Imperfect Data and Optimal Allocation of Data-collection Resources*

Guilherme Duarte
gjduarte@upenn.edu

Dean Knox
dcknox@upenn.edu

August 17, 2023

Abstract

Complications in applied work often prevent researchers from obtaining unique point estimates of target quantities using cheaply available data—at best, ranges of possibilities, or sharp bounds, can be reported. To make progress, researchers frequently collect more information by (1) re-cleaning existing datasets, (2) gathering secondary datasets, or (3) pursuing entirely new designs. Common examples include manually correcting missingness, recontacting attrited units, validating proxies with ground-truth data, finding new instrumental variables, and conducting follow-up experiments. These auxiliary tasks are costly, forcing tradeoffs with (4) larger samples from the original approach. Researchers' data-collection strategies, or choices over these tasks, are often based on convenience or intuition. In this work, we show how to provably identify the most cost-efficient data-collection strategy for a given research problem. We quantify the quality of existing data using the width of the confidence regions on the sharp bounds, which captures two sources of uncertainty: statistical uncertainty due to finite samples of the variables measured, and fundamental uncertainty because some variables are not measured at all. We then show how to compute the expected information gain, defined as the expected amount by which each data-collection task will narrow these bounds by addressing one or both sources of uncertainty. Finally, we select the task with the greatest information efficiency, or gain per unit cost. Leveraging recent advances in automatic bounding, we show that efficiency is computable for essentially any discrete causal system, estimand, and auxiliary data task. Based on this theoretical framework, we develop a method for optimal adaptive allocation of data-collection resources. Users first input a causal graph, estimand, and past data. They then enumerate distributions from which future samples can be drawn, fixed and per-sample costs, and any prior beliefs. Our method automatically derives and sequentially updates the optimal data-collection strategy.

Keywords: causal inference, research design, adaptive designs, partial identification, constrained optimization

*Guilherme Jardim Duarte is a Ph.D. student in the Operations, Information and Decisions Department, the Wharton School of the University of Pennsylvania. Dean Knox is an Andrew Carnegie Fellow and an assistant professor in the Operations, Information and Decisions Department, the Wharton School of the University of Pennsylvania We gratefully acknowledge financial support from AI for Business and the Analytics at Wharton Data Science and Business Analytics Fund. This research was made possible in part by grants from the Carnegie Corporation of New York and Arnold Ventures. The statements made and views expressed are solely the responsibility of the authors.

Contents

1 Introduction

Complications in applied often prevent point identification of causal estimands using cheaply available data—at best, sharp bounds containing ranges of possibilities can be reported. To make progress, researchers frequently collect more information by (1) re-cleaning existing datasets, (2) gathering secondary datasets, or (3) pursuing entirely new designs. Common examples include manually correcting missingness, recontacting attrited units, validating proxies with ground-truth data, finding new instrumental variables, and conducting follow-up experiments. These auxiliary tasks are costly, forcing tradeoffs with (4) larger samples from the original approach.

In this paper, we demonstrate how analysts can determine the optimal strategy for future data collection about a causal system, represented by a graph that describes causal relationships between a set of *main variables* that could in theory be observed, as well as a set of disturbances that are fundamentally unobservable. We will suppose that analysts possess a set of existing datasets and must decide between a set of candidate data tasks. Each existing dataset consists of a sample of observations taken from a marginal or joint distribution of the main variables. Similarly, each candidate task represents a future sample that could be taken from a marginal or joint distribution of the main variables. We will further suppose that each candidate task has a known per-observation cost and that analysts have a fixed data-collection budget to allocate among tasks.

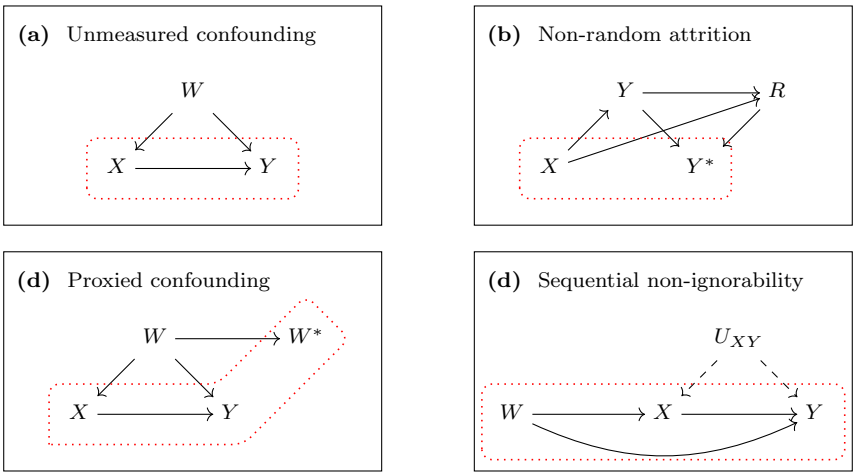
We define each task's efficiency as the expected information gain per unit cost. Gain is formalized as the narrowing of the confidence region on sharp bounds, capturing two kinds of benefits: point-identifying new aspects of the causal system and reducing statistical uncertainty. Leveraging recent advances in automatic bounding (Duarte et al., 2022), we prove efficiency is computable for essentially any discrete causal system, estimand, and auxiliary data task.

We propose a method for optimal adaptive allocation of data-collection resources. Users first input a causal graph, estimand, and past data. They then enumerate distributions from which future samples can be drawn, fixed and per-sample costs, and any prior beliefs. Our method automatically derives and sequentially updates the optimal data-collection strategy.

The unmeasured confounding scenario of Figure 1(a) will be used to introduce the idea of auxiliary data collection tasks that help improve upon these initial bounds, along with notation and key concepts. Figure 1(b) depicts non-random attrition, or selective dropout from a study that is influenced by treatment assignment and outcome; as a result, outcome data contains missingness (Kaplan and Atkins, 1987; Weuve et al., 2012). Figure 1 depicts proxy confounding, in which analysts are unable to adjust for an unobserved

common cause of treatment and outcome, but possess a proxy for this confounder that is subject to measurement error. The Figure 1(b–c) cases will be used to illustrate the idea of collecting auxiliary information on conditional distributions, building upon the marginal-distribution data-collection tasks that will be introduced in the Figure 1(a) scenario (Whitman et al., 2020). Finally, Figure 1(d) depicts a mediation design in which analysts are able to collect initial data on all main variables—treatment, mediator, and outcome—but nonetheless cannot identify causal mediation effects due to the unobservable correlation between mediator and outcome errors (Imai et al., 2013; Gaesser et al., 2020). This case is used to introduce the idea of auxiliary experiments that gather additional information by indirectly manipulating the mediator.

