Proceedings of Machine Learning Research vol 275:1347–1364, 2025 4th Conference on Causal Learning and Reasoning

Temporal Inverse Probability Weighting for Causal Discovery in Controlled Before–After Studies: Discovering ADEs in Generics

Aubrey Barnard BARNARD@CS.WISC.EDU
Obstetrics & Gynecology, University of Wisconsin–Madison, Madison, WI, USA

Peggy Peissig PEISSIG.PEGGY@GMAIL.COM
Marshfield Clinic Research Institute, Marshfield, WI, USA

David Page DAVID.PAGE@DUKE.EDU
Biostatistics & Bioinformatics, Computer Science, Duke University, Durham, NC, USA

Editors: Biwei Huang and Mathias Drton

Abstract

Adverse drug events (ADEs) cost society lives and an estimated $30 billion per year in the USA alone. Their prevalence has led to the public losing trust in the safety of drugs, especially generics (e.g., Eban, 2019). These concerns have motivated the wide study of methods for general ADE discovery, but discovering ADEs in generic drugs challenges causal discovery methods with a scenario of multiple treatments over time, a scenario which presents new problems and opportunities for machine learning. In response, this research develops methods for causal discovery based on analyzing controlled before–after studies with differential prediction and temporal inverse probability weighting. These methods are easy to realize by employing off-the-shelf machine learning classifiers. Experiments on both synthetic and real electronic health records demonstrate the ability of the methods to control for confounding, discover generic-specific ADEs in synthetic data, and hypothesize brand–generic differences in real-world data that agree with known ones. These are the abilities that causal discovery methods need for helping establish the facts of generic drug safety.

Keywords: causal discovery, inverse probability weighting, controlled before–after studies, time-varying confounding, adverse drug events, generic drugs

1. Introduction

Due to our intuition that reasoning about the world fundamentally relies on understanding causality, causal discovery has been a technical research area within artificial intelligence for a long time (e.g., Pearl, 1988) and continues to draw substantial attention, particularly in applications to healthcare. In 2008, the FDA's Sentinel Initiative (US FDA, 2008) helped direct this attention towards adverse drug events (ADEs) in response to their high societal impact: ADEs cost many lives and an estimated $30 billion per year in the USA alone (Sultana et al., 2013). The FDA initiated a series of programs for computational postmarketing surveillance ("pharmacovigilance" or "pharmacosurveillance") that spurred the development of many methods for ADE discovery. Most of those methods targeted ADEs in general, but generic drugs raise unique considerations, such as subtle efficacy differences, patient choice, and time-varying confounding, which motivate the development of methods that address the special challenges and opportunities of pharmacosurveillance of generic drugs.

Were a generic to have unique effects, one challenge is not knowing what those effects might be before the generic enters the market. For approval, manufacturers must certify that a generic has the same amounts of the same active ingredients and demonstrate bioequivalence through studies in vivo, among other requirements (US FDA, 2017). This similarity actually means that any effect

specific to a generic is unlikely to have been suspected based on evidence from the brand, although clinical trials for the brand often hint at issues even if they do not have the power to confirm them. While generics do not need to undergo clinical trials, they need to undergo bioequivalence studies. These studies may surface similar hints of issues, but they are much more limited than clinical trials, often testing on only a few tens of young, healthy individuals (Lewek and Kardas, 2010). This makes them unlikely to discover differences in patient outcomes that are subtle or involve complex medical contexts, leaving possible ADEs underexplored. Existing methods for ADE detection, such as those methods studied by OMOP or its successor OHDSI (§2.1), assume the task is to match drugs with a finite set of predefined ADEs such as kidney injury, liver failure, or myocardial infarction (heart attack). Given the unknown nature of possible generic-specific effects, such existing methods for ADE detection are not appropriate for the crucial step of *hypothesizing* ADEs. This work proposes a general machine learning approach that can hypothesize ADEs rather than requiring possible ADEs to be predefined.

