The Medical Deconfounder: Assessing Treatment Effects with Electronic Health Records (EHRs)

Linying Zhang LINYING.ZHANG@COLUMBIA.EDU
Department of Biomedical Informatics, Columbia University
New York, New York, USA

Yixin Wang YIXIN.WANG@COLUMBIA.EDU
Department of Statistics, Columbia University
New York, New York, USA

Anna Ostropolets ANNA.OSTROPOLETS@COLUMBIA.EDU
Department of Biomedical Informatics, Columbia University
New York, New York, USA

Jami J. Mulgrave JAMI.MULGRAVE@COLUMBIA.EDU
Department of Biomedical Informatics, Columbia University
New York, New York, USA

David M. Blei DAVID.BLEI@COLUMBIA.EDU
Department of Statistics, Department of Computer Science, Columbia University
New York, New York, USA

George Hripcsak GEORGE.HRIPCSAK@COLUMBIA.EDU
Department of Biomedical Informatics, Columbia University
New York, New York, USA

Abstract

The treatment effects of medications play a key role in guiding medical prescriptions. They are usually assessed with randomized controlled trials (RCTs), which are expensive. Recently, large-scale electronic health records (EHRs) have become available, opening up new opportunities for more cost-effective assessments. However, assessing a treatment effect from EHRs is challenging: it is biased by *unobserved confounders*, unmeasured variables that affect both patients' medical prescription and their outcome, e.g. the patients' social economic status. To adjust for unobserved confounders, we develop the *medical deconfounder*, a machine learning algorithm that unbiasedly estimates treatment effects from EHRs. The medical deconfounder first constructs a substitute confounder by modeling which medications were prescribed to each patient; this substitute confounder is guaranteed to capture all multi-medication confounders, observed or unobserved (Wang and Blei, 2018). It then uses this substitute confounder to adjust for the confounding bias in the analysis. We validate the medical deconfounder on two simulated and two real medical data sets. Compared to classical approaches, the medical deconfounder produces closer-to-truth treatment effect estimates; it also identifies effective medications that are more consistent with the findings in the medical literature.

1. Introduction

The treatment effect of medications plays a key role in guiding medical prescriptions. Usually, a treatment effect is assessed with a randomized controlled trial (RCT): each patient is randomly assigned to the treatment or the control group; only the treatment group receives the medication. The treatment effect is then assessed by comparing the average outcome of the two groups. RCTs are considered the gold standard for treatment effect assessment (Concato et al., 2000). Their randomized treatment assignments make RCTs immune to confounding bias and amenable to classical statistical tests of significance (Byar et al., 1976; Suresh, 2011). But, though theoretically sound, RCTs have substantial limitations: they are expensive, labor-intensive, and time-consuming. Moreover, they also do not always generalize to the real patient population (Deaton and Cartwright, 2018; Sanson-Fisher et al., 2007).

Electronic health records (EHRs) have recently emerged as an appealing alternative data source to RCTs for estimating treatment effects (Schuemie et al., 2018; Levine et al., 2018; Tannen et al., 2009). EHRs contain large-scale observational data about the medical history of patients, such as patient demographics, diagnosis, medications, and laboratory tests. In particular, patients' medication records and their lab tests can serve as evidence for medications' treatment effect: we can view their medication records as the treatment assignments and their lab tests as the outcome. This view of EHRs opens up new opportunities for more cost-effective ways of estimating treatment effects.

How can we use EHRs to estimate a treatment effect? A naive approach is to compare, for each medication, the outcome of the treated and the untreated patients. However, this approach leads to a biased assessment of the treatment effect; the treated and untreated population may not be comparable. For example, the two populations may be different in their age distributions, and this difference in age can lead to a difference in their health outcomes. Hence, naively comparing the outcomes between the treated and the untreated does not lead to correct treatment effect estimate of the medication. In causal inference terms, age is a *confounder*; it affects both whether a patient is treated and her outcome. When confounders are *observed*, we can adjust for them using classical causal inference methods like matching, subclassification and inverse probability weighting (Imbens and Rubin, 2015; Lopez et al., 2017; McCaffrey et al., 2013; Zanutto et al., 2005; Rassen et al., 2011; Lechner, 2001).

However, in EHRs many confounders are *unobserved*. For example, a patient's social economic status (SES) can influence both what medications she receives and her health condition. However, SES is an integrated measure of a person's sociological (e.g., occupation and education level) and economical (e.g., income) position in the society; it is typically not recorded in EHR systems. Such unobserved confounders challenge traditional causal inference methods; these methods assume all confounders are observed (Hernan and Robins, 2019).

To tackle this challenge, we develop the *medical deconfounder*, a machine learning approach that unbiasedly assesses treatment effects from EHRs. The medical deconfounder takes in patients' medication records (as the treatment) and lab tests (as the outcome) from EHRs; it outputs a set of medications that are deemed effective. To adjust for unobserved confounders, the medical deconfounder first models patients' medication records using a probabilistic factor model. It then constructs a substitute confounder based on this probabilistic factor model; this substitute confounder is guaranteed to capture all multi-medication confounders, both observed and unobserved (Wang and Blei, 2018). The medical deconfounder finally fits an outcome model. This outcome model describes how the lab test (outcome) depends on both the medications prescribed and the substitute confounder. The dependence on medications in the outcome model reflects the treatment effect of the medications.

Why might the medical deconfounder work? The key idea is to infer unobserved confounders by modeling how medications are prescribed together. For example, consider a cohort of patients with type 2 diabetes mellitus. We are interested in which of the medications taken by diabetic patients have an effect on their hemoglobin A1c (HbA1c). One confounder is body mass index (BMI), which affects both the medical prescription and the outcome HbA1c. If a patient is overweight or obese (i.e. has a high BMI), they are often prescribed with both diabetic medications and weight-lowering medications; overweight or obese patients also have higher HbA1c. Moreover, BMI is not recorded for all patient visits in the EHRs, rendering it an *unobserved confounder*. However, we can infer this unobserved confounder—BMI—by looking at which medications are prescribed together. If a patient is prescribed with both diabetic medications and weight-lowering medications, she probably has a high BMI. This is precisely what the medical deconfounder does; it constructs a substitute for unobserved confounders by modeling which medications are prescribed together.

In the next sections, we set up the treatment effect assessment problem in causal inference notations. We then describe the medical deconfounder and evaluate it on both simulation studies and real case studies. We apply the medical deconfounder to four datasets: two simulated and two real on distinct types of diseases. Across datasets, the medical deconfounder produces closer-to-truth treatment effect estimates than classical methods; it also identifies effective medications that are more consistent with the medical literature.

Technical Significance We propose the medical deconfounder, a machine learning approach to treatment effect estimation from EHRs. The medical deconfounder leverages probabilistic factor models to improve treatment effect estimates from EHRs. Between the two most popular options of probabilistic factor models (i.e. Poisson matrix factorization (PMF) (Schmidt et al., 2009; Gopalan et al., 2015) and deep exponential family (DEF) (Ranganath et al., 2015)), we find DEF helps to recover closer-to-truth treatment effects than PMF.

Clinical Relevance Assessing treatment effects is an important task that guides medical prescription. However, this task is challenging when the data comes from observational EHRs as opposed to randomized experiments. The presence of multiple medications further complicates the task. In this work, we propose the medical deconfounder as a solution to treatment effect assessment with EHRs. The medical deconfounder can adjust for unobserved confounders in EHRs and identify medications that causally affects the clinical outcome of interest.

Related work This work draws on two threads of related work.

The first body of related work is on probabilistic modeling for causal inference. Probabilistic models excel at capturing hidden patterns of high-dimensional data; examples include latent Dirichlet allocation (LDA) (Blei et al., 2003) and Poisson matrix factorization (PMF) (Schmidt et al., 2009; Gopalan et al., 2015). Recently, probabilistic modeling has been applied to causal inference. For example, Louizos et al. (2017) use variational autoencoders to infer unobserved confounders from proxy variables. Kocaoglu et al. (2017) and Ozery-Flato et al. (2018) connect generative adversarial network (GAN) and causal inference. Tran and Blei (2017), Wang and Blei (2018), and Ranganath and Perotte (2018) leverage probabilistic models for estimating unobserved confounders of multiple causes. The medical deconfounder in this work extends their use of probabilistic models into assessing the treatment effect of medications.

