Partial Identifiability in Discrete Data With Measurement Error

Noam Finkelstein[1] **Roy Adams**[2] **Suchi Saria**[1,3] **Ilya Shpitser**[1]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA
[2]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University, Baltimore, Maryland, USA
[3]Bayesian Health, New York, New York, USA

Abstract

When data contains measurement errors, it is necessary to make modeling assumptions relating the error-prone measurements to the unobserved true values. Work on measurement error has largely focused on models that fully identify the parameter of interest. As a result, many practically useful models that result in *bounds* on the target parameter – known as partial identification – have been neglected. In this work, we present a method for partial identification in a class of measurement error models involving discrete variables. We focus on models that impose linear constraints on the target parameter, allowing us to compute partial identification bounds using off-the-shelf LP solvers. We show how several common measurement error assumptions can be composed with an extended class of instrumental variable-type models to create such linear constraint sets. We further show how this approach can be used to bound causal parameters, such as the average treatment effect, when treatment or outcome variables are measured with error. Using data from the Oregon Health Insurance Experiment, we apply this method to estimate bounds on the effect Medicaid enrollment has on depression when depression is measured with error.

1 INTRODUCTION

Measurement error is a problem in fields ranging from machine learning to medicine to the social and behavioral sciences. In machine learning, mislabeled training data may lead to degraded model performance [Frénay and Verleysen, 2013]. In medical research, incomplete patient histories or misdiagnosed conditions may lead to biased estimates of treatment–outcome relationships [Rothman et al., 2008, Brakenhoff et al., 2018]. In fields such as economics and political science, which rely heavily on survey data, factors such as poor question design, social stigma, and recall bias can all lead to spurious responses [Molinari, 2008, Imai and Yamamoto, 2010]. Inferences drawn from such data may differ in a systematic way from the truth, leading to **measurement error bias**. As messy observational data is increasingly used to make decisions and inform policy, it is critical that we develop methods to account for measurement error bias in a wide range of settings.

In both statistics and machine learning, methods that account for measurement error either (a) do not give formal bias guarantees, (b) rely on validation data containing both true values and error-prone measurements, or (c) rely on domain-specific assumptions about how measurement errors occur. As formal guarantees are desirable and validation data is frequently unavailable or costly, we focus on the latter of these options and refer to the measurement error assumptions, collectively, as the **measurement error model**. In statistics, many such measurement error models have been proposed including the classical, Berkson's, and mean independent error models, all of which apply to continuous variables (see Carroll et al. [2006] and Gustafson [2003] for reviews of such models). In machine learning, the majority of work on measurement error has focused on classification in the presence of label errors [Frénay and Verleysen, 2013]. In this setting, common measurement error models include bounded error probabilities [Liu and Tao, 2015, Natarajan et al., 2013], label independent errors [Angluin and Laird, 1988, Ghosh et al., 2017], perfect separability of the true class labels [Ghosh et al., 2017], or some combination thereof.

Both the machine learning and statistics literatures have focused primarily on measurement error models that guarantee **full identification**. That is, in the limit of infinite data, these models guarantee that the target parameter(s) can be narrowed to a single value. While full identifiability is a desirable property, many common measurement error settings, especially those involving discrete variables, do not

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:1798–1808.

satisfy the necessary assumptions. In such settings, it may instead be possible, using the available assumptions, to identify *bounds* on the target parameter, referred to as **partial identification** [Manski, 1990]. Unfortunately, partially identifiable measurement error models remain under-studied.

We address this gap by proposing an easy to implement method to account for measurement error in a class of partially identifiable discrete variable models, many of which cannot be handled using current methods. Our approach, which is similar to that of Balke and Pearl [1993] and Imai and Yamamoto [2010], is to encode the measurement error model as a set of constraints on the target parameter and to calculate bounds by maximizing and minimizing the target parameter over this constraint set. We focus primarily on measurement error models that produce *linear* constraints, allowing us to write this optimization problem as a linear program. This approach allows a practitioner to mix and match any combination of such modeling assumptions with little effort. This flexibility means that it is trivial to compute bounds under new measurement error models, enabling sensitivity analysis to different modeling choices.

In this work, we propose a general method for partial identification in measurement error models where the modeling assumptions can be encoded as linear constraints on the target parameter. Our primary contribution is to define useful classes of modeling assumptions that can be written as linear constraints and, thus, are amenable to this approach. In Section 3, we show that this includes several common measurement error assumptions arising in settings where, for example, credible bounds on error rates are known, errors may only occur in one direction, or errors are more likely between certain categories than others. Additionally, in certain settings we may have access to auxiliary variables that give additional information about the error process, resulting in improved bounds. In our main result (Section 4), we show how an extended class of instrumental variable-type models produce linear constraints, allowing us to incorporate a variety of such variables. This extended class allows us to use classic instruments – i.e. variables that affect the measurement only via the true value – as well as informative variables that are not classic instruments. This includes variables that affect the measurement directly, such as question order in a survey or surveyor gender, as well as variables that are confounded with (rather than a cause of) the true value.

In Section 5, we show how our method can be applied to causal parameters, such as the average treatment effect (ATE), when outcome, treatment, or confounding variables are measured with error. Finally, using data from the Oregon Health Insurance Experiment [Finkelstein et al., 2012], we demonstrate our approach by estimating bounds on the effect of Medicaid enrollment on depression when using an error prone measurement for the outcome variable. We use our approach to test how sensitive the observed effect is to varying degrees of measurement error under several different error models. In the following section, we introduce the basic measurement error problem and the linear programming approach.

2 MEASUREMENT ERROR AND THE LP FORMULATION

Suppose that we are interested in the distribution of a discrete random variable X with support in $\mathcal{X}$, but instead of observing X directly, we can only observe a discrete error-prone measurement, denoted Y, with support in $\mathcal{Y}$. Without any assumptions about the measurement distribution, $P(Y \mid X)$, we cannot say anything about the distribution of X. On the opposite extreme, if $P(Y \mid X)$ is known and invertible, then $P(X)$ is fully identifiable from observations of Y [Rothman et al., 2008, Kuroki and Pearl, 2014]. Our interest is in between these two extremes, where we assume certain properties of $P(X, Y)$, but not the whole distribution. For example, we might assume that the measurement is more often correct than incorrect, or that errors can only occur in one direction. We refer to these assumptions collectively as the **measurement error model**. Our goal is to use a measurement error model, along with observations of Y, to bound a scalar function η of the data distribution, which we refer to as the **parameter of interest**. Common parameters of interest include marginals or moments of P, such as $\mathbb{E}[X]$.