Another challenge of ADE discovery in generics is the large difference in time between when the brand version debuts and when the generic version debuts, at which time it often happens that health insurance providers will require that patients switch from brand to generic. Together, these circumstances mean that any observational study of brand versus generic versions of a drug will face study groups that are exclusive in both time and treatment, making the groups less comparable than in a typical observational study. However, a key characteristic of ADE discovery in generics is that *patients are on the generic version for the same reasons they are on the brand version*. This effectively matches on risk factors and indications which helps make the groups more comparable again. In many cases, the patients are even the same, having switched from brand to generic. Through self-controlled studies, ADE discovery in generics offers an opportunity to reduce the especially difficult problem of *unobserved* confounders, confounders that are not included in the data and also may not be included as latent variables in any models. Nevertheless, the large time gap remains a difficulty because of the potential for *temporal* or *time-varying* confounders. For example, when generic gabapentin became available in 2005, the healthcare system studied here switched patients from brand to generic, but also switched to electronic prescriptions around the same time. Thus, the feature "prescription transmitted electronically" is the best discriminator between brand and generic when the study does not control for changes over time. Indeed, we have repeatedly encountered similar situations: if one doesn't control for temporal differences, all of the most important features are just proxies for time passing. Responding to these characteristics, this work proposes an approach specifically designed to take advantage of the similarity between study groups and control for temporal confounding.

With the unique challenges and opportunities of generic drug ADE discovery in mind, this work proposes a new approach to causal discovery from observational data that analyzes a controlled before–after study with general machine learning classifiers and temporal inverse probability weighting. The study design takes advantage of the brand versus generic setting where (1) the treated groups have similarities, like sharing risk factors and indications, or involving the same patients at different times, (2) the treatments are sequential, making temporal confounding a problem, and (3) all of the possible effects of treatment are not precisely defined nor even suspected before the analysis. While born of the brand versus generic setting, the proposed study design and analysis apply to any comparison of two treatments separated in time, but evaluating its performance beyond generic drug pharmacosurveillance is left to future work. Within this scope of evaluation, the proposed approach is found to be more accurate at identifying the true generic-specific ADEs

in synthetic data than differential prediction, and, when analyzing real EHR data, it hypothesizes differences between brand and generic that agree with known differences without false discovery.[1]

2. Adverse Drug Event Discovery

2.1. Adverse Drug Events

ADEs are estimated to account for up to 30% of hospital admissions and at least $30 billion in annual healthcare costs in the USA (Sultana et al., 2013). Although the U.S. Food and Drug Administration (FDA) and its counterparts elsewhere have preapproval processes for drugs that are rigorous and involve randomized controlled clinical trials, such processes cannot possibly uncover everything about a drug. While a clinical trial might use only a thousand patients, once a drug is released on the market it may be taken by millions of patients (Stang et al., 2010). As a result, additional risks often come to light after a drug is released on the market to a larger, more diverse population.

While generic drugs are expected to act the same as brand drugs in general,[2] and studies generally show equivalence (e.g., Desai et al., 2019; Kharasch et al., 2019), some of these additional risks might be specific to generic drugs. Rightly or wrongly, concerns have been raised because generic drugs may have differences in inactive ingredients, pharmacokinetic profiles, or manufacturing processes, so differences in safety or efficacy could theoretically occur. Leclerc et al. (2017) claimed evidence for differences in ADE profiles of brand versus generic ACE inhibitors, and the FDA found differences in efficacy of brand versus generic versions of both methylphenidate and bupropion (US FDA, 2016, 2012).

Due to the risks of ADEs to patient safety, the FDA and other USA government agencies made pharmacovigilance a high national research priority. In response, the FDA, National Institutes of Health, and PhARMA formed the Observational Medical Outcomes Partnership (OMOP) (Stang et al., 2010) to develop and compare methods for ADE detection, work that continues under its successor, the Observational Health Data Sciences and Informatics (OHDSI) program (Hripcsak et al., 2015). Their contributions include a benchmark ADE identification task, standardized data models, and tools for computational epidemiology.

2.2. Existing Methods for ADE Discovery in EHRs

Causal discovery has been studied for years within artificial intelligence (e.g., Pearl, 1988) and statistics (e.g., Good, 1961), but has only more recently been applied to ADE discovery. OMOP evaluated the ability of various methods to rediscover known ADEs from data in EHR and insurance claims databases (Madigan and Ryan, 2011). One such method, disproportionality analysis (Zorych et al., 2011), constructs a 2×2 table of the treatment and response, and asks if a measure of association, such as relative risk or odds ratio, is higher than would be expected by chance. This exemplifies the prototypical setup of many observational studies, which tend to have trouble controlling for confounding. Another approach, multiple self-controlled case series (MSCCS) (Simpson et al., 2013), handles confounding better by using a self-controlled study design and estimating