The second body of related work is on multiple causal inference with unobserved confounding. Tran and Blei (2017) and Heckerman (2018) focus on genome-wide association studies (GWAS); they

consider single-nucleotide polymorphisms (SNPs) as the multiple causes and estimate their effects on a trait of interest (e.g., height). Wang and Blei (2018) develop the deconfounder algorithm for multiple causal inference; it leverages probabilistic factor models to infer unobserved multi-cause confounders from the assignments of the multiple causes. Multiple causal inference with unobserved confounding was also studied in Ranganath and Perotte (2018) with an information-theoretic approach; their method is applied to estimate the causal effect of multiple lab measurements on the length of stay in the ICU. More recently, Bica et al. (2019) extend the deconfounder algorithm to time series data; they use recurrent neural network (RNN) to infer time-dependent unobserved confounders for multiple causal inference. The medical deconfounder presents another extension of the deconfounder algorithm; it extends the deconfounder to assess causal effect of multiple medications in EHRs.

2. The medical deconfounder

We frame treatment effect assessment as a multiple causal inference (Wang and Blei, 2018; Ranganath and Perotte, 2018) and describe the medical deconfounder.

2.1. Treatment effect assessment as a multiple causal inference

We first set up notation. Consider a dataset of N patients and D ($D > 1$) medications. Denote $\mathbf{A}_i$ as the medication record of patient i, $i = 1, \ldots, n$; it is a binary vector of length D that describes whether patient i has taken each of the D medications $\mathbf{A}_i = (A_{i1}, \ldots, A_{iD}) \in \{0, 1\}^D$. For example, the medication record of patient i is $\mathbf{A}_i = (0, 1, 0, \ldots, 0)$ if she has only taken the second medication. Each patient also has an outcome Y_i. For example, it can be the difference of pre-treatment and post-treatment lab measurements of patient i. For each patient, we observe both her medication records and her outcome

$$\{(\mathbf{A}_i, Y_i) : i = 1, \ldots, n\}.$$

The goal of treatment effect assessment is to identify the medications that (causally) affect the clinical outcome. In other words, all else being equal, the clinical outcome of a patient should be different if she had (or had not) taken the effective medication. We formulate this goal as a (multiple) causal inference problem (Imbens and Rubin, 2015; Rubin, 1974, 2005; Wang and Blei, 2018; Ranganath and Perotte, 2018). Denote $Y_i(\mathbf{a})$ as the potential outcome of patient i if she were assigned with treatment $\mathbf{a}$ of the medications. Either factual or counterfactual, this treatment $\mathbf{a}$ is a D-dimensional binary vector of medications: $\mathbf{a} \in \{0, 1\}^D$. Then the jth medication causally affects the outcome if the expected potential outcome of a patient is different had she taken (or not taken) the jth medication:

$$\mathbb{E}\left[Y_i(A_{i1}, \ldots, A_{ij-1}, 1, A_{ij+1}, \ldots, A_D) - Y_i(A_{i1}, \ldots, A_{ij-1}, 0, A_{ij+1}, \ldots, A_D)\right] \neq 0. \quad (1)$$

While treatment effects depend on all the potential outcomes $\{Y_i(\mathbf{a}) : \mathbf{a} \in \{0, 1\}^D\}$, we only observe one of them—the one that corresponds to the patient's medication record: $Y_i = Y_i(\mathbf{A}_i)$. To infer treatment effects from only the observed data, we develop the *medical deconfounder* by extending the deconfounder algorithm for multiple causal inference (Wang and Blei, 2018). The deconfounder algorithm can unbiasedly estimate $\mathbb{E}[Y_i(\mathbf{a})] - \mathbb{E}[Y_i(\mathbf{a}')]$ for all $\mathbf{a}$ and $\mathbf{a}'$ (and hence the left hand side of Equation (1)). It assumes "no unobserved single-cause confounders", i.e. no unmeasured variables can affect the outcome and *only one* medication (Wang and Blei, 2018).

The idea of the deconfounder is to construct a substitute confounder Z_i by fitting a probabilistic factor model to the medication records $\{\mathbf{A}_i : i = 1, \ldots, n\}$. This constructed substitute confounder Z_i satisfies ignorability (Rosenbaum and Rubin, 1983; Imai and Van Dyk, 2004)

$$Y_i(\mathbf{a}) \perp\!\!\!\perp \mathbf{A}_i \,|\, Z_i, \tag{2}$$

assuming "no unobserved single-cause confounders." This ignorability given Z_i (Equation (2)) greenlights causal inference. We can treat the substitute confounder Z_i as if it were an observed confounder and proceed with causal inference (Imbens and Rubin, 2015)

$$\mathbb{E}\left[Y_i(\mathbf{a})\right] = \mathbb{E}_Z\left[\mathbb{E}_Y\left[Y_i \,|\, Z_i, \mathbf{A}_i = \mathbf{a}\right]\right]. \tag{3}$$

Equation (3) lets us conclude treatment effects from EHRs and evaluate whether each medication is causally effective via Equation (1).

The medical deconfounder extends the deconfounder into medical settings. It operates in two steps. First, we fit a probabilistic factor model to all the medication records $\mathbf{A}_i$. This step lets us construct a substitute confounder Z_i for each patient. We then fit an outcome model treating this substitute confounder Z_i as an observed confounder. The fitted outcome model leads to treatment effect estimates of medications. We discuss the details of these two steps in the next sections.

2.2. The medical deconfounder

We describe the medical deconfounder in details. We first discuss how to construct a substitute confounder from prescription records in EHRs. Then we discuss how to assess the treatment effect of medications with an outcome model.

2.2.1. CONSTRUCTING THE SUBSTITUTE CONFOUNDER

The medical deconfounder constructs a substitute confounder Z_i by fitting a probabilistic factor model of the medication records $\{\mathbf{A}_i : i = 1, \ldots, N\}$. This probabilistic factor model needs to capture the observed distribution of the medication records $p(\mathbf{A}_i)$. We study three options of the probabilistic factor model for the medical deconfounder: probabilistic principal component analysis (PPCA), Poisson matrix factorization (PMF), and deep exponential family (DEF). Figure 1 shows the graphical representations of the three probabilistic factor models.

Probabilistic principal component analysis (PPCA) PPCA is a probabilistic formulation of PCA using a Gaussian latent variable model (Tipping and Bishop, 1999). For each patient i, their medication record $\mathbf{A}_i = (A_{i1}, \ldots, A_{iD})$ is modeled as a normal random variable; its mean is an inner product of a K-dimensional latent variable Z_i and some $(K \times D)$-dimensional parameter θ; we posit a standard normal prior on each Z_i: for $i = 1, \ldots, N$, and $j = 1, \ldots, D$, we have

$$\begin{aligned} Z_i &\sim \mathcal{N}(0, \lambda^2), \\ A_{ij} \,|\, Z_i &\sim \mathcal{N}(z_i^T \theta_j, \sigma^2). \end{aligned}$$

The latent variable Z_i will serve as the substitute confounder in the medical deconfounder.

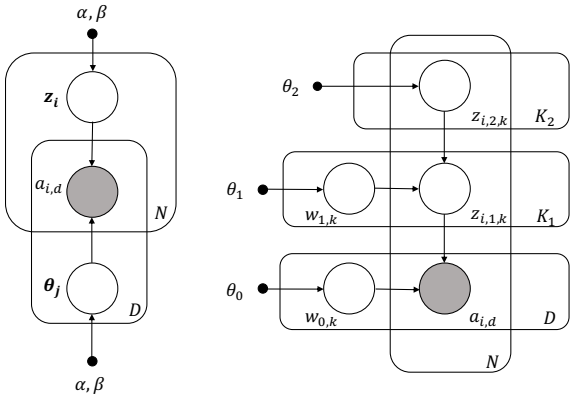

Figure 1: Graphical representation of PPCA (left), PMF (left) and a two-layer DEF (right). Random variables are represented by circles (shaded: observed; unobserved: hollow). Priors are represented by solid dots. In PPCA, z_i, θ_j and a_{id} are modeled with normal distributions. In PMF, z_i and θ_j are modeled with Gamma distribution, and a_{id} are modeled with Poisson distribution.

Poisson matrix factorization (PMF) PMF is a probabilistic factor model specific to modeling binary or count data (Schmidt et al., 2009; Gopalan et al., 2015). Because each medication treatment A_{ij} is binary—a patient either takes or does not take a medication, we can model the patients' medication records $\mathbf{A}_i$ with PMF. PMF is a similar factor model to PPCA except in its distributional assumptions; PMF models each medication treatment with a Poisson distribution and posits Gamma priors on the latent variables Z_i: for $i = 1, \ldots, N$, and $j = 1, \ldots, D$,

$$\begin{aligned} Z_i &\sim \text{Gamma}(\alpha, \beta), \\ \theta_j &\sim \text{Gamma}(\alpha, \beta), \\ A_{ij} \mid Z_i &\sim \text{Poisson}(Z_i^T \theta_j). \end{aligned}$$

In PMF, Z_i is the patient-specific latent variable for patient i; θ_j is medication-specific latent variable for medication j; A_{ij} indicates whether patient i took medication j. Both Z_i and θ_j are K-dimensional random variables. The latent variable Z_i will serve as the substitute confounder downstream.