Figure 1: **Common data challenges in applied research.** Panel (a) depicts an unmeasured confounding scenario in which analysts possess initial data on treatment X and outcome Y, as indicated by the dotted red region, but cannot identify the average treatment effect due to the lack of information on a common cause W that confounds them. Panel (b) depicts a nonrandom attrition scenario in which analysts possess initial data on treatment X and recorded outcome Y^*, which can differ from true outcome Y. A unit's response R is influenced by both treatment X and true outcome Y: if $R = 1$ then recorded outcome $Y^* = Y$, but if $R = 0$ then $Y^* = \texttt{NA}$. Panel (c) depicts a scenario in which analysts possess initial data on treatment X and outcome Y, as well as a noisy proxy W^* of the true confouncer W that is measured with error. Panel (d) depicts a mediation scenario in which analysts possess initial data on treatment W, mediator X, and outcome Y, but cannot identify causal mediation effects due to X-Y confounding by the fundamentally unmeasurable disturbance U_{XY}.



2 A motivating problem

To fix ideas, we first introduce notation and key concepts with a concrete example of unmeasured confounding, a common obstacle in applied research. We follow the convention that bold letters denote matrices and arrows denote vectors. Uppercase and lowercase letters denote random variables and their realizations, respectively.

Figure 1(a) depicts a structured causal system in which a common cause $W \in \{w_0, w_1\}$, such as disease status, influences both a treatment $X \in \{x_0, x_1\}$, such as receiving a drug, and an outcome $Y \in \{y_0, y_1\}$, such as six-month survival. For clarity of exposition, we discuss the case in which random variables are binary; note that our theoretical results hold for categorical or ordinal variables of finite cardinality.

Observable data on each unit consists of an independent and identically distributed (i.i.d.) draw of the random vector $[W, X, Y]^\top$. However, the data that is observed in practice may be a subset of the data that is observable in principle. Suppose that an initial dataset contains information only on $\vec{D}^{(1)} = [X, Y]^\top$, so that the confounder W is unobserved; this dataset is depicted in Figure 2(a). Let $n^{(1)} = 100$ indicate the number of samples available in this dataset, and let $\boldsymbol{D}^{(1)} \equiv \left[\vec{D}_1^{(1)}, \ldots, \vec{D}_{n^{(1)}}^{(1)}\right]^\top$ collect the samples. Subsequently, we will use $\boldsymbol{D}^{(s)}$ to denote the s-th collected dataset, containing $n^{(s)}$ samples of $\vec{D}^{(s)}$.

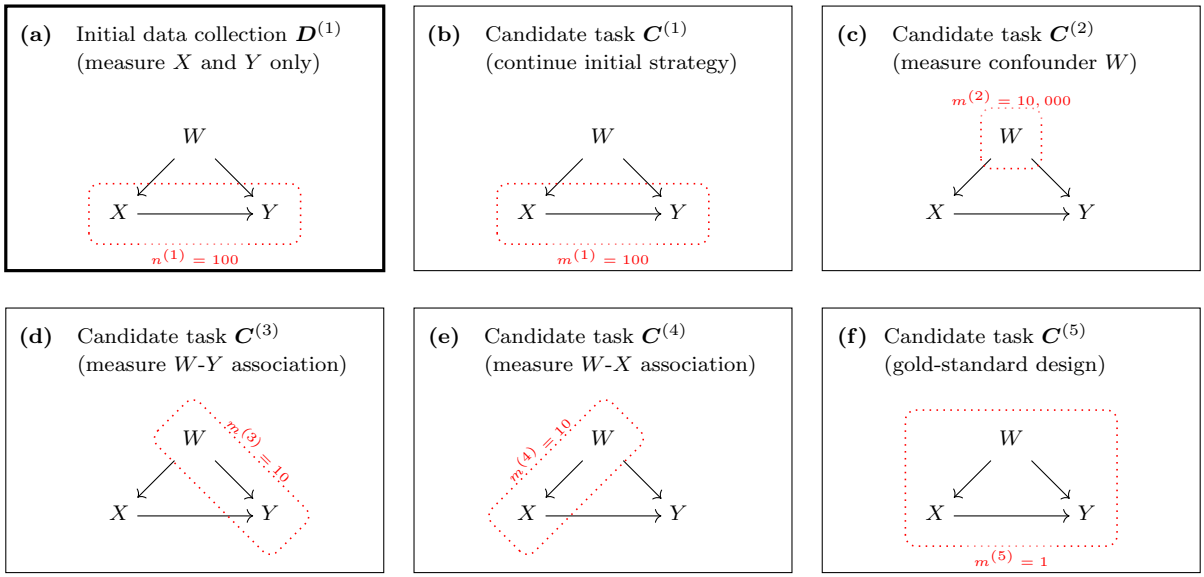
Next, we turn to the quantity of interest, which often involve the causal effect of manipulating a causally prior variable, e.g. by setting X to some value x, on a downstream variable. Potential-outcomes notation allows us to formally define these effects. The potential outcome $Y(x)$ indicates the value of Y that would have realized if—potentially contrary to fact—the treatment x had been received. We will refer to variables such as Y and $Y(x)$ as *factual* and *counterfactual* variables, respectively. For a review of potential-outcomes notation in graphical models, see Richardson and Robins (2013b) and Shpitser (2018).

More specifically, suppose that the quantity of interest is the average treatment effect (ATE) of manipulating the treatment, $\mathbb{E}[Y(x_1) - Y(x_0)]$. It is well known that $\boldsymbol{D}^{(1)}$, which contains data on X and Y alone, is insufficient to identify the ATE. This data does not permit analysts to adjust for the confounder W; as a result, there are many possible ATE values that are consistent with the observed data and therefore cannot be rejected.

We are now ready to state the motivating problem. In Figure 2, we consider take stock of options for addressing this problem in a cost-effective manner, given that a budget B is available for data collection. For simplicity, we suppose the analyst will fully expend the budget B on this second round of data collection, which will produce $\boldsymbol{D}^{(2)}$. A succession of candidate tasks are depicted in subsequent panels. The t-th candidate task involves measuring $\vec{C}^{(2,t)}$; after expending fixed costs, the remaining budget permits collection of $m^{(2,t)}$ observations. We will refer to the hypothetical dataset that would be obtained from candidate task t with the $m^{(2,t)} \times |\vec{C}^{(2,t)}|$ random matrix, $\boldsymbol{C}^{(2,t)}$.

