Ivy: Instrumental Variable Synthesis for Causal Inference

Zhaobin Kuang[†], Frederic Sala, Nimit Sohoni, Sen Wu,
Aldo Córdova-Palomera, Jared Dunnmon, James Priest, and Christopher Ré
Stanford University

Abstract

A popular way to estimate the causal effect of a variable x on y from observational data is to use an *instrumental variable* (IV): a third variable z that affects y only through x. The more strongly z is associated with x, the more reliable the estimate is, but such *strong IVs* are difficult to find. Instead, practitioners combine more commonly available *IV candidates*—which are not necessarily strong, or even valid, IVs—into a single "summary" that is plugged into causal effect estimators in place of an IV. In genetic epidemiology, such approaches are known as *allele scores.* Allele scores require strong assumptions—independence and validity of all IV candidates—for the resulting estimate to be reliable. To relax these assumptions, we propose Ivy, a new method to combine IV candidates that can handle correlated and invalid IV candidates in a robust manner. Theoretically, we characterize this robustness, its limits, and its impact on the resulting causal estimates. Empirically, we show that Ivy can correctly identify the directionality of known relationships and is robust against false discovery (median effect size ≤ 0.025) on three real-world datasets with no causal effects, while allele scores return more biased estimates (median effect size ≥ 0.118).

1 Introduction

A goal of causal inference is to ascertain the causal relationship between a pair of variables (the *risk factor* x and the *outcome* y) from observational data. This is difficult because causal relationships can be distorted by *confounders*: common causes of the risk factor and the outcome that may be unobserved. To address this difficulty, a third variable, called an *instrumental variable* (IV), can be used to estimate causal effect. Informally, an IV only affects the outcome through its effect on the risk factor. IV methods are widely used in practice (Angrist and Krueger, 1991; Mokry et al., 2015; Walker et al., 2017; Millwood et al., 2019). In particular, we are motivated by Mendelian randomization (MR) (Burgess and Thompson, 2015), a representative use case in which genetic markers serve as IVs to infer causation among clinical variables.

IV methods are most reliable when the IV z is strongly associated with the risk factor x, but such *strong IVs* are often difficult to identify in practice. Instead, practitioners typically rely on more readily available *IV candidates.* These variables may not be strong, or even valid, IVs, but can be used in lieu of an unavailable strong IV. To this end, a two-phase approach can be used: first, *synthesize*: combine the IV candidates into a summary variable, and secondly, *estimate*: plug the summary variable into a causal effect estimator.

In MR, a popular, state-of-the-art approach for the synthesis phase is *allele scores.* The summary variables generated by allele scores are meant to reduce bias in causal estimates (Angrist and Pischke, 2008; Davies et al., 2015). In the words of Burgess et al. (2017), allele scores are a "recent innovation" in MR and are a "recommend[ed]" way to utilize plentiful IV candidates—but with the caveat that if an IV candidate is not actually a valid IV, allele scores may lead to "potentially misleading estimates." Indeed, allele score methods suffer two main weaknesses: they implicitly assume that the IV candidates (1) are *all* valid IVs and (2) are independent conditioned on the summary variable (Sebastiani et al., 2012). When these assumptions are not met, as often happens in practice, the resulting estimate may turn out to be unreliable.[1]

To improve robustness against invalidity and dependencies among the IV candidates while still reaping the benefits of the two-phase approach (e.g., modularity and bias reduction), we propose Ivy, a novel way to synthesize a summary IV from IV candidates. Ivy

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

[1] See Appendix A for an extended discussion.

produces a summary IV by modeling it as a latent variable, and inferring its value based on the statistical dependencies among the IV candidates. Ivy is inspired by recent advances in the theory of weak supervision, leveraging results on structure learning (Varma et al., 2019). Ivy targets the synthesis phase and is orthogonal to the effect estimation phase: the summary IV it generates can directly be plugged into IV-based causal effect estimators, whether they are classical (Wald, 1940; Angrist et al., 1996), robust (Bowden et al., 2016; Kang et al., 2016), or modern (Hartford et al., 2017; Athey et al., 2019).

We provide theoretical bounds on the robustness of our approach against invalidity or dependencies among the IV candidates. Specifically,

- We analyze the parameter estimation error for Ivy. Under weaker assumptions than allele scores, and with sufficiently many samples, Ivy's error scales as $O(1/\sqrt{n})$ for n samples. Even outside of this regime, when Ivy may fail to identify all invalid IVs or dependencies, the resulting error is mild (scaling linearly in the number of misspecified dependencies and undetected invalid IVs).
- We translate the error in the parameter estimation into bounds for a downstream parametric causal effect estimator —the Wald estimator—which is a commonly used estimator in MR.
- We further adapt our analysis to show how, in contrast to Ivy, allele scores may produce unreliable estimates in the presence of invalidity or dependency among IV candidates.

Empirically, we show that Ivy can more reliably estimate causal effects compared to allele score methods, even with low-quality uncurated IV candidates with potential dependencies and invalidity. On three real-world datasets with no causal effects, Ivy yields median effect size less than 0.025, while allele scores return more biased estimates (median effect size $\geq$ 0.118). This result aligns with our theoretical insights into Ivy and allele scores.

2 Background

We consider a two-phase approach to estimating causal effects with IV candidates. First, the IV candidates are combined to form a summary (the synthesis phase). Second, in the effect estimation phase, this summary is plugged into an estimator, along with the risk factor and outcome, to produce an effect. Our approach tackles the first phase, and is orthogonal to the second phase. We give background on these ideas below.

We seek to infer the causal relationship between a risk factor x and an outcome y. This relationship may be distorted by a confounder c, which is a common cause of both x and y. To handle confounding, an instrumental variable z may be used. z directly induces a change in x independent of c. This change will alter the value of y only through the causal link between x and y, enabling us to measure the causal link (Figure 1a). We focus on the setting where x, y, c, and z are binary, although our procedure can be extended to handle continuous x, y, and c. A valid IV is a variable satisfying Definition 1; otherwise, it is invalid.

Definition 1 (Burgess and Thompson (2015))**.** An instrumental variable z satisfies (i) *Relevance*: z is not independent of the risk factor, i.e. $z \not\perp x$; (ii) *Exclusion Restriction*: z can only influence the outcome through x, i.e. $z \perp y \mid x, c$; (iii) *Unconfoundedness*: z is independent of the confounder, i.e. $z \perp c$.