Deep exponential family (DEF) A DEF is a flexible probabilistic factor model that has multiple layers of latent variables as in neural networks (Ranganath et al., 2015). We focus on a two-layer DEF; it has the following structure:

$$\begin{aligned} W_{l,k} &\sim \text{Gamma}(\alpha, \beta), \\ Z_{i,2,k} &\sim \text{Gamma}(\alpha, \beta), \\ Z_{i,1,k} \mid Z_{i,2,k} &\sim \text{Gamma}(\alpha, g(W_{1,k} Z_{i,2,k})), \\ A_{i,d} \mid Z_{i,1} &\sim \text{Poisson}(g(W_0 Z_{i,1})). \end{aligned}$$

The variable $Z_{i,l,k}$ corresponds to the kth latent variable in the l-th layer for patient i. The variable $W_{l,k}$ is a K-dimensional weight vector in the l-th layer. The variable A_{id} is a binary indicator of whether patient i is prescribed with medication j.

In all three probabilistic factor models, the latent variable Z_i will serve as the substitute confounder in downstream treatment effect estimation. Specifically, we will fit the probabilistic model, i.e. infer θ,

using Markov chain Monte Carlo methods (Robert and Casella, 2005) or variational inference (Jordan et al., 1999; Blei et al., 2017). We then compute the posterior expectation of Z_i given the inferred $\hat{\theta}$,

$$\hat{Z}_i \triangleq \mathbb{E}_Z \left[Z_i \mid \mathbf{A}_i, \hat{\theta} \right].$$

If the probabilistic factor model fits the data well, then we can use the constructed substitute confounder $\hat{Z}_i$ in the downstream treatment effect assessment.

To assess the adequacy of the probabilistic factor model, we follow Wang and Blei (2018) to perform a predictive check (Gelman et al., 1996). For each patient i, we randomly hold out $s\%$ entries of her medication record $\mathbf{A}_i$. The predictive check then proceeds in three steps:

1. Generate replicated datasets for the heldout entries based on the inferred posterior $p(Z_i \mid \mathbf{A}_i, \hat{\theta})$.

2. Compare the value of a test statistic on the replicated datasets to that of the observed dataset. The test statistic is the expected log-likelihood of the heldout entries

   $$t(\mathbf{X}_{\text{heldout}}) \triangleq \mathbb{E}_{\mathbf{Z},\theta}[\log p(\mathbf{X}_{\text{heldout}} \mid \mathbf{Z}, \hat{\theta}) \mid \mathbf{X}_{\text{obs}}].$$

   We compute this test statistic on both the observed dataset $\mathbf{X}_{\text{heldout}}$ and each replicated dataset $\mathbf{X}^{\text{rep}}_{\text{heldout}}$.

3. Conclude the probabilistic factor model is adequate if the predictive score is close to 0.5. The predictive score is defined as

   $$\text{predictive score} \triangleq p\left(t(\mathbf{X}^{\text{rep}}_{\text{heldout}}) < t(\mathbf{X}_{\text{heldout}})\right). \tag{4}$$

   A close-to-0.5 predictive score indicates neither under-fitting nor over-fitting of the data (Wang and Blei, 2018). Otherwise, the probabilistic factor model is inadequate.

If a probabilistic factor model is deemed inadequate by the predictive check, we must choose a different factor model. We repeat the construction of the substitute confounder $\hat{Z}_i$ until one constructed $\hat{Z}_i$ passes the predictive check.

2.2.2. FITTING A BAYESIAN LINEAR REGRESSION OUTCOME MODEL

After constructing substitute confounder $\hat{Z}_i$, the medical deconfounder adjusts for it as if it were an observed confounder in causal inference. Specifically, we fit a Bayesian regression model to Y_i against both the medication record $\mathbf{A}_i$ and the substitute confounder $\hat{Z}_i$

$$Y_i \sim \mathcal{N}(\sum_{j=1}^{D} \beta_j A_{ij} + \sum_{k=1}^{K} \gamma_k \hat{Z}_{ik}, \sigma^2),$$

where K is the dimension of the substitute confounder Z_i. For studies with more than two medications, we posit an isotropic Gaussian prior $\mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$ on all coefficients β_j and γ_k.

We estimate the regression coefficients $\beta_j, j = 1, \ldots, D$ with mean-field variational inference. They indicate the average treatment effect of each medication:

$$\beta_j = \mathbb{E}\left[Y_i(A_{i1}, \ldots, A_{ij-1}, 1, A_{ij+1}, \ldots, A_{iD}) - Y_i(A_{i1}, \ldots, A_{ij-1}, 0, A_{ij+1}, \ldots, A_{iD})\right]. \tag{5}$$

When the coefficient β_j is significantly different from zero, we conclude that medication j causally affects the clinical outcome of interest.

3. Simulation studies

We first study the medical deconfounder on two simulation studies. The two simulation studies are of distinct nature: one has only two causes; the other has many causes. Below we first describe the evaluation metrics and the baseline method we compare, and then discuss the details of the two simulation studies.

3.1. Performance metrics and baseline methods

In both simulation studies, we evaluate the performance of the medical deconfounder by the closeness-to-truth of its causal estimates. We then compare these estimates with classical methods that do not adjust for unobserved confounders.

Performance metrics As a measure of closeness-to-truth in simulations, we compute the root mean square error (RMSE) between the estimated treatment effects and the true effects. The RMSE is defined as

$$\text{RMSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \sqrt{\frac{1}{D}\sum_{j=1}^{D}(\hat{\beta}_j - \beta_j)^2},$$

where $\hat{\boldsymbol{\beta}}$ is the estimated treatment effect and $\boldsymbol{\beta}$ is the true effect.

We also evaluate the posterior distribution of the treatment effect by " % coverage," i.e. how often the estimated 95% credible interval (CI) covers the true treatment effect. We derive the 95% credible interval (CI) from the posterior distribution of the outcome model, and compute the % coverage by

$$\%\text{ coverage} = \frac{\text{Number of CI covers the truth}}{\text{Number of total treatments}} \times 100\%.$$

Baseline methods We compare the medical deconfounder with classical methods that do not adjust for unobserved confounders. These methods simply model the outcome as a function of the medical records only; they do not adjust for any confounders. We call them "the unadjusted model." Specifically, they fit the following Bayesian regression model

$$Y_i \sim \mathcal{N}(\sum_{j=1}^{D}\beta_j A_{ij}, \sigma^2).$$

They then take the β coefficients as the effect size of each medication.

In addition to the unadjusted model, we also compare the medical deconfounder to an oracle model. The oracle model has access to the true unobserved confounders C_i; it fits a Bayesian regression model to both the medical records $\mathbf{A}_i$ (medications) and the true confounders

$$Y_i \sim \mathcal{N}(\sum_{j=1}^{D}\beta_j A_{ij} + \sum_{k=1}^{K}\gamma_k C_{ik}, \sigma^2).$$

We emphasize that these unobserved confounders C_i are not available in practice. The oracle model illustrates the best possible performance in assessing treatment effects.

Computation We fit probabilistic factor models using black box variational inference (Ranganath et al., 2014) as implemented in Edward (Tran et al., 2016, 2017). We then draw 1000 samples from the inferred posterior and fit the outcome model using automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2017) as implemented in the rstanarm package (Carpenter et al., 2017) of R (R Core Team, 2013).

3.2. Simulation study I: A two-medication simulation

The first simulation study of the medical deconfounder is on a toy example of only two medications. Under unobserved confounding, the medical deconfounder is able to tell the causal (i.e. causally-effective) medication from the non-causal (i.e. non-causally-effective) medication. By contrast, the unadjusted model returns both medications as causal.

Experimental setup We experiment the medical deconfounder in two setups. In both, there is an unobserved confounder C_i and two medications A_{i1} and A_{i2} for each patient i. The unobserved confounder C_i is multi-medication; both medications A_{i1} and A_{i2} are linearly dependent on the unobserved confounder C_i. We then simulate a continuous outcome Y_i that is also linearly dependent on the confounder C_i. We consider two setups of the outcome. In the first setup, neither of the causes is causal. In the second, one of the causes is causal. (Figure 2 illustrates the two settings with graphical models.)

