Adjusting for Indirectly Measured Confounding Using Large-Scale Propensity Score

Linying Zhang[a], **Yixin Wang**[b], **Martijn J. Schuemie**[c], **David M. Blei**[d,e], **George Hripcsak**[a,f]

[a]Department of Biomedical Informatics, Columbia University Irving Medical Center, 622 W. 168th Street, PH20, New York, 10032, NY, USA

[b]Department of Statistics, University of Michigan, 1085 S University Ave, Ann Arbor, 48109, MI, USA

[c]Janssen Research and Development, 1125 Trenton-Harbourton Road, Titusville, 08560, NJ, USA

[d]Department of Statistics, Columbia University, 1255 Amsterdam Ave, New York, 10027, NY, USA

[e]Department of Computer Science, Columbia University, 500 West 120 Street, Room 450 MC0401, New York, 10027, NY, USA

[f]Medical Informatics Services, New York-Presbyterian Hospital, 622 W. 168th Street, PH20, New York, 10032, NY, USA

Abstract

Confounding remains one of the major challenges to causal inference with observational data. This problem is paramount in medicine, where we would like to answer causal questions from large observational datasets like electronic health records (EHRs) and administrative claims. Modern medical data typically contain tens of thousands of covariates. Such a large set carries hope that many of the confounders are directly measured, and further hope that others are indirectly measured through their correlation with measured covariates. How can we exploit these large sets of covariates for causal inference? To help answer this question, this paper examines the performance of the large-scale propensity score (LSPS) approach on causal analysis of medical data. We demonstrate that LSPS may adjust for indirectly measured confounders by including tens of thousands of covariates that may be correlated with them. We present conditions under which LSPS removes bias due to indirectly measured confounders, and we show that LSPS may avoid bias when inadvertently adjusting for variables (like colliders) that otherwise can induce bias. We demonstrate the performance of LSPS with both simulated medical data and real medical data.

Keywords

1. Introduction

Causal inference in the setting of unmeasured confounding remains one of the major challenges in observational research. In medicine, electronic health records (EHRs) have become a popular data source for causal inference, where the goal is to estimate the causal effect of a treatment on a health outcome (e.g., the effect of blood-pressure medicine on the probability of a heart attack). EHRs typically contain tens of thousands of variables, including treatments, outcomes, and many other variables, such as patient demographics, diagnoses, and measurements.

Causal inference on these data is often carried out using propensity score adjustment [1]. Researchers first select confounders among the many observed variables, either manually (based on medical knowledge) or empirically. Then they estimate a propensity model using those selected variables and employ the model in a standard causal inference method that adjusts for the propensity score (the conditional probability of treatment). While this strategy is theoretically sound, in practice researchers may miss important confounders in the selection process, which leads to confounding bias, or may include variables that induce other types of bias (e.g., a "collider" or a variable that induces "M-bias").

In this paper, we study a closely related, but different, technique, known as large-scale propensity score (LSPS) adjustment [2]. LSPS fits an L1-regularized logistic regression with all pre-treatment covariates to estimate the propensity model. LSPS then uses standard causal inference methods, with the corresponding propensity scores, to estimate the causal effect. For example, LSPS might be used with matching [3, 4, 5] or subclassification [6].

In contrast to the traditional approach of explicitly selecting confounders, LSPS is a "kitchen-sink" approach that includes all of the covariates in the propensity model. While the L1-regularization might lead to a sparse propensity model, it is not designed to select the confounders in particular. Instead, it attempts to create the most accurate propensity model based on the available data, and LSPS diagnostics (described below) use covariate balance between treatment and control groups (i.e., that covariates are distributed similarly in the two groups) to assess whether *all* covariates are in fact adjusted-for in the analysis regardless of their L1-regularization coefficient.

The discussion over how many covariates to include in a propensity model is an old one [7, 8, 9, 10, 11, 12, 13] and considers–in the setting of imperfect information about variables– the tradeoff between including all measured confounders versus including variables that may increase bias and variance [14, 15]. To address this issue, LSPS uses only pre-treatment covariates to avoid bias from mediators and simple colliders, and it uses diagnostics and domain knowledge to avoid variables highly correlated with the treatment but uncorrelated with outcomes (known as "instruments"). Including such variables can increase the variance of the estimate [14, 15, 16, 17] and can amplify bias [18, 19, 20, 21, 22, 23].

In medicine, empirical studies of the performance of LSPS have shown it to be superior to selecting confounders [2, 24, 25, 26]. Consequently, LSPS has been used in a number of studies, both clinical [27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37] and methodological [38, 39, 40, 41, 42, 43].

Further, researchers have studied whether LSPS may also address indirectly measured confounders [34, 40, 44]. The hope behind these studies is that when we adjust for many covariates, we are likely to be implicitly adjusting for the confounders that are not directly measured but are correlated with existing covariates. Hripcsak et al.[34] and Schuemie et al. [40] used LSPS to estimate the causal effect of anti-hypertension drugs, adjusting for about 60,000 covariates. An important confounder, baseline blood pressure, was not contained in most of the data sources. In the one source that did contain blood pressure, adjusting for all the other covariates but no blood pressure resulted in (nearly) balancing blood pressure between propensity-score-stratified cohorts; the resulting causal inference was identical to the one obtained when including blood pressure in the propensity model.

Based on this observation, Chen et al.[44] studied the effect of dropping large classes of variables from the LSPS analysis, using balance of the covariates between the treatment and control groups as a metric for successful adjustment (i.e., is every covariate in the propensity model balanced between the cohorts). If all the variables of one type were eliminated from the propensity model (e.g., medical diagnoses), then the inclusion of a large number of other variables (e.g., medications, procedures) resulted in the complete balancing of the missing variables. Even more striking, if all variables related to one medical area like cardiology were dropped from the model (e.g., all cardiology-related diagnoses, procedures, medications, etc.), then the rest of the covariates still balanced the dropped cardiology covariates. Yet if too few covariates were included, such as just demographics, then balance was not achieved on the other covariates. Based on these studies, LSPS appears to be adjusting for variables that are not included but correlated with the included covariates.

In this paper, we explore conditions under which LSPS can adjust for indirectly measured confounders. In particular, we provide some theoretical assumptions under which LSPS is robust to some indirectly measured confounders. They are based on the "pinpointability" assumption used in Wang and Blei [45, 46]. A variable is pinpointed by others if it can be expressed as a deterministic function of them, though the function does not need to be known. In the context of causal inference from EHR data, we show that if confounders that are indirectly measured but can be pinpointed by the measured covariates, then LSPS implicitly adjusts for them. For example, if high blood pressure could be conceivably derived from the many other covariates (e.g., diagnoses, medicines, other measurements) then LSPS implicitly adjusts for high blood pressure even though it is not directly measured.

From a theoretical perspective, pinpointability is a strong and idealized assumption. But in practice, several empirical observations showed that important confounders that are not directly measured often appear to undergo adjustment when LSPS is used. Therefore, there might be hope that some of the indirectly measured confounders are capturable by the existing covariates. We do not assert LSPS as a magical solution to unmeasured confounding —the assumption is strong—but as an attempt to better understand the empirical success of LSPS in adjusting for indirectly measured confounders. To explore this phenomenon, we use synthetic data to empirically study the sensitivity of LSPS to the degree to which pinpointability is violated. We find that under perfect pinpointability, adjusting for measured covariates removes the bias due to indirectly measured confounding. As the data deviates from pinpointability, adjusting for the measured covariates becomes less adequate.

Finally, we study real-world medical data to compare LSPS to a traditional propensity score method based on previously used manually selected confounders. We find that removing a known confounder has a bigger impact on a traditional propensity score method than on LSPS, presumably because it is indirectly measured. This finding suggests that including large-scale covariates with LSPS provides a better chance of correcting for confounders that are not directly measured.

The paper is organized as follows. Section 2 describes the LSPS algorithm, the pinpointability assumption, and the effect of pinpointability on M-structure colliders, instruments, and near-instruments. Section 3 studies the impact of violations of pinpointability on the fidelity of the estimated causal effects. Section 4 presents empirical studies comparing LSPS to classical propensity-score adjustment (with manually selected covariates), and methods that do not adjust. Section 5 compares LSPS to other approaches to adjusting for indirectly measured confounding and makes connection to other related work. Section 6 concludes the paper.

2. The Large Scale Propensity Score Algorithm

In this section, we summarize the LSPS algorithm, describe an assumption under which LSPS will adjust for indirectly measured confounding and potentially mitigate the effect of adjusting for unwanted variables, and make some remarks on the assumption.