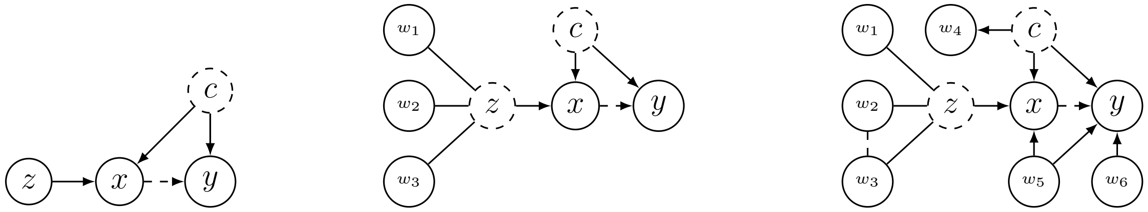
Figure 1a depicts the setting where a valid IV is observable. The dashed confounder node c indicates that IV methods can deal with unobserved confounders between x and y. By contrast, estimating effects without accounting for confounding may lead to failure in distinguishing between spurious correlation and causation. The following is a well-known example of spurious correlation in epidemiology, dismissed by a careful use of IVs.

Example 1. *The concentration of high-density lipoprotein (HDL) is negatively correlated with the occurrence of coronary artery disease (CAD) and thus appears protective, but recent studies suggest that there is no causal link. The correlation is spurious due to confounders such as the concentration of other lipid species (Rye and Ong, 2015). Nonetheless, the strength of this spurious correlation led to a hypothesized causal link, but drugs developed to raise HDL levels failed to prevent CAD (Schwartz et al., 2012). This spurious correlation was later dismissed by a series of MR studies (Voight et al., 2012; Holmes et al., 2014; Rader and Hovingh, 2014).*

2.1 IV Synthesis

The more strongly a valid IV is associated with the risk factor, the more reliable the resulting causal effect estimate. However, finding such strong IVs is challenging in practice. Instead, practitioners often combine more widely available IV candidates—variables that are weakly associated with the risk factors, intercorrelated, or even invalid IVs—into a summary IV. One way to view this procedure is that the summary IV is a prediction of a latent variable that, while unobserved, can serve as a strong IV.

Allele Scores The use of unweighted/weighted allele scores (UAS/WAS) to synthesize a summary IV



(a) Classic IV Setup (b) Ivy (Simple Setting) (c) Ivy (with Invalid and Correlated IVs)

Figure 1: IV method settings (unobserved variables are dashed; the dashed arrow between x and y is the causal relationship we seek to estimate, dashed edges are dependencies that we seek to infer): (a) the traditional setting with observed strong IV z, (b) a simple setting where we do not see z, but see noisy weak IV candidates w_1, w_2, w_3 independent conditioned on z, (c) a more challenging setting that Ivy can handle where some IV candidates have dependencies (w_2, w_3), others are invalid (w_4 violates unconfoundedness, w_5 and w_6 violate exclusion restriction, and w_6 violates relevance).

is a popular leading approach in MR (Burgess and Thompson, 2013; Davies et al., 2015; Burgess et al., 2016). UAS weights each IV candidate equally while WAS weights them based on their associations to the risk factor. While allele scores can mitigate bias induced by weak IV candidates, they assume that these IV candidates are all valid and independent conditioned on the summary (Figure 1b). Thus, dependencies (Sebastiani et al., 2012) or invalidity (Burgess et al., 2017) in IV candidates (Figure 1c) can still result in unreliable effect estimates when using the summary variable. Our proposed approach, Ivy, can be viewed as a *generalization* of allele scores to lessen these issues.

2.2 Effect Estimation

In the effect estimation phase, the risk factor x, the outcome y, and the summary (or, when available, the strong IV) z are used in an estimation procedure to obtain an estimate of the causal effect of x on y.

In MR, the standard estimator is the Wald ratio β_{zy}/β_{zx}, where β_{zx} and β_{zy} are the logistic regression coefficients of predicting x and y using z, respectively. While Ivy can be plugged into other estimators, we analyze the estimation phase for the commonly used Wald estimator in MR.

3 IV Synthesis With Ivy

We describe the Ivy framework for instrumental variable synthesis. We begin with our problem setup and assumptions. Then we present Ivy (Algorithm 1) and its components. Next, we theoretically characterize the model parameter estimation error in Ivy due to invalid IV candidates, misspecified dependencies, and sampling noise. Finally, we bound the impact of this error on downstream causal effect estimation.

3.1 Problem Setup

We seek to use a valid, but unobserved IV $z \in \{-1, 1\}$ to infer the causal relationship between the risk factor $x \in \{-1, 1\}$ and the outcome $y \in \{-1, 1\}$. This causal relationship is obscured by potentially unobserved confounders $c \in \{-1, 1\}^d$. The data generation process among x, y, z, and c follows some probability distribution $\mathcal{D}$. Although we do not directly observe z, we do observe m IV candidates $w \in \{-1, 1\}^m$. Only some of these m IV candidates are valid.

If the IV z could be observed, we could directly plug it into a causal effect estimator; unfortunately, z is rarely known in practice. Thus, the primary challenge is to reliably infer z from w, i.e. to estimate the distribution $\mathrm{P}(z \mid w)$, and to characterize how this impacts the reliability of downstream causal inference.

Notation We use "IV candidate" and "candidate" interchangeably. We call candidates that are valid/invalid IVs "valid/invalid candidates". We denote the index set of the valid candidates as $V \subseteq [m]$, where $[m] := \{1, 2, \ldots, m\}$. We use w_V to represent the subvector of the vector w indexed by V (i.e. the subvector corresponding to the valid candidates). When the subscript is omitted, $\|\cdot\|$ denotes the 2-norm.

Inputs and Outputs We have access to data $\left\{(x^{(i)}, y^{(i)}, w^{(i)})\right\}_{i=1}^{n}$: n samples each of the risk x, the outcome y, and the m IV candidates. Our goal is to produce a causal effect estimate $\hat{\alpha}_{x \to y}$ of x on y.

3.2 Assumptions

We explain the assumptions made by Ivy, in particular comparing to those made by allele scores. These are described in further depth in Section B.2.

First, we describe assumptions on validity. We assume the majority of IV candidates are valid IVs, and for the invalid candidates ($i \notin V$), $w_i \perp z$. These assumptions weaken those of allele scores, which assume that all candidates are valid IVs.