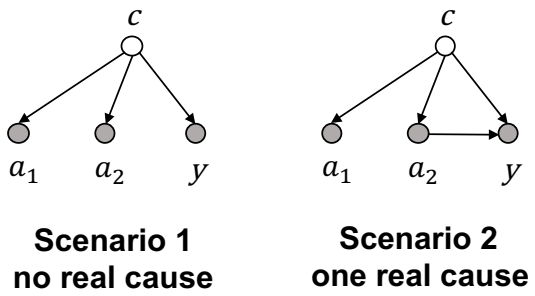

Figure 2: Causal graph of the two setups of the two-medication simulation study. Scenario 1 includes no real cause, while setup 2 has one real cause a_2. The confounder c is a multi-medication confounder in both setups.

Specifically, for each patient i, we simulate her confounder C_i and the medication records $\mathbf{A}_i$ as

$$\begin{aligned} C_i &\sim \mathcal{N}(0, 1), \\ A_{1i} &= 0.3C_i + \epsilon_i, \\ A_{2i} &= 0.4C_i + \epsilon_i, \end{aligned}$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$. In the first setup, the outcome is simulated as

$$Y_i = 0.5C_i + \epsilon_i.$$

In the second, it is simulated as

$$Y_i = 0.5C_i + 0.3A_{2i} + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$. In both setups, we simulate a sample size of $N = 1,000$.

Deconfounding with the medical deconfounder We use probabilistic principal component analysis (PPCA) with latent dimensionality $K = 1$ as the probabilistic factor model. To assess the model fit, we perform the predictive check for the factor model by randomly holding out 20% of the data. The fitted model returns a predictive score close to 0.5; it passes the predictive check.

	medication 1		medication 2	
	coef (std err)	**p-value**	**coef (std err)**	**p-value**
Truth	0.000	1	0.000	1
Oracle	0.025 (0.039)	0.522	-0.022 (0.040)	0.594
Unadjusted	0.125 (0.042)	0.003	0.141 (0.041)	0.001
Med. Dcf.	0.020 (0.081)	**0.803**	0.052 (0.071)	**0.461**

Table 1: Estimated treatment effects in the two-medication simulation with no real cause. The p-value for each medication tests the null hypothesis that the coefficient is equal to zero (no causal effect). The medical deconfounder ("Med. Dcf.") returns closer-to-truth p-values of the coefficients than the baseline method.

	medication 1		medication 2	
	coef (std err)	**p-value**	**coef (std err)**	**p-value**
Truth	0.000	1	0.300	0
Oracle	0.058 (0.038)	0.132	0.329 (0.039)	0.000
Unadjusted	0.181 (0.040)	0.000	0.469 (0.040)	0.000
Med. Dcf.	0.069 (0.063)	**0.272**	0.333 (0.072)	**0.000**

Table 2: Estimated treatment effects in the two-medication simulation with one real cause. The p-value for each medication tests the null hypothesis that the coefficient is equal to zero (no causal effect). The medical deconfounder ("Med. Dcf.") returns closer-to-truth p-values of the coefficients than the baseline method.

Results Table 1 and Table 2 present the regression coefficients and p-values of the three models in the two experimental setups. We compare the unadjusted model (no control), the medical deconfounder (control for the substitute confounder), and the oracle model (control for the true unobserved confounder). In both setups, the unadjusted model leads to biased causal coefficient estimates. The medical deconfounder reduces the bias of estimates, and returns causal coefficients that are nearly the same as those from the oracle.

Moreover, the medical deconfounder is able to identify the true causal medication in the second setup. After adjusting for the substitute confounder, the coefficient of the true causal medication stays significant while the non-causal one becomes insignificant. Their p-values are consistent with whether they are causal. In contrast, the unadjusted model returns statistically significant coefficients for both medications; it leads to a wrong conclusion that both medications are causal. In rare runs, the medical deconfounder did not adjust the raw coefficient estimates significantly, but even then, it increased the variance of the estimate of the non-causal medication so that it can still correctly classify medications as causal or non-causal.

3.3. Simulation study II: A multi-medication simulation

We next evaluate the medical deconfounder on a multi-medication simulated dataset. As in the first simulation, the medical deconfounder improves the effect size estimates for the medications; the confidence interval of treatment effect estimates also covers the truth more often than classical methods.

Experimental setup We simulate a dataset of $D = 50$ medications and $N = 5,000$ patients. The medication record $\mathbf{A}_i$ of each patient is influenced by a ten-dimensional multi-medication unobserved confounder C_i. A real-valued outcome is simulated as a function of the confounder C_i and the medication record $\mathbf{A}_i$. The simulated dataset is at a similar scale to the dataset we use in the empirical studies.

We simulate each multi-medication confounder C_{ik} from a standard normal distribution,

$$C_{ik} \sim \mathcal{N}(0, 1), \quad k = 1, \ldots, 10.$$

Then we simulate the medication record of each patient i from a Bernoulli distribution,

$$A_{ij} \sim \text{Bern}(\sigma(\sum_{k=1}^{K} \lambda_{kj} C_{ik})), \quad j = 1, \ldots, 50,$$

where $\sigma(\cdot)$ is the sigmoid function and $\lambda_{kj} \sim \mathcal{N}(0, 0.5^2)$. Finally, we simulate a continuous outcome Y_i as a function of both the confounder and the medication record,

$$Y_i = \sum_{j=1}^{D} \beta_j A_{ij} + \sum_{k=1}^{K} \gamma_k C_{ik} + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$, $\beta_j \sim \mathcal{N}(0, 0.25^2)$, and $\gamma_k \sim \mathcal{N}(0, 0.25^2)$. To mimic the sparsity of causal medications in practice, we randomly select 80% of the medications and set their coefficients β_j to zero, therefore, only 10 medications are causal.

Deconfounding with the medical deconfounder We implement two probabilistic factor models PMF and DEF for the medical deconfounder. The PMF passes the predictive check with $K = 450$; the DEF passes the predictive check with 30 and 4 latent variables in each layer. Both factor models yield predictive scores close to 0.5.

	RMSE	**% Coverage**		
		All	**Causal**	**Non-causal**
Oracle	0.05	78	50	85
Unadjusted	0.14	38	30	40
Med. Dcf. (PMF)	**0.12**	38	30	40
Med. Dcf. (DEF)	0.13	**48**	**40**	**50**

Table 3: RMSE and % coverage of CI of the multi-medication simulation. (Lower RMSE is better; higher % coverage is better.) The medical deconfounder produces closer-to-truth causal estimates than the unadjusted model. The CI of estimates from DEF covers more true effects than the unadjusted.

Results Table 3 summarizes the causal estimation results of the oracle model, the unadjusted model, and the medical deconfounder with PMF and DEF as probabilistic factor models. The medical deconfounder with both probabilistic factors produce less biased effect estimates compared to the unadjusted model. Also, 48% of the CI's from DEF covers the truth, higher than the 38% from the unadjusted model. The increase of % coverage by DEF is a consequence of both correctly identifying more causal treatments, and decreasing the false positives.

4. Case studies

We apply the medical deconfounder to two case studies on real datasets of distinct disease cohorts. In both studies, the medical deconfounder identifies causal medications that are consistent with the medical literature. Below we discuss the two disease cohorts and present the empirical results.

4.1. Cohort extraction and evaluation methods

In both case studies, we extract patient cohorts from the Columbia University Medical Center database. The database contains de-identified electronic health records standardized and stored according to the Observational Health Data Science and Informatics (OHDSI) format (Hripcsak et al., 2016). We apply the medical deconfounder to each cohort. Medical experts then perform literature reviews and evaluate the results returned by the medical deconfounder.

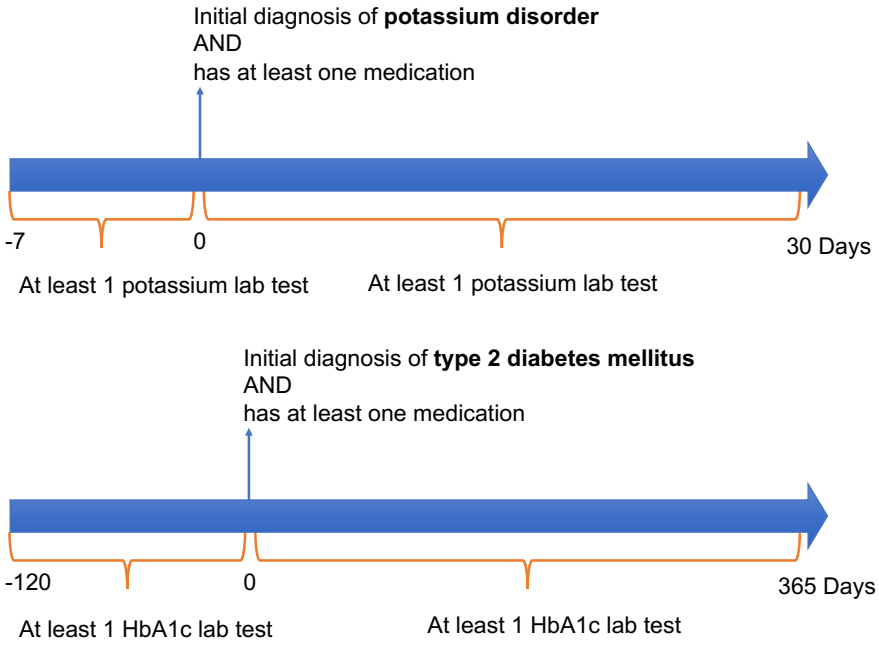

Figure 3: The diagram of cohort definition for potassium disorders (top) and type 2 diabetes mellitus (bottom). Patients meeting all criteria in the diagram are included in the cohort. Lines and arrows represent required intervals between events.

