008). Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification. *Journal of Educational and Behavioral Statistics*, 33(1), 111–130.

Grimmer, J. (2015). We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, 48(1), 80–83.

Grimmer, J., Messing, S., & Westwood, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4), 413–434.

Groeling, T. (2013). Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, 16.

Groseclose, T. & Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, 120(4), 1191–1237.

Gu, X. S. & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4), 405–420.

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, 22(1), 1–30.

Hansen, B. B. & Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of computational and Graphical Statistics*, 15(3), 609–627.

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Stampfer, M. J., Willett, W. C., Manson, J. E., & Robins, J. M. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology (Cambridge, Mass.)*, 19(6), 766.

Hernan, M. A., Brumback, B., & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 11(5), 561–570.

Hirano, K. & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3), 259–278.

Ho, D. E., Quinn, K. M., et al. (2008). Measuring explicit political positions of media. *Quarterly Journal of Political Science*, 3(4), 353–377.

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

Horowitz, J. L. & Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American statistical Association*, 95(449), 77–84.

Huitfeldt, A., Kalager, M., Robins, J. M., Hoff, G., & Hernán, M. A. (2015). Methods to estimate the comparative effectiveness of clinical strategies that administer the same intervention at different times. *Current epidemiology reports*, 2(3), 149–161.

Iacus, S. M., King, G., Porro, G., & Katz, J. N. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, (pp. 1–24).

Ichino, A., Mealli, F., & Nannicini, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics*, 23, 305–327.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A*, 171(2), 481–502.

Imai, K. & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 243–263.

Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1), 443–470.

Imai, K. & Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, 19(1), 1–19.

Imbens, G. (2003). Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review,*, 93, 126–132.

Imbens, G. & Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62, 467–476.

Imbens, G. & Rubin, D. (1997a). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25(1), 305–327.

Imbens, G. W. & Rubin, D. B. (1997b). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, (pp. 305–327).

Imbens, G. W. & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jin, H. & Rubin, D. (2008a). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103, 101–111.

Jin, H. & Rubin, D. B. (2008b). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103(481), 101–111.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 160035.

Kempthorne, O. (1952). *Design and Analysis of Experiments*. Wiley; New York.

Kennedy, E. H., Taylor, J. M., Schaubel, D. E., & Williams, S. (2010). The effect of salvage therapy on survival in a longitudinal study with treatment by indication. *Statistics in medicine*, 29(25), 2569–2580.

Kim, C., Daniels, M., Hogan, J., Choirat, C., & Zigler, C. (2019). Bayesian methods for multiple mediators: Relating principal stratification and causal mediation in the analysis of power plant emission controls. *arXiv preprint arXiv:1902.06194*.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14 (pp. 1137–1145).: Montreal, Canada.

Kroeger, M. A. (2016). Plagiarizing policy: Model legislation in state legislatures. *Princeton typescript*.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. In *International Conference on Machine Learning* (pp. 957–966).

Le, Q. & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning* (pp. 1188–1196).

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3), 337–346.

Lee, D. & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 485, 281–355.

Lee, D. S. & Lemieuxa, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2), 281–355.

Levine, M. N. & Julian, J. A. (2008). Registries that show efficacy: good, but not good enough.

Li, F., Mattei, A., Mealli, F., et al. (2015). Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, 9(4), 1906–1931.

Li, Y., Taylor, J., & Elliott, M. (2011). Causal assessment of surrogacy in a metanalysis of colorectal cancer trials. *Biostatistics*, 12, 478–492.

Li, Y. P., Propert, K. J., & Rosenbaum, P. R. (2001). Balanced risk set matching. *Journal of the American Statistical Association*, 96(455), 870–882.

Little, R. J. & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health*, 21(1), 121–145.

Little, R. J. & Rubin, D. B. (2014). *Statistical analysis with missing data*, volume 333. John Wiley & Sons.

MacLean, D. L. & Heer, J. (2013). Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *Journal of the American Medical Informatics Association*, 20(6), 1120–1127.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Marko, N. F. & Weil, R. J. (2010). The role of observational investigations in comparative effectiveness research. *Value in Health*, 13(8), 989–997.

Marsh, L. C. & Cormier, D. R. (2001). *Spline regression models*, volume 137. Sage.

Mason, W. & Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1), 1–23.