Our approach to bounding these parameters is to translate all modeling assumptions into a set of distributions $\mathcal{M}$ that we assume contains $P(X, Y)$. We can then bound the parameter of interest by finding its maximum and minimum over the set $\mathcal{M}$. Formally, let Δ^d be the d-dimensional simplex, let $\eta : \Delta^{|\mathcal{X}|\times|\mathcal{Y}|} \to \mathbb{R}$ be the parameter of interest, and let $\mathcal{M} \subseteq \Delta^{|\mathcal{X}|\times|\mathcal{Y}|}$ be the set of distributions allowed under the modeling assumptions. Then, assuming that $\mathcal{M}$ contains the true distribution P, we can bound $\eta(P)$ as

$$\theta_L = \inf_{\substack{Q\in\mathcal{M} \\ Q(Y)=P(Y)}} \eta(Q) \leq \eta(P) \leq \sup_{\substack{Q\in\mathcal{M} \\ Q(Y)=P(Y)}} \eta(Q) = \theta_U. \tag{1}$$

The interval $[\theta_L, \theta_U]$ is referred to as the **partial identification bounds**. If $\theta_L = \theta_U$, we say η is **fully identified** under model $\mathcal{M}$. Note that we explicitly require that Q and P match on the observed variables – i.e. $Q(Y) = P(Y)$ – which we refer to as the **observed data constraints**. In finite samples, we will instead constrain Q to match the empirical distribution; however, we defer details on estimation to Section 1 of the supplementary materials. Until then, we will assume that the distribution of the observed variables is known. Additionally, constraining Q to be in $\mathcal{M}$ enforces the constraint that Q is a proper distribution (i.e. is non-negative and sums to one) which we refer to as the **probability constraints**.

In this work, we focus entirely on parameters η that are linear in P. This includes many common parameters such as all marginals and uncentered moments of P, notably $P(X)$ and $\mathbb{E}[X]$. If, moreover, the *model* $\mathcal{M}$ can be written as a set of linear constraints, then the upper and lower bounds in Equation 1 form linear programs (LPs) which can be solved using any off-the-shelf LP solver. By the intermediate value theorem, it follows that if $\mathcal{M}$ can be linearly expressed, the bounds obtained through these linear programs are sharp, i.e. no tighter bounds can be achieved without making additional assumptions. In the remainder of the paper, we define and operationalize several useful classes of modeling assumptions that can be written as linear constraints and are amenable to the LP approach.

3 ERROR ASSUMPTIONS

We focus first on assumptions that directly constrain the parameters of $P(X, Y)$. Such assumptions, which we refer to as **measurement error constraints**, arise in a number of settings. For example, in settings where X represents a sensitive or stigmatized characteristic, such as drug use, and the measurement Y is obtained through participant self-report, it may be reasonable to assume that Y is never a false positive [Adams et al., 2019]. In the binary setting, this assumption translates to the constraint $P(X = 0, Y = 1) = 0$ which, importantly, is linear in $P(X, Y)$. In this section, we describe several such constraints and give an example linear program based on a combination thereof.

For notational simplicity, let $\phi_{x,y} = P(X = x, Y = y)$ be the (unknown) joint distribution of X and Y. As in Section 2, ϕ must satisfy the probability and observed data constraints. That is, $\sum_{x,y} \phi_{x,y} = 1$, $\phi_{x,y} \geq 0$ for all x and y, and $\sum_x \phi_{x,y} = P(Y = y)$ for all y. Below, we provide several useful linear measurement error constraints[1].

(A0) Bounded error proportion: $\sum_{x \neq y} \phi_{x,y} \leq \epsilon$

(A1) Unidirectional errors: $\sum_{y<x} \phi_{x,y} = 0$

(A2) Symmetric error probabilities: $\phi_{x,y} = \phi_{x,y'} \ \forall \ |x - y| = |x - y'|$

(A3) Error probabilities decrease by distance: $\phi_{x,y} \geq \phi_{x,y'} \ \forall \ |x - y| > |x - y'|$

(A4) Equal expectations: $\sum_{x,y}(x - y)\phi_{x,y} = 0$

Assumption (A0) may be reasonable when there is sufficient previous literature to specify a range of plausible error rates, but not the exact measurement distribution (e.g., see discussion of sensitivity analysis in Rothman et al. [2008]). Assumption (A1) may be used to represent positive label only data, which is common in areas such as ecology [Sólymos et al., 2012] and public health [Adams et al., 2019].

[1]Each of these constraints can be softened by adding a slack parameter which can be fixed or varied in a sensitivity analysis.

Assumption (A2) represents a generalization of the zero-mean measurement error assumption to discrete variables, commonly made in settings where the errors are due to imprecision of the instrument. Assumption (A3) may be reasonable in settings where the values of X are ordinal and small errors are more likely than large ones. Finally, (A4) applies when Y represents a noisy, but unbiased measurement of X. As with (A2), this may be reasonable when errors are due to imprecision.

We can now use any combination of these assumptions to define the measurement error model $\mathcal{M}$. As an example, suppose we are interested in bounding $\mathbb{E}[X]$ subject to assumptions (A0) and (A2). The resulting LP is shown below.

$$\begin{aligned} \text{objective:} \quad & \sum_{x,y} x\phi_{x,y} && (2) \\ \text{s.t.} \quad & \sum_x \phi_{x,y} = P(Y = y) && \forall\ y \\ & \phi_{x,y} \geq 0 && \forall\ x, y \\ & \sum_{x \neq y} \phi_{x,y} \leq \epsilon \\ & \phi_{x,y} \geq \phi_{x,y'} && \forall\ |x - y| < |x - y'| \end{aligned}$$

Note that we can easily add or remove constraints from $\mathcal{M}$, or vary parameters like ϵ, as a form of sensitivity analysis (e.g., see the sensitivity analysis in Section 6).