Case study I: Potassium disorders cohort. Patients who meet the following criteria are included in the potassium disorder cohort:

- was diagnosed with hypokalemia or hyperkalemia with continuous observation of at least 7 days before and 30 days after initial diagnosis (index date);
- has at least 1 measurement of potassium in serum/blood within 7 days prior to the first diagnosis;
- has at least 1 measurement of potassium in serum/blood within 30 days after the first diagnosis;
- has at least 1 medication exposure on the same day of initial diagnosis.

After data preprocessing, there are 6185 patients and 33 unique medications included in this cohort.

Case study II: Type 2 diabetes cohort. Patients who meet the following criteria are included in the type 2 diabetes cohort:

- was diagnosed with type 2 diabetes with continuous observation of at least 30 days before and 30 days after the initial diagnosis (index date);
- has at least 1 measurement of HbA1C 120 days prior to the first diagnosis;
- has at least 1 measurement of HbA1C within 365 days after the first diagnosis;
- has at least 1 medication exposure on the same day of initial diagnosis.

After data preprocessing, there are 5564 patients and 30 unique medications included in this cohort.

Data preprocessing For both cohorts, patients' medication records on the index date and their lab measurements immediately before and after the index date are extracted from the database using the OHDSI Atlas interface (OHDSI team, 2019). All medications are mapped to ingredients and dosage is ignored. To reduce the sparsity of the patient-medication matrix, we remove the 5% least frequent ingredients from downstream analysis.

Evaluation methods Due to the unavailability of true treatment effects in real datasets, we compare the medical deconfounder estimates with the findings reported in the medical literature. Medical experts perform literature review for all the medications appeared in the studies; they look for evidence indicating the presence or absence of causal relationships between the medications and the outcome of interest.

4.2. Case study I: Potassium disorders

We apply the medical deconfounder to the patient cohort of potassium disorders. Consider all the medications taken by the cohort of patients with potassium disorders. The goal is to identify which of these medications have causal effects on the serum potassium level. We find that the medications identified to be causal by the medical deconfounder are in concordance with the evidence from the medical literature.

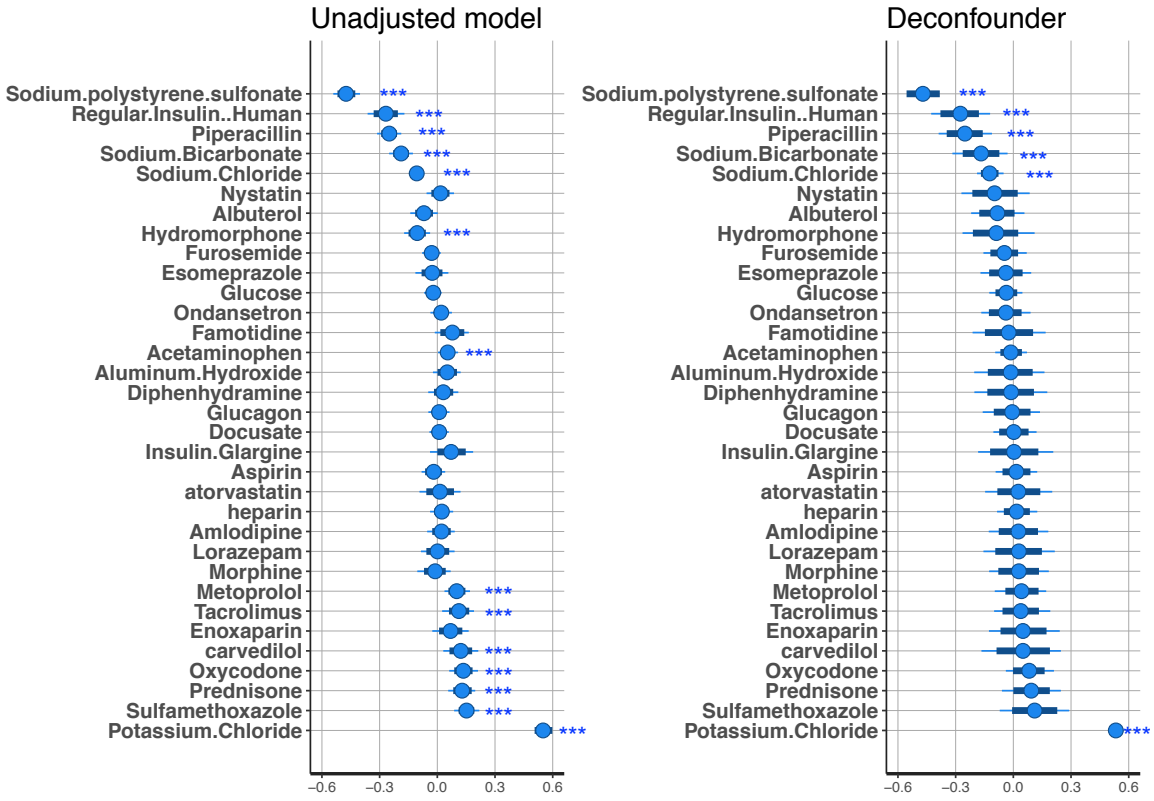

Figure 4: Treatment effects of medications in the potassium disorders cohort estimated by the unadjusted model (left) and the medical deconfounder (right). The medical deconfounder returns causal medications that are more consistent with the medical literature. The mean, 80% credible interval, and 95% credible interval of the estimated coefficients are indicated by the circle, the horizontal bar, and the solid line respectively. A medication is determined *causal* if its 95% credible interval excludes zero and is marked with "***". A positive coefficient means that the medication increases serum potassium.

Results Figure 4 shows the coefficients estimated by the medical deconfounder (control for the substitute confounder) and the unadjusted model (no control). The medical deconfounder reduces false positive discoveries while the true causal medications remain significant after adjustment. (A medication is determined *causal* if its 95% credible interval excludes zero.)

Five medications are found to be causal by both models with well-supported medical literature on their physiological mechanisms: (1) sodium polystyrene sulfonate is a potassium-binding resin commonly used to treat hyperkalemia by increasing the excretion of potassium in stool (Batterink et al., 2015); (2) insulin lowers serum potassium by internalizing potassium intracellularly (McNicholas et al., 2018); (3) piperacillin (often prescribed with tazobactam) is a commonly used antibiotic for various infections and are report to cause hypokalemia in a series of case report (Zaki and Lad, 2011; Hussain et al., 2010; Polderman and Girbes, 2002); (4) sodium bicarbonate raises systemic pH, a process accompanied by potassium movement into the cells to maintain electroneutrality, leading to decrease of potassium in the blood (Abuelo, 2018; Burnell et al., 1956); (5) potassium chloride is commonly administered to replenish potassium in patients with low serum potassium.

Twenty-seven medications are identified as non-causal by the medical deconfounder, including eight medications changing from causal to non-causal after deconfounding. For most of these medications, we can not find evidence in the medical literature that suggests their influence on potassium, although a few medications may require more detailed evaluation. Among them, one medication albuterol is reported to have a potassium-lowering effect in patients with renal failure (Montoliu et al., 1987), but neither the unadjusted model nor the medical deconfounder identifies it as a causal medication. We hypothesize that this is because the cohort of renal failure patients in this dataset is not large enough for this effect to be detected. The other medication, furosemide, which is a diuretic used to reduce extra fluid in the body, has a delayed effect on potassium compared to other medications with immediate effect (e.g., sodium polystyrene sulfonate and regular insulin). Given this study uses the potassium measurement immediately after medications are prescribed to assess the treatment effect of all medications, there may not be enough time for the effect of furosemide to appear (Mushiyakh et al., 2011; Stason et al., 1966).