1. Over-flagging, or low precision, is a major risk to any approach, especially given already lowered public trust (Rahman et al., 2017).
2. The FDA requires that manufacturing processes for a generic consistently produce the correct drug, which must be pharmaceutically equivalent to the brand (US FDA, 2017), but compliance with these regulations can be hard to enforce, especially when a manufacturer is not in the USA.

patient-specific baseline risks. Subsequent methods have extended this idea by modeling risks that vary over time for a single patient (Kuang et al., 2017), or by combining patient-specific baselines with probabilistic graphical model learning (Geng et al., 2018).

The foundation of modern causal inference in both randomized and nonrandomized studies is the Rubin causal model (Rubin, 1974), which compares the outcomes (responses) in treated and control groups. In nonrandomized (observational) studies, confounders may obscure the treatment effect, but one way to lessen confounding is to balance the study groups on their propensity for treatment (Rosenbaum and Rubin, 1983), perhaps by inverse probability weighting (IPW) (Robins et al., 1994), before estimating the treatment effect. Differential prediction[3] (Linn, 1978; Radcliffe and Surry, 1999) extends the approach of the Rubin causal model by building models of response in each of the treated and control groups and then comparing those models. While differential prediction was developed for marketing and standardized testing, it has been used for causal inference (Gutierrez and Gérardy, 2017), for example by Robins (1994), Vansteelandt and Goetghebeur (2003), and Nassif et al. (2012).

The above methods estimate the causal relationship between only two variables at a time, a treatment and a response. By contrast, structural causal modeling (Spirtes et al., 2000; Pearl, 2009) estimates all of the direct causal relationships among a set of variables at once. In this framework, a structural equation model or causal Bayesian network represents the causal system (the "laws of nature"), and the goal is to learn the model or just its parameters from data (e.g., Loh and Wainwright, 2013; Barnard and Page, 2018). Thereafter, the model can be queried about the effects of interventions or counterfactual situations.

All of the work reviewed so far assumes that possible ADEs have been identified and precisely defined before the analysis, which does not apply to *de novo* ADE discovery where possible effects need to be hypothesized. Page et al. (2012) proposed a method for hypothesizing ADEs that finds logical clauses to distinguish between cases (on the drug) and controls based on events after starting the drug. While this hypothesizes many events, the study design did not adequately control for confounding.

3. Methods for Finding Differential Effects of Two Treatments

This work addresses the following novel task, generic adverse drug event (ADE) discovery:

> Given a database of clinical records, discover effects caused by taking the generic version of a drug that are different than the effects caused by the brand version.

To help make this task tractable, it is assumed that (1) an effect can be represented by some combination of features available in the data and (2) any effect worth discovering occurs frequently enough to be distinguishable from noise given the number of patients on the brand and generic versions of a drug. Nevertheless, this task poses two major challenges: hypothesizing effects that are causally reasonable (Hill, 1965) and controlling for confounding.

To address these challenges, this work proposes an approach, *causal discovery machine learning*, that analyzes controlled observational studies with machine learning methods. While ML methods do not normally produce models that are causally "reasonable," by combining them with appro-

3. Differential prediction is also known as uplift modeling, difference in differences, or structural mean models.

priate study designs and standard causal assumptions[4] they become instruments of causal inference. This combination produces a general approach to causal discovery that applies equally well to any two treatments over time as it does to brand and generic versions of a drug.

3.1. Hypothesizing Effects Using Causal Discovery Machine Learning

Since the possible effects of a generic drug are unknown, the first challenge is hypothesizing them. Rather than modeling $\mathbb{P}(E \mid X)$ for a known effect (ADE) E based on covariates X, the methods proposed herein hypothesize effects by modeling $\mathbb{P}(T \mid X)$, using data after the start of each drug. This ensures that any important features in the model occur after the drug and so are possible effects of the drug. Then, having generic-takers ($T = g$) be positive examples and brand-takers ($T = b$) be negative examples focuses the possible effects on differences between brand and generic, which can be inspected after fitting. While this setup is similar to that of Page et al. (2012), it uses arbitrary classifiers and a study design that controls for temporal confounding.