Were multiple measurements $\mathbf{Y} \equiv \{Y_1, ..., Y_K\}$ observed with no assumptions made about the relationship between them, then the observed data constraint would be expressed on $P(\mathbf{Y})$ and the target parameter would be expressed as a linear function of the full distribution $P(X, \mathbf{Y})$. Each Y_k would potentially be subject to its own measurement error constraints, depending on what knowledge is available about the measurement process. These constraints would be expressed on the marginals $P(X, Y_k)$, which maintains linearity. In the following section, we consider scenarios where auxiliary variables, such as instrumental variables, are also available.

4 AUXILIARY VARIABLES

In the previous section, we relied only on domain knowledge about the distribution $P(X, Y)$ to partially identify the target parameter. In some cases, we may have auxiliary variables, such as additional measurements or sources of variation in the measurement process, that give information about $P(X, Y)$ [Carroll et al., 2006]. Under certain assumptions, such variables may be used to obtain tighter partial identification bounds on the target parameter. For example, suppose that a patient's age can only affect the results of a medical test through the true (unobserved) value. Then variability in test results that is explained by patient age must be attributable to variability in the true value, which

gives us information about $P(X)$ and, potentially, η. This example illustrates one particularly important type of auxiliary variable known as an **instrumental variable** (IV). In this section, we show how the linear programming approach can be used when auxiliary variables obey the classic IV model, and then we generalize this to a broader class of IV-type models, allowing us to incorporate other types of auxiliary variables.

4.1 THE CLASSIC IV MODEL

In the classic IV model[2], shown as a Bayesian network in Figure 1, the observed variable A is referred to as an instrument for the relationship between X and Y which is, in turn, confounded by the unobserved variable Λ [Balke and Pearl, 1993]. Like X and Y, we will assume that A is discrete. In contrast to the typical IV setting, we assume X to be unobserved; however, we will call this model the classic IV model as the assumed conditional independencies remain unchanged. As described above, if we believe that the age of a study participant may only affect Y through X, then participant age may be a valid IV.

Previous work has shown that the constraints imposed by this model on the conditional distribution $P(X, Y \mid A)$ are linear [Fine, 1982, Balke and Pearl, 1993, Bonet, 2001, Swanson et al., 2018]. Thus, recalling the $P(A)$ is known, we are able to include the classic IV model as part of $\mathcal{M}$. To express these constraints *explicitly* as part of an LP, we will rely on **potential outcomes**. The potential outcome variable $X(a)$ represents the value X would have taken had we intervened to set $A = a$ [Rubin, 2005]. Let $\tilde{X} = \{X(a) : a \in \mathcal{A}\}$ represent the set of potential outcome variables for X under different interventions on its parent A and let $\tilde{Y} = \{Y(x) : x \in \mathcal{X}\}$ be similarly defined. The variable sets $\tilde{X}$ and $\tilde{Y}$ are referred to as **response function variables** [Balke, 1995] since $\tilde{X}$ can be thought of as a random function mapping values of A to values of X. Finally, let ψ be the joint distribution over $\tilde{X}$ and $\tilde{Y}$ such that $\psi_{\tilde{x},\tilde{y}} = P(\tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y})$. We will now compute bounds for η by solving an LP parameterized by ψ rather than ϕ. As observed in Balke and Pearl [1993] and Bonet [2001], all independencies in the classic IV model are now given by $A \perp \tilde{X}, \tilde{Y}$ which can be written

$$P(A = a, \tilde{X} = \tilde{x}, \tilde{Y} = \tilde{y}) = P(A = a)\psi_{\tilde{x},\tilde{y}}. \quad (3)$$

Under the causal consistency assumption [Pearl, 2009], which can be concisely phrased as $A = a \wedge X(a) = x \implies X = x$, and the independence assumption in Equation (3), ψ is connected to the conditional distribution $P(X, Y \mid A)$ by the linear map

$$\begin{aligned} P(X{=}x, Y{=}y \mid A{=}a) &= P(X(a){=}x, Y(x){=}y) \quad (4) \\ &= \sum_{\tilde{x},\tilde{y}} \mathbb{I}(\tilde{y}(x){=}y, \tilde{x}(a){=}x)\psi_{\tilde{x},\tilde{y}}. \end{aligned}$$

We can now enforce all constraints of the classic IV model by replacing $P(A, X, Y)$ with Equation 4 in the target parameter η and the observed data constraints $P(A, Y) = \sum_x P(A, X = x, Y)$, which maintains linearity in ψ. Additionally, all measurement error constraints, which are expressed on the marginal $P(X, Y) = \sum_a P(A = a, X, Y)$, can now be expressed as linear constraints on ψ by substituting ψ into these constraints according to Equation 4. As an example, suppose $\mathcal{M}$ combines the classic IV model with Assumption (A0). Then we can obtain bounds on $\mathbb{E}[X]$ using the following LP:

$$\begin{aligned} \text{objective: } & \sum_{a,\tilde{x},\tilde{y}} x(a)P(A{=}a)\psi_{\tilde{x},\tilde{y}} \quad (5) \\ & \text{s.t. } \psi_{\tilde{x},\tilde{y}} \geq 0 \\ & \sum_{x \neq y,a,\tilde{x},\tilde{y}} \mathbb{I}(\tilde{y}(x){=}y, \tilde{x}(a){=}x)\psi_{\tilde{x},\tilde{y}}P(A{=}a) \leq \epsilon \\ & \sum_{x,\tilde{x},\tilde{y}} \mathbb{I}(\tilde{y}(x){=}y, \tilde{x}(a){=}x)\psi_{\tilde{x},\tilde{y}} = P(Y{=}y \mid A{=}a). \end{aligned}$$

While the classic IV model has proven useful in a wide range of applications, there remain many useful auxiliary variables that are not covered by this model. For example, in survey settings, the question order or perceived gender of the surveyor can affect the observed responses, but both are independent of the true value [Catania et al., 1996, Huddy et al., 1997]. Question order and surveyor gender are not classic IVs, but they can still provide information about the relationship between X and Y. We now generalize the linearity result for the classical IV model to a class of IV-type models that includes this setting, as well as several others.