Next, we continue with assumptions on dependencies. To allow for dependencies, we model the candidates and z via an Ising model (the canonical binary maximum-entropy distribution with pairwise dependencies). We write the density of the model as

$$\tfrac{1}{\mathcal{Z}}\exp(\theta_z^* z + \Sigma_{i\in V}\theta_i^* w_i z + \Sigma_{(i,j)\in E}\theta_{ij}^* w_i w_j), \tag{1}$$

where $\mathcal{Z}$ is a normalization constant, E is the set of pairwise dependencies between valid IVs, and the θ^* terms are the model parameters. While allele scores require the maximal level of sparsity in the model (no dependencies, so that E is empty), our assumptions are weaker: we only require that for each valid IV candidate w_i there are at least two others that are independent of w_i and each other conditioned on z, and, conversely, that candidates that are dependent (i.e., in E) are all mutually dependent. Lastly, we require that on average, valid IV candidates agree with z more often than not. We discuss identifiability of causal effects in Appendix B.3.

3.3 Algorithmic Framework

We describe the Ivy framework (Algorithm 1). First, because our data may include both valid and invalid IV candidates, and because even the valid candidates may have dependencies, we learn a set of valid candidates and dependencies directly from our data (Algorithm 2). Next, we learn the mean parameters of the joint distribution of our estimated valid w_i's and z, without observing z (Algorithm 3). Concretely, (μ^*, O^*), the true mean parameters[2], are $\mathbb{E}[wz]$ and $\mathbb{E}[ww^T]$ (where $\mathbb{E}[wz]$ is a vector with entries $\mathbb{E}[w_i z]$). We observe the w's, so we can easily estimate O^* by $\hat{O}$. More challenging is to estimate μ^*, since we do not observe z; we use our learned dependencies and validity to estimate μ^* by $\hat{\mu}$. Finally, in Algorithm 4 we use $\hat{\mu}$ and $\hat{O}$ to form an estimate $\hat{z}$ of z. We also describe how to use $\hat{z}$ in a generic IV-based estimator F to get a causal effect estimate (the estimation phase). We describe the components of Algorithm 1 in detail.

Algorithm 1 Ivy Algorithmic Framework

```
Input: Data $\{(w^{(i)}, x^{(i)}, y^{(i)})\}_{i=1}^n$.
$\hat{V}, \hat{E} \leftarrow$ StructureLearn (data, $\lambda, \gamma, T_1, T_2$)
$\hat{\mu} \leftarrow$ ParamLearn (data, $\hat{V}, \hat{E}$)
$\hat{\alpha}_{x\to y} \leftarrow$ CausalEst (Estimator, data, $\hat{V}, \hat{\mu}$)
Output: Causal effect estimate $\hat{\alpha}_{x\to y}$.
```

Algorithm 2 Valid IV and Dependency Learning (StructureLearn)

```
Input: Data $\{w^{(i)}\}_{i=1}^n$, params. $\lambda$, $\gamma$, $T_1$, and $T_2$.
1: Compute sample covariance matrix $\hat{\Sigma}$ from $w^{(i)}$'s.
2: $(\hat{S}, \hat{L}) \leftarrow \operatorname{argmin}_{L\succeq 0,\, S-L\succ 0} \mathcal{L}(S - L, \hat{\Sigma}) + \lambda_n(\gamma\|S\|_1 + \|L\|_*)$,
   where $\mathcal{L}$ is a loss function.
3: $\hat{\ell} \leftarrow \operatorname{argmin}_{\ell} \|\hat{L} - \ell\ell^T\|_F$
4: $\hat{V} \leftarrow \{j : |(\hat{\Sigma}\hat{\ell})_j| \geq T_1\}$
5: $\hat{E} \leftarrow \left\{(i,j) : i,j \in \hat{V}, i < j, \hat{S}_{i,j} > T_2\right\}$
Output: Estimated valid IV candidate set $\hat{V}$, estimated dependency set $\hat{E}$.
```

Step 1: Identify Valid IV Candidates and their Dependencies. *Inputs*: data and hyperparameters. *Outputs*: estimated set of valid candidates $\hat{V}$ and estimated dependency set $\hat{E}$ of $\hat{V}$. Our method for learning the valid IVs and their dependencies is an application of recent approaches for structure learning (Varma et al., 2019) in graphical models. The main challenge is that without observing z, all of the valid IV candidates will appear to be correlated, although may be independent conditioned on z. Meanwhile, the valid and invalid candidates form mutually-independent components. We recover both the graph structure and the covariances between the IV candidates (valid and invalid) and z via a robust PCA approach. This enables us to estimate which IVs are valid and their statistical dependencies. The procedure is given in Algorithm 2.

Concretely, the identification of the valid candidates and their dependencies translates to decomposing a rank-one matrix and a sparse matrix from their sum (Line 2 of Algorithm 2). Here, the candidate validity ends up being encoded in the rank-one component $\hat{L}$ and the dependencies are encoded in the sparse component $\hat{S}$. Thus, we can threshold the vector corresponding to the rank-one matrix $\hat{L}$ to obtain the valid IVs and then threshold the corresponding submatrix of $\hat{S}$ containing valid IVs to obtain the dependencies. There are several choices of loss functions. For our analysis, we use $\mathcal{L}(S-L, \hat{\Sigma}) = \frac{1}{2}\mathrm{tr}((S-L)\hat{\Sigma}(S-L)) - \mathrm{tr}(S-L)$.

Step 2: Estimate Parameters of the Candidate Model. *Inputs*: data, $\hat{G} := (\hat{V}, \hat{E})$. *Outputs*: estimated parameters $\hat{O}, \hat{\mu}$. In Algorithm 3, we learn the mean parameters. We leverage conditional independencies encoded in our estimated dependency structure to obtain these parameters without ever observing z, via the agreements and disagreements of the IV candidates. We adapt Ratner et al. (2019).

Specifically, we set $a_j := w_j z$ for all $j \in \hat{V}$. Then

[2] These are expectations of the sufficient statistics in (1). $\mathbb{E}[z]$ is also a parameter; we assume it is known, but it can also be estimated (see, for example, Ratner et al. 2019).