The first candidate task, shown in panel (b), is to continue the original strategy, collecting another $m^{(2,1)} = 100$ samples from $\vec{C}^{(2,1)} = [X, Y]^\top$. A second candidate, in panel (c), is to measure the prevalence of the confounder alone, $\vec{C}^{(2,2)} = [Z]$. Intuitively, if the disease turns out to be extremely rare—i.e.,

Figure 2: **Possible data-collection tasks in a study with initially unmeasured confounding.** Panel (a) depicts an initial dataset, $\boldsymbol{D}^{(1)}$, that contains $n^{(1)}$ samples from the marginal distribution of X and Y. Subsequent panels depict candidate tasks for a second round of data collection.



if $\Pr(W = w_1) \ll \Pr(X = x_0)$ and $\Pr(W = w_1) \ll \Pr(X = x_1)$—then the confounder cannot drive much selection into treatment, and treatment-control comparisons will already be mostly within disease-free units with the same $W = w_0$ value. Often, administrative data can provide this information at relatively low per-sample costs, e.g. identifying $m^{(2,2)} = 10,000$ samples from $\vec{C}^{(2,2)}$ in data already collected by the Centers for Disease Control. Panels (d–e) depict still more candidate tasks with differing per-sample costs: collecting $m^{(2,3)} = 10$ samples from $\vec{C}^{(2,3)} = [W, Y]^\top$ or $m^{(2,4)} = 10$ samples from $\vec{C}^{(2,4)} = [W, X]^\top$. These capture the conventional wisdom that excluded variables are unproblematic if they either (i) do not correlate with the outcome or (ii) do not correlate with the treatment. These tasks can be substantially more expensive, as they may require laboratory testing to measure W and either a lengthy delay (to measure Y) or obtaining access to medical records for a new patient (to measure X). Finally, panel (f) represents the gold-standard approach of collecting data from the full joint distribution $\vec{C}^{(2,5)} = [W, X, Y]^\top$, which can be prohibitively costly—here, only $m^{(2,5)} = 1$ sample can be collected with the same budget.

Existing datasets:	$\boldsymbol{D}^{(1)}$,	past draw of	$n^{(1)} = 100$ samples of	$\vec{D}^{(1)}$	$= [X, Y]^\top$
Candidate tasks:	$\boldsymbol{C}^{(2,1)}$,	future draw of	$m^{(2,1)} = 100$ samples of	$\vec{C}^{(2,1)}$	$= [X, Y]^\top$
	$\boldsymbol{C}^{(2,2)}$,	future draw of	$m^{(2,2)} = 10,000$ samples of	$\vec{C}^{(2,2)}$	$= [W]$
	$\boldsymbol{C}^{(2,3)}$,	future draw of	$m^{(2,3)} = 10$ samples of	$\vec{C}^{(2,3)}$	$= [W, Y]^\top$
	$\boldsymbol{C}^{(2,4)}$,	future draw of	$m^{(2,4)} = 10$ samples of	$\vec{C}^{(2,4)}$	$= [W, X]^\top$
	$\boldsymbol{C}^{(2,5)}$,	future draw of	$m^{(2,5)} = 1$ sample of	$\vec{C}^{(2,5)}$	$= [W, X, Y]^\top$

3 Preliminaries

3.1 Notation

We now define notation and key concepts more formally to permit generalization to additional settings. Let $\vec{V} = [V_1, \ldots, V_J]^\top$ be a random vector containing *main variables* that can in principle be measured by analysts, and let $\vec{U} = [U_1, \ldots, U_K]^\top$ represent fundamentally *unobservable disturbances*, or random errors. Throughout this paper, we will assume that main variables are discrete and of finite cardinality; we place no restriction on the disturbances, which can be continuous or even multidimensional.

We will suppose that analysts possess (i) a canonical directed acyclic graph (DAG) $\mathcal{G}$, representing the theoretically possible causal relationships among $\vec{V}$ and $\vec{U}$.[1] Figure 1(a) depicts one such DAG, representing the motivating problem first introduced in Section 2. Here, the main variables are $\vec{V} = [W, X, Y]^\top$. Each of the unmeasurable random disturbances $\vec{U} = [U_W, U_X, U_Y]^\top$ influences its corresponding main variable. By convention, these disturbances are left implicit in the graph; disturbances will only be drawn explicitly if they influence more than one main variable.[2]

We further suppose that analysts possess (ii) the sample space of each main variable and (iii) a causal estimand or quantity of interest τ that is a functional involving factual or counterfactual versions of any main variable, such as $\mathbb{E}[Y(x_1) - Y(x_0)]$, the average treatment effect (ATE). Optionally, analysts may also supply (iv) assumptions justified by domain expertise.[3]

Finally, we suppose that analysts possess (v) a possibly empty collection of measured datasets, $\{\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}\}$,

[1] Throughout, we will work with the nonparametric structural equation models with independent errors (NPSEM-IE); however, our results are applicable to the model of Robins (1986) and Richardson and Robins (2013a) as well. A DAG is in canonical form if (i) no disturbance U_k has a parent in $\mathcal{G}$; and (ii) there exists no pair of disturbances, U_k and $U_{k'}$, such that U_k influences a subset of variables influenced by $U_{k'}$. See Evans (2016) and Duarte et al. (nd) for additional discussion and an algorithm to canonicalize arbitrary DAGs.

[2] An example of a multi-variable disturbance—also referred to as a set of correlated errors—is U_{XY} in Figure 1(d), which creates confounding between mediator X and outcome Y in that graph. In contrast to the Figure 1(a) problem in which confounding is induced by the main variable W, the Figure 1(d) confounding induced by U_{XY} fundamentally cannot be measured and adjusted for.

[3] Common examples include monotonicity assumptions (i.e., that an encouragement must have a weakly positive effect on treatment uptake) or disabling assumptions (i.e., that a treatment cannot have an effect in the absence of some necessary condition).

indexed by s. We will denote the variables measured in the s-th dataset, or the columns of $\boldsymbol{D}^{(s)}$, as $\vec{D}^{(s)} \subset \vec{V}$, and $i \in \{1, \ldots, n^{(s)}\}$ will index i.i.d. rows of $\boldsymbol{D}^{(s)}$. For example, in the motivating problem from Section 2, analysts initially possess a single dataset $\boldsymbol{D}^{(1)}$, consisting of $n^{(1)}$ rows that are each a sample from $\vec{D}^{(1)} = [X, Y]^\top$.