3.2. Reducing Confounding Using Self-Controlled Studies

Since causal discovery from observational data faces the possibility of confounding, the second challenge posed by generic ADE discovery is reducing confounding, especially temporal confounding. The proposed approach tackles this in three ways: by taking advantage of the similarities between brand and generic, by setting up a self-controlled study, and by employing temporally-matched control groups. First, patients taking brand or generic versions of a drug are taking it for the same reasons, the same indications, thus controlling for confounding by indication. Nevertheless, other variables could be confounders, especially when hypothesizing effects, which can pick up on any differences between treatments, spurious or real. To address this, note that many patients switch from brand to generic, which means they can serve as their own controls. So, second, these patients are enrolled in both treatment groups, making the study self-controlled. The self-control via switchers and the similarities in patient histories between brand- and generic-takers are both mechanisms that serve to match on observed and even unobserved variables, thereby controlling for confounding by those variables. However, self-control cannot help with changes over time, so, third, the proposed approach employs temporally-matched control groups to combat temporal (time-varying) confounding. It is especially important to control for temporal differences because otherwise the hypothesized effects tend to be just proxies for time passing.

3.3. Controlled Before–After Studies

While a typical study compares a treated group with a control group, the setting of brand versus generic needs something different because it has three study groups, two treatments over time and controls. Putting these two treatments together with the three control mechanisms described above results in the study design in Figure 1, a type of controlled before–after study (Shadish et al., 2002). It contains two treatments T, before (B) and after (A) a threshold in time t, which can be chosen globally or per unit (patient). Each treated unit has a control unit that corresponds in time. The treated groups establish the effects and the temporally-matched control groups provide a baseline for comparison and reduce confounding.

4. The standard assumptions for causal inference are exchangeability, positivity, consistency, and models that are suitably expressive (Hernán and Robins, 2020).

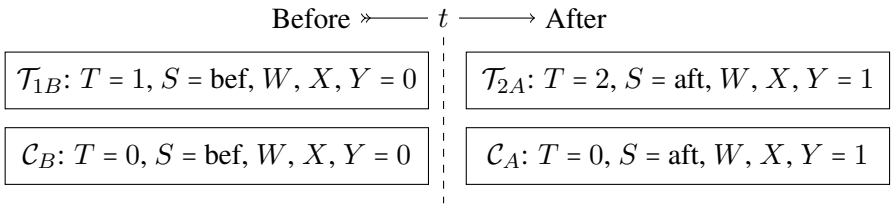


Figure 1: Controlled before–after study for two treatments T, with time spans S, unit weight W, covariates X, and outcome Y. The outcome is only for illustrating existing methods.

To create the treated and control groups, patients were enrolled through three enrollment scenarios in order to accommodate enrolling as many brand- and generic-takers as possible, increasing the power and robustness of the study. The three scenarios were: (1) brand to generic switchers ($\mathcal{T}_{1B} \to \mathcal{T}_{2A}$) were matched with generic to brand switchers ($\mathcal{T}_{2B} \to \mathcal{T}_{1A}$) as in a crossover design, (2) leftover switchers were matched with never-takers ($\mathcal{C}_B \to \mathcal{C}_A$), and (3) brand-only-takers were matched with generic-only-takers and served as each other's controls: $(\mathcal{T}_{1B} \to \mathcal{C}_A) \veebar (\mathcal{C}_B \to \mathcal{T}_{2A})$.[5] Accordingly, the time threshold was chosen per matched pair. Matching attributes included demographics, the date when the drug was started, and measures of interaction with the health system.

How can one analyze such a study with all of its treatments and controls? Let f be some outcome measure of each group. Then the difference between treated groups $f(\mathcal{T}_{2A}) - f(\mathcal{T}_{1B})$ is the effect, the difference between control groups $f(\mathcal{C}_A) - f(\mathcal{C}_B)$ models the changes over time, and the difference in differences is the temporally-adjusted effect,

$$(f(\mathcal{T}_{2A}) - f(\mathcal{T}_{1B})) - (f(\mathcal{C}_A) - f(\mathcal{C}_B)). \tag{1}$$

The typical analysis approach in fields such as statistics or epidemiology would be to estimate Equation 1—for example, with a regression model—but we desire an approach that will work in general with many machine learning models, so we treat it as a binary classification task by taking the signs of the terms as the class labels: $+\mathcal{T}_{2A}$, $-\mathcal{T}_{1B}$, $-\mathcal{C}_A$, $+\mathcal{C}_B$. This design controls for temporal and other differences because each classification group includes both before and after units, and both treated and control units. Furthermore, by setting up analyses to discover differences between these groups based on data after treatment, this design adapts the analyses to hypothesize effects and do causal discovery machine learning. (See §B for an example.)