Mattei, A., Li, F., & Mealli, F. (2013). Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. *The Annals of Applied Statistics*, 7(4), 2360–2013.

Mattei, A. & Mealli, F. (2007). Application of the prinicipal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics*, 63(2), 437–446.

Mattei, A. & Mealli, F. (2011). Augmented designs to assess principal strata effects. *Journal of the Royal Statistical Society - Series B*, 73, 729–752.

Mattei, A. & Mealli, F. (2016). Regression discontinuity designs as local randomized experiments. *Observational Studies*, 2, 156–173.

Mattei, A., Mealli, F., & Pacini, B. (2014). Identification of causal effects in the presence of nonignorable missing outcome values. *Biometrics*, 70(2), 278–288.

Mealli, F., Imbens, G. W., Ferro, S., & Biggeri, A. (2004). Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics*, 5(2), 207–222.

Mealli, F. & Mattei, A. (2012). A refreshing account of principal stratification. Commentary on "Principal Stratification - a goal or a tool?" by Judea Pearl. *The International Journal of Biostatistics*, 8(1), Article 8.

Mealli, F. & Pacini, B. (2013a). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108(503), 1120–1131.

Mealli, F. & Pacini, B. (2013b). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108, 1120–1131.

Mercatanti, A., Li, F., & Mealli, F. (2014). Improving inference of Gaussian mixtures using auxiliary variables. *Statistical Analysis and Data Mining*, 8(1), 34–48.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mozer, R. (2019). Replication Data for: Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality.

Mozer, R. & Glickman, M. E. (2019). Estimating the effects of medical interventions: A framework for the design and analysis of longitudinal studies with treatment by indication. *Working paper*.

Mozer, R., Kessels, R., & Rubin, D. B. (2017). Disentangling treatment and placebo effects in randomized experiments using principal stratification—an introduction. In *The Annual Meeting of the Psychometric Society* (pp. 11–23).: Springer.

Mozer, R. & Mealli, F. (2019). Causal inference with complex data: A guide for the modern statistician. *Working paper*.

Mozer, R., Miratrix, L., Kaufman, A. R., & Anastasopoulos, L. J. (2019). Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Forthcoming at Political Analysis*.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4), 465–472.

Pashley, N. E. & Miratrix, L. W. (2017). Insights on variance estimation for blocked and matched pairs designs. *arXiv preprint arXiv:1710.10342*.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence* (pp. 411–420).: Morgan Kaufmann Publishers Inc.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Peterson, A. & Spirling, A. (2018). Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis*, 26(1), 120–128.

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124: Vienna, Austria.

Poses, R. M., Smith, W. R., McClish, D. K., & Anthony, M. (1995). Controlling for confounding by indication for treatment: Are administrative data equivalent to clinical data? *Medical care*, (pp. AS36–AS46).

Prentice, R. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8, 431–440.

Ratkovic, M. & Tingley, D. (2017). Causal inference through the method of direct estimation. *arXiv preprint arXiv:1703.05849*.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016a). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515), 988–1003.

Roberts, M. E., Stewart, B. M., & Nielsen, R. A. (2018). Adjusting for confounding with text matching. *arXiv preprint*.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016b). Navigating the local modes of big data. *Computational Social Science*, 51.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429), 106–121.

Rosenbaum, P. (1987). Model-based direct adjustment. *Journal of the Royal Statistical Society: Series B*, 82, 387–394.

Rosenbaum, P. (2002a). *Observational Studies*. New York: Springer.

Rosenbaum, P. & Rubin, D. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B*, 45(2), 212–218.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408), 1024–1032.

Rosenbaum, P. R. (2002b). Observational studies. In *Observational studies* (pp. 1–17). Springer.

Rosenbaum, P. R. (2010). *Design of observational studies*. Springer.

Rosenbaum, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics*, 21(1), 57–71.

Rosenbaum, P. R. & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Rosenbaum, P. R. & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33–38.

Rubin, D. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(1), 161–170.

Rubin, D. (2006a). Causal inference through potential outcomes and principal stratification: application to studies with censoring due to death. *Statistical Science*, 91, 299–321.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics*, (pp. 159–183).

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, (pp. 185–203).

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5), 688.

Rubin, D. B. (1978a). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, (pp. 34–58).