Algorithm 3 Parameter Learning (ParamLearn)

```
Input: Data $\left\{w^{(i)}\right\}_{i=1}^{n}$, $\hat{G} = (\hat{V}, \hat{E})$ where $\hat{V}$ are estimated valid candidates, $\hat{E}$ are edges among them.
  Form estimated matrix $\hat{O} \leftarrow \frac{1}{n}\sum_{i=1}^{n} w_{\hat{V}}^{(i)}(w_{\hat{V}}^{(i)})^T$.
  $\hat{\Omega} \leftarrow \{(i,j) : w_i, w_j \text{ are disconnected in } \hat{G} \setminus \{z\}\}$
  Form matrix $M_{\hat{\Omega}}$ and vector $\hat{q}$ from $\hat{O}$
  $\hat{\ell} \leftarrow \text{argmin}_{\ell} \|M_{\hat{\Omega}}\ell - \hat{q}\|$
  $|\hat{\mu}| \leftarrow \exp(\hat{\ell}/2)$
  Recover $\text{sgn}(\hat{\mu})$
Output: Estimated model mean parameters $\hat{O}, \hat{\mu}$.
```

Algorithm 4 Synthesis & Causal Effect Estimation (CausalEst)

```
Input: Data $\left\{(w^{(i)}, x^{(i)}, y^{(i)})\right\}_{i=1}^{n}$, estimated parameters $\hat{O}, \hat{\mu}$, $\hat{V}$, $\hat{E}$, causal effect estimator $F(\cdot)$.
  for $i \in [n]$ do                                        ▷ Synthesize
      $\hat{z}^{(i)} \leftarrow \mathrm{P}_{\hat{\mu},\hat{O}}\left(z \mid w_{\hat{V}}^{(i)}\right)$
  end for
  $\hat{\alpha}_{x\to y} \leftarrow F\left(\left\{(\hat{z}^{(i)}, x^{(i)}, y^{(i)})\right\}_{i=1}^{n}\right)$.          ▷ Estimate
Output: Causal effect estimate $\hat{\alpha}_{x\to y}$.
```

the mean parameter $\mu_i := \mathbb{E}[a_i] = \mathbb{E}[w_i z]$. Since $z^2 = 1$, $\mathbb{E}[a_i a_j] = \mathbb{E}[(w_i z)(w_j z)] = \mathbb{E}[w_i w_j]$. We can estimate $\mathbb{E}[w_i w_j]$ from data. Moreover, if w_i and w_j are independent conditioned on z (i.e. (i,j) is an edge in $\hat{\Omega}$), then $\mathbb{E}[a_i a_j] = \mathbb{E}[a_i]\mathbb{E}[a_j]$, which means $\log \mathbb{E}^2[a_i] + \log \mathbb{E}^2[a_j] = \log \mathbb{E}^2[w_i w_j]$. We form a system of equations $M_{\hat{\Omega}}\ell = q$, with q the vector of $\log \mathbb{E}^2[w_i w_j]$ terms and ℓ the vector of $\log \mathbb{E}^2[a_i]$ terms. The matrix $M_{\hat{\Omega}}$ is formed by taking each $(i,j) \notin \hat{\Omega}$ and adding a row with a 1 in positions i and j and 0's elsewhere. We solve this to get estimates $\hat{\mu}_i$ of $\mathbb{E}[a_i]$ up to sign; using the assumption that valid candidates agree with z the majority of the time, we recover the signs. This gives $\hat{\mu}$ (and $\hat{O}$ was estimated earlier).

Step 3: Synthesize IV and Estimate Causal Effect *Inputs*: data, $\hat{O}, \hat{\mu}$, $\hat{V}$, $\hat{E}$, and causal effect estimator $F(\cdot)$. *Outputs*: causal effect estimate $\hat{\alpha}_{x\to y}$. Finally, in Algorithm 4, we generate a probabilistically synthesized version of z called $\hat{z}$ from our model parameterized by $\hat{O}, \hat{\mu}$. We obtain samples of z based on these to account for the uncertainty in the synthesized summary IV, concluding synthesis. We then feed these samples, along the risk factor and the outcome, to a causal effect estimator in the estimation phase, producing a causal effect estimate.

3.4 Theoretical Analysis

We theoretically analyze Ivy and provide bounds on its parameter estimation error. We further analyze the error in downstream causal effect estimation using the Wald estimator—a common estimator of causal effects in MR—as a proof-of-concept. We focus on the scaling with respect to the number of samples n and the number of IV candidates m. We present a simplified bound that explains the conceptual result, and provide a more general version in Appendix B.4.

Parameter Estimation Bound We show how the gap between the parameters μ^* of (1) and our estimated $\hat{\mu}$ decays with the number of samples.[3] We fix $R_{\min}$, the lowest correlation between valid candidates, and $C_{\min}$, the lowest accuracy for a valid candidate. Then, let c_0, c_1 be constants and d be the largest degree of a valid IV candidate in G.

Theorem 1. *Let $\hat{\mu}$ be the result of Algorithm 1 run on n samples of m IV candidates, where $m > c_0$. Denote μ^* to be the mean parameter of* (1). *If $n > c_1 d^2 m$, then with probability at least $1 - \frac{1}{m}$,*

$$\mathbb{E}[\|\hat{\mu} - \mu^*\|] \leq \frac{16m^{\frac{5}{2}}}{R_{\min}}\|M^{\dagger}\|\sqrt{\frac{2\pi}{n}}.$$

Remark The bound on the estimation error goes to 0 as $O(1/\sqrt{n})$, while it scales as $O(m^{5/2})$ in the number of IV candidates. The bound also depends on the smallest correlation between a pair of valid IVs; the smaller this term, the more samples we need to accurately estimate μ^*. $\|M^{\dagger}\|$ is the largest singular value of the pseudoinverse of $M := M_{\Omega}$, i.e., the true M formed with the edges from G; it indicates the cost of solving our problem (which is independent of n).

Under the assumptions in Section 3.2, Ivy can handle invalid candidates and dependencies in G. This is because with sufficiently many samples (the requirement $n > c_1 d^2 m$), the structure learning component correctly identifies valid candidates and the correct dependencies among them, with high probability. The more dependencies that have to be estimated (that is, the larger the number of sources m and degree d), the more samples we need. However, once we pass a threshold, we are operating only over valid IVs and a correct model, enabling the estimation error to go to zero. In Appendix B.4, we present a more technical result, applicable to the low-sample regime. In that case, the structure learning component may not identify all invalid IVs and may leave some edges, and we bound the impact of these unidentified invalid IVs and misspecified dependencies.