3.4. Analysis Methods

The observational studies herein were constructed according to the study design in Figure 1 and then analyzed with classification, differential prediction, and a method developed here, *temporal* inverse probability weighting (IPW), all listed in Table 1. The classification method applied binary classifiers directly to the positives and negatives from the study design (Equation 1), which turns out to already be a form of differential prediction, accomplished on regular data by flipping the labels of the control groups (Jaśkowski and Jaroszewicz, 2012). It will be called differential classification.

The second analysis method was differential prediction using SVMs (with linear kernels) modified to maximize uplift (Kuusisto et al., 2014). (Uplift is a measure of differences between groups

5. Throughout, "$\veebar$" means "versus" and indicates comparing or contrasting groups.

Method	Model	Before–After Study Classification Setup
DFP, 2-model	$\mathbb{P}(Y = 1 \mid T = 1, X) \underline{\vee} \mathbb{P}(Y = 1 \mid T = 0, X)$	$\mathcal{M}_{\text{DFP2}}(\mathcal{M}_{\mathcal{C}}(-\mathcal{C}_B \underline{\vee} +\mathcal{C}_A) \underline{\vee} \mathcal{M}_{\mathcal{T}}(-\mathcal{T}_{1B} \underline{\vee} +\mathcal{T}_{2A}))$
DFP, 1-model	$\mathbb{P}(Z \mid X)$ where $Z = (T = 0)(Y = 0) + (T = 1)(Y = 1)$	$\mathcal{M}_{\text{DFP1}}(-\mathcal{C}_A \underline{\vee} -\mathcal{T}_{1B} \underline{\vee} +\mathcal{C}_B \underline{\vee} +\mathcal{T}_{2A})$
DFC	$\mathbb{P}(Z \mid X, Y)$ where $Z = (T = 0)(S = b) + (T = 1)(S = a)$	$\mathcal{M}_{\text{DFC}}((-\mathcal{C}_A \cup -\mathcal{T}_{1B}) \underline{\vee} (+\mathcal{C}_B \cup +\mathcal{T}_{2A}))$
temporal IPW	$\mathbb{P}(S \mid T = 0, W, X) \overset{\text{IPW}}{\rightarrow} \mathbb{P}(T \mid T \neq 0, W', X)$	$\mathcal{M}_{\text{TIPW}}(\mathcal{M}_{\mathcal{C}}(-\mathcal{C}_B \underline{\vee} +\mathcal{C}_A) \overset{\text{IPW}}{\rightarrow} \mathcal{M}_{\mathcal{T}}(-\mathcal{T}'_{1B} \underline{\vee} +\mathcal{T}'_{2A}))$

Table 1: Analysis methods. DFP: differential prediction, DFC: differential classification, $\underline{\vee}$: versus.

analogous to Equation 1.) Differential prediction seeks to predict whether units will respond to treatment. It builds a model of response to treatment, builds a separate model of response despite no treatment, and then compares the two models, although some methods model the difference in responses with a single, combined model (Table 1). Because the setting of brand versus generic has two treatments and an unknown response, the standard differential prediction setting, $\mathbb{P}(Y \mid T, X)$, must be adapted, which this work does by modeling $\mathbb{P}(S \mid T, X)$ instead. This reuses the same study groups, but sets them up for causal discovery machine learning (allowing hypothesizing responses).

3.4.1. TEMPORAL INVERSE PROBABILITY WEIGHTING

In addition to the differential methods above, the studies were analyzed using a new method that builds on inverse probability weighting (IPW) (Rosenbaum and Rubin, 1983; Imbens and Rubin, 2015), adapting it to the temporal setting of controlled before–after studies. Whereas typical IPW corrects for a patient's propensity for treatment, making the treated and control groups more comparable, the proposed method uses IPW to remove temporal trends in the data, making the before and after periods more comparable. This is sufficient to address all sources of confounding because temporal changes are the only confounders left after using a patient as their own control, which elegantly controls for everything else, even unmeasured confounders. As a result, compared to the standard assumption of no unmeasured confounders, temporal IPW needs only a weaker assumption of no unmeasured temporal confounders.