Because all main variables are discrete, each row $\vec{D}_i^{(s)}$ is a draw from a categorical distribution over a vector-valued alphabet representing possible joint values of the measured variables, with event probabilities that we denote $\vec{\theta}_{\vec{D}^{(s)}}$. In our running example, this alphabet is $\left\{[x_0, y_0]^\top, \ldots, [x_1, y_1]^\top\right\}$, and event probabilities are $\vec{\theta}_{XY} = [\ p(x_0, y_0), \ldots, p(x_1, y_1)\ ]^\top$. For convenience, we will define the counting operation $\#\boldsymbol{D}^{(s)} \equiv \sum_{i=1}^{n^{(s)}} \left[\mathbb{1}\left\{\vec{D}_i^{(s)} = \vec{d}^{(s)}\right\}\right]$, for example, $\sum_{i=1}^{n^{(s)}} \left[\mathbb{1}\left\{\vec{D}_i^{(s)} = [x_0, y_0]^\top\right\}, \ldots, \mathbb{1}\left\{\vec{D}_i^{(s)} = [x_1, y_1]^\top\right\}\right]$. It can be seen that $\#\boldsymbol{D}^{(s)}$ follows a multinomial distribution.

As a technical condition, we require (iii–v) to be defined in terms of elementary arithmetic functionals[4] of the full data law of $\vec{V}$, i.e., the joint distribution over all factual and counterfactual versions of every variable in $\vec{V}$.[5] This is an mild regularity condition that, for (iii), permits essentially all standard estimands including additive effects, risk ratios, and odds ratios. Monotonic transformations of elementary arithmetic functionals, such as log odds ratios, are also straightforwardly handled.[6] It accommodates total effects, conditional effects, and mediated effects, including so-called "cross-world" quantities, such as controlled direct effects, that involve contradictory hypothetical scenarios that are not observable under any experimental design.[7] Similarly, on (v), our results accommodate any joint, marginal, or conditional information obtained from observational and experimental data collection, including data subject to missingness, mismeasurement, or selection. An example of a non-elementary arithmetic functional is the estimand $\mathbb{1}\{\text{ATE is rational}\}$.

3.2 Prior work on automatic sharp bounding

When analysts possess datasets $\left\{\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}\right\}$ that provide imperfect information on observed quantities $\left\{\vec{\theta}_{\vec{D}^{(1)}}, \ldots, \vec{\theta}_{\vec{D}^{(S)}}\right\}$, a target quantity τ may not be uniquely identified. This can remain true even as the number of observations in each dataset grows large, because issues such as omitted variables, selection,

[4] Elementary arithmetic operations are addition, subtraction, multiplication, and division.

[5] For example, the full data law in Section 2 is the joint distribution over all of the following: factual values of W; factual $X = X(W)$ in which the input W takes its natural value; counterfactuals of the form $X(w)$ that would occur if the input were manipulated to $w \in \{w_0, w_1\}$, factual $Y = Y(X, W)$; and various counterfactuals of the form $Y(x, W)$, $Y(X, w)$, and $Y(x, w)$ for $w \in \{w_0, w_1\}$ and $x \in \{x_0, x_1\}$. In other words, it is $\Pr[W = w, X = x, X(w_0) = x', X(w_1) = x'', Y = y, Y(x_0, W) = y', Y(x_1, W) = y'', Y(X, w_0) = y''', Y(X, w_1) = y'''', Y(x_0, w_0) = y''''', \ldots]$.

[6] This is because we will construct bounds on the estimand by computing the minimum and maximum value that is consistent with available information (i.e., subject to constraints that represent assumptions and observed data), and bounds on a mononotic transform of x can be obtained by first bounding x and then applying the transform.

[7] A simple example of a cross-world estimand is the probability of causation, $\Pr[Y(x_0) = y_0 \mid Y(x_1) = y_1]$, such as the proportion of individuals sickened by chemical exposure who would not have fallen ill if they had not been exposed. This is the elementary arithmetic functional $\Pr[Y(x_0) = y_0, Y(x_1) = y_1] \,/\, \Pr[Y(x_1) = y_1]$

mismeasurement, and missingness can mean there exists more than one possible τ^* value that is consistent with the available information. In such cases, the best that analysts can do is report *sharp bounds*, or an interval of the form $[\underline{\tau}, \overline{\tau}]$ in which endpoints represent the minimum and maximum τ^* values consistent with available information. Sharpness is a desirable property that essentially guarantees that the reported bounds are the narrowest possible valid bounds, i.e. that they have extracted all possible information from the available data. Notably, when available data is sufficient to uniquely point-identify the true value τ, sharpness ensures that the bounds collapse on this point, i.e. $\underline{\tau} = \overline{\tau} = \tau$.

Duarte et al. (nd) prove that for every possible setting described in Section 3.1—for any target quantity, under any DAG, given any collection of potentially imperfect information—the problem of obtaining sharp causal bounds can be reduced to *polynomial programming*, or the minimization and maximization of a polynomial objective function subject to polynomial equality and inequality constraints. Intuitively, this transformation proceeds by defining a collection of optimization variables that represent the principal strata sizes for every variable in the graph. Any statement involving an event $p(\vec{d}^{\,s})$ can then be re-expressed in terms of the addition and multiplication of these strata sizes. Duarte et al. (nd) provide a series of algorithms, collectively referred to as `autobounds`, which conduct this principal stratification, transform causal inference problems into constrained optimization problems, and solve these problems to obtain sharp bounds. We rely heavily on these prior results to obtain

$$[\underline{\tau}, \overline{\tau}] = \texttt{autobounds}\left(\vec{\theta}_{\vec{D}^{(1)}}, \ldots, \vec{\theta}_{\vec{D}^{(S)}}\right), \text{ the population bounds, and}$$
$$[\hat{\underline{\tau}}, \hat{\overline{\tau}}] = \texttt{autobounds}\left(\hat{\vec{\theta}}_{\vec{D}^{(1)}}, \ldots, \hat{\vec{\theta}}_{\vec{D}^{(S)}}\right), \text{ the estimated bounds.}$$

3.3 Statistical inference on bounds

In what follows, we will rely heavily on the notion of an *fail-to-reject* region of possible target-quantity values that combines both (1) *fundamental uncertainty*, which arises e.g. from the fact that even infinite data on X and Y is insufficient to uniquely identify the average treatment effect of $X \to Y$ in the presence of confounding by W; and (2) *statistical uncertainty*, which arises e.g. from the fact that when only finite samples are available, $\vec{\theta}_{XY}$ is likely to be estimated with substantial error. To quantify both sources of uncertainty, we will utilize the *confidence bounds*, $[\underline{\tau}_\alpha, \overline{\tau}_\alpha]$, which are designed to widen the estimated bounds by some amount to ensure that the population bounds will be fully contained within the confidence bounds in at least $1 - \alpha$ of repeated samples.