[3] In Appendix B.4 we bound $\mathbb{E}[\|\hat{O}-O^*\|]$ with Lemma 1.

Application to Allele Scores UAS implicitly follows the conditionally independent model above. Our framework helps obtain new insights on its behavior. Specifically, when the ground truth model is *not* conditionally independent, we can explain the approximation error in the parameters estimated by UAS.

As long as there is at least one misspecified dependency, the parameter error in UAS cannot go to zero. Specifically, let $n \to \infty$ and suppose there is a dependency between w_1 and w_2, but we miss it. Then, we do not have conditional independence, so $\mathbb{E}[w_1 w_2] \neq \mathbb{E}[a_1]\mathbb{E}[a_2]$. Form q' with $\mathbb{E}[a_1]\mathbb{E}[a_2]$ and q with $\mathbb{E}[w_1 w_2]$. We can write $q' - q = \delta e_1$ for some $\delta \neq 0$, since q is only incorrect in one position. Then, $\|\ell' - \ell\| = \|M^\dagger(q' - q)\| = \|M^\dagger(\delta e_1)\| = |\delta|\|M^\dagger e_1\| \geq \frac{|\delta|}{\|M\|}$, which is a lower bound that is independent of n. Thus we obtain that $\mathbb{E}[\|\mu' - \mu^*\|] > 0$.

Causal Effect Estimation Error Next, we bound the causal effect estimation error when using Ivy's synthesized IV. We bound the mean squared error $\mathbb{E}[(\hat{\alpha}_{x\to y} - \alpha^*_{x\to y})^2]$ between the effect with Ivy's version of z and that with the true z, as a function of the parameter error $\mathbb{E}[\|\hat{\mu} - \mu^*\|]$ we obtained in Theorem 1.

We use the popular Wald estimator as an example. Let β^*_{zx} and β^*_{zy} be the population-level coefficients of z from the logistic regressions to predict x and y under $\mathcal{D}$, and $\hat{\beta}_{\hat{z}y}, \hat{\beta}_{\hat{z}x}$ the corresponding regression coefficients of $\hat{z}$. Define $\alpha^*_{x\to y} := \beta^*_{zy}/\beta^*_{zx}$ as the population-level Wald estimator. Suppose that the population-level logistic loss of $\mathcal{D}$ satisfies Lemma 3 in Appendix B.5, so that it is λ-strongly convex. Again suppose $m > c_0, n > c_1 d^2 m$ and large enough such that for some $\kappa \in (0,1)$, $\max\{|\hat{\beta}_{\hat{z}y} - \beta^*_{zy}|, |\hat{\beta}_{\hat{z}x} - \beta^*_{zx}|\} \leq \kappa \beta^*_{zx}$, and let c_2 be a constant.

Theorem 2. *Run Algorithm 1 on n samples of m IV candidates to synthesize $\hat{z}$'s that are plugged into the Wald estimator to obtain the causal effect estimate $\hat{\alpha}_{x\to y}$. Then, the error in the estimate $\hat{\alpha}_{x\to y}$ compared to the true effect $\alpha^*_{x\to y}$ is bounded as follows:*

$$\mathbb{E}[(\hat{\alpha}_{x\to y} - \alpha^*_{x\to y})^2] \leq \sqrt{\frac{1}{n}} \cdot \frac{6000 c_2 m^{\frac{5}{2}} (\beta^*_{zx} + \beta^*_{zy})^2 (1 + \|M^\dagger\|)}{R_{\min} \lambda (1-\kappa)^2 \beta^{*4}_{zx}}.$$

Theorem 2 quantifies how the estimation error of z propagates to the downstream Wald estimator. The error goes to 0 as $1/\sqrt{n}$, suggesting that, under the conditions we described, we can indeed perform reliable causal inference from weak IV candidates. Our final observation is that model misspecification may lead to nonzero error in the causal estimates (see Section B.7): with even one misspecified dependency, $\mathbb{E}[\|\mu' - \mu^*\|] > 0$ with positive probability. We can lower bound $(\hat{\alpha}_{x\to y} - \alpha^*_{x\to y})^2$ in terms of $\mathbb{E}[\|\mu' - \mu^*\|]$, concluding that $(\hat{\alpha}_{x\to y} - \alpha^*_{x\to y})^2 > 0$ for such cases.

4 Experiments

We empirically validate that the summary IVs synthesized by Ivy lead to reliable causal effect estimates when plugged into standard causal effect estimators on real-world healthcare datasets. Specifically,

- In Section 4.1, we show, in clinically-motivated scenarios where only uncurated (potentially dependent or invalid) IV candidates are available, that Ivy can synthesize a summary IV that leads to more reliable effect estimates than allele scores.
- In Section 4.2, in scenarios with hand-picked curated (putatively valid and conditionally independent) IV candidates, we show that Ivy performs comparably well to allele scores.
- In Section 4.3, we evaluate the Ivy framework on synthetic data and further focus on its robustness against violation of key assumptions.

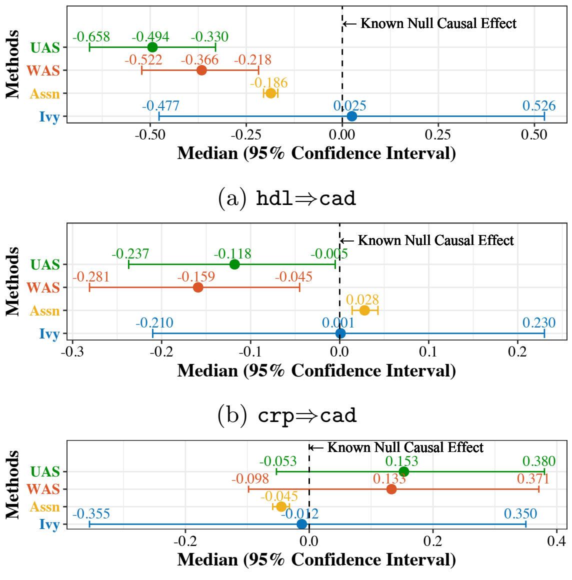
We describe the datasets, methods, and evaluation metrics and then report our primary findings.[4]

Datasets In collaboration with cardiologists, we selected real-world health data collected from the UK Biobank (Sudlow et al., 2015) for a variety of cardiac conditions. Because heart diseases are a major class of conditions affected by many factors, we examined five factors (for instance, we study the LDL-CAD link, as in Burgess et al. 2016). The most challenging aspect of selecting datasets for causal inference is the lack of ground truth effects. As a result, we have three desiderata for our dataset choices:

- We need some risk-outcome pairs where strong clinical evidence exists to support that there is *no causal relationship*, while for other pairs, there is strong evidence of a positive relationship;
- We require standard pairs that have previously been tested against in the MR literature;
- To evaluate performance in the favorable setting where IV candidates are valid and conditionally independent, we need access to curated sets of candidates.