Temporal IPW for controlled before–after studies works as follows. First, a model of temporal trends is built by training a classifier to classify control units as before or after. Next, that classifier predicts before (treatment 1) or after (treatment 2) for each of the treated units. The units that the classifier predicts correctly exhibit similar temporal trends to those that exist in the controls. The units that the classifier predicts incorrectly cannot be distinguished based on temporal trends, so their distinguishing characteristics have to do with the treatments (which are the only other differences except those due to confounding, for which the study design controls). Then, each treated unit is reweighted by the inverse of the probability that the model assigns to its correct label. This downweights units that exhibit mainly temporal trends and upweights units that do not, thereby controlling for temporal trends and focusing on differences between treatments. Finally, a second classifier is trained on the reweighted treated units to discover the differences between them. This process is detailed in Algorithm 1.

Algorithm 1: Temporal IPW for analyzing before–after studies.

Input: sets of data $\mathcal{C}_B, \mathcal{C}_A, \mathcal{T}_{1B}, \mathcal{T}_{2A}$ from a controlled before–after study
Output: model $\mathcal{M}_\mathcal{T}$ that discriminates treatments while controlling for changes over time

$$\texttt{ipw}(m, t, x) = 1/\mathbb{P}_{\mathcal{M}=m}(T = t \mid X = x) \tag{2}$$

$\mathcal{M}_\mathcal{C} \leftarrow \texttt{fit}(-\mathcal{C}_B \veebar +\mathcal{C}_A)$; `// Fit model of temporal trends`
for $\mathcal{T}$ **in** $[\mathcal{T}_{1B}, \mathcal{T}_{2A}]$ **do** `// Remove trends by reweighting treateds`
  $\mathcal{T}' \leftarrow [(t, s, x, w \cdot \texttt{ipw}(\mathcal{M}_\mathcal{C}, t, x))$ **for** $(T = t, S = s, X = x, W = w)$ **in** $\mathcal{T}]$;
$\mathcal{M}_\mathcal{T} \leftarrow \texttt{fit}(-\mathcal{T}'_{1B} \veebar +\mathcal{T}'_{2A})$; `// Fit model of treatments`

The temporal IPW algorithm can be understood as searching for an event E that maximizes the relative risk between two treatments T, where E is some function f of the data X.

$$\frac{\mathbb{P}(E \mid T = 2)}{\mathbb{P}(E \mid T = 1)} = \frac{\mathbb{P}(T = 2 \mid E)\mathbb{P}(E)/\mathbb{P}(T = 2)}{\mathbb{P}(T = 1 \mid E)\mathbb{P}(E)/\mathbb{P}(T = 1)} = \frac{\mathbb{P}(T = 2 \mid E)}{\mathbb{P}(T = 1 \mid E)} \frac{\mathbb{P}(T = 1)}{\mathbb{P}(T = 2)} \tag{3}$$

Expanding the relative risk on the left in terms of Bayes rule and simplifying leads to the two terms on the right of Equation 3. The first term corresponds to the classification objective for $\mathcal{M}_\mathcal{T}$: learn a function $E = f(X)$ that discriminates between treatments. The second term is the temporal inverse probability weighting, because $T = 1$ corresponds to before ($S = b$) and $T = 2$ to after ($S = a$) in the before–after study design:

$$\frac{\mathbb{P}(T = 1)}{\mathbb{P}(T = 2)} = \frac{\mathbb{P}(S = b)}{\mathbb{P}(S = a)} = 1 \Big/ \frac{\mathbb{P}(S = a)}{\mathbb{P}(S = b)} \,. \tag{4}$$

The term on the right of Equation 4 is equivalent to Equation 2 because $\mathbb{P}(S = a)/\mathbb{P}(S = b)$ corresponds to the classification objective for $\mathcal{M}_\mathcal{C}$, which discriminates between before and after. These correspondences show how learning $\mathcal{M}_\mathcal{C}$, reweighting, and learning $\mathcal{M}_\mathcal{T}$ effectively finds a model of an event E whose relative risk between treatments is maximized.