4.2 GENERAL IV MODELS

Before defining our general class of IV-type models, we establish some notation: Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a latent variable Bayesian network where $\mathbf{V}$ and $\mathbf{E}$ represent the vertices and edges of the network, respectively. For a variable $V \in \mathbf{V}$, let $Pa(V)$ be the parents of V in $\mathcal{G}$ and $Ch(V)$ be the children of V in $\mathcal{G}$. Equipped with this notation, we define the following class of general IV models:

Definition 4.1 (General IV model)**.** The latent variable Bayesian network $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a **general IV model** if there exists an unobserved variable $\Lambda \in \mathbf{V}$ such that: (1) all descendants of Λ, denoted $\mathbf{B}$, are children of Λ, (2) all *observed* non-descendants of Λ, denoted $\mathbf{A}$, have at most

[2] Several variations of the classic IV model have been proposed; however, we refer to the version used in [Balke and Pearl, 1993]. For a review of others, see Swanson et al. [2018].

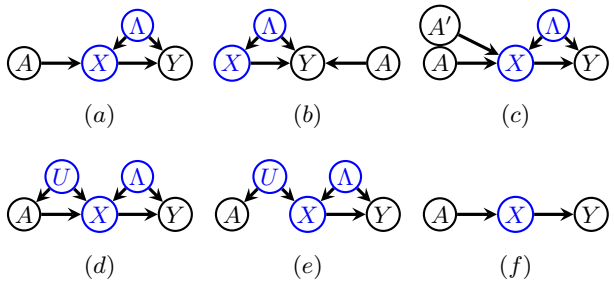


Figure 1: Model (a) represents the classic IV model, (b) - (e) represent general IV models, and (f) represents a simple model not covered by Definition 1. In all graphs, black nodes represent observed variables and blue nodes represent unobserved variables.

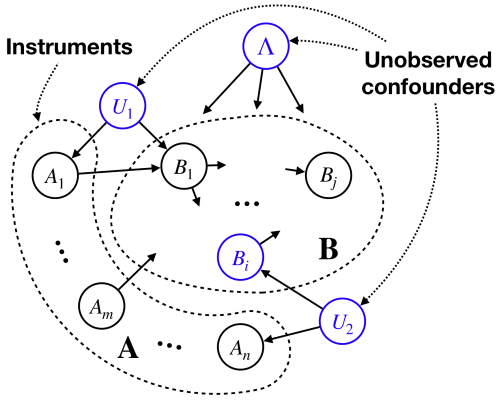


Figure 2: Illustration of a general IV model. Unobserved variables are shown in blue. The variables in $\mathbf{B}$ are all children of Λ. The variables in $\mathbf{A}$ are referred to as instruments, must be observed, and cannot be children of Λ.

one child and that child is in $\mathbf{B}$, and (3) all *unobserved* non-descendants of Λ are parents of exactly one variable in $\mathbf{B}$, denoted B, and one variable in $\mathbf{A}$, denoted A and, if A has a child, it must be B.

In such graphs, let $\mathbf{B}$ be the children of Λ and let $\mathbf{A}$ be the observed non-descendants of Λ, which we refer to as the **instruments**. An illustration of a general IV model is shown in Figure 2. This class of graphs trivially includes the classic IV model in Figure 1 (a) with $\mathbf{B} = \{X, Y\}$ and $\mathbf{A} = \{A\}$, but also extends the classic IV model in several important ways. First, we can now include instruments for Y as shown in Figure 1 (b). As described above, this may occur when some aspect of the measurement process is randomized, such as the order of responses in a survey or the gender of an in-person surveyor (e.g., see [Catania et al., 1996, Huddy et al., 1997]). Second, we can include multiple independent instruments as shown in Figure 1 (c) (e.g., see [Angrist and Keueger, 1991, Poderini et al., 2020]). Finally, this extends the classic IV model to allow for instruments in $\mathbf{A}$ to be confounded with variables in $\mathbf{B}$. This includes models such as those shown in Figures 1 (d) and (e). In particular, Figure 1 (e) can be used to represent a model where A is a proxy for the true unobserved IV, U.

By Proposition 5 in Evans [2016], for any general IV graph $\mathcal{G}$, it is possible to find a new graph $\mathcal{G}'$ that represents the same model over $\{\mathbf{A}, \mathbf{B}\}$ such that every instrument in $\mathcal{G}'$ is randomized, according to the following procedure:

Procedure 1 (Randomize Instruments)**.** *Remove each unobserved variable with children $A \in \mathbf{A}$ and $B \in \mathbf{B}$, and make A a parent of B (if it is not already).*

The new graph $\mathcal{G}'$ is more convenient to reason about, as the distribution of $P(\mathbf{B}(\mathbf{a}))$ is identified as $P(\mathbf{B} \mid \mathbf{A} = \mathbf{a})$. Since we are interested in parameters of $P(\mathbf{A}, \mathbf{B})$, we can without loss of generality reason about the modified graph $\mathcal{G}'$ rather than about the original graph $\mathcal{G}$. In Section 5, we will see that differences between $\mathcal{G}$ and $\mathcal{G}'$ become relevant when we consider causal parameters.

We now give our main result, which allows us to use LPs for partial identification in general IV models with measurement error. Let $\mathbf{B_O} \subseteq \mathbf{B}$ and $\mathbf{B_U} \subseteq \mathbf{B}$ be the set of observed and unobserved variables in $\mathbf{B}$, respectively. Then we have the following proposition:

Proposition 1. *For any graph $\mathcal{G}$ in the class of general IV graphs, the constraints imposed by the observed data $P(\mathbf{A}, \mathbf{B_O})$ and the model $\mathcal{G}$ on $P(\mathbf{A}, \mathbf{B})$ can be represented linearly.*

Recalling that bounds can be obtained by minimizing and maximizing a linear function of $P(\mathbf{A}, \mathbf{B})$ over constraints imposed by the observed data and the model, this leads directly to the corollary:

Corollary 1. *For any model $\mathcal{M}$ comprised of a discrete general IV model $\mathcal{G}$ and any number of linear measurement error constraints, sharp bounds may be obtained for any linear function of $P(\mathbf{A}, \mathbf{B})$ by solving an LP.*