Rubin, D. B. (1978b). Bias reduction using mahalanobis metric matching. *ETS Research Report Series*, 1978(2).

Rubin, D. B. (1980a). Comment. *Journal of the American Statistical Association*, 75(371), 591–593.

Rubin, D. B. (1980b). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371), 591–593.

Rubin, D. B. (2006b). *Matched sampling for causal effects*. Cambridge University Press.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1), 20–36.

Rubin, D. B. (2010). On the limitations of comparative effectiveness research. *Statistics in medicine*, 29(19), 1991–1995.

Salton, G. (1991). Developments in automatic text retrieval. *Science*, (pp. 974–980).

Salton, G. & McGill, M. J. (1986). *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Sarndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer.

Schwartz, S., Li, F., & Mealli, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association*, 31(10), 949–962.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6), 546–555.

Shortreed, S. M. & Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4), 1111–1122.

Silber, J. H., Rosenbaum, P. R., Ross, R. N., Ludwig, J. M., Wang, W., Niknam, B. A., Mukherjee, N., Saynisch, P. A., Even-Shoshan, O., Kelz, R. R., et al. (2014). Template matching for auditing hospital cost and quality. *Health Services Research*, 49(5), 1446–1474.

Slaughter, J. L., Reagan, P. B., Newman, T. B., & Klebanoff, M. A. (2017). Comparative effectiveness of nonsteroidal anti-inflammatory drug treatment vs no treatment for patent ductus arteriosus in preterm infants. *JAMA pediatrics*, 171(3), e164354–e164354.

Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, 27(1), 325–353.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254–263).: Association for Computational Linguistics.

Sox, H., Greenfield, S., Cassel, C., et al. (2009). Committee on comparative effectiveness research prioritization; institute of medicine. initial national priorities for comparative effectiveness research.

Spertus, J. V. & Normand, S.-L. T. (2018). Bayesian propensity scores for high-dimensional causal inference: A comparison of drug-eluting to bare-metal coronary stents. *Biometrical Journal*, 60(4), 721–733.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.

Steiner, D. F., MacDonald, R., Liu, Y., Truszkowski, P., Hipp, J. D., Gammage, C., Thng, F., Peng, L., & Stumpe, M. C. (2018). Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American journal of surgical pathology*, 42(12), 1636–1646.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1).

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503), 755–770.

Tanner, M. & Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398), 528–540.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 267–288).

Van der Laan, M. J. & Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media.

VanderWeele, T. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters*, 78, 2957–2962.

Wager, S. & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.

Wang, C., Dominici, F., Parmigiani, G., & Zigler, C. M. (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics*, 71(3), 654–665.

Zeng, Q. T., Tse, T., Divita, G., Keselman, A., Crowell, J., Browne, A. C., Goryachev, S., & Ngo, L. (2007). Term identification methods for consumer health vocabulary development. *Journal of medical Internet research*, 9(1).

Zhang, J. & Rubin, D. (2003a). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(1), 353–358.

Zhang, J., Rubin, D., & Mealli, F. (2008). Evaluating the effects of job training programs on wages through principal stratification. In D. Millimet, J. Smith, & E. Vytlacil (Eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (pp. 117–145). Elsevier.

Zhang, J., Rubin, D., & Mealli, F. (2009). Likelihood-based analysis of the causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, 104, 166–176.

Zhang, J. L. & Rubin, D. B. (2003b). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4), 353–368.

Zigler, C. & Belin, T. (2012). A Bayesian approach to improved estimation of causal effect predictiveness for a principal surrogate endpoint. *Biometrics*, 68, 922–932.

Zigler, C., Dominici, F., & Wang, Y. (2012). Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics*, 13(2), 289–302.

Zubizarreta, J. & Kilcioglu, C. (2016). designmatch: Construction of optimally matched samples for randomized experiments and observational studies that are balanced by design. *R package version 0.1*, 1, 187.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500), 1360–1371.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511), 910–922.

Zubizarreta, J. R., Paredes, R. D., Rosenbaum, P. R., et al. (2014a). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics*, 8(1), 204–231.

Zubizarreta, J. R., Small, D. S., & Rosenbaum, P. R. (2014b). Isolation in the construction of natural experiments. *The Annals of Applied Statistics*, (pp. 2096–2121).