Two medications, changing from causal to non-causal after deconfounding, are found to have an effect on potassium level in the literature. One medication is tacrolimus, which is an immunosuppressive medication prescribed for patients with organ transplant to lower the risk of organ rejection. Tacrolimus can increase serum potassium concentration due to reduced efficiency of urinary potassium excretion (Lee and Kim, 2007). The other medication is sulfamethoxazole, which is an antibiotic to treat infection. It is found to reduce renal potassium excretion through the competitive inhibition of epithelial sodium channels when co-administered with trimethoprim (Velazquez et al., 1993; Antoniou et al., 2010). These two medications are prescribed to patients with relatively complicated health problems, and thus more scientific study may be necessary to understand the mechanism. Even though the medical deconfounder does not identify these two medications to be causal but the unadjusted model does, the medical deconfounder still identifies effective medications that are more consistent with the medical literature (six medications identified as causal only by the unadjusted model lack evidence for an effect on potassium).

4.3. Case study II: Type 2 diabetes mellitus

We next study the medical deconfounder on a patient cohort of type 2 diabetes mellitus. The goal is to identify medications that causally affect hemoglobin A1c (HbA1c). HbA1c measures the percentage

of a protein called hemoglobin in the bloodstream that is bound by glucose; it is a key indicator of the average blood glucose over the previous two to three months (Sherwani et al., 2016). In contrast to the first case study where the treatment effect is immediate, HbA1c reflects the long-term effect of medications on regulating blood glucose. This long-term effect poses additional challenges in treatment effect assessments.

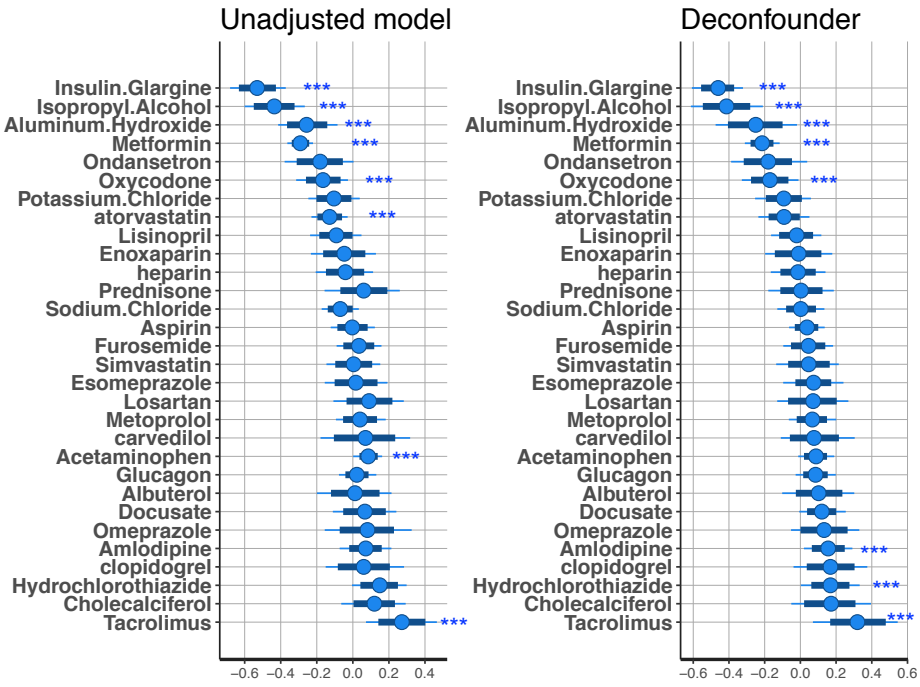

Figure 5: Treatment effects of medications in the diabetes cohort estimated by the unadjusted model (left) and the medical deconfounder (right). The medical deconfounder returns causal medications that are more consistent with the medical literature. The mean, 80% credible interval, and 95% credible interval of the estimated treatment effect are indicated by the circle, the horizontal bar and the solid line respectively. A medication is determined *causal* if its 95% credible interval excludes zero and is marked with "***". A negative treatment effect means that the medication down-regulates HbA1c, and a positive treatment effect means the medication up-regulates HbA1c.

Figure 5 shows the treatment effects estimated by the medical deconfounder (control for the substitute confounder) and the unadjusted model (no control).

The medical deconfounder returns three causal medications with positive coefficients. Among the three, tacrolimus is the only medication that is causal in both the medical deconfounder and the unadjusted model. Both of the other two medications only appear significant in the medical deconfounder. These two medications—amlodipine and hydrochlorothiazide—are medications for treating high blood pressure, a common comorbidity of diabetes. They have been found to induce hyperglycemia in non-diabetic patients with essential hypertension in several comparative studies (Fukao et al., 2011; Cooper-DeHoff et al., 2010). These findings in the literature are consistent with the positive treatment effect estimates by the medical deconfounder. Moreover, both of the causal

medications are the first line recommended therapies for hypertension, so the finding that the two medications can cause hyperglycemia are important to guide the treatment decision of hypertension.

In more details, one of the medication amlodipine can induce hyperglycemia likely because it blocks the calcium channels that inhibits the release of insulin from β cells in the pancreas (Sandozi, 2010). The other medication hydrochlorothiazide is a thiazide diuretics, a class of medications that are known to promote hyperglycemia and in some cases contribute to the new onset of diabetes (Cooper-DeHoff et al., 2010; Gress et al., 2000). The exact mechanism is unknown, but it is postulated to involve worsening of insulin resistance, inhibition of glucose uptake, and decreased insulin release, among other pathways.

Two medications, acetaminophen and atorvastatin, are identified causal by the unadjusted model, but are deemed non-causal in the medical deconfounder. We do not find any evidence of causal relationship between acetaminophen and blood glucose, except a few reports about its interference on blood glucose sensors (Zyoud et al., 2011; Tierney et al., 2000). Atorvastatin is reported to increase the incidence of diabetes by decreasing insulin sensitivity and increase ambient glycemia in hypercholesterolemic patients (Koh et al., 2010). Its estimated effect by the unadjusted model is negatively causal. Although the medical deconfounder is not able to identify this medication to be causal with positive effect, the estimated treatment effect is more positive after deconfounding, a change in the direction consistent with its potential influence on increasing glucose.

The same five medications with a negative effect on HbA1c are returned by both models. These include two well-known medications for treating type 2 diabetes, insulin and metformin (Rojas and Gomes, 2013; Hirst et al., 2012; Swinnen et al., 2009). Isopropyl alcohol is not a medication but an ingredient in alcohol-based sanitizers that are commonly used to clean patients' skin before a blood test. A few studies were found addressing concerns about the interference of isopropyl alcohol on the accuracy of blood glucose test, but results are inconsistent among the studies (Mahoney et al., 2011; Dunning et al., 1994). There exists little literature about aluminum hydroxide and oxycodone on their association with blood glucose. These could be novel findings for further investigations.

5. Discussion

In this paper, we propose the medical deconfounder, a machine learning algorithm for assessing treatment effects of medications with EHRs. For a cohort of patients, the medical deconfounder works with multiple relevant medications simultaneously and adjusts for unobserved multi-medication confounders. The medical deconfounder then identifies medications that causally affect the clinical outcome of interest. We study the medical deconfounder on four datasets, two simulated and two real. Across datasets, the medical deconfounder improves the treatment effect estimates; it also identifies causal medications that are more consistent with the medical literature than existing methods. These empirical results show that the medical deconfounder can yield insights around medication efficacy and adverse medication reactions.

As venues of future work, the medical deconfounder can be extended to longitudinal settings, which will allow us to accommodate disease progression and estimate time-dependent treatment effects of the medications. We can also conduct sensitivity analyses of the treatment effect estimates on the probabilistic factor model and the outcome model. These analyses will allow us to understand how the modeling choices in the medical deconfounder affect its treatment effect estimates.

Acknowledgments

This work was supported by NIH R01LM006910, NIH U01HG008680, ONR N00014-17-1-2131, ONR N00014-15-1-2209, NIH 1U01MH115727-01, NSF CCF-1740833, DARPA SD2 FA8750-18-C-0130, IBM, 2Sigma, Amazon, NVIDIA, and Simons Foundation.

References

J. G. Abuelo. Treatment of Severe Hyperkalemia: Confronting 4 Fallacies. *Kidney Int Rep*, 3(1): 47–55, Jan 2018.

T. Antoniou, T. Gomes, D. N. Juurlink, M. R. Loutfy, R. H. Glazier, and M. M. Mamdani. Trimethoprim-Sulfamethoxazole-Induced Hyperkalemia in Patients Receiving Inhibitors of the Renin-Angiotensin System: A Population-Based StudyTrimethoprim-Sulfamethoxazole-Induced Hyperkalemia. *Archives of Internal Medicine*, 170(12):1045–1049, 06 2010.