2.1. The LSPS algorithm

We summarize the LSPS algorithm [2], including the heuristics and diagnostics that normally surround it (e.g., Weinstein [25]). Consider a study where a very large number of covariates are available (e.g., over 10,000) and the problem of estimating the causal effect of a treatment. Rather than selecting confounding covariates and adjusting for them, LSPS adjusts for all of the available covariates. It uses only pre-treatment covariates to avoid adjusting for mediators and simple colliders (which induce bias), and it uses diagnostics and domain knowledge to avoid "instruments," variables that are correlated with the treatment but do not affect the outcome. (Such variables increase the variance of the causal estimate.)

By design, LSPS includes all measured confounders. The hope is that in real-world data, such as in medicine, adjusting for all the other non-confounder variables would not impart bias, and empirical comparisons to traditional propensity approaches seem to bear that out [2, 24, 25, 26]. The further hope is that by balancing on a large number of covariates, other indirectly measured factors would also become balanced, and this is what we address in Section 2.2.

The inputs to LSPS are observed pre-treatment covariates $\mathbf{X}$ and binary treatment T. The output is the estimated causal effect $\hat{\nu}$. LSPS works in the following steps.

1. Remove "instruments."—Remove covariates that are highly correlated with the treatment and are unlikely to be causally related to the outcome. Univariate correlation to treatment is checked numerically, and domain expertise is used to determine if the highly correlated variables are not causally related to the outcome; if the relationship is unclear,

then the variable is not removed. Note these covariates are commonly called "instruments," and used in instrumental variable analysis [47]. LSPS, however, does not do instrumental variable analysis, and removes these variables to reduce downstream variance.

2. Fit the propensity model and calculate propensity scores.—Given the remaining covariates, fit an L1-regularized logistic regression [48] to estimate propensity scores $p(t \mid \mathbf{x})$. The regression is

$$p(t \mid \mathbf{x}) = \frac{1}{1 + e^{-\theta^\top \mathbf{x}}},$$

where θ is the vector of the regression parameters. L1-regularized logistic regression minimizes

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} -t_i \log(p(t_i \mid \mathbf{x}_i; \theta)) - (1 - t_i)\log(1 - p(t_i \mid \mathbf{x}_i; \theta)) + \lambda \sum_{j=1}^{M} |\theta_j|,$$

where λ is the tuning parameter that controls the strength of the L1 penalty.

LSPS uses cross-validation to select the best regularization parameter λ. It then refits the regression model on the entire dataset with the selected regularization parameter. Finally, it uses the resulting model to extract the propensity scores for each datapoint.

3. Check the equipoise of the propensity model.—In this step, LSPS assesses whether the conditional distribution of assignment given by the propensity model is too certain, i.e., whether the treatment and control groups are too easily distinguishable. The reason is that a propensity model that gives assignments probabilities close to zero or one leads to high-variance estimates [15], e.g., because it is difficult to match datapoints or create good subclasses.

To assess this property of the propensity model, LSPS performs the diagnostic test of Walker et al. [49]. This diagnostic assesses the overlapping support of the distribution of the preference score, which is a transformation of the propensity score[1], on the treatment and control groups. If there is overlapping support (at least half the mass with preference between 0.3 and 0.7) then the study is said to be in *equipoise*. If a study fails the diagnostic, then the analyst considers if an instrument has been missed and removes it, or interprets the results with caution.

4. Match or stratify the dataset based on propensity scores and then check covariate balance.—Matching [3, 4, 5] or subclassification [6] on propensity scores can

---

[1]Define the preference score *Pref* as a transformation of the propensity score $p(t \mid \boldsymbol{x})$ that adjusts for the probability of treatment $p(t = 1)$.

$$\ln[Pref/(1 - Pref)] = \ln[p(t \mid \mathbf{x})/(1 - p(t \mid \mathbf{x}))] - \ln[p(t = 1)/(1 - p(t = 1))].$$

be used to create groups of individuals who are similar. The details about the two methods are provided in Supplementary S1. The remaining of the algorithm is explained in the context of 1-to-1 matching. Once the matched groups are created, balance is assessed by computing the standardized mean difference (SMD) of each covariate between the treated group and the control group from the matched dataset

$$\text{SMD} = \frac{\bar{x}_{t\,=\,1} - \bar{x}_{t\,=\,0}}{\sqrt{\left(\sigma_{t\,=\,1}^2 + \sigma_{t\,=\,0}^2\right)/2}},$$

where $\bar{x}_{t\,=\,1}$ and $\bar{x}_{t\,=\,0}$ are the mean of the covariate in the treated and the control group respectively, and $\sigma_{t\,=\,1}^2$ and $\sigma_{t\,=\,0}^2$ are the variance of the covariate in the treated and the control group respectively. Following Austin[50], if any covariate has a SMD over 0.1 [50], then the comparison is said to be out of balance, and the study needs to be discarded (or interpreted with caution).

5.   Estimate the causal effect.—The last step is to use the matched data to estimate the causal effect. In the simulations in Section 3, the causal effect of interest is the average treatment effect

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)],$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes for a subject under treatment and under control.

To estimate the ATE using matched data, a linear regression is fitted on the matched data with a treatment indicator variable. The coefficient of the treatment indicator is the average treatment effect. When subclassification is used to create balanced subclasses, the effect is estimated within each subclass and then aggregated across subclasses. The weight for each subclass is proportional to the total number of individuals in each subclass.

In the empirical studies of Section 4, the causal effect of interest is hazard ratio (the outcome Y is time to event). When matching is used, we fit a Cox proportional hazards model [51] on the matched dataset to estimate the hazards ratio. When subclassification is used, we fit a Cox model within each subclass and then weigh the conditional hazards ratio by the size of the subclass to obtain the marginal hazards ratio. More details are provided in Supplementary S2.

2.2.   Adjusting for indirectly measured confounders

As noted in the introduction, LSPS has been found to adjust for known but indirectly measured confounders [34, 40, 44]. We describe here an assumption under which LSPS will adjust for indirectly measured confounding.

Consider the causal graph in Fig.1 for an individual i (the subscript is omitted in the graph), where $T_i \in \{0, 1\}$ is a binary treatment, Y_i is the outcome (either binary or continuous), $\mathbf{X}_i \in \{0, 1\}^M$ is a high-dimensional vector of observed pre-treatment covariates with length

M (which includes observed confounders and other variables), and $U \in \mathbb{R}$ is the indirectly measured confounder. The goal is to estimate the causal effect of treatment T on the outcome Y. To do so, we need to adjust for both pre-treatment covariates $\mathbf{X}$ (including directly measured confounders) and U (indirectly measured confounder). We assume that there are no other unmeasured confounders.

In the following sections, we will demonstrate that LSPS can still produce unbiased causal estimates even in the presence of indirectly measured confounders. We first introduce Assumption 1, which indicates the relationship between measured covariates and indirectly measured confounders.

Assumption 1. *(Pinpointability of indirectly measured confounder) An unmeasured confounder U is said to be pinpointed by measured covariates if*

$$p(u \mid \mathbf{x}) = \delta(f(\mathbf{x})) \tag{1}$$

where $\delta(\cdot)$ denotes a point mass at $f(\cdot)$.

In other words, the indirectly measured confounder U can be represented by a deterministic function f of the measured covariates $\mathbf{X}$. Theorem 1, building upon Assumption 1, formally states the conditions that LSPS needs in order to obtain unbiased causal estimates by only conditioning on measured covariates. We use the potential outcome framework by Rubin [52]. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcome under treatment and under control respectively for an individual i.

Theorem 1. *The treatment and the potential outcomes are independent conditioning on all confounders, both the directly measured ($\mathbf{X}$) and indirectly measured (U),*

$$T \perp\!\!\!\perp Y(1), Y(0) \mid \mathbf{X}, U . \tag{2}$$

Under the pinpointability assumption, the above conditional independence can be reduced to only conditioning on the measured covariates,

$$T \perp\!\!\!\perp Y(1), Y(0) \mid \mathbf{X} . \tag{3}$$

In other words, the causal effect of the treatment on the outcome is identifiable by only adjusting for the measured covariates $\mathbf{X}$. We do not need to know the indirectly measured confounders U or its functional form $f(\cdot)$.

PROOF. Theorem 1 relies on the marginalization over U in computing the propensity score using high-dimensional measured covariates,

$$p(t \mid \mathbf{x}) = \int p(t \mid u, \mathbf{x}) p(u \mid \mathbf{x}) \mathrm{d}u \tag{4}$$

$$= p(t \mid u^*, \mathbf{x}), \tag{5}$$

where $u* = f(\mathbf{x})$.