To explicitly express the set of linear constraints imposed by a general IV model, we can generalize the response function approach described for the classic IV model. For each variable $B \in \mathbf{B}$, we will define the set of potential outcomes of B under different joint settings of its parents in $\mathbf{A} \cup \mathbf{B}$ – i.e., the response function variable for B. Formally, for each variable $B \in \mathbf{B}$, let $\tilde{B} = \{B(p) : p \in \text{support}(Pa(B) \cap (\mathbf{A} \cup \mathbf{B}))\}$ where support$(\mathbf{X})$ is the joint support of variables in $\mathbf{X}$. Further, let $\tilde{\mathbf{B}} = \{\tilde{B} : B \in \mathbf{B}\}$ be the collection of all $\tilde{B}$. As before, let $\psi_{\tilde{\mathbf{b}}} = P(\tilde{\mathbf{B}} = \tilde{\mathbf{b}})$ denote the joint distribution over all such response function variables. Because $P(\mathbf{A})$ is assumed to be known, the only relevant independency imposed by the graph is given

by $\mathbf{A} \perp \tilde{\mathbf{B}}$. This is a generalization of the independence constraint captured in Equation (3), and can be written

$$P(\mathbf{A} = \mathbf{a}, \tilde{\mathbf{B}} = \tilde{\mathbf{b}}) = \mathbf{P}(\mathbf{A} = \mathbf{a})\psi_{\tilde{\mathbf{b}}} \quad \forall\, \mathbf{a}, \tilde{\mathbf{b}}. \tag{6}$$

Finally, we must linearly link ψ to the observed full data distribution $P(\mathbf{A}, \mathbf{B})$. Under causal consistency and the assumption that all instruments have been randomized according to Procedure 1, we have $P(\mathbf{B}(\mathbf{a}) = \mathbf{b}) = P(\mathbf{B} = \mathbf{b} \mid \mathbf{A} = \mathbf{a})$ which links the full data distribution to a potential outcome distribution. This potential outcome distribution, $P(\mathbf{B}(\mathbf{a}))$ is, in turn, a linear function of ψ. In the classic IV model, this link is relatively transparent as demonstrated in Equation 4. We will use the following proposition to extend Equation 4 to general IV models.

Proposition 2. *Consider a set of outcome variables $\mathbf{Z} \subseteq \mathbf{B}$ and a set of treatment variables $\mathbf{T} \subseteq \{\mathbf{A}, \mathbf{B}\} \setminus \mathbf{Z}$ in a general IV model with randomized instruments. Then*

$$\begin{aligned} &P(\mathbf{Z}(\mathbf{t}) = \mathbf{z}) \\ &= \sum_{\mathbf{v}:\mathbf{v}_\mathbf{Z}=\mathbf{z}} P(\mathbf{A} = \mathbf{v}_\mathbf{A}) \\ &\qquad \cdot P\left(B(\mathbf{t}_{Pa(B)}, \mathbf{v}_{Pa(B)\setminus\mathbf{T}}) = \mathbf{v}_B : B \in \mathbf{B}\right) \end{aligned} \tag{7}$$

where for the set of values $\mathbf{v}$ and set of variables $\mathbf{S}$, $\mathbf{v}_\mathbf{S}$ is the values of $\mathbf{S}$ in $\mathbf{v}$ and

$$\begin{aligned} &P\left(B(\mathbf{t}_{Pa(B)}, \mathbf{v}_{Pa(B)\setminus\mathbf{T}}) = \mathbf{v}_B : B \in \mathbf{B}\right) \\ &\quad = \sum_{\tilde{\mathbf{b}}} \psi_{\tilde{\mathbf{b}}} \prod_{B \in \mathbf{B}} \mathbb{I}\left(\tilde{b}(\mathbf{t}_{Pa(B)}, \mathbf{v}_{Pa(B)\setminus\mathbf{T}}) = \mathbf{v}_B\right). \end{aligned} \tag{8}$$

Using this proposition, we can linearly map ψ to the full data distribution according to

$$\begin{aligned} &P(\mathbf{B} = \mathbf{b} \mid \mathbf{A} = \mathbf{a}) \\ &\qquad = \sum_{\tilde{\mathbf{b}}} \psi_{\tilde{\mathbf{b}}} \prod_{B \in \mathbf{B}} \mathbb{I}\left(\tilde{b}(\mathbf{a}_{Pa(B)}, \mathbf{b}_{Pa(B)\setminus\mathbf{A}}) = \mathbf{b}_B\right). \end{aligned}$$

As in the classic IV model, we can enforce the constraints of the general IV model by substituting ψ for $P(\mathbf{A}, \mathbf{B})$ in the target parameter, observed data constraints, and measurement error constraints according to this equation. For the purposes of illustration, Section 4 of the supplementary materials contains linear programs for the models shown in Figures 1 (b) - (e).

In Section 2 of the supplementary materials, we present a procedure to obtain non-sharp partial identification bounds under any graphical model $\mathcal{G}$, by relaxing the model until it is in the class of general IV models. In addition, we provide sharp bounds for the important case of Figure 1 (f), which is not a general IV model.

5 CAUSAL PARAMETERS AND CONSTRAINTS

Until now, we have focused on bounding functions of the full data distribution $P(\mathbf{A}, \mathbf{B})$. In this section, we extend the linear programming approach to *causal* parameters involving the distribution of one or more potential outcome variables. Suppose we are interested in the effect of treatment $\mathbf{T}$ on outcome $\mathbf{Z}$, i.e. in parameters of the potential outcome distribution $P(\mathbf{Z}(\mathbf{t}))$. Measurement error on the outcome, treatment, or observed confounders can all lead to biased parameter estimates if unaccounted for [Rothman et al., 2008]. In this section, we show how the constraints presented in the previous sections can be used to bound causal parameters in the presence of measurement error and introduce additional linear constraints that apply specifically to causal inference settings.