J. Batterink, J. Lin, S. H. Au-Yeung, and T. Cessford. Effectiveness of Sodium Polystyrene Sulfonate for Short-Term Treatment of Hyperkalemia. *Can J Hosp Pharm*, 68(4):296–303, 2015.

I. Bica, A. M. Alaa, and M. van der Schaar. Time Series Deconfounder: Estimating Treatment Effects over Time in the Presence of Hidden Confounders. *arXiv e-prints*, Feb 2019.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent (d)irichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

J. M. Burnell, B. H. Scribner, B. T. Uyeno, and M. F. Villamir. The effect in humans of extracellular pH change on the relationship between serum potassium concentration and intracellular potassium. *J. Clin. Invest.*, 35(9):935–939, Sep 1956.

D. P. Byar, R. M. Simon, W. T. Friedewald, J. J. Schlesselman, D. L. DeMets, J. H. Ellenberg, M. H. Gail, and J. H. Ware. Randomized clinical trials. Perspectives on some recent ideas. *N. Engl. J. Med.*, 295(2):74–80, Jul 1976.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

J. Concato, N. Shah, and R. I. Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, 342(25):1887–1892, 2000.

R. M. Cooper-DeHoff, S. Wen, A. L. Beitelshees, I. Zineh, J. G. Gums, S. T. Turner, Y. Gong, K. Hall, V. Parekh, A. B. Chapman, E. Boerwinkle, and J. A. Johnson. Impact of abdominal obesity on incidence of adverse metabolic effects associated with antihypertensive medications. *Hypertension*, 55(1):61–68, Jan 2010.

A. Deaton and N. Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2 – 21, 2018.

P. Dunning, C. Rantzau, and G. Ward. Effect of alcohol swabbing on capillary blood glucose measurements. *Practical Diabetes International*, 11(6):251–254, 1994.

K. Fukao, K. Shimada, M. Hiki, T. Kiyanagi, K. Hirose, A. Kume, H. Ohsaka, R. Matsumori, T. Kurata, T. Miyazaki, and H. Daida. Effects of calcium channel blockers on glucose tolerance, inflammatory state, and circulating progenitor cells in non-diabetic patients with essential hypertension: a comparative study between azelnidipine and amlodipine on glucose tolerance and endothelial function - a crossover trial (agent). *Cardiovascular Diabetology*, 10(1):79, Sep 2011.

A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, pages 733–760, 1996.

P. Gopalan, J. M. Hofman, and D. M. Blei. Scalable recommendation with hierarchical Poisson factorization. In *Uncertainty in Artificial Intelligence*, 2015.

T. W. Gress, F. J. Nieto, E. Shahar, M. R. Wofford, and F. L. Brancati. Hypertension and antihypertensive therapy as risk factors for type 2 diabetes mellitus. *New England Journal of Medicine*, 342 (13):905–912, 2000.

D. Heckerman. Accounting for hidden common causes when inferring cause and effect from observational data. *arXiv preprint arXiv:1801.00727*, 2018.

M. Hernan and J. Robins. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. 2019.

J. A. Hirst, A. J. Farmer, R. Ali, N. W. Roberts, and R. J. Stevens. Quantifying the effect of metformin treatment and dose on glycemic control. *Diabetes Care*, 35(2):446–454, Feb 2012.

G. Hripcsak, P. B. Ryan, J. D. Duke, N. H. Shah, R. W. Park, V. Huser, M. A. Suchard, M. J. Schuemie, F. J. DeFalco, A. Perotte, J. M. Banda, C. G. Reich, L. M. Schilling, M. E. Matheny, D. Meeker, N. Pratt, and D. Madigan. Characterizing treatment pathways at scale using the ohdsi network. *Proceedings of the National Academy of Sciences*, 113(27):7329–7336, 2016.

S. Hussain, S. Syed, and K. Baloch. Electrolytes imbalance: a rare side effect of piperacillin/ tazobactam therapy. *J Coll Physicians Surg Pak*, 20(6):419–420, Jun 2010.

K. Imai and D. A. Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

G. Imbens and D. Rubin. *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, Nov. 1999.

M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017.

K. K. Koh, M. J. Quon, S. H. Han, Y. Lee, S. J. Kim, and E. K. Shin. Atorvastatin causes insulin resistance and increases ambient glycemia in hypercholesterolemic patients. *J. Am. Coll. Cardiol.*, 55(12):1209–1216, Mar 2010.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

M. Lechner. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labor Market Policies*, pages 43–58. Springer, 2001.

C. H. Lee and G. H. Kim. Electrolyte and Acid-base disturbances induced by clacineurin inhibitors. *Electrolyte Blood Press*, 5(2):126–130, Dec 2007.

M. E. Levine, D. J. Albers, and G. Hripcsak. Methodological variations in lagged regression for detecting physiologic drug effects in EHR data. *J Biomed Inform*, 86:149–159, Oct 2018.

M. J. Lopez, R. Gutman, et al. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, 32(3):432–454, 2017.

C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6449–6459, 2017.

J. J. Mahoney, J. M. Ellison, D. Glaeser, and D. Price. The effect of an instant hand sanitizer on blood glucose monitoring results. *J Diabetes Sci Technol*, 5(6):1444–1448, Nov 2011.

D. F. McCaffrey, B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414, 2013.

B. A. McNicholas, M. H. Pham, K. Carli, C. H. Chen, N. Colobong-Smith, A. E. Anderson, and H. Pham. Treatment of Hyperkalemia With a Low-Dose Insulin Protocol Is Effective and Results in Reduced Hypoglycemia. *Kidney Int Rep*, 3(2):328–336, Mar 2018.

J. Montoliu, X. M. Lens, and L. Revert. Potassium-Lowering Effect of Albuterol for Hyperkalemia in Renal Failure. *Archives of Internal Medicine*, 147(4):713–717, 04 1987.

Y. Mushiyakh, H. Dangaria, S. Qavi, N. Ali, J. Pannone, and D. Tompkins. Treatment and pathogenesis of acute hyperkalemia. *J Community Hosp Intern Med Perspect*, 1(4), 2011.

OHDSI team. Atlas repository. http://www.ohdsi.org/web/atlas, 2019. Accessed: 2019-07-30.

M. Ozery-Flato, P. Thodoroff, and T. El-Hay. Adversarial Balancing for Causal Inference. *arXiv e-prints*, art. arXiv:1810.07406, Oct 2018.

K. H. Polderman and A. R. Girbes. Piperacillin-induced magnesium and potassium loss in intensive care unit patients. *Intensive Care Med*, 28(4):520–522, Apr 2002.

R Core Team. R: A language and environment for statistical computing. 2013.

R. Ranganath and A. Perotte. Multiple causal inference with latent confounding. *arXiv preprint arXiv:1805.08273*, 2018.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

R. Ranganath, L. Tang, L. Charlin, and D. Blei. Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771, 2015.

J. A. Rassen, D. H. Solomon, R. J. Glynn, and S. Schneeweiss. Simultaneously assessing intended and unintended treatment effects of multiple treatment options: a pragmatic "matrix design". *Pharmacoepidemiology and Drug Safety*, 20(7):675–683, 2011.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387212396.

L. B. Rojas and M. B. Gomes. Metformin: an old but still the best treatment for type 2 diabetes. *Diabetol Metab Syndr*, 5(1):6, Feb 2013.

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

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

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

T. Sandozi. Study of effect of amlodipine on blood sugar level. *Asian Journal of Medical Sciences*, 1 (1):4–5, May 2010.

R. W. Sanson-Fisher, B. Bonevski, L. W. Green, and C. D'Este. Limitations of the randomized controlled trial in evaluating population-based health interventions. *American Journal of Preventive Medicine*, 33(2):155 – 161, 2007.

M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In *International Conference on Independent Component Analysis and Signal Separation*, pages 540–547. Springer, 2009.

M. J. Schuemie, G. Hripcsak, P. B. Ryan, D. Madigan, and M. A. Suchard. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences*, 115(11):2571–2577, 2018.

S. I. Sherwani, H. A. Khan, A. Ekhzaimy, A. Masood, and M. K. Sakharkar. Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients. *Biomark Insights*, 11:95–104, 2016.

W. B. Stason, P. J. Cannon, H. O. Heinemann, and J. H. Laragh. Furosemide. A clinical evaluation of its diuretic action. *Circulation*, 34(5):910–920, Nov 1966.

K. Suresh. An overview of randomization techniques: An unbiased assessment of outcome in clinical research. *J Hum Reprod Sci*, 4(1):8–11, Jan 2011.

S. G. Swinnen, J. B. Hoekstra, and J. H. DeVries. Insulin therapy for type 2 diabetes. *Diabetes Care*, 32 Suppl 2:S253–259, Nov 2009.