When U is weakly pinpointed by $\mathbf{X}$, or in other words, $f(\mathbf{X})$ measures U with error, the average treatment effect is not point identifiable. Assuming identification holds conditional on the unmeasured confounder, Ogburn and VanderWeele [53] show that the average treatment effect adjusting for the noisy measured confounder is between the unadjusted and the true effects, under some monotonicity assumptions. We extend the work by Ogburn and VanderWeele [53] to conditions where additional confounders and covariates exist.

Theorem 2. *Let T be a binary treatment, Y be an outcome, $\mathbf{X}_C$ be all measured confounders, U be an ordinal unmeasured confounder, and U' be the noisy measurement of U. Assume that the measurement error of U is tapered and nondifferential with respect to T and Y conditional on $\mathbf{X}_C$, and that $\mathbb{E}[Y \mid T, U, \mathbf{X}_C = \mathbf{x}_C]$ and $\mathbb{E}[T \mid U, \mathbf{X}_C = \mathbf{x}_C]$ are monotonic in U for each value $\mathbf{x}_C$ in the support of $\mathbf{X}_C$. Then the average treatment effect adjusting for the measured covariates lies between the true effect and effect adjusting for only the measured confounders, that is, $ATE_{true} \leq ATE_{cov} \leq ATE_{conf}$ or $ATE_{true} \geq ATE_{cov} \geq ATE_{conf}$, where $ATE_{true} = \mathbb{E}_{X_C, U}[Y(1)] - \mathbb{E}_{X_C, U}[Y(0)]$, $ATE_{cov} = \mathbb{E}_{X_C, U'}[Y \mid T = 1] - \mathbb{E}_{X_C, U'}[Y \mid T = 0]$, and $ATE_{conf} = \mathbb{E}_{X_C}[Y \mid T = 1] - \mathbb{E}_{X_C}[Y \mid T = 0]$.*

Theorem 2 states that the average treatment effect adjusting for all measured covariates (i.e., by LSPS) is bounded between the true effect and the effect adjusting for only the measured confounders. Nondifferential misclassification error means that $p(U = u' \mid U = u, T = t, Y = y, \mathbf{X}_C = \mathbf{x}_c) = p(U = u' \mid U = u, T = t', Y = y', \mathbf{X}_C = \mathbf{x}'_c)$ for all $y, y' \in \mathscr{Y}$, $t, t' \in \{0, 1\}$, and $\mathbf{x}_c, \mathbf{x}'_c \in \mathscr{X}_\mathscr{C}$. Tapered misclassification error means that the misclassification probabilities $p_{ij} = p(U' = i \mid U = j)$, $i, j \in \{1, \ldots, K\}$ is nonincreasing in both directions away from the true unmeasured confounder.

We show that Theorem 1 and Theorem 2 hold in the simulations with various degrees of pinpointability. Proof of Theorem 2 is an extension of the work by Ogburn and VanderWeele [53], and is given in the Supplement.

2.3. Effect of pinpointing on instruments and M-bias

Because LSPS uses a large number of covariates, there is a concern that adjusting for these covariates will induce bias due to M-structure colliders, instrumental variables (IVs), and near instrumental variables (near-IVs). As noted above, our goal is not to do instrumental variable analysis but rather to remove their potential effect of increasing variance and amplifying bias. IVs are addressed in part by domain knowledge and diagnostics, but some IVs may remain. In this section, we discuss how LSPS in the setting of pinpointing may address them.

2.3.1. Effect on IV and near-IV—Instrumental variables [47] may persist despite LSPS's procedures. In the setting of unmeasured confounding, IV can cause bias amplification as shown numerically [54, 55] and proved theoretically in various scenarios [18, 19, 20, 21, 22, 23]. Insofar as pinpointing adjusts for indirectly measured confounding

(Fig.2a), even if there are IVs in the propensity score model, they will not produce bias amplification [22].

Near-instrumental variables (near-IVs) [15], which are weakly related to the outcome and strongly related to the treatment, may also lead to bias amplification [15, 18, 22], and the bias amplification or the confounding may dominate. Just as for IVs, pinpointing (Fig.2b) may reduce bias amplification by reducing indirectly measured confounding [22], while the confounding is eliminated by adjusting for the near-IV.

2.3.2.   Addressing M-bias—Despite the use of pre-treatment variables, bias through colliders is still possible due to causal structures like the one in Fig. 2c, known as an M-structure, causing M-bias. In this case, two unobserved underlying causes create a path from T to Y via a collider that can precede T in time. If the collider is included in the many covariates, then this can induce bias. LSPS may be able to address M-bias in the following way. If the common cause between the treatment and the collider (Z_1) can be pinpointed by the measured covariates, then this will block the back-door path from T to Y. Similarly, the common cause between the outcome and the collider (Z_2) could be pinpointed, also blocking the path. The assertion that one or both of these common causes is pinpointed is similar to the assertion that U is pinpointed.

3.   Simulations

We use the simulation to show that, under the assumption of pinpointability, LSPS can adjust for the indirectly measured confounding, and as the condition deviates from pinpointability, bias in LSPS increases, and the estimate by LSPS is between the true effect and the effect adjusting for only measured confounders. In this simulation, we assume that the large number of covariates $\mathbf{X}$ are derived from a smaller number of underlying latent variables $\mathbf{V}$. This data generating process induces dependencies among the measured covariates, mimicking the dependencies observed in EHR.

3.1.   Simulation setup

Each simulated data set contains $N = 5,000$ patients, $M = 100$ measured covariates (including 10 measured confounder), 1 indirectly measured confounder, 10 latent variables, a treatment and an outcome. The data set is $\mathscr{D} = \{\mathbf{v}_i, \mathbf{x}_i, u_i, t_i, y_i\}_{i=1}^{N}$, where $\mathbf{v}_i$ is a vector of latent variables, $\mathbf{v}_i = (v_{i1}, \ldots, v_{i10})$; $\mathbf{x}_i$ is a vector of measured covariates, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iM})$; u_i, t_i and y_i are all scalar, representing the indirectly measured confounder, treatment and outcome respectively.

Below are the steps to simulate data for patient i.

1.   Simulate the latent variable $\mathbf{v}_i$ as

$$\mathbf{v}_i \sim \text{Bernoulli}(0.5)^K.$$

2.   Simulate measured covariates $\mathbf{x}_i$ as

$$\mathbf{x}_i \sim \text{Bernoulli}\left(\text{sigmoid}\left(\mathbf{v}_i^\top \boldsymbol{\beta}_x\right)\right),$$

where $\boldsymbol{\beta}_x \sim \mathcal{N}(0, 0.1)^{K \times M}$.

3. Simulate the indirectly measured confounder u_i as

$$u_i = \mathbf{x}_i^\top \boldsymbol{\beta}_u$$

where $\boldsymbol{\beta}_u \sim \mathcal{N}(0, 1)^M$. Notice that u is a deterministic function of $\mathbf{x}$. To allow only a small subset of the covariates pinpoint u, we randomly select 90% of the $\boldsymbol{\beta}_u$ and set their value to 0.

4. Simulate the treatment t_i as

$$t_i \sim \text{Bernoulli}\left(\text{sigmoid}\left(\mathbf{x}_i^\top \gamma_x + u_i \gamma_u\right)\right),$$

where the effect of the indirectly measured confounder on the treatment $\gamma_u = 1$, $\gamma_x \sim \mathcal{N}(0.5, 1)$ for the 10% of covariates that serve as measured confounders. The rest of the γ_x are set to 0.

5. Simulate the outcome y_i as

$$y_i \sim \mathcal{N}\left(\mathbf{x}_i^\top \boldsymbol{\eta}_x + u_i \eta_u + t_i v, 0.1\right),$$

where the true causal effect $v = 2$, the effect of the indirectly measured confounder on the outcome $\eta_u = 1$, $\boldsymbol{\eta}_x \sim \mathcal{N}(0.5, 1)$ for the covariates that serve as measured confounders and 0 otherwise.

The above steps illustrate the simulation under pinpointability. To increase the deviation from pinpointability, we add an increasing amount of random noise to the indirectly measured confounder. To do so, we modify the simulation of u_i in Step 3 to be

$$u_i = \mathbf{x}_i^\top \boldsymbol{\beta}_u + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. To increase deviation from pinpointability, we increase σ^2 from 10^{-4} to 10^4. At each pinpointability level, the Gaussian noise is non-differential with respect to the treatment, outcome, and the measured covariates, and decreasing in both directions away from the true unmeasured confounder. We simulate 50 datasets at each pinpoitability level.