Compared to the other methods, temporal IPW is set up to better learn the differences between treatments because it has an optimization objective that cannot be gamed and has greater statistical efficiency. Specifically, differential prediction can appear to do well by gaming its objective: by learning a model that has an especially *low* accuracy on the *controls* rather than an especially *high* accuracy on the *treateds*. While differential classification does not have this weakness, its statistical efficiency is less than that of temporal IPW because fully half of its training examples (the controls) are not actually relevant to characterizing the true target, $(T = 1) \veebar (T = 2)$. The result is that differential classification must do better at classifying the controls, $+\mathcal{C}_B \veebar -\mathcal{C}_A$, than temporal IPW does at classifying the treateds, $-\mathcal{T}_{1B} \veebar +\mathcal{T}_{2A}$, for differential classification to have a higher overall true positive rate than temporal IPW, despite that we expect there to be relatively fewer differences in the controls than in the treateds. Proposition 1 formalizes this notion. (§A contains the proof.)

Proposition 1 (TIPW dominates DFC) *Suppose temporal IPW has true positive rate TPR_{TIPW} and differential classification has TPR_{DFC}, but that they share the same TPR $\alpha_\mathcal{T}$ on the treateds, $-\mathcal{T}_{1B} \veebar +\mathcal{T}_{2A}$. Let DFC have TPR $\beta_\mathcal{C}$ on the controls, $+\mathcal{C}_B \veebar -\mathcal{C}_A$. Then,*

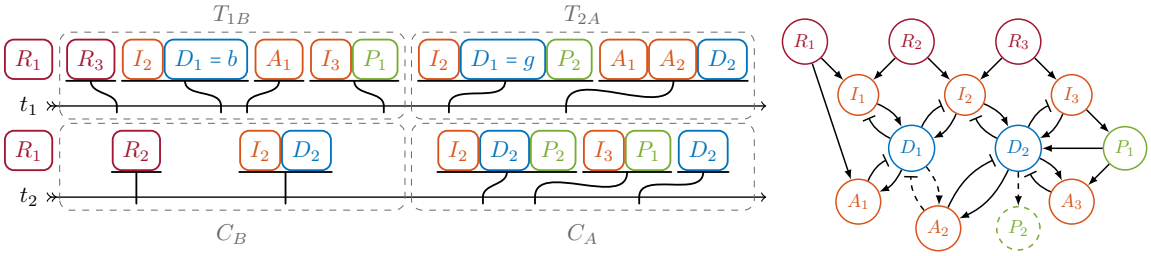
$$\alpha_\mathcal{T} \geq \beta_\mathcal{C} \iff TPR_{TIPW} \geq TPR_{DFC}. \tag{5}$$

4. Experiments and Results

The methods from §3 were first evaluated on synthetic data where the ground truth ADEs specific to the generic version of an artificial drug were known. Then the same methods were applied to actual EHR data, using real-world brand–generic drug pairs that have had widespread use.

4.1. Electronic Health Records Data

The data used in the experiments came from electronic health records (EHR) databases. Typical EHR data is kept in a relational database and consists of multiple tables for information like demographics, diagnoses, drugs, procedures, measurements such as lab tests and vitals, etc. Each row in a table can be considered an event if it has a timestamp (e.g., D_1 in Figure 2); otherwise it can be considered a fact (e.g., R_1). Viewed from the perspective of a single patient, all the facts and events pertaining to that patient form a sequence of events that is that patient's history or timeline, as shown in Figure 2(*a*). All of the data was analyzed in the form of feature vectors extracted from patient histories: the relevant study period was selected from the patient's history, the events during the period were counted, and the counts formed a feature vector along with demographic facts.

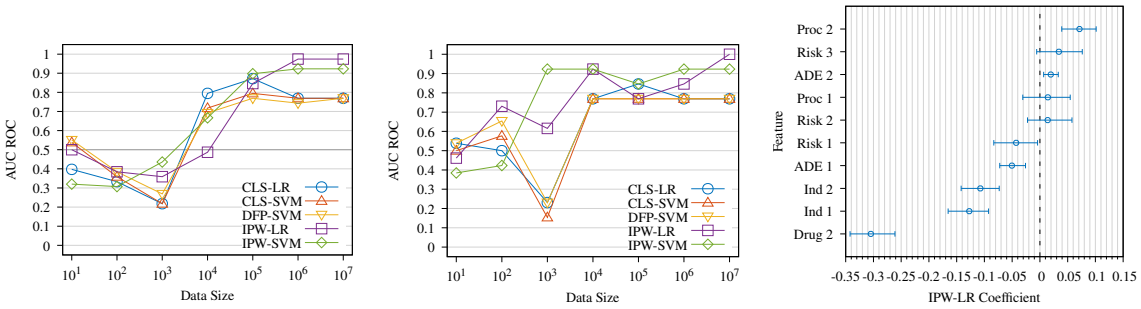


(*a*) Example timelines in electronic health records (*b*) CTBN for EHR timelines