Because Procedure 1 *does* alter the causal model of a graph, we cannot use *any* general IV graph, as we did in the previous section. Instead, to make use of Proposition 2, we limit our attention in this section to general IV graphs *with randomized instruments*. By Proposition 2, $P(\mathbf{Z}(\mathbf{t}))$ is linear in ψ, and thus any linear function of $P(\mathbf{Z}(\mathbf{t}))$ can be bounded by employing the constraints on ψ described in the previous sections. This leads directly to the following corollary:

Corollary 2. *For any model $\mathcal{M}$ comprised of a discrete general IV model $\mathcal{G}$ with randomized instruments and any number of linear measurement error constraints, sharp bounds may be obtained for any linear function of $P(\mathbf{Z}(\mathbf{t}))$ by solving an LP.*

A number of important causal parameters can be written as linear functions of $P(\mathbf{Z}(\mathbf{t}))$, most notably the average treatment effect (ATE) defined as $\mathbb{E}[\mathbf{Z}(\mathbf{t}) - \mathbf{Z}(\mathbf{t}')]$ and the probability of a non-zero treatment effect defined as $P(\mathbf{Z}(\mathbf{t}) \neq \mathbf{Z}(\mathbf{t}'))$.

Remark 1. *Corollary 2 makes no mention of whether the variables in $\mathbf{Z}$ and $\mathbf{T}$ are observed. As a result, sharp bounds on causal parameters like the ATE can be obtained even when treatment, outcome, or both are subject to measurement error.*

In addition to the graphical and measurement error assumptions discussed so far, it often makes sense to encode further **causal assumptions** into the model. One especially important causal assumption is the causal **monotonicity assumption**, which relates potential outcomes under different interventions. For intervention variable T and potential outcome $Z(t)$, the general monotonicity assumption can be written as

(A5) Monotonicity:
$P(Z(t) = z, Z(t') = z') = 0 \;\forall t' > t, z' < z.$

Assumption (A5) can be applied to cases where it is believed that receiving a binary treatment cannot decrease the outcome; however, it can also be applied to the measurement error setting to encode the assumption that increasing the true value cannot lead to a decrease in the measurement. Additional causal constraints – such as limits on the effect size or the proportion affected, or the assumption of decreasing returns of increases in an ordinal treatment value – may be similarly imposed. As with the measurement error assumptions, equality constraints can be relaxed by specifying that the sums are bounded from above, rather than identically equal to zero.

6 EMPIRICAL EXAMPLE: THE OREGON HEALTH INSURANCE EXPERIMENT

To demonstrate the LP approach, we analyzed the effect of Medicaid enrollment on mental health outcomes using public data from the Oregon Health Insurance Experiment (OHIE) [Finkelstein et al., 2012][3]. In 2008, the state of Oregon expanded Medicaid coverage using a lottery to determine who would become eligible for enrollment. This randomization created a natural experiment, allowing researchers to study the effects of Medicaid coverage on healthcare usage and health outcomes. For complete details on the OHIE, see Finkelstein et al. [2012]. In one such study, Baicker et al. [2018] found, among other things, that Medicaid enrollment reduced depression as measured by the Patient Health Questionnaire (PHQ-9) taken approximately two years after the lottery.

The PHQ-9 is a nine question survey measuring various depressive symptoms on a 0 to 3 point scale. The total of these points is frequently used as a measure of overall depression, with a score above 10 serving as a cutoff for moderate to severe depression [Kroenke et al., 2001]. However, as a measurement of diagnosable depression, PHQ-9 scores are subject to various forms of measurement error. For example, Kroenke et al. [2001] acknowledged that scores between 10 and 15 represent a "gray zone", with much lower precision than scores above 15. Torous et al. [2015] found that observed PHQ-9 scores were sensitive to the way the survey was administered, with average scores reported on a mobile app 3 points higher than average scores reported to a live surveyor. In this section, we estimate bounds on the effect of Medicaid enrollment on depression and use our method to test the sensitivity of these estimates to different combinations of measurement error assumptions.

[3]The full dataset can be found at `https://www.nber.org/programs-projects/projects-and-centers/oregon-health-insurance-experiment`.

6.1 TARGET PARAMETER AND MODELING ASSUMPTIONS

Our target parameter is the ATE of Medicaid enrollment on the presence of moderate to severe depression as measured by the PHQ-9 score. We restrict our analysis to single person households resulting in a sample of size $N = 9,599$ lottery enrollees. Following Kroenke et al. [2001], we categorized the PHQ-9 scores into 5 bins representing no (0-4), mild (5-9), moderate (10-14), moderately severe (15-20), and severe (> 20) depression. As in Baicker et al. [2018], we treat winning the enrollment lottery as a binary instrumental variable for Medicaid enrollment. Let $A \in \{0, 1\}$ represent winning the lottery, let $T \in \{0, 1\}$ represent Medicaid enrollment, let $X \in \{1, ..., 5\}$ represent a person's true depression category, and let $Y \in \{1, ..., 5\}$ represent the measured PHQ-9 category. Then, the ATE is given by $P(X(T{=}1) > 2) - P(X(T{=}0) > 2)$.

For all analyses, we assume the general IV model shown in Figure 4 where Λ represents and unobserved confounder. Additionally, we make the following monotonicity assumption reflecting the belief that enrolling in Medicaid cannot increase depressive symptoms

$$P(X(T{=}0) = x, X(T{=}1) = x') = 0 \ \forall \ x < x'. \quad (9)$$

Even with no measurement error, the ATE is only partially identified under these assumptions. Our goal is to test the sensitivity of partial identification bounds on the ATE to various combinations of the following measurement error assumptions:

(EB) $P(|X(a) - Y(a)| > 0) \leq \epsilon \quad \forall \ a$

(Exp) $\mathbb{E}[X(a)] = \mathbb{E}[Y(a)] \quad \forall \ a$

(Sym) $P(X(a){=}x, Y(a){=}y) = P(X(a){=}x, Y(a){=}y')$ $\forall \ a, |x - y| = |x - y'|$

(Mon) $P(X(a){=}x, Y(a){=}y) > P(X(a){=}x, Y(a){=}y')$ $\forall \ a, |x - y| < |x - y'|$

All of these assumptions are versions of the measurement error assumptions listed earlier in Section 3. Assumption (EB) is version (A1) and says that the total proportion of errors is less than ϵ. The sensitivity parameter ϵ can be thought of as a total error budget, and we vary ϵ across the grid $[0.00, 0.001, ..., 0.04]$ to test how large ϵ can be before the partial identification set includes zero. Assumption (Exp) is a version of (A2) and says that the measured PHQ-9 category is an unbiased measurement of the true depression category. Assumption (Sym) is a version of (A2) and says that an error of magnitude k upwards is as likely as an error of magnitude k downwards. Finally, assumption (Mon) is a version of (A5) and says that small errors are more likely than large errors. Note that all assumptions were applied under both settings of the instrumental variable A, reflecting the belief that they apply regardless of whether the person