3.2. Statistical analysis

We demonstrate LSPS's capacity in adjusting for unmeasured confounding relative to other methods under varied degree of pinpointability. Specifically, we compared the following five methods:

- unadjusted: no covariate was adjusted for.

- manual without U: adjust for all confounders not including U

- manual (oracle): adjust for all confounders including U

- LSPS without U: adjust for all measured covariates not including U

- LSPS: adjust for all covariates including U

Notice that manual with U coincides with oracle in this simulation, because manual with U adjusts for nothing but the confounders, both measured and unmeasured. In practice, because the confounding structure is rarely known, it is unlikely a manual method captures all the confounders.

For the four methods that adjust for confounders, we estimated propensity scores with L1-regularized logistic regression (and selected the regularization parameter with cross-validation). We then used 1:1 matching and subclassification (results not shown) to create a balanced dataset. To estimate the average treatment effect, we fit a linear regression model on the balanced dataset. We then calculated the mean, 95% confidence interval, and root-mean squared error (RMSE) of the effect estimates.

The RMSE, defined as follows, can be calculated because the true treatment effect is known in the simulation ($\nu = 2$). This metric is not applicable to the empirical study because the true effect of medications is unknown in practice.

$$\text{RMSE} = \sqrt{\frac{1}{S} \sum_s (\hat{v}_s - v)^2},$$

where $\hat{v}_s$ is the effect estimate at simulation s, and we simulate a total of $S = 50$ datasets at each given pinpointability condition.

3.3. Results

Fig. 4 shows the results of the simulation. When pinpointability holds reasonably well, the unmeasured confounder has a bigger impact on the manual method than on LSPS. As pinpointability gets weaker, adjusting for large-set of covariates becomes less adequate for accounting for the unmeasured confounder, and the estimated treatment effect approaches the estimate from the method adjusting for only the measured confounders.

Under strong pinpointability, the estimates from the two large-scale approaches (LSPS without U and LSPS) have almost the same bias, variance, and RMSE compared to the estimate from the oracle. The manual approach (manual without U) does not benefit from pinpointability because it does not include covariates that assist pinpointing. As the condition deviates from strong pinpointability, the large-scale approach (LSPS without U) becomes increasingly biased, approaches and eventually overlaps the estimate from the manual without U approach. This simulation result matches Theoreom 2 that ATE_{cov} is between ATE_{conf} and ATE_{true} under certain monotoniticty assumptions. In the simulation,

ATE_{conf} is given by manual without U, ATE_{cov} is given by LSPS without U, and ATE_{true} is given by manual with U.

4. Empirical studies

We now use real data to compare LSPS to the traditional propensity-score adjustment (with manually selected covariates) to adjusting for confounding. With an EHR database, we compared the effect of two anti-hypertension drugs, hydrochlorothiazide and lisinopril, on two clinical outcomes, acute myocardial infarction (AMI) and chronic kidney disease (CKD). For both outcomes, type 2 diabetes mellitus (T2DM) is a known confounder. Thus, by including or excluding T2DM in an adjustment model while keeping other covariates the same, we can assess a method's capacity in adjusting for a known confounder that is not directly measured but may be correlated with measured covariates.

4.1. Cohort and covariates

We used a retrospective, observational, comparative cohort design [56]. We included all new users of hydrochlorothiazide monotherapy or lisinopril monotherapy and defined the index date as the first observed exposure to either medication. We excluded patients who had less than 365 days of observation prior, a prior hypertension treatment, initiated another hypertension treatment within 7 days, or had the outcome prior to index date. We followed patients until their end of continuous exposure, allowing for maximum gaps of 30-days, or their end of observation in the database, whichever came first.

For the LSPS-based approach, we used more than 60,000 covariates in the EHR database, including demographics, all medications in the 365 days prior to index date, all diagnoses in the 365 days prior to index date, and the Charlson Comorbidity Index score, as baseline covariates in the propensity model.

For traditional PS adjustment, covariates were selected by experts for inclusion in related hypertension drug studies [57, 58, 59, 60, 61], including T2DM, anti-glycemic agent, age groups, female, index year, coronary artery disease, myocardial infarction, asthma, heart failure, chronic kidney disease, atrial fibrillation, Charlson index - Romano adaptation, Platelet aggregation inhibitors excl. heparin, Warfarin, corticosteroids for systemic use, dipyridamole, non-steroidal anti-inflammatory drugs(NSAIDS), proton-pump inhibitors (PPIs), statins, estrogens, progestogens, body mass index (BMI), chronic obstructive pulmonary disease(COPD), liver disease, dyslipidemia, valvular heart disease, drug abuse, cancer, HIV infection, smoking and stroke.

Most covariates (e.g., diagnoses, medications, procedures) were encoded as binary, that is, 1 indicates the code is present in the patient's medical history prior to treatment, and 0 otherwise. For some variables that are often considered as continuous (e.g., lab tests), LSPS does not impute the value because imputation could do more harm than good when missing mechanism is not known. Instead, LSPS encodes the lab ordering pattern as binary variables. Residual error appears as measurement error.

The study was run on the Optum© de-identified electronic health record database of aggregated electronic health records.

4.2. Statistical analysis

We examined a method's capacity in adjusting for confounding that is not directly measured by comparing the effect estimates from each method with or without access to the confounder. We excluded all variables related to T2DM, including diagnoses and anti-glycemic medications in models without access to indirectly measured confounders. We included an unadjusted method as a baseline for comparison. Specifically, we studied the following five methods (analogous to the five methods in the simulation):

- unadjusted: no covariate was adjusted for.

- manual: adjust for a list of manually selected confounders.

- manual without T2DM: adjust for a list of manually selected confounders without T2DM-related confounders.

- LSPS: adjust for all pre-treatment covariates in the database.

- LSPS without T2DM: adjust for all pre-treatment covariates in the database without T2DM-related confounders.

For the four methods that adjust for confounders, we estimated propensity scores with L1-regularized logistic regression (and selected the regularization parameter with cross-validation). We used subclassification and stratified the dataset into 10 subclasses. To estimate the treatment effect, we fit a Cox proportional-hazards model [51] to estimate the hazard ratio (HR). We then calculated the mean and 95% confidence interval of the HR.

4.3. Results

Fig. 5 shows the results of empirical studies. These results show that the T2DM had a bigger impact on the manual methods than on the LSPS methods. The impact was determined by comparing the absolute difference in effect estimates between the two manual models versus the two LSPS-based models. In the CDK study, the absolute difference between the two manual methods was 0.09 (Manual without T2DM: HR 0.77 [95% CI, 0.71–0.83]; Manual: HR 0.86 [95% CI, 0.79–0.93]), higher than the absolute difference between the two LSPS-based methods, which was 0.05 (LSPS without T2DM: HR 0.84 [95% CI, 0.77–0.92]; LSPS: HR 0.89 [95% CI, 0.82–0.97]). In fact, the estimates for manual with T2DM, LSPS with T2DM, and LSPS without T2DM were all closer to each other than to manual without T2DM. Therefore, whether manual with T2DM or LSPS with T2DM is actually closer to ground truth, LSPS without T2DM is closer to either one than is manual without T2DM.

This finding suggests that by including large-scale covariates, one has a better chance of correcting for confounders that are not directly measured. The pinpointability assumption is more likely to hold when there are many measured covariates.

5. Discussion

We have illustrated conditions under which LSPS adjusts for indirectly measured confounding and the impact of violations of such conditions on effect estimation. We have found in previous practice, in our current simulations, and in our current real-world study that indirectly measured (or unused) confounding can be adjusted for in LSPS, apparently working better than smaller, manually engineered sets of covariates that are also missing the confounder.

Even though pinpointing in the current simulation is achieved by generating the unmeasured confounder from a function of the measured covariates, this does not suggest the causal direction between the unmeasured confounder and the measured covariates. In medicine, it is likely that an unmeasured confounder (e.g., a disease such as T2DM) induces dependencies among large-scale clinical covariates (e.g., medications for treating the disease, laboratory tests for monitoring the disease, and other diseases that often co-occur with the disease can be correlated). In other words, the unmeasured confounder could be a latent variable in a factor model, and the strength of pinpointing depends on the number of measured covariates and the degree of dependency among the covariates.

We describe here methods that are related to LSPS. These methods are related to LSPS in different ways. Section 5.1 compares and contrasts LSPS to other methods that also address unmeasured confounding in causal effect estimation. Section 5.2 compares LSPS to another propensity score-based method that also uses large-scale covariates. Section 5.3 draws similarity between LSPS and another method for causal effect estimation in the presence of indirectly measured confounding where pinpointability is a required assumption.