The five risk factors we use are high-density lipoprotein (HDL), low-density lipoprotein (LDL), systolic blood pressure (SBP), C-reactive protein (CRP), and vitamin D (VTD). The outcome is occurrence of coronary artery disease (CAD). Single-nucleotide polymorphisms (SNPs) associated with these factors are used as IV candidates. These pairs are well-understood by clinicians, enabling us to use these pairs as proxies to the ground truth (Collaboration, 2011; Lieb et al., 2013; Holmes et al., 2014; Manousaki et al., 2016). Us-

[4] In Appendix C, we give further details about our setup and additional experiments.



(a) hdl⇒cad

(b) crp⇒cad

(c) vtd⇒cad

Figure 2: Wald ratios in estimating the causal effects of three risk factors (HDL, C-reactive protein, and vitamin D) to the occurrence of coronary artery disease using uncurated IVs. Goal: 0 causal effect.

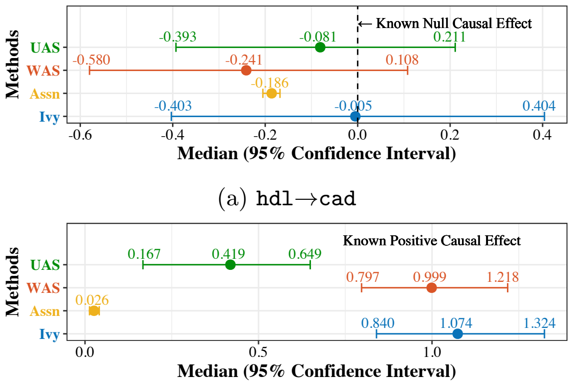
ing the risk factors, outcome, and IV candidates, we extract 11 datasets from the UK Biobank for our experiments (details in Table A.2).

Methods Ivy produces a summary in the synthesis phase, so we compare to allele scores—UAS and WAS—in Sections 4.1-4.3, as they also produce a summary IV. Additionally, we report results of logistic regression (Assn), which is a proxy for the confounded association between the risk factor and the outcome.

Metric After the synthesis phase, we use the summary IV in the estimation phase by plugging it into a causal effect estimator, along with the risk factor and the outcome. In all experiments, we use the Wald ratio to estimate effects. We report the median Wald ratio and its 95% confidence interval (CI). In MR, a CI that covers the origin is interpreted as no causal effect, while strictly positive/negative CIs indicate positive/negative causal effects.

4.1 MR with Uncurated IVs

We first use the summary variable synthesized by Ivy to draw causal inference in common clinical scenarios where only low-quality IV candidates are available. As shown in Figure 2, Ivy dismisses *known spurious* correlations on all three of the real-world datasets (median effect size ≤ 0.025); in comparison, allele scores yield more biased estimates (median effect size ≥ 0.118).



(a) hdl→cad

(b) ldl→cad

Figure 3: Wald ratios in estimating the causal effect between high(low)-density lipoprotein and coronary artery disease using curated IVs. Goal in (a): 0 causal effect. Goal in (b): positive causal effect.

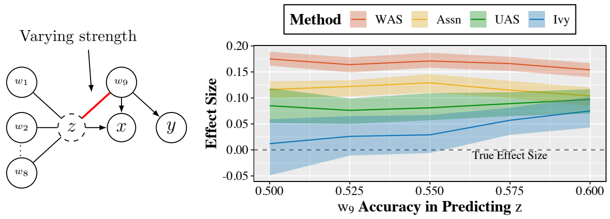
Specifically, we test spurious relationships between three potential risk factors (HDL, CRP, and VTD) and CAD: these are known to be noncausal, so the true effect size is 0. We compare Ivy with UAS, WAS, and Assn. Results are in Figure 2. Both UAS and WAS return negative causal effects for HDL (UAS median: -0.494; WAS median: -0.366; Figure 2a) and CRP (UAS median: -0.118; WAS median: -0.159; Figure 2b) with negative CIs. By contrast, Ivy does not identify a causal effect (Ivy median: 0.025 and 0.001 for HDL and CRP, respectively), with CIs covering the origin. In Figure 2c, the CIs of all three methods cover the origin, indicating successful dismissal. Nonetheless, the median estimates of UAS (0.153) and WAS (0.133) are skewed towards the positive direction, while Ivy's is very close to the origin (-0.012).

Ivy tends to have a wider confidence interval compared to allele scores, as it selects only a subset of IV candidates. Allele scores make use of all candidates regardless of their validity, and may be hurt by one or more being invalid. In all cases, Association (Assn) fails to dismiss spurious correlation, highlighting the importance of the use of IVs for debiasing causal estimates.

4.2 MR with Curated IVs

Next, we use a summary IV using a set of curated (putatively valid and conditionally independent) candidates with both known non-causal and known causal pairs. While all methods work, for the positive LDL-CAD relationship, Ivy retains the positive performance of WAS over UAS. The results are in Figure 3.

Concretely, since we are now in the fortunate (but rarer) setting in which the IV candidates are "good,"



(a) Causal model (b) Causal effect estimation

Figure 4: Dismissing spurious correlations when z is invalid. As the invalidity of z, i.e., the accuracy of w_9 in predicting z, increases, all methods eventually fail. However, Ivy is the most robust.

we expect that both Ivy and allele scores provide reasonable estimates. We use the known noncausal relationship between HDL and CAD (Example 1) and the known positive causal relationship between LDL and CAD. Ivy is compared with UAS, WAS, and Assn. In terms of dismissing spurious correlation (Figure 3a), the 95% CIs of all three IV-based methods (Ivy, UAS, WAS) cover the origin, indicating successful dismissal. Notably, the median estimate of Ivy is closest to the origin (-0.005) compared to other methods (UAS median: -0.081; WAS median: -0.241), suggesting a potentially less biased estimate from Ivy. Again, Assn fails to dismiss spurious correlation even in this "easier" setting.