Figure 2: Example EHR timelines and CTBN for generating them. (R)isk factor, (I)ndication, (D)rug, (P)rocedure, (A)DE. P_2 is introduced at the same time as generic D_1, midway through the timelines. Generic D_1 causes A_2 whereas brand does not. The dashed lines indicate these temporal differences. Perpendicular arrowheads ⊣ mark inhibitors.

4.2. Experiments on Synthetic Data

The synthetic data was generated by a continuous time Bayesian network (CTBN) (Nodelman et al., 2002). A network was designed with representative structure that involved risk factors, indications, drugs, procedures, and adverse drug events (Figure 2(*b*)). The temporal differences were the availability of the generic version of drug D_1 and a distracter, the introduction of procedure P_2. These were introduced midway through the samples and are indicated by dashed lines in Figure 2(*b*). The difference between brand and generic was an extra ADE: both brand and generic D_1 caused A_1, but only generic D_1 caused A_2. To make the synthetic data realistically difficult, D_1 caused A_2 with an incidence of 5.5 occurrences per 100 patients per year, which agrees with the literature (e.g., Gurwitz et al., 2003). Samples were drawn from the CTBN to produce a data set with 10M (10^7) patients, of which 1.8M ended up being cases. In order to create learning curves, subsets were formed by taking the first n patients, where n ranged from 10^1 to 10^7. Within each subset, cases



(*a*) Default (*b*) Best ADE Discovery (*c*) Top 10 IPW-LR Features

Figure 3: Results of experiments on the synthetic data: learning curves and top features from "default" IPW-LR on data size 10^5. Positive coefficients favor generic; negative coefficients favor brand. CLS: differential classification, DFP: differential prediction, IPW: temporal IPW, LR: logistic regression, SVM: support vector machine with a linear kernel.

were matched 1-to-1 with controls and enrolled in a before–after study as explained in §3.3, with D_1 = brand and D_1 = generic being the two treatments.

To test the methods, each one was applied to the study built from each data subset and then evaluated by how well it identified the true generic-specific ADE A_2, as measured by the AUC ROC of ranking all the features by their scores from the model, where the feature scores were just the coefficients of those features. While any binary classifier could work as a model, ones that could produce suitable feature scores were chosen, namely logistic regression (LR) and support vector machines (SVMs). (Here, suitable scores are those that are able to separate features into positive effects, favoring generic, and negative effects, favoring brand, all while ranking ADEs above other events.) Logistic regression was chosen because it is commonly used in causal inference, and support vector machines with linear kernels were chosen because they were also used for differential prediction. Other kinds of classifiers (naïve Bayes, tree models, SVMs with other kernels, neural networks) were tried, and they either did not classify well or had issues ranking features, such as needing some unknown function to successfully translate model parameters into suitable feature scores. For tuning, the standard notions do not directly apply, because the training task (distinguishing between brand-takers and generic-takers in a before–after study encoded as a binary classification task) is different from the evaluation task (discovering generic-specific ADEs). Thus, experiments were run with two ways of picking the regularization strength hyperparameters: default (SVM: $C = 1$, LR: $\lambda = 1$), and best ADE discovery (averaged over data size), as if an oracle provided the best-performing hyperparameters.

Figures 3(*a*)–3(*b*) show the results of these experiments. One can see that IPW-LR eventually distinguishes the ADE from the temporal distracter and confounders given enough data, whereas the differential classification and prediction methods plateau without being able to find the ADE. IPW-SVM promises to do better than IPW-LR but then also plateaus. After inspecting the results, we think this is because the SVM is *too expressive* and so is able to somewhat undo the IPW, allowing confounding to creep back in. Figure 3(*c*) shows the 10 features with the largest mean coefficient magnitudes and their 99% confidence intervals from applying IPW-LR with default hyperparameters to 1000 bootstrap samples from the 10^5 data (17.7k cases). One can see that the temporal distracter