The confidence bounds represent an interval of possible answers to the causal query that currently cannot

be ruled out. When analysts refer to the inadequacy of currently available data, a natural interpretation is that this fail-to-reject region is too wide to test a hypothesis with the desired power—for example, when the fail-to-reject region crosses zero, analysts cannot distinguish whether the treatment is beneficial or harmful. In what follows, we will develop techniques for targeting data collection toward parts of a system that contribute most to this overall uncertainty.

However, a key challenge is that existing methods for constructing confidence bounds are problematic. We briefly review two general methods before developing an alternative that addresses their limitations.

Duarte et al. (nd) offers a frequentist approach that quantifies the uncertainty in estimates of the observed quantities, $\widehat{\vec{\theta}}_{XY}$. It constructs a hypercube that is guaranteed to contain the true $\vec{\theta}_{XY}$ in at least $1-\alpha$ of repeated samples. Finally, it finds the most extreme possible values of the target quantity that can arise from any $\vec{\theta}^{*}_{XY}$ within this region. A problem with this approach is that it is highly conservative—in simulations, Duarte et al. (nd) found that 95% confidence bounds constructed with this approach achieved a 100% coverage rate. As a result, the amount of statistical uncertainty will tend to be overstated, potentially leading to misallocation of data-collection resources.

An alternative, fully Bayesian approach was first proposed by Chickering and Pearl (1996) and generalized by Zhang and Bareinboim (2021) to non-geared discrete causal systems. In this approach, analysts place a uniform Dirichlet prior over principal strata sizes, rather than the observed quantities. The bounds then act as an asymptotically rectangular likelihood function. Outside the bounds, the likelihood goes to zero as sample sizes grow large, because data is inconsistent with certain combinations of principal strata sizes; inside the bounds, the likelihood inside the bounds is flat, because data is uninformative. While intuitively appealing, Richardson et al. (2011) document severe issues with this approach: seemingly uninformative priors on principal strata often translate into highly informative implicit priors on the quantity of interest, posterior inferences are highly sensitive to small perturbations of priors, and confidence bounds constructed with this procedure can be almost surely invalid over repeated samples. At its core, these issues arise due to fundamental challenges with Bayesian inference on unidentified quantities, where typical Bernstein-von Mises guarantees do not apply.

To address this limitation, we develop a new technique based on the *transparent reparameterization* proposed by Richardson et al. (2011), which distinguishes between causal parameters that are identified and unidentified. Given a collection of datasets $\{\boldsymbol{D}^{(1)},\ldots,\boldsymbol{D}^{(S)}\}$, we first collect consider the set of quantities that are or could be identified—namely, $\vec{\theta}_{\boldsymbol{V}}$. For these quantities only, we impose a uniform prior; this ensures that the posterior $p\big(\vec{\theta}_{\boldsymbol{D}^{(1)}},\ldots,\vec{\theta}_{\boldsymbol{D}^{(S)}}|\boldsymbol{D}^{(1)},\ldots,\boldsymbol{D}^{(S)}\big)$ is asymptotically normal as $n^{(1)},\ldots,n^{(S)}$

grow large. For example, in the motivating problem where $\boldsymbol{D}^{(1)}$ consists of $n^{(1)} = 100$ samples from $[X, Y]^\top$, we impose e.g. a uniform prior $\vec{\theta}_{WXY} \sim \text{Dirichlet}(1)$ and obtain the posterior $(\vec{\theta}_{XY}|\boldsymbol{D}^{(1)}) \sim \text{Dirichlet}(\#\text{D}^{(1)} + 1)$, where $\#\boldsymbol{D}$ denotes the number of times that each possible row $\vec{d} \in \mathcal{S}(\vec{D})$ occurs in a dataset $\boldsymbol{D}$. We jointly sample $\vec{\theta}^*_{\boldsymbol{D}^{(1)}}, \ldots, \vec{\theta}^*_{\boldsymbol{D}^{(S)}}$ from the posterior $p\big(\vec{\theta}_{\boldsymbol{D}^{(1)}}, \ldots, \vec{\theta}_{\boldsymbol{D}^{(S)}}|\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}\big)$ before turning to the partially identified principal-strata parameters. Because the available data may impose little to no restriction on many of these parameters, we use `autobounds` to engage in best- and worst-case reasoning about the values that they may take on. Because this is a deterministic algorithm, we compute a single posterior sample of the bounds by $[\underline{\tau}^*, \overline{\tau}^*] = \texttt{autobounds}\Big(\vec{\theta}^*_{\vec{D}^{(1)}}, \ldots, \vec{\theta}^*_{\vec{D}^{(S)}}\Big)$, obtaining a single posterior sample of the best- and worst-case bounds. By taking the $\frac{\alpha}{2}$-th and $(1 - \frac{\alpha}{2})$-th quantiles of the lower- and upper-bound samples, respectively, we obtain credible intervals that possess asymptotic frequentist guarantees on e.g. nominal coverage rates. For compactness, we will refer to this procedure as $[\underline{\tau}_\alpha, \overline{\tau}_\alpha] = \texttt{autobounds}_\alpha\left(\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}\right)$.