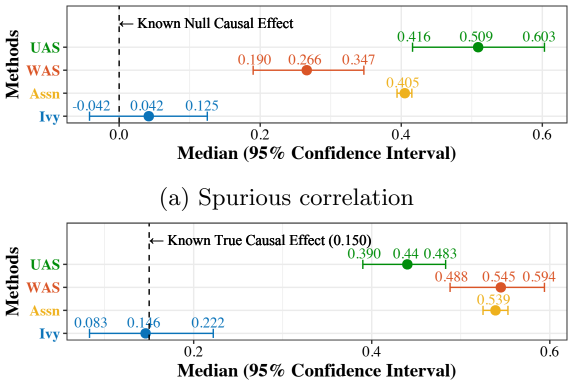
In terms of identifying a true causal relationship (Figure 3b), all three IV-based methods correctly identify the direction of the causal relationship (UAS median: 0.419; WAS median: 0.999; Ivy median: 1.074), as indicated by the positive CIs of the causal estimates. The lengths of the CIs of the three IV-based methods are also comparable to each other. On this dataset, Ivy yields an estimate most similar to that of WAS—matching the property that Ivy mimics allele scores in the setting where IV-candidates are high-quality.

4.3 Synthetic Experiments

Now we use synthetic data, controlling candidate properties and the ground-truth. We validate the robustness of Ivy and compare the effect to the ground-truth.

Robustness We investigate how robust Ivy is to an important violation of our main assumptions (that all the invalid candidates are independent of z). Then, the summary z itself may be an invalid IV. We show that Ivy yields a causal estimate that is more robust to this case compared to allele scores. Of course, when the invalidity is sufficiently strong, eventually Ivy also fails to dismiss a spurious correlation (Figure 4b).

We use the spurious correlation model in Figure 4a. The candidate w_9 serves as a confounder between the risk factor and the outcome. Here z is invalid because z is associated with w_9, and we increase this association strength (red edge) to force more invalidity. We expect Ivy to downweight the influence of w_9 while UAS and WAS may not. Indeed, Ivy performs well when z is nearly valid (i.e., nearly independent of w_9), and gradually degrades (blue curve), while allele scores immediately struggle. Eventually, increasing the amount of invalidity causes Ivy to fail as well.



(a) Spurious correlation

(b) True causal effect=0.150

Figure 5: Wald ratios in causal effect estimation using synthetic data. The true causal effects are 0 and 0.15.

Dismissing Spurious Correlations Next, we generate synthetic data with no causal effect along with valid and invalid IVs and adding dependencies. The results are in Figure 5a. Ivy recovers the dependency structure and identifies the invalid candidates. As a result, Ivy can successfully dismiss the spurious correlation by identifying no causal effects (Ivy median: 0.042) while both UAS and WAS fail to do so by yielding estimates that are consistent with the direction of the spurious correlation (UAS median: 0.266, WAS median: 0.509).

Positive Causal Effects We use synthetic data with positive effects and dependent, partially invalid IV candidates. Experimental results are reported in Figure 5b. Ivy provides a median estimate (0.146) that is closest to the true effect (0.150) while both UAS (0.440) and WAS (0.545) return median estimates that are biased towards the observational association.

5 Conclusion

We introduce Ivy, a framework that synthesizes from IV candidates a summary IV used for downstream causal inference. Through theoretical analysis and empirical studies, we demonstrate the robustness and limitation of Ivy in handling invalidity and dependencies among IV candidates.

Acknowledgement

The authors thank David Amar, Jason Fries, Sinong Geng, Beliz Gunel, Vitor Hadad, Ramtin Keramati, Allison Koenecke, Sharon Li, Avner May, Tong Mu, Ethan Steinberg, Anna Thomas, Steve Yadlowsky, and Jiaming Zeng for helpful conversations and feedback.

We gratefully acknowledge the support of DARPA under Nos. FA87501720095 (D3M), FA86501827865 (SDH), and FA86501827882 (ASED); NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML); ONR under No. N000141712266 (Unifying Weak Supervision); the Moore Foundation, NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, American Family Insurance, Google Cloud, Swiss Re, and members of the Stanford DAWN project: Teradata, Facebook, Google, Ant Financial, NEC, VMWare, and Infosys. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of DARPA, NIH, ONR, or the U.S. Government.

References

Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montreal, Canada, 2018.

Joshua D. Angrist and Alan B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 11 1991. ISSN 0033-5533. doi: 10.2307/2937954. URL `https://doi.org/10.2307/2937954`.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricists companion*. Princeton University Press, 2008.

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

Joshua David Angrist, Guido W Imbens, and Alan B Krueger. Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67, 1999.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *arXiv preprint arXiv:1710.10251*, 2018.

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

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92 (439):1171–1176, 1997.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *arXiv preprint arXiv:1905.12495*, 2019.

Blai Bonet. Instrumentality tests revisited. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 48–55. Morgan Kaufmann Publishers Inc., 2001.

John Bound, David A. Jaeger, and Regina M. Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430): 443–450, 1995.

Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology*, 44(2):512–525, 2015.

Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314, 2016.

Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2018.

Stephen Burgess and Jeremy A Labrecque. Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates. *European journal of epidemiology*, 33(10):947–952, 2018.

Stephen Burgess and Simon G Thompson. Use of allele scores as instrumental variables for mendelian randomization. *International journal of epidemiology*, 42(4):1134–1144, 2013.

Stephen Burgess and Simon G. Thompson. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. Chapman and Hall/CRC Press, 1st edition, 2015.

Stephen Burgess, Frank Dudbridge, and Simon G Thompson. Combining information on multiple instrumental variables in mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine*, 35(11):1880–1906, 2016.

Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, 26(5):2333–2355, 2017.

Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40(4): 1935–1967, 2012.

C Reactive Protein Coronary Heart Disease Genetics Collaboration. Association between C reactive protein and coronary heart disease: Mendelian randomisation analysis based on individual participant data. *Bmj*, 342:d548, 2011.

Alexander D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In *AISTATS 2019*, Okinawa, Japan, 2019.

Neil M Davies, Stephanie von Hinke Kessler Scholder, Helmut Farbmacher, Stephen Burgess, Frank Windmeijer, and George Davey Smith. The many weak instruments problem and mendelian randomization. *Statistics in Medicine*, 34(3):454–468, 2015.

P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '06*, 2006.