5.1. Relation to proxy variable, multiple imputation and residual bias detection

Studies such as those by Kuroki and Pearl [62], Miao et al.[63], and Tchetgen Tchetgen et al.[64] have shown that causal effects can be identified by observing proxy variables of confounders that are not directly measured. In this case, the confounder is known but not measured, there is sufficient knowledge of the structural causal model such that proxies can be selected, and there is the knowledge that there are no other unmeasured confounders. In contrast, LSPS does not require explicit knowledge of the causal model.

Another approach is to use measured covariates to explicitly model unmeasured confounders using multiple imputation [65, 66]. While it differs from our approach, it exploits the same phenomenon, that some covariates contain information about unmeasured confounders. This is in contrast to LSPS, where there is no explicit model of unmeasured confounders; adjusting for measured covariates should be effective for causal inference as long as the measured covariates pinpoint the unmeasured confounders.

Given the need to assume no additional unmeasured confounding—additional in the sense of not being pinpointed or not having proxies—a complementary approach is to estimate the degree of residual bias, potentially including additional unmeasured confounding. Large-scale use of negative and synthetic positive controls [42, 41] can detect residual bias and

can additionally be used to calibrate estimates to assure appropriate coverage of confidence intervals. LSPS is usually coupled with such empirical calibration [43, 42].

5.2. Relation to high-dimensional propensity score adjustment

LSPS adjusts for all available pre-treatment covariates. In practice, because the sample size is limited, regularized regression selects a subset of variables to represent the information contained in the whole set of covariates, but the goal is to represent all the information nonetheless. Therefore, LSPS diagnostics [42, 2] test balance not just on the variables that regularized regression included in the model, but on all the covariates. All covariates are retained because even those that are not direct confounders may still contribute to the pinpointing of the unobserved confounders. Therefore, LSPS is not a confounder selection technique.

LSPS is distinct from techniques that attempt to select confounders empirically [67]. Some of these techniques also start with large numbers of covariates, but they attempt to find the subset that are confounders using information about the treatment and outcome. They then adjust for the selected covariates. As long as all confounders are observed and then selected, adjusting for them should eliminate confounding. It may not, however, benefit from the pinpointing that we identify in this paper. Unlike LSPS, confounder selection techniques are dependent on the outcomes, and the outcome rates are often very low in medical studies, potentially leading to variability in selection. Empirical studies [44] show that adjusting for a small number of confounders does not successfully adjust for unobserved confounders, and an empirical comparison of the methods favored LSPS [2].

5.3. Relation to the deconfounder

LSPS and the deconfounder [45, 46] are distinct but share several features. The deconfounder is a causal inference algorithm that estimates unbiased effects of multiple causes in the presence of unmeasured confounding. Under the pinpointability assumption (unmeasured confounders are pinpointable by multiple causes), the deconfounder can infer unmeasured confounders by fitting a probabilistic low-rank model to capture the dependencies among multiple causes. The deconfounder has been applied to EHR data for treatment effect estimation in the presence of unmeasured confounding [68]. Both methods thus can be shown to address unmeasured confounders when there is pinpointing.

6. Conclusions

In summary, LSPS is a confounding adjustment approach that includes large-scale pre-treatment covariates in estimating propensity scores. It has previously been demonstrated that LSPS balances unused covariates and can adjust for indirectly measured confounding. This paper contributes to understanding conditions under which LSPS adjusts for indirectly measured confounders, and how causal effect estimation by LSPS is impacted when such conditions are violated. We demonstrated the performance of LSPS on both simulated and real medical data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

[1]. Rosenbaum PR, Rubin DB, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55.

[2]. Tian Y, Schuemie MJ, Suchard MA, Evaluating large-scale propensity score performance through real-world and synthetic data experiments, International Journal of Epidemiology 47 (6) (2018) 2005–2014. [PubMed: 29939268]

[3]. Rubin DB, Matching to Remove Bias in Observational Studies, Biometrics. Journal of the International Biometric Society 29 (1) (1973) 159.

[4]. Rubin DB, The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies, Biometrics. Journal of the International Biometric Society 29 (1) (1973) 185.

[5]. Stuart EA, Matching Methods for Causal Inference: A Review and a Look Forward, Statistical Science 25 (1) (2010) 1–21. [PubMed: 20871802]

[6]. Rosenbaum PR, Rubin DB, Reducing Bias in Observational Studies Using Subclassification on the Propensity Score, Journal of the American Statistical Association 79 (387) (1984) 516.

[7]. Rubin DB, The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials, Statistics in Medicine 26 (1) (2007) 20–36. [PubMed: 17072897]

[8]. Shrier I, Re: The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials, Statistics in Medicine 27 (14) (2008) 2740–2741. [PubMed: 18069729]

[9]. Rubin DB, Author's reply re: The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials, Statistics in Medicine 27 (14) (2008) 2741–2742.

[10]. Shrier I, Propensity scores, Statistics in Medicine 28 (8) (2009) 1317–1318. [PubMed: 19288531]

[11]. Sjölander A, Propensity scores and M-structures, Statistics in Medicine 28 (9) (2009) 1416–1420. [PubMed: 19340845]

[12]. Pearl J, Remarks on the method of propensity score, Statistics in Medicine 28 (9) (2009) 1415–1416. [PubMed: 19340847]

[13]. Rubin DB, Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups?, Statistics in Medicine 28 (9) (2009) 1420–1423.

[14]. Rubin DB, Estimating causal effects from large data sets using propensity scores, Annals of Internal Medicine 127 (8 Pt 2) (1997) 757–763. [PubMed: 9382394]

[15]. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ, Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates, American Journal of Epidemiology 174 (11) (2011) 1213–1222. [PubMed: 22025356]

[16]. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T, Variable selection for propensity score models, American Journal of Epidemiology 163 (12) (2006) 1149–1156. [PubMed: 16624967]

[17]. Austin PC, Grootendorst P, Anderson GM, A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study, Statistics in Medicine 26 (4) (2007) 734–753. [PubMed: 16708349]

[18]. Pearl J, On a Class of Bias-Amplifying Variables that Endanger Effect Estimates, in: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, 2010, pp. 417–424.

[19]. Pearl J, Invited Commentary: Understanding Bias Amplification, American Journal of Epidemiology 174 (11) (2011) 1223–1227. [PubMed: 22034488]

[20]. Pearl J, Linear Models: A Useful "Microscope" for Causal Analysis, Journal of Causal Inference 1 (1) (2013) 155–170.

[21]. Wooldridge JM, Should instrumental variables be used as matching variables?, Research in Economics 70 (2) (2016) 232–237.

[22]. Steiner PM, Kim Y, The Mechanics of Omitted Variable Bias: Bias Amplification and Cancellation of Offsetting Biases, Journal of Causal Inference 4 (2) (2016) 20160009.

[23]. Ding P, Vanderweele TJ, Robins JM, Instrumental variables as bias amplifiers with general outcome and confounding, Biometrika 104 (2) (2017) 291–302. [PubMed: 29033459]

[24]. Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D, Empirical Performance of a New User Cohort Method: Lessons for Developing a Risk Identification and Analysis System, Drug Safety 36 (Suppl 1) (2013) 59–72.

[25]. Weinstein RB, Ryan P, Berlin JA, Matcho A, Schuemie M, Swerdel J, Patel K, Fife D, Channeling in the Use of Nonprescription Paracetamol and Ibuprofen in an Electronic Medical Records Database: Evidence and Implications, Drug Safety 40 (12) (2017) 1279–1292. [PubMed: 28780741]

[26]. Weinstein RB, Ryan PB, Berlin JA, Schuemie MJ, Swerdel J, Fife D, Channeling Bias in the Analysis of Risk of Myocardial Infarction, Stroke, Gastrointestinal Bleeding, and Acute Renal Failure with the Use of Paracetamol Compared with Ibuprofen, Drug Safety 43 (9) (2020) 927–942. [PubMed: 32500272]

[27]. Lane JCE, Weaver J, Kostka K, Duarte-Salles T, Abrahao MTF, Alghoul H, Alser O, Alshammari TM, Biedermann P, Banda JM, Burn E, Casajust P, Conover MM, Culhane AC, Davydov A, DuVall SL, Dymshyts D, Fernandez-Bertolin S, Fišter K, Hardin J, Hester L, Hripcsak G, Kaas-Hansen BS, Kent S, Khosla S, Kolovos S, Lambert CG, Lei J. v. d.,Lynch KE, Makadia R, Margulis AV, Matheny ME, Mehta P, Morales DR, Morgan-Stewart H, Mosseveld M, Newby D, Nyberg F, Ostropolets A, Park RW, Prats-Uribe A, Rao GA, Reich C, Reps J, Rijnbeek P, Sathappan SMK, Schuemie M, Seager S, Sena AG, Shoaibi A, Spotnitz M, Suchard MA, Torre CO, Vizcaya D, Wen H, Wilde M. d., Xie J, You SC, Zhang L, Zhuk O, Ryan P, Prieto-Alhambra D, O.-C.-. consortium, Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study, The Lancet Rheumatology 2 (11) (2020) e698–e711. [PubMed: 32864627]