4 A decision-theoretic framework for data collection

We are now ready to develop a general procedure for identifying an optimal follow-up data-collection task from a set of T candidate tasks $\left\{\boldsymbol{C}^{(S+1,1)}, \ldots, \boldsymbol{C}^{(S+1,T)}\right\}$ with costs $B^{(S+1,1)}, \ldots, B^{(S+1,T)}$, based on a possibly empty set of S datasets that have already been collected $\left\{\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}\right\}$. First, recall that after S rounds of data collection, the current fail-to-reject region at confidence level α is

$$\left[\underline{\tau}^{(S)}_\alpha, \overline{\tau}^{(S)}_\alpha\right] = \texttt{autobounds}_\alpha\left(\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}\right).$$

Next, observe that if task t were implemented to reveal $\boldsymbol{C}^{(S+1,t)}$, the fail-to-reject region would become

$$\left[\underline{\tau}^{(S+1,t)}_\alpha, \overline{\tau}^{(S+1,t)}_\alpha\right] = \texttt{autobounds}_\alpha\left(\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}, \boldsymbol{C}^{(S+1,t)}\right).$$

Note that $\underline{\tau}^{(S+1,t)}_\alpha$ and $\overline{\tau}^{(S+1,t)}_\alpha$ are deterministic functions of the random matrix $\boldsymbol{C}^{(S+1,t)}$ and are therefore themselves random variables. The task that achieves the maximal expected information efficiency—i.e. allows analysts to rule out the largest swath of possible values for the target quantity per unit cost—is

$$\arg\min_t \; \tfrac{1}{B^{(S+1,t)}} \mathbb{E}_{\boldsymbol{C}^{(S+1,t)}}\left[\overline{\tau}^{(S+1,t)}_\alpha - \underline{\tau}^{(S+1,t)}_\alpha \mid \boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}\right]$$

which can be numerically approximated to arbitrary precision in three steps. First, we will simulate a hypothetical dataset $\boldsymbol{c}^{(S+1,t)}$ from the current posterior over plausible datasets, $p\big(\boldsymbol{c}^{(S+1,t)} \mid \boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}\big)$; we will describe how to do so below. Second, we append that hypothetical dataset to the existing actual datasets $\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}$, apply `autobounds`, note the width of the fail-to-reject region that would be obtained if the t-th data-collection task had produced $\boldsymbol{c}^{(S+1,t)}$, and average this over many hypothetical datasets to numerically compute the expected fail-to-reject region width. Finally, we repeat the preceding steps for every candidate task and select the one that yields the narrowest expected fail-to-reject region.

Hypothetical draws from $p(\boldsymbol{c}^{(S+1,t)} \mid \boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)})$ are generated in two stages, based on the factorization $p(\boldsymbol{c}^{(S+1,t)}) = p\left(\boldsymbol{c}^{(S+1,t)} | \vec{\theta}_{\vec{C}^{(S+1,t)}}\right) p\left(\vec{\theta}_{\vec{C}^{(S+1,t)}} \mid \boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}\right)$. We will describe these in turn. In the first stage, by definition, each row is an i.i.d. draw of the vector-valued categorical random variable $\vec{C}^{(S+1,t)}$, with category proportions $\vec{\theta}_{\vec{C}^{(S+1,t)}}$. For example, consider task $t = 2$ in Figure 2(c), in which $\vec{C}^{(S+1,2)} = [W]$. Each row in this $m^{(S+1,2)} \times 1$ random matrix is an i.i.d. draw from $\{[w_0], [w_1]\}$, with probabilities $\vec{\theta}_W = [p(w_0), p(w_1)]^\top$. For task $t = 5$ in Figure 2(f), in which $\vec{C}^{(S+1,5)} = [W, X, Y]^\top$, each row in this $m^{(S+1,5)} \times 8$ random matrix is an i.i.d. draw from $\{[w_0, x_0, y_0], \ldots, [w_1, x_1, y_1]\}$ with probabilities $\vec{\theta}_{WXY} = [p(w_0, x_0, y_0), \ldots, p(w_1, x_1, y_1)]^\top$.

The second deals with the fact that the correct category proportions, $\vec{\theta}_{\vec{C}^{(S+1,t)}}$, are unknown—we can only draw posterior samples of plausible values based on the data already collected, $\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(S)}$. As in Section 3.3, we place a uniform Dirichlet prior on the joint distribution of observable values—in the running example, $\vec{\theta}_{WXY}$—then incorporate the observed data. Because the likelihood can involve count data from disparate margins, analytic computation of the posterior can be complex: for example, $\boldsymbol{D}^{(1)}$ may consist of samples of $[X, Y]^\top$, while $\boldsymbol{D}^{(2)}$ might involve samples of $[W, X]^\top$. To carry out posterior sampling, we therefore conduct Markov chain Monte-Carlo using the Stan software package (Team, 2023).

Note that this step can involve inferences that are entirely unsupported by data. In the confounding example, initial data may only consist of for $[X, Y]^\top$ samples, meaning that the evaluation of task $t = 2$ in Figure 2(c)—collection of W samples—relies solely on the analyst's prior on $\vec{\theta}_W = [p(w_0), p(w_1)]^\top$. By the same token, inferences can be only partially supported by data. A concrete example is the case discussed above, where analysts have collected samples from $[X, Y]^\top$ and $[W, X]^\top$, then seek to evaluate the effectiveness of candidate data collection on $[W, X, Y]^\top$. In this case, inferences about the marginal distribution of each variable is informed by data, but analysts rely solely on the prior to guess what dependence between W and Y might be revealed by this candidate data-collection task. Crucially, our procedure guarantees that this information from the prior is never used as the basis of a scientific claim. This is because our proce-

dure distinguishes between what can be thought of as an exploratory phase (the forward-looking question of what data to collect next, where cost-benefit evaluation inherently relies on beliefs about what new patterns might be observed) and the confirmatory or hypothesis-testing phase (the backward-looking question of what claims are supported by the data that has now been collected). In the former phase, we rely pragmatically on Bayesian methods to speculate about causal parameters that may be currently uninformed or partially informed by data, but with the assurance that we will either (1) then collect precisely the data needed to measure that parameter, if the follow-up candidate task is selected; or (2) discard the speculation, instead conducting best- and worst-case reasoning about the parameter's possible values. In the latter phase, we use Bayesian methods solely for uncertainty quantification, and solely only to the extent that they are guaranteed to be asymptotically equivalent to frequentist alternatives. For this reason, our procedure can be thought of as analogous to power calculation, in that it rests on beliefs or assumptions that will subsequently be verified empirically.

Applying this procedure iteratively will yield a greedy sequential data collection strategy. The dynamic programming solution to the problem of sequential data collection remains the subject of ongoing work.