R. L. Tannen, M. G. Weiner, and D. Xie. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. *British Medical Journal*, 338, 2009.

M. J. Tierney, S. Garg, N. R. Ackerman, S. J. Fermi, J. Kennedy, M. Lopatin, R. O. Potts, and J. A. Tamada. Effect of acetaminophen on the accuracy of glucose measurements obtained with the GlucoWatch biographer. *Diabetes Technol. Ther.*, 2(2):199–207, 2000.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

D. Tran and D. M. Blei. Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*, 2017.

D. Tran, A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, and D. M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.

D. Tran, R. Ranganath, and D. M. Blei. Deep and hierarchical implicit models. *arXiv preprint arXiv:1702.08896*, 2017.

H. Velazquez, M. A. Perazella, F. S. Wright, and D. H. Ellison. Renal Mechanism of Trimethoprim-induced Hyperkalemia. *Annals of Internal Medicine*, 119(4):296–301, 08 1993.

Y. Wang and D. M. Blei. The Blessings of Multiple Causes. *arXiv e-prints*, art. arXiv:1805.06826, May 2018.

S. A. Zaki and V. Lad. Piperacillin-tazobactam-induced hypokalemia and metabolic alkalosis. *Indian J Pharmacol*, 43(5):609–610, Sep 2011.

E. Zanutto, B. Lu, and R. Hornik. Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, 30(1):59–73, 2005.

S. H. Zyoud, R. Awang, S. A. Sulaiman, and S. W. Al-Jabi. Impact of serum acetaminophen concentration on changes in serum potassium, creatinine and urea concentrations among patients with acetaminophen overdose. *Pharmacoepidemiology Drug Safety*, 20(2):203–208, Feb 2011.

Appendix A. Tables of estimated treatment effects from empirical studies

	Unadjusted model			Deconfounder		
	Estimate	2.5% CI	97.5% CI	Estimate	2.5% CI	97.5% CI
Sodium.polystyrene.sulfonat	**-0.48**	**-0.54**	**-0.40**	**-0.47**	**-0.60**	**-0.34**
Regular.Insulin..Human	**-0.27**	**-0.36**	**-0.17**	**-0.27**	**-0.43**	**-0.12**
Piperacillin	**-0.25**	**-0.31**	**-0.19**	**-0.25**	**-0.39**	**-0.11**
Sodium.Bicarbonate	**-0.19**	**-0.25**	**-0.13**	**-0.16**	**-0.32**	**-0.03**
Sodium.Chloride	**-0.11**	**-0.13**	**-0.08**	**-0.12**	**-0.19**	**-0.05**
Nystatin	0.02	-0.06	0.09	-0.10	-0.27	0.08
Albuterol	-0.07	-0.14	0.00	-0.08	-0.22	0.06
Hydromorphone	**-0.11**	**-0.17**	**-0.04**	-0.08	-0.26	0.11
Furosemide	-0.03	-0.08	0.02	-0.05	-0.16	0.07
Esomeprazole	-0.03	-0.11	0.06	-0.04	-0.17	0.09
Glucose	-0.02	-0.07	0.02	-0.04	-0.13	0.05
Ondansetron	0.02	-0.04	0.08	-0.04	-0.17	0.09
Famotidine	0.08	-0.01	0.16	-0.02	-0.21	0.17
Acetaminophen	**0.05**	**0.00**	**0.11**	-0.01	-0.09	0.07
Aluminum.Hydroxide	0.05	-0.02	0.12	-0.01	-0.20	0.16
Diphenhydramine	0.03	-0.05	0.11	-0.01	-0.20	0.18
Glucagon	0.01	-0.05	0.06	-0.01	-0.16	0.14
Docusate	0.01	-0.04	0.06	0.00	-0.10	0.12
Insulin.Glargine	0.07	-0.04	0.19	0.00	-0.18	0.21
Aspirin	-0.02	-0.08	0.04	0.02	-0.09	0.12
atorvastatin	0.01	-0.09	0.12	0.02	-0.15	0.20
heparin	0.02	-0.04	0.08	0.02	-0.08	0.12
Amlodipine	0.02	-0.05	0.09	0.03	-0.13	0.18
Lorazepam	0.00	-0.08	0.09	0.03	-0.16	0.22
Morphine	-0.01	-0.10	0.07	0.03	-0.13	0.18
Metoprolol	**0.10**	**0.04**	**0.17**	0.04	-0.10	0.17
Tacrolimus	**0.11**	**0.03**	**0.19**	0.04	-0.10	0.19
Enoxaparin	0.07	-0.03	0.16	0.05	-0.13	0.24
carvedilol	**0.12**	**0.03**	**0.21**	0.05	-0.17	0.25
Oxycodone	**0.13**	**0.06**	**0.21**	0.08	-0.04	0.21
Prednisone	**0.13**	**0.06**	**0.20**	0.09	-0.06	0.25
Sulfamethoxazole	**0.15**	**0.09**	**0.22**	0.11	-0.07	0.29
Potassium.Chloride	**0.55**	**0.48**	**0.62**	**0.53**	**0.43**	**0.64**

Table A1: Treatment effects of medications in the potassium disorder cohort estimated by the unadjusted model and the medical deconfounder. The mean, lower and upper bound of 95% credible interval of the estimated coefficients are included. Causal medications found by each model are in bold; their 95% credible intervals exclude zero. A positive coefficient means that the medication increases serum potassium and vice versa.

	Unadjusted model			Deconfounder		
	Estimate	2.5% CI	97.5% CI	Estimate	2.5% CI	97.5% CI
Insulin.Glargine	**-0.53**	**-0.68**	**-0.37**	**-0.46**	**-0.61**	**-0.32**
Isopropyl.Alcohol	**-0.44**	**-0.60**	**-0.27**	**-0.42**	**-0.61**	**-0.21**
Aluminum.Hydroxide	**-0.26**	**-0.42**	**-0.08**	**-0.25**	**-0.47**	**-0.02**
Metformin	**-0.29**	**-0.36**	**-0.22**	**-0.21**	**-0.31**	**-0.12**
Ondansetron	-0.17	-0.38	0.00	-0.18	-0.39	0.04
Oxycodone	**-0.17**	**-0.31**	**-0.03**	**-0.17**	**-0.33**	**-0.01**
Potassium.Chloride	-0.11	-0.25	0.04	-0.09	-0.25	0.06
atorvastatin	**-0.13**	**-0.23**	**-0.03**	-0.09	-0.24	0.05
Lisinopril	-0.09	-0.24	0.05	-0.02	-0.16	0.12
Enoxaparin	-0.05	-0.23	0.13	-0.01	-0.20	0.18
heparin	-0.04	-0.21	0.11	-0.01	-0.17	0.14
Prednisone	0.06	-0.16	0.26	0.00	-0.18	0.19
Sodium.Chloride	-0.07	-0.17	0.03	0.00	-0.13	0.13
Aspirin	-0.01	-0.12	0.12	0.04	-0.06	0.14
Furosemide	0.04	-0.09	0.16	0.05	-0.10	0.18
Simvastatin	0.00	-0.15	0.15	0.05	-0.14	0.21
Esomeprazole	0.02	-0.16	0.19	0.07	-0.10	0.24
Losartan	0.09	-0.11	0.28	0.07	-0.13	0.27
Metoprolol	0.04	-0.09	0.18	0.07	-0.06	0.20
carvedilol	0.07	-0.18	0.32	0.07	-0.11	0.30
Acetaminophen	**0.08**	**0.01**	**0.16**	0.09	-0.01	0.19
Glucagon	0.02	-0.08	0.13	0.09	-0.03	0.20
Albuterol	0.01	-0.20	0.21	0.10	-0.10	0.30
Docusate	0.07	-0.11	0.24	0.12	-0.01	0.25
Omeprazole	0.08	-0.16	0.33	0.13	-0.05	0.33
Amlodipine	0.07	-0.07	0.21	**0.16**	**0.02**	**0.29**
clopidogrel	0.06	-0.15	0.28	0.16	-0.04	0.37
Hydrochlorothiazide	0.15	-0.01	0.30	**0.17**	**0.01**	**0.33**
Cholecalciferol	0.12	-0.06	0.29	0.18	-0.05	0.40
Tacrolimus	**0.27**	**0.07**	**0.47**	**0.32**	**0.07**	**0.55**

Table A2: Treatment effects of medications in the diabetes cohort estimated by the unadjusted model and the medical deconfounder. The mean, lower and upper bound of 95% credible interval of the estimated coefficients are included. Causal medications found by each model are in bold; their 95% credible intervals exclude zero. A negative coefficient means that the medication decreases HbA1c and vice versa.