[28]. Duke JD, Ryan PB, Suchard MA, Hripcsak G, Jin P, Reich C, Schwalm M, Khoma Y, Wu Y, Xu H, Shah NH, Banda JM, Schuemie MJ, Risk of angioedema associated with levetiracetam compared with phenytoin: Findings of the observational health data sciences and informatics research network, Epilepsia 58 (8) (2017) e101–e106. [PubMed: 28681416]

[29]. Morales DR, Conover MM, You SC, Pratt N, Kostka K, Duarte-Salles T, Fernández-Bertolín S, Aragón M, DuVall SL, Lynch K, Falconer T, Bochove K. v., Sung C, Matheny ME, Lambert CG, Nyberg F, Alshammari TM, Williams AE, Park RW, Weaver J, Sena AG, Schuemie MJ, Rijnbeek PR, Williams RD, Lane JCE, Prats-Uribe A, Zhang L, Areia C, Krumholz HM, Prieto-Alhambra D, Ryan PB, Hripcsak G, Suchard MA, Renin–angiotensin system blockers and susceptibility to COVID-19: an international, open science, cohort analysis, The Lancet Digital Health 3 (2) (2021) e98–e114. [PubMed: 33342753]

[30]. Burn E, Weaver J, Morales D, Prats-Uribe A, Delmestri A, Strauss VY, He Y, Robinson DE, Pinedo-Villanueva R, Kolovos S, Duarte-Salles T, Sproviero W, Yu D, Speybroeck MV, Williams R, John LH, Hughes N, Sena AG, Costello R, Birlie B, Culliford D, O'Leary C, Morgan H, Burkard T, Prieto-Alhambra D, Ryan P, Opioid use, postoperative complications, and implant survival after unicompartmental versus total knee replacement: a population-based network study, The Lancet Rheumatology 1 (4) (2019) e229–e236.

[31]. Wilcox MA, Villasis-Keever A, Sena AG, Knoll C, Fife D, Evaluation of disability in patients exposed to fluoroquinolones, BMC Pharmacology and Toxicology 21 (1) (2020) 40. [PubMed: 32493505]

[32]. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, Reich CG, Duke J, Madigan D, Hripcsak G, Ryan PB, Comprehensive comparative effectiveness and safety of

first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis, The Lancet 394 (10211) (2019) 1816–1826.

[33]. You SC, Jung S, Swerdel JN, Ryan PB, Schuemie MJ, Suchard MA, Lee S, Cho J, Hripcsak G, Park RW, Park S, Comparison of First-Line Dual Combination Treatments in Hypertension: Real-World Evidence from Multinational Heterogeneous Cohorts, Korean Circulation Journal 50 (1) (2019) 52–68. [PubMed: 31642211]

[34]. Hripcsak G, Suchard MA, Shea S, Chen R, You SC, Pratt N, Madigan D, Krumholz HM, Ryan PB, Schuemie MJ, Comparison of Cardiovascular and Safety Outcomes of Chlorthalidone vs Hydrochlorothiazide to Treat Hypertension, JAMA Internal Medicine 180 (4) (2020) 542–551. [PubMed: 32065600]

[35]. Kim Y, Tian Y, Yang J, Huser V, Jin P, Lambert CG, Park H, You SC, Park RW, Rijnbeek PR, Zandt MV, Reich C, Vashisht R, Wu Y, Duke J, Hripcsak G, Madigan D, Shah NH, Ryan PB, Schuemie MJ, Suchard MA, Comparative safety and effectiveness of alendronate versus raloxifene in women with osteoporosis, Scientific Reports 10 (1) (2020) 11115.

[36]. You SC, Rho Y, Bikdeli B, Kim J, Siapos A, Weaver J, Londhe A, Cho J, Park J, Schuemie M, Suchard MA, Madigan D, Hripcsak G, Gupta A, Reich CG, Ryan PB, Park RW, Krumholz HM, Association of Ticagrelor vs Clopidogrel With Net Adverse Clinical Events in Patients With Acute Coronary Syndrome Undergoing Percutaneous Coronary Intervention, JAMA : the Journal of the American Medical Association 324 (16) (2020) 1640–1650. [PubMed: 33107944]

[37]. Vashisht R, Jung K, Schuler A, Banda JM, Park RW, Jin S, Li L, Dudley JT, Johnson KW, Shervey MM, Xu H, Wu Y, Natrajan K, Hripcsak G, Jin P, Zandt MV, Reckard A, Reich CG, Weaver J, Schuemie MJ, Ryan PB, Callahan A, Shah NH, Association of Hemoglobin A 1c Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin: Analysis From the Observational Health Data Sciences and Informatics Initiative, JAMA Network Open 1 (4) (2018) e181755.

[38]. Schuemie MJ, Weinstein R, Ryan PB, Berlin JA, Quantifying bias in epidemiologic studies evaluating the association between acetaminophen use and cancer, Regulatory Toxicology and Pharmacology 120 (2021) 104866.

[39]. Schuemie MJ, Cepede MS, Suchard MA, Yang J, Schuler YTA, Ryan PB, Madigan D, Hripcsak G, How Confident Are We About Observational Findings in Health Care: A Benchmark Study, Harvard Data Science Review 2 (1) (2020).

[40]. Schuemie MJ, Ryan PB, Pratt N, Chen R, You SC, Krumholz HM, Madigan D, Hripcsak G, Suchard MA, Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study, Journal of the American Medical Informatics Association 27 (8) (2020) 1268–1277. [PubMed: 32827027]

[41]. Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA, Improving reproducibility by using high-throughput observational studies with empirical calibration, Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences 376 (2128) (2018) 20170356.

[42]. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA, Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data, Proceedings of the National Academy of Sciences 115 (11) (2018) 2571–2577.

[43]. Schuemie MJ, Ryan PB, Pratt N, Chen R, You SC, Krumholz HM, Madigan D, Hripcsak G, Suchard MA, Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND), Journal of the American Medical Informatics Association 27 (8) (2020) 1331–1337. [PubMed: 32909033]

[44]. Chen R, Schuemie M, Suchard M, Ostropolets A, Zhang L, Hripcsak G, Evaluation of large-scale propensity score modeling and covariate balance on potential unmeasured confounding in observational research (abstract), in: Proceedings of the AMIA Symposium, 2020.

[45]. Wang Y, Blei DM, The Blessings of Multiple Causes, Journal of the American Statistical Association 114 (528) (2019) 1574–1596.

[46]. Wang Y, Blei DM, Towards Clarifying the Theory of the Deconfounder, ArXiv (2020).

[47]. Hernán MA, Robins JM, Instruments for Causal Inference, Epidemiology (Cambridge, Mass.) 17 (4) (2006) 360–372. [PubMed: 16755261]

[48]. Hastie T, Tibshirani R, Friedman J, The elements of statistical learning: data mining, inference and prediction, 2nd Edition, Springer, 2009.

[49]. Walker A, Patrick, Lauer, Hornbrook, Marin, Platt, Roger V, Stang, Schneeweiss, A tool for assessing the feasibility of comparative effectiveness research, Comparative Effectiveness Research Volume 3 (2013) 11–20.

[50]. Austin PC, Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research, Communications in Statistics - Simulation and Computation 38 (6) (2009) 1228–1234.

[51]. Cox DR, Regression Models and Life-Tables, JSTOR 34 (2) (1972) 187–220.

[52]. Rubin DB, Estimating causal effects of treatments in randomized and nonrandomized studies, Journal of Educational Psychology 66 (5) (1974) 688.

[53]. Ogburn EL, Vanderweele TJ, Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders, Biometrika 100 (1) (2013) 241–248. doi:10.1093/biomet/ass054. [PubMed: 24014285]

[54]. Bhattacharya J, Vogt WB, Do Instrumental Variables Belong in Propensity Scores?, Working Paper, National Bureau of Economic Research (2007).

[55]. Middleton JA, Scott MA, Diakow R, Hill JL, Bias Amplification and Bias Unmasking, Political Analysis 24 (3) (2016) 307–323.