4.1 Simulations

4.2 Confounding simulation

Figures 3–4 depict the application of our proposed method to the simple confounding scenario. In Figure 3, we suppose that an analyst has completed initial data collection and is assessing the five candidate follow-up data-collection tasks outlined above. Our proposed method conducts this evaluation by Monte-Carlo simulation. For each task, 200 hypothetical datasets that could be observed are drawn from the multinomial-Dirichlet posterior over all possible datasets that could result from that task. Estimated bounds and the 95% fail-to-reject region are then computed for each hypothetical dataset. Finally, the expected gain width of the resulting fail-to-reject region is computed for each task by averaging over the 200 hypothetical datasets.

Next, in Figure 4, we show how this procedure can be iteratively applied to obtain a greedy algorithm for sequential data collection. At each stage, researchers identify the data-collection task that is expected to produce the narrowest fail-to-reject region, leading to the greatest gain in information about the target quantity.

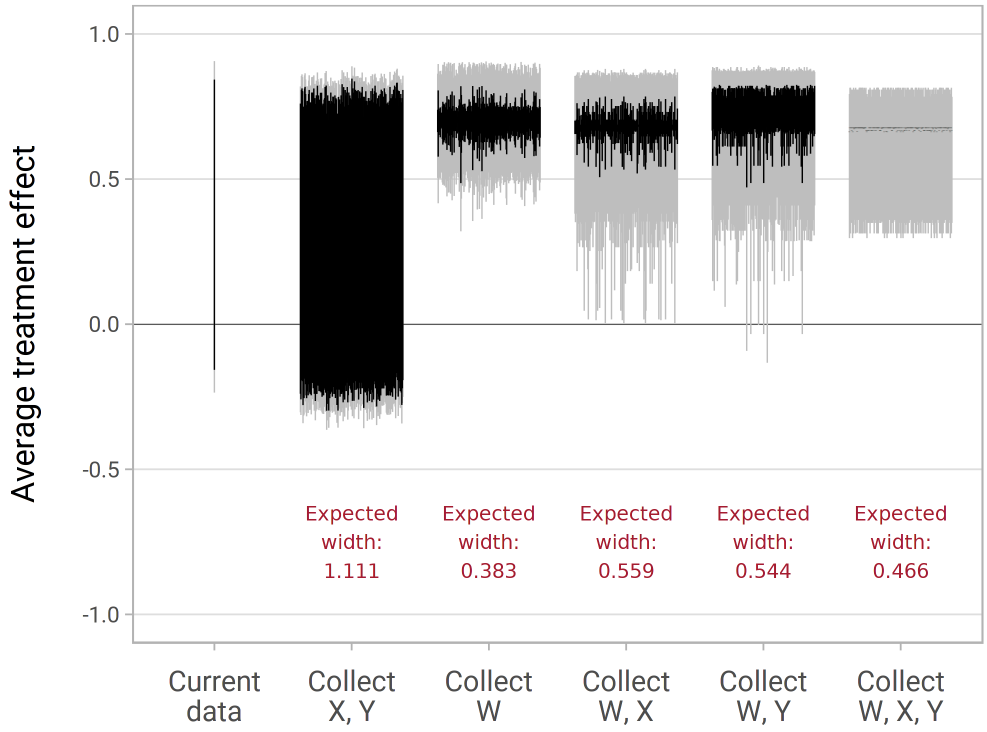


Figure 3: **Evaluation of one-shot data-collection candidate tasks**. Leftmost error bar depicts current estimated bounds on the average treatment effect (black) and current 95% fail-to-reject region (gray) based on initial data depicted in Figure 2(a). The current width of the fail-to-reject region represented by the error bar is 1.142. Each subsequent cluster of error bars depicts a candidate follow-up data-collection task in Figure 2(b–f). For each candidate task, hypothetical datasets are sampled from a Dirichlet-multinomial posterior over possible datasets that might be observed if researchers carried out the task. Each error bar within a cluster represents the fail-to-reject region that would be obtained if that new data had been collected. Red text below each cluster represent the expected width of the fail-to-reject region after conducting a follow-up data-collection task, holding total budget fixed. For example, the expected width after allocating this budget entirely to samples from $p(x, y)$ is 1.111, representing an information gain (i.e. narrowing of the fail-to-reject region) of only 0.031. In contrast, allocating this same budget to collecting less-expensive samples of $p(w)$ would result in an expected width of 0.383, for a gain of 0.759. Collecting confounder data is therefore more efficient.

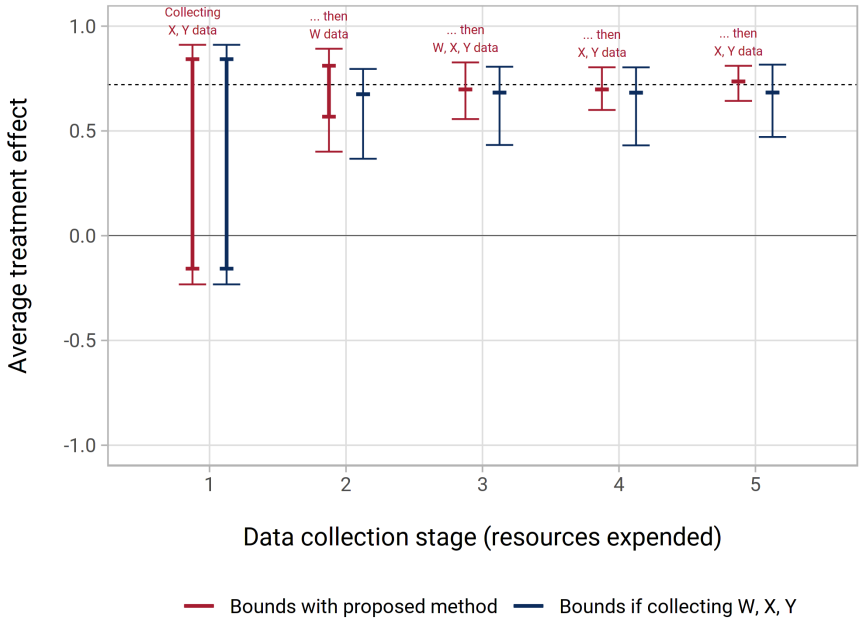
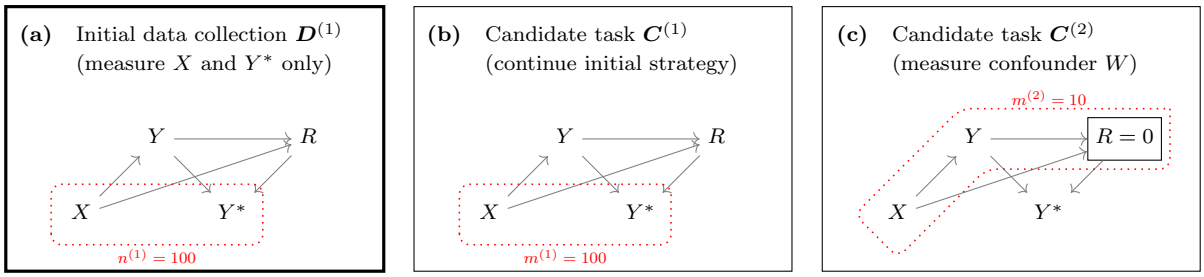