Georg B Ehret, Patricia B Munroe, Kenneth M Rice, Murielle Bochud, Andrew D Johnson, Daniel I Chasman, Albert V Smith, Martin D Tobin, Germaine C Verwoert, Shih-Jen Hwang, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103, 2011.

Byron Ellis and Wing Hung Wong. Learning causal bayesian network structures from experimental data. *Journal of the American Statistical Association*, 103(482):778–789, 2008.

G. Freeman, B. J. Cowling, and C. M. Schooling. Power and sample size calculations for mendelian randomization studies using one genetic instrument. *Int. Journal Epidemiology*, 42(4):1157–1163, 2013.

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montreal, Canada, 2018.

Chirok Han. Detecting invalid instruments using l1-gmm. *Economics Letters*, 101(3):285–287, 2008.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, Sydney, Australia, 2017.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

David Heckerman. A bayesian approach to learning causal networks. In *Proceedings of the Eleventh conference on Uncertainty in Artificial Intelligence (UAI 1995)*, pages 285–295, Montreal, Canada, 1995.

Michael V Holmes, Folkert W Asselbergs, Tom M Palmer, Fotios Drenos, Matthew B Lanktree, Christopher P Nelson, Caroline E Dale, Sandosh Padmanabhan, Chris Finan, Daniel I Swerdlow, et al. Mendelian randomization of blood lipids for coronary heart disease. *European heart journal*, 36 (9):539–550, 2014.

Jean Honorio. Lipschitz parametrization of probabilistic graphical models. *arXiv preprint arXiv:1202.3733*, 2012.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. Cost-optimal learning of causal graphs. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, Sydney, Australia, 2017.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 1st edition, 2009.

Jason D Lee and Trevor J Hastie. Learning the structure of mixed graphical models. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 388–396, Scottsdale, AZ, USA, 2013.

Wolfgang Lieb, Henning Jansen, Christina Loley, Michael J Pencina, Christopher P Nelson, Christopher Newton-Cheh, Sekar Kathiresan, Muredach P Reilly, Themistocles L Assimes, Eric Boerwinkle, et al. Genetic predisposition to higher blood pressure increases coronary artery disease risk. *Hypertension*, 61(5):995–1001, 2013.

Po-Ling Loh and Martin J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Annals of Statistics*, 41(6):3022–3049, 2013.

Despoina Manousaki, Lauren E Mokry, Stephanie Ross, David Goltzman, and J Brent Richards. Mendelian randomization studies do not support a role for vitamin D in coronary artery disease. *Circulation: Cardiovascular Genetics*, 9(4):349–356, 2016.

Iona Y Millwood, Robin G Walters, Xue W Mei, Yu Guo, Ling Yang, Zheng Bian, Derrick A Bennett, Yiping Chen, Caixia Dong, Ruying Hu, Gang Zhou, Bo Yu, Weifang Jia, Sarah Parish, Robert Clarke, George Davey Smith, Rory Collins, Michael V Holmes, Liming Li, Richard Peto, and Zhengming Chen. Conventional and genetic evidence on alcohol and vascular disease aetiology: a prospective study of 500000 men and women in China. *The Lancet*, 2019. ISSN 01406736. doi: 10.1016/S0140-6736(18)31772-0.

Lauren E Mokry, Omar Ahmad, Vincenzo Forgetta, George Thanassoulis, and J Brent Richards. Mendelian randomisation applied to drug development in cardiovascular disease: a review. *Journal of Medical Genetics*, 52(2):71–79, 2015. ISSN 0022-2593. doi: 10.1136/jmedgenet-2014-102438. URL `https://jmg.bmj.com/content/52/2/71`.

Judea Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 435–443. Morgan Kaufmann Publishers Inc., 1995.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Daniel J Rader and G Kees Hovingh. Hdl and cardiovascular disease. *The Lancet*, 384(9943):618–625, 2014.

A. J. Ratner, Christopher M. De Sa, Sen Wu, Daniel Selsam, and C. Ré. Data programming: Creating large training sets, quickly. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, 2016.

A. J. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, 2019.

Kerry-Anne Rye and Kwok Leung Ong. Hdl function as a predictor of coronary heart disease events: time to re-assess the hdl hypothesis? *The Lancet Diabetes & Endocrinology*, 3(7):488–489, 2015.

Gregory G Schwartz, Anders G Olsson, Markus Abt, Christie M Ballantyne, Philip J Barter, Jochen Brumm, Bernard R Chaitman, Ingar M Holme, David Kallend, Lawrence A Leiter, et al. Effects of dalcetrapib in patients with a recent acute coronary syndrome. *New England Journal of Medicine*, 367(22):2089–2099, 2012.

Paola Sebastiani, Nadia Solovieff, and Jenny Sun. Naïve bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: not so different after all! *Frontiers in genetics*, 3:26, 2012.

Amit Sharma. Necessary and probably sufficient test for finding valid instrumental variables. *arXiv preprint arXiv:1812.01412*, 2018.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

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.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *arXiv preprint arXiv:1004.4389*, 2011.

P. Varma, F. Sala, A. He, A. J. Ratner, and C. Ré. Learning dependency structures for weak supervision models. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.

Benjamin F Voight, Gina M Peloso, Marju Orho-Melander, Ruth Frikke-Schmidt, Maja Barbalic, Majken K Jensen, George Hindy, Hilma Hólm, Eric L Ding, Toby Johnson, et al. Plasma hdl cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet*, 380(9841):572–580, 2012.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

A. Wald. The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, 11(3):284–300, 1940.

Venexia M Walker, George Davey Smith, Neil M Davies, and Richard M Martin. Mendelian randomization: a novel approach for the prediction of adverse drug events and drug repurposing opportunities. *International journal of epidemiology*, 46(6): 2078–2089, 2017.

Yixin Wang and David M Blei. The blessings of multiple causes. *arXiv preprint arXiv:1805.06826*, 2018.

Yixin Wang and David M Blei. Multiple causes: A causal graphical view. *arXiv preprint arXiv:1905.12793*, 2019.

Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George Davey Smith. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 2018.

Philip G Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

Changjing Wu, Hongyu Zhao, Huaying Fang, and Minghua Deng. Graphical model selection with latent variables. *Electronic Journal of Statistics*, 11: 3485–3521, 2017.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, Stockholm, Sweden, 2018.

Chong Ho Yu. Exploratory data analysis. *Methods*, 2: 131–160, 1977.

Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.