[56]. Hernán MA, Robins JM, Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available, American Journal of Epidemiology 183 (8) (2016) 758–764. doi:10.1093/aje/kwv254. [PubMed: 26994063]

[57]. Chien S-C, Ou S-M, Shih C-J, Chao P-W, Li S-Y, Lee Y-J, Kuo S-C, Wang S-J, Chen T-J, Tarng D-C, Chu H, Chen Y-T, Comparative Effectiveness of Angiotensin-Converting Enzyme Inhibitors and Angiotensin II Receptor Blockers in Terms of Major Cardiovascular Disease Outcomes in Elderly Patients, Medicine 94 (43) (2015) e1751. [PubMed: 26512568]

[58]. Hicks BM, Filion KB, Yin H, Sakr L, Udell JA, Azoulay L, Angiotensin converting enzyme inhibitors and risk of lung cancer: population based cohort study, BMJ (Clinical Research Ed.) 363 (2018) k4209.

[59]. Ku E, McCulloch CE, Vittinghoff E, Lin F, Johansen KL, Use of Antihypertensive Agents and Association With Risk of Adverse Outcomes in Chronic Kidney Disease: Focus on Angiotensin-Converting Enzyme Inhibitors and Angiotensin Receptor Blockers, Journal of the American Heart Association 7 (19) (2018) e009992.

[60]. Magid DJ, Shetterly SM, Margolis KL, Tavel HM, O'Connor PJ, Selby JV, Ho PM, Comparative Effectiveness of Angiotensin-Converting Enzyme Inhibitors Versus Beta-Blockers as Second-Line Therapy for Hypertension, Circulation: Cardiovascular Quality and Outcomes 3 (5) (2010) 453–458. [PubMed: 20716714]

[61]. Hasvold LP, Bodegård J, Thuresson M, Stålhammar J, Hammar N, Sundström J, Russell D, Kjeldsen SE, Diabetes and CVD risk during angiotensin-converting enzyme inhibitor or angiotensin II receptor blocker treatment in hypertension: a study of 15 990 patients, Journal of Human Hypertension 28 (11) (2014) 663–669. [PubMed: 25211055]

[62]. Kuroki M, Pearl J, Measurement bias and effect restoration in causal inference, Biometrika 101 (2) (2014) 423–437.

[63]. Miao W, Geng Z, Tchetgen Tchetgen EJ, Identifying causal effects with proxy variables of an unmeasured confounder, Biometrika 105 (4) (2018) 987–993. [PubMed: 33343006]

[64]. Tchetgen Tchetgen EJ, Ying A, Cui Y, Shi X, Miao W, An Introduction to Proximal Causal Learning, ArXiv (2020).

[65]. Albogami Y, Winterstein AG, Plasmode simulation of multiple imputation performance using internal validation data to adjust for unmeasured confounders, in: Pharmacoepidemiology and Drug Safety, Vol. 29, 2020, pp. 414–414.

[66]. Albogami Y, Cusi K, Daniels MJ, Wei Y-JJ, Winterstein AG, Glucagon-like peptide 1 receptor agonists and chronic lower respiratory disease exacerbations among patients with type 2 diabetes, Diabetes Care (2021).
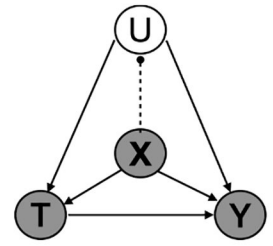
[67]. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA, High-dimensional propensity score adjustment in studies of treatment effects using health care claims data, Epidemiology 20 (4) (2009) 512. [PubMed: 19487948]

[68]. Zhang L, Wang Y, Ostropolets A, Mulgrave JJ, Blei DM, Hripcsak G, The Medical Deconfounder: Assessing Treatment Effects with Electronic Health Records, in: Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR, 2019, pp. 490–512.

[69]. Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D, Massive parallelization of serial inference algorithms for complex generalized linear models, ACM Transactions on Modeling and Computer Simulation 23 (2013) 10.
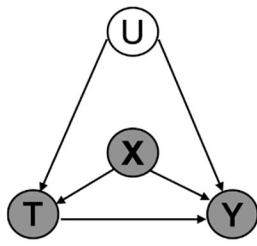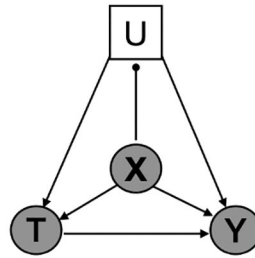
(a) No pinpointability   (b) Perfect pinpointability   (c) Weak pinpointability

Figure 1:
Causal graphs to estimate the treatment effect of T on the outcome Y. (a) Causal graph under no pinpointability. The unmeasured confounder U is not pinpointed by $\mathbf{X}$. (b) Causal graph under perfect pinpointability. The unmeasured confounder U is a deterministic function of the measured covariates $\mathbf{X}$. (c) Causal graph under weak pinpointability. The unmeasured confounder is only partially pinpointed by $\mathbf{X}$. Random variables are represented with circles, deterministic variables are represented with squares, measured variables are shaded, indirectly measured or unmeasured are not shaded, strong pinpointing is presented with a solid line, and weak pinpointing is presented with a dash line.
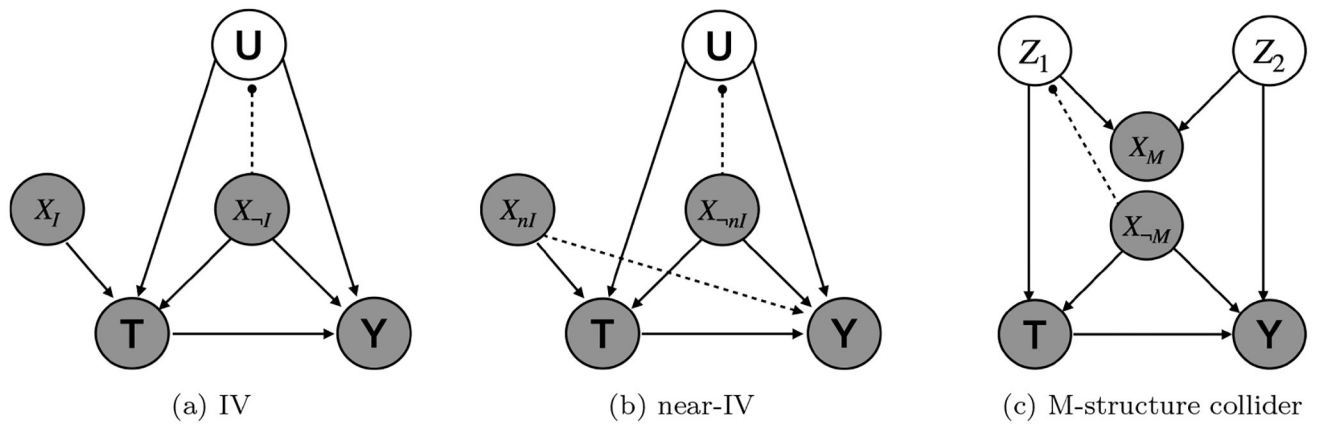
(a) IV (b) near-IV (c) M-structure collider