Figure 4: **Performance of greedy sequential data-collection candidate task selection for confounding**. Red error bars depict estimated bounds (thick lines) and 95% fail-to-reject regions (thin lines) on the average treatment effect after collecting data in successive stages. At each stage, the locally optimal task—obtained following the approach of Figure 3—is described in red text. Blue error bars depict the estimated bounds and 95% fail-to-reject regions that would result if analysts had followed a textbook strategy of expending the full budget on samples from $p(w, x, y)$ instead. The rightmost pair of error bars shows that this textbook approach produces a fail-to-reject region that is more than twice as wide.

5 Missingness simulation

Figure 5 depicts the available options for the nonrandom outcome missingness problem. In this scenario, researchers possess an initial sample from the distribution $p(x, y^*)$, shown in panel (a), in which 10% of respondents nonrandomly attrit from the study in ways that are correlated with treatment outcome and the true, unobserved outcome.

Figure 5: **Possible data-collection tasks in a study with nonrandom attrition.** Panel (a) depicts an initial dataset, $\boldsymbol{D}^{(1)}$, that contains $n^{(1)}$ samples from the marginal distribution of X and Y^*. Subsequent panels depict candidate tasks for a second round of data collection.



We conduct a simulation that is designed to mimic a challenging attrition problem that can arise in randomized experiments. The true average treatment effect in this simulation is -0.01, but if researchers select the task shown in panel (b)—i.e., continuing the initial strategy—they will be unable to correctly determine even the sign of this quantity. In our simulation, ignoring the issue and simply dropping missing values will in expectation produce a biased estimate of $+0.01$, precisely the opposite of the true value. However, merely computing Manski (1990) bounds will yield a fail-to-reject region that always spans zero, meaning both positive and negative effects cannot be ruled out. To extract further information, researchers must pursue the approach of panel (c), selecting a random sample of 10 units with missing outcomes and expending resources to reveal their true outcome, much as canvassers may knock on doors to contact subjects that fail to respond to initial phone calls. Perhaps surprisingly, Figure 6 shows that relatively little of this follow-up is required to obtain precise estimates of the treatment effect; nearly all resources are expended on the primary task, which improves estimates of the average outcome among responders; out of the first 50 data-acquisition stages, only two are spent on follow-up with non-responders. Intuitively, this is because non-responders represent only 10% of the population, and therefore even noisy estimates of the average outcome among this group can rapidly narrow down the range of possible effect values.

For illustrative purposes, the initial results presented above cover two challenges that are widespread in observational and experimental data analysis, respectively. In the full paper, the method is illustrated

Figure 6: **Performance of greedy sequential data-collection candidate task selection for non-random missingness**. In the upper panel, red error bars depict estimated bounds (thick lines) and 95% fail-to-reject regions (thin lines) on the average treatment effect after collecting data greedily in successive stages using our proposed method. In the lower panel, blue error bars depict the inferences that could be drawn if analysts had instead pursued the common approach of expending the full budget on samples from $p(w, y^*)$. Note that under this common approach, the sign of the average treatment effect will never be identified.

with simulations for each scenario from Figure 1. In ongoing work, we are extending this method in two directions: (1) the evaluation of mixed data-collection strategies, in which a budget is split over two or more tasks; and (2) dynamic programming approaches to the sequential data-collection problem.

References

Chickering, D. M. and J. Pearl (1996). A clinician's tool for analyzing non-compliance. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 1269–1276.

Duarte, G., N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser (n.d.). An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association (Theory & Methods)*.

Evans, R. J. (2016). Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics 43*(3), 625–648.

Gaesser, B., Y. Shimura, and M. Cikara (2020). Episodic simulation reduces intergroup bias in prosocial intentions and behavior. *Journal of Personality and Social Psychology 118*(4), 683.

Imai, K., D. Tingley, and T. Yamamoto (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 176*(1), 5–51.

Kaplan, R. M. and C. J. Atkins (1987). Selective attrition causes overestimates of treatment effects in studies of weight loss. *Addictive Behaviors 12*(3), 297–302.

Manski, C. (1990). Nonparametric bounds on treatment effects. *The American Economic Review 80*(2), 319–323.

Richardson, T. S., R. J. Evans, and J. M. Robins (2011). Transparent parameterizations of models for potential outcomes. *Bayesian Statistics 9*, 569–610.

Richardson, T. S. and J. M. Robins (2013a). Single world intervention graphs: A primer. In *Second UAI workshop on causal structure learning, Bellevue, Washington*. Citeseer.

Richardson, T. S. and J. M. Robins (2013b). Single world intervention graphs (SWIGs) : A unification of the counterfactual and graphical approaches to causality. *Working Paper, Center for Stat. & Soc. Sci., U. Washington 128*(30).

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling 7*(9-12), 1393–1512.

Shpitser, I. (2018). Identification in graphical causal models. In M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright (Eds.), *Handbook of Graphical Models*. CRC Press.

Team, S. D. (2023). Stan modelling language users guide and reference manual.

Weuve, J., E. J. T. Tchetgen, M. M. Glymour, T. L. Beck, N. T. Aggarwal, R. S. Wilson, D. A. Evans, and C. F. M. de Leon (2012). Accounting for bias due to selective attrition: the example of smoking and cognitive decline. *Epidemiology (Cambridge, Mass.) 23*(1), 119.

Whitman, J. D., J. Hiatt, C. T. Mowery, B. R. Shy, R. Yu, T. N. Yamamoto, U. Rathore, G. M. Goldgof, C. Whitty, J. M. Woo, et al. (2020). Test performance evaluation of sars-cov-2 serological assays. *Nature biotechnology 38*(10), 1174.

Zhang, J. and E. Bareinboim (2021). Non-parametric methods for partial identification of causal effects. Technical report, Causal Artificial Intelligence Lab, Columbia University.