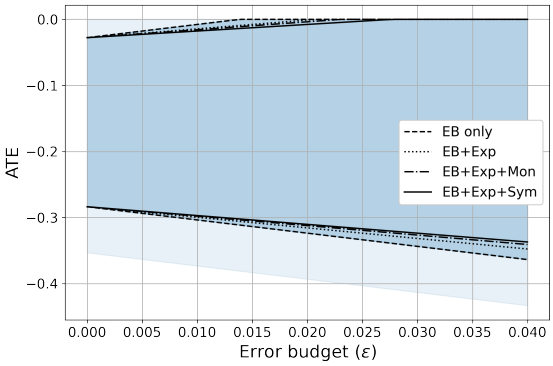
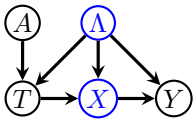


Figure 3: Partial identification bounds on the ATE of Medicaid enrollment on depression under different measurement error assumptions and error budgets ϵ. The dark and light blue regions represent the point estimate and 95% confidence interval, respectively, for the partial identification bounds under the weakest set of assumptions (EB only). Confidence intervals for the other models were omitted to avoid cluttering the plot.



Model	Max. budget
EB only	1.4%
EB+Exp	2.1%
EB+Exp+Mon	2.4%
EB+Exp+Sym	2.8%

Figure 4: The figure on the left shows the assumed genral IV model. The table on the right shows the maximum error budgets for each measurement error model.

wins the lottery or not. Importantly, our goal in this work is not to argue that any particular combination of these assumptions is "correct" in the context of the OHIE, a task we leave to subject-matter experts. Instead, our goal is to demonstrate how these assumptions may be flexibly applied to test the sensitivity of one's conclusions.

6.2 RESULTS

The estimated partial identification bounds under different measurement error models and error budgets ϵ are shown in Figure 3. The method used to estimate confidence intervals is described in Section 1 of the supplementary materials. In all cases, as ϵ grows, the upper bound approaches and eventually reaches zero. A quantity of interest is the ϵ at which the upper bound *hits* zero, which we refer to as the **maximum error budget**. The maximum error budget for each measurement error model is shown in the table in Figure 4. Unsurprisingly, using only the (EB) assumption resulted in the lowest maximum error budget whereas combining the (EB), (Exp), and (Sym) assumptions doubled this maximum error budget. However, no measurement error model resulted in a maximum error budget of more than 2.8% highlighting the potential sensitivity of these results to measurement error.

7 RELATED WORK

Measurement error occurs in many scientific settings and there is substantial literature on identification spread across a number of different methodological sub-disciplines. Much of this work concerns full identification in parametric models and we refer the interested reader to Carroll et al. [2006] and Gustafson [2003] for full treatments of these topics. In the machine learning literature, work on measurement error has primarily focused on measurement error models that allow for full identification and we again refer the interested reader to these works [Natarajan et al., 2013, Xiao et al., 2015, Liu and Tao, 2015, Ghosh et al., 2017, Adams et al., 2019, Shankar et al., 2019], in particular, the excellent review by Frénay and Verleysen [2013]. In this section, we review results on partial identification in measurement error and related settings.

Several works, particularly in econometrics, have presented partial identifiability results under various measurement error models. Horowitz and Manski [1995] considered the setting presented in Section 2, deriving sharp bounds on the distribution of the true value under a particular error model where data is "contaminated" by data from another, unknown, distribution. Molinari [2008] considered the same setting, presenting a procedure for *verifying* whether a particular distribution is in the partially identified set under a wide range of assumptions about the error distribution, including some non-linear assumptions. Henry et al. [2014] considered partial identifiability in a class of finite mixture models which includes, as a special case, the Markov chain model shown in Figure 1 (f), similarly proposing a method for verifying if a distribution is in the identified set. Our work differs from Molinari [2008] and Henry et al. [2014] in two important ways. First, the models covered by these methods have some overlap with those presented in this work, but, notably, do not cover the class of models described in Section 4. Conversely, by focusing on verification, Molinari [2008] and Henry et al. [2014] are able to cover certain non-linear models not covered by our approach. Second, computing partial identification bounds based on these methods requires performing guess-and-check which can be costly in high-dimensional spaces. For models where the methods do overlap with our approach, they will all produce the same sharp bounds; however, bounds can potentially be computed much faster via the linear programming method.

The linear programming approach we used to compute partial identification bounds is inspired by the approach used by Balke and Pearl [1993] to derive bounds on causal effects in trials with partial compliance. This approach was

first applied to the measurement error problem by Imai and Yamamoto [2010] to partially identify the ATE in a randomized trial under measurement error on the treatment variable; however, Imai and Yamamoto [2010] consider only a specific measurement error model which does *not* fall into the class of models considered here. A future direction of research may be to unify the model presented in Imai and Yamamoto [2010] with general IV models. Our work is also related to efforts to enumerate constraints on margins of the full data distribution implied by a latent variable Bayesian Network [Wolfe et al., 2019, Evans, 2012, 2018]. In such works, unobserved variables are not of primary interest and do not have known cardinality, so no attempt is made to bound functionals of their distribution. However, as indicated by our use of results from Bonet [2001], constraints on the observed data law can be used to derive restrictions on unobserved variables of known cardinality.

8 DISCUSSION

In this work, we presented an approach for computing bounds on distributional and causal parameters involving one or more discrete variables which are subject to measurement error. At the heart of this approach is the encoding of the target parameter and modeling constraints as linear functions of the joint distribution of all variables in the model. The target parameter can then be maximized and minimized, with respect to this distribution and subject to the the modeling constraints, to produce sharp bounds for any observed data distribution. In particular, we presented a class of graphical models that can be linearly expressed, and a procedure for finding a linear relaxation of models outside this class. We applied our approach to data from the Oregon Health Insurance Experiment, testing the sensitivity of conclusions drawn from this data to various measurement error assumptions.