Figure 2:
Causal graph of a) instrumental variable, b) near-instrumental variable and c) M-structure collider. The dashed line with a solid dot means that the variable by the solid dot can be pinpointed by the measured covariates. We use different subscripts to distinguish the measured covariates playing different roles in the causal graph.
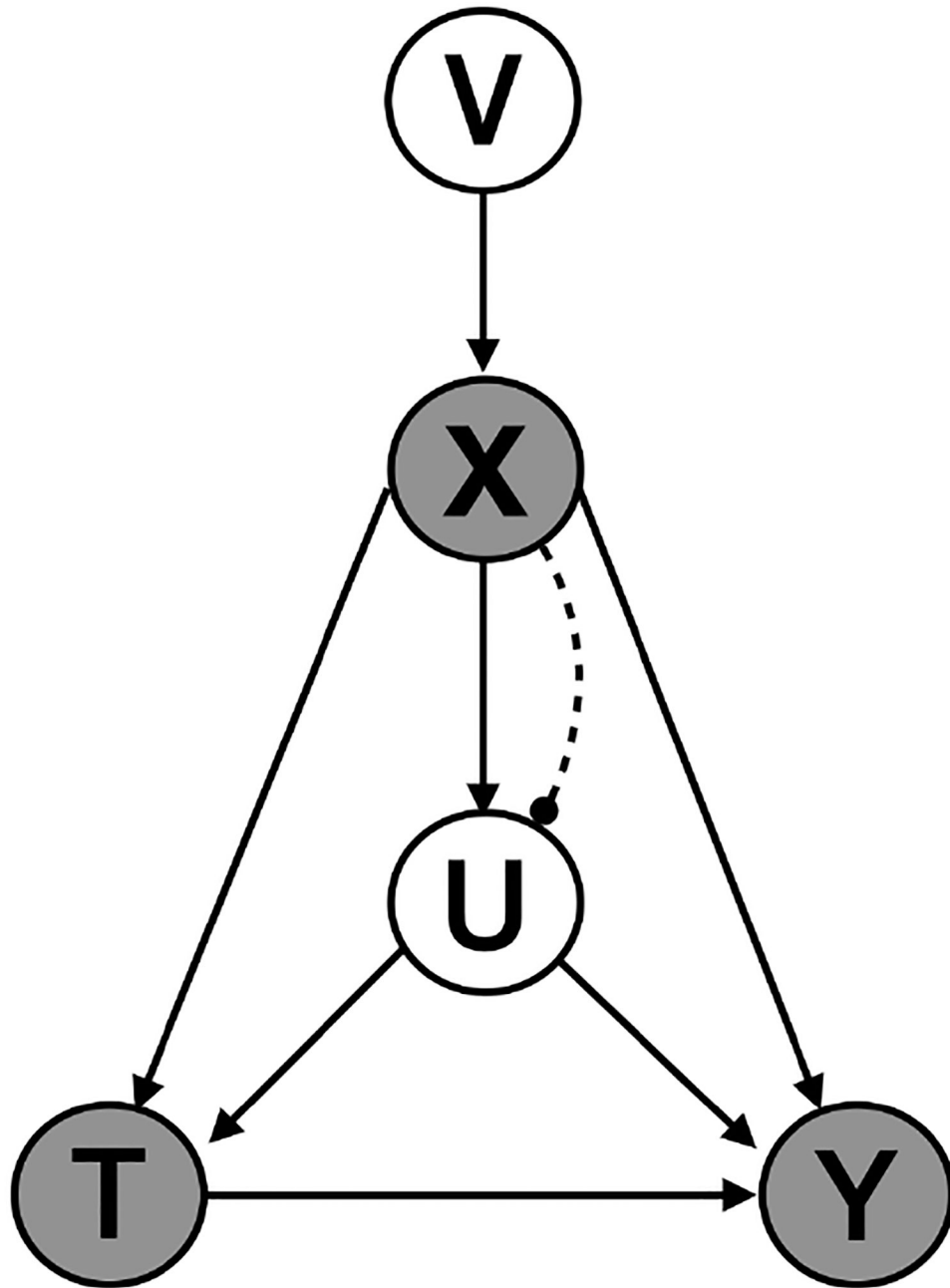
Figure 3:
Causal diagram of the simulation to estimate the effect of the treatment T on the outcome Y. The high-dimensional measured covariates **X** are induced by a low-dimensional latent variable **V**. The unmeasured confounder U is simulated as a function of the measured covariates **X**. When the function is deterministic, U is pinpointed by **X**. When the function is stochastic, U is weakly pinpointed by **X**. The degree of pinpointability is varied by adding varied amount of noise into the function.
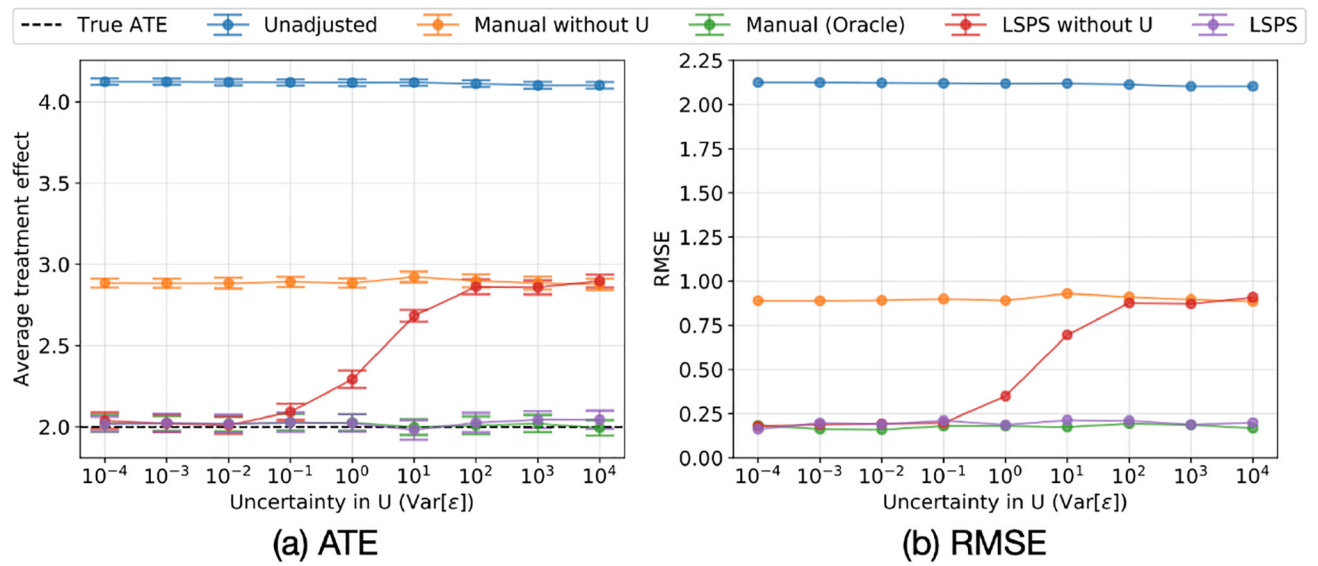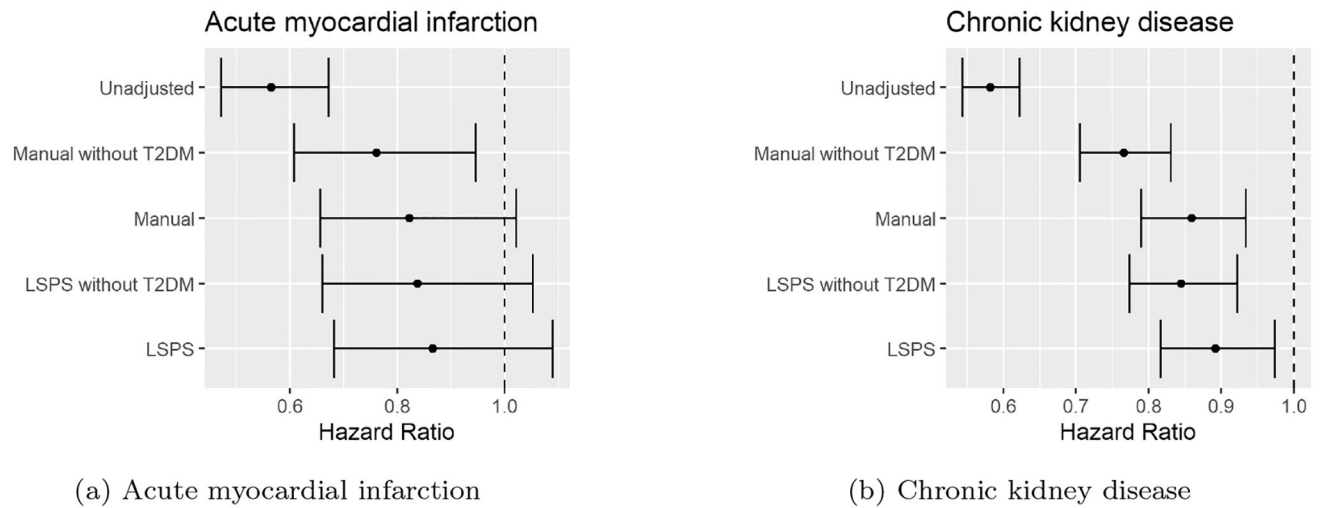
Figure 4:

Sensitivity analysis of pinpointability in Simulations 1. As pinpointability of the indirectly measured confounder decreases, LSPS's ability to adjust for the indirectly measured confounder decreases. (a) The mean and 95% CI of the estimated average treatment effect. (b) The RMSE of the estimated average treatment effect.

(a) Acute myocardial infarction

(b) Chronic kidney disease

Figure 5:
Comparison of hazard ratio from the unadjusted model and four models adjusting for confounders. The indirectly measured (or unused) confounder T2DM had a bigger impact on the HR estimated by manual models than by LSPS. (a) HR of the two anti-hypertensive medications on AMI. (b) HR of the two anti-hypertensive medications on CKD.