As is generally the case, the validity of the bounds computed using this method depend on the validity of the measurement error model and an incorrect model may lead to biased bounds. This is true regardless of the method used to *compute* the bounds. Without validation data, it is not generally possible to test the validity of the measurement error constraints presented in Section 3; however, the models described in Section 4 imply certain inequalities on the observed data distribution that can be used to falsify a model [Bonet, 2001]. In fact, the LP approach provides a simple way to perform this test: If the LP is infeasible, then the model is inconsistent with the observed data distribution. In cases where the model is not falsified, we recommend testing the sensitivity of one's conclusions to each individual assumption as demonstrated in Section 6.

This work suggests several future lines of inquiry. We described a class of graphical models that result in linear modeling constraints; however, this class is certainly not exhaustive and work is needed to characterize which models result in non-trivial bounds for which target parameters. In particular, in Sections 5 and 6, we focused primarily on settings where the treatment or outcome variables are measured with error. Additional work is needed to extend this approach to settings with mismeasured confounders (e.g., see [Kuroki and Pearl, 2014]). Finally, this work focused on discrete data and additional work is needed to extended this approach to continuous distributions (e.g., to bound moments of these distributions, as in Henry et al. [2014]).

Author Contributions

Noam Finkelstein and Roy Adams had equal contributions to this paper.

Acknowledgements

This work was supported by funding from ONR grant number N00014-18-1-2760, NSF CAREER grant number 1942239, NSF grant numbers 1939675 and 1840088, and NIH R01 grant number AI127271-01A1. This information or content and conclusions are those of the authors and should not be construed as the official position or policy of, nor should any endorsements be inferred by ONR, NSF, NIH, or the U.S. Government.

References

Roy Adams, Yuelong Ji, Xiaobin Wang, and Suchi Saria. Learning models from data with measurement error: Tackling underreporting. In *International Conference on Machine Learning*, pages 61–70, 2019.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Joshua D. Angrist and Alan B. Keueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

Katherine Baicker, Heidi L. Allen, Bill J. Wright, Sarah L. Taubman, and Amy N. Finkelstein. The effect of medicaid on management of depression: evidence from the oregon health insurance experiment. *The Milbank Quarterly*, 96 (1):29–56, 2018.

Alexander Balke. *Probabilistic counterfactuals: semantics, computation, and applications*. University of California, Los Angeles, 1995.

Alexander Balke and Judea Pearl. Nonparametric bounds on causal effects from partial compliance data. *Journal of the American Statistical Association*, 1993.

Blai Bonet. Instrumentality tests revisited. In Jack S. Breese and Daphne Koller, editors, *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 48–55. Morgan Kaufmann, 2001.

Timo B. Brakenhoff, Marian Mitroiu, Ruth H. Keogh, Karel G.M. Moons, Rolf H.H. Groenwold, and Maarten van Smeden. Measurement error is often neglected in medical literature: a systematic review. *Journal of clinical epidemiology*, 98:89–97, 2018.

Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

Joseph A. Catania, Diane Binson, Jesse Canchola, Lance M. Pollack, Walter Hauck, and Thomas J. Coates. Effects of interviewer gender, interviewer choice, and item wording on responses to questions concerning sexual behavior. *Public Opinion Quarterly*, 60(3):345–375, 1996.

Robin J. Evans. Graphical methods for inequality constraints in marginalized dags. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2012.

Robin J. Evans. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 2016.

Robin J. Evans. Margins of discrete Bayesian networks. *Annals of Statistics*, 46(6A):2623–2656, 2018.

Arthur Fine. Hidden variables, joint probability, and the bell inequalities. *Physics Review Letters*, 48, 1982.

Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. The Oregon Health Insurance Experiment: Evidence from the First Year*. *The Quarterly Journal of Economics*, 127(3): 1057–1106, 07 2012.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

Aritra Ghosh, Himanshu Kumar, and P.S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Paul Gustafson. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press, 2003.

Marc Henry, Yuichi Kitamura, and Bernard Salanié. Partial identification of finite mixtures in econometric models. *Quantitative Economics*, 5(1):123–144, 2014.

Joel L. Horowitz and Charles F. Manski. Identification and robustness with contaminated and corrupted data. *Econometrica*, 63(2):281–302, 1995.

Leonie Huddy, Joshua Billig, John Bracciodieta, Lois Hoeffler, Patrick J. Moynihan, and Patricia Pugliani. The effect of interviewer gender on the survey response. *Political Behavior*, 19(3):197–220, 1997.

Kosuke Imai and Teppei Yamamoto. Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis. *American Journal of Political Science*, 54(2):543–560, 2010.

Kurt Kroenke, Robert L. Spitzer, and Janet B.W. Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613, 2001.

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

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

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

Francesca Molinari. Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, 144(1):81 – 117, 2008.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, volume 26, pages 1196–1204, 2013.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Davide Poderini, Rafael Chaves, Iris Agresti, Gonzalo Carvacho, and Fabio Sciarrino. Exclusivity graph approach to instrumental inequalities. In *Uncertainty in Artificial Intelligence*, pages 1274–1283. PMLR, 2020.

Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008.

Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Shiv Shankar, Daniel Sheldon, Tao Sun, John Pickering, and Thomas G. Dietterich. Three-quarter sibling regression for denoising observational data. In *Proceedings of the Twenty-Eighth International Joint Conference on*

Artificial Intelligence, IJCAI-19, pages 5960–5966. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

Péter Sólymos, Subhash Lele, and Erin Bayne. Conditional likelihood approach for analyzing single visit abundance survey data in the presence of zero inflation and detection error. *Environmetrics*, 23(2):197–205, 2012.

Sonja A. Swanson, Miguel A. Hernán, Matthew Miller, James M. Robins, and Thomas S. Richardson. Partial identification of the average treatment effect using instrumental variables: Review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947, 2018.

John Torous, Patrick Staples, Meghan Shanahan, Charlie Lin, Pamela Peck, Matcheri Keshavan, and Jukka-Pekka Onnela. Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (phq-9) depressive symptoms in patients with major depressive disorder. *JMIR mental health*, 2(1):e8, 2015.

Elie Wolfe, Robert W Spekkens, and Tobias Fritz. The inflation technique for causal inference with latent variables. *Journal of Causal Inference*, 7(2), 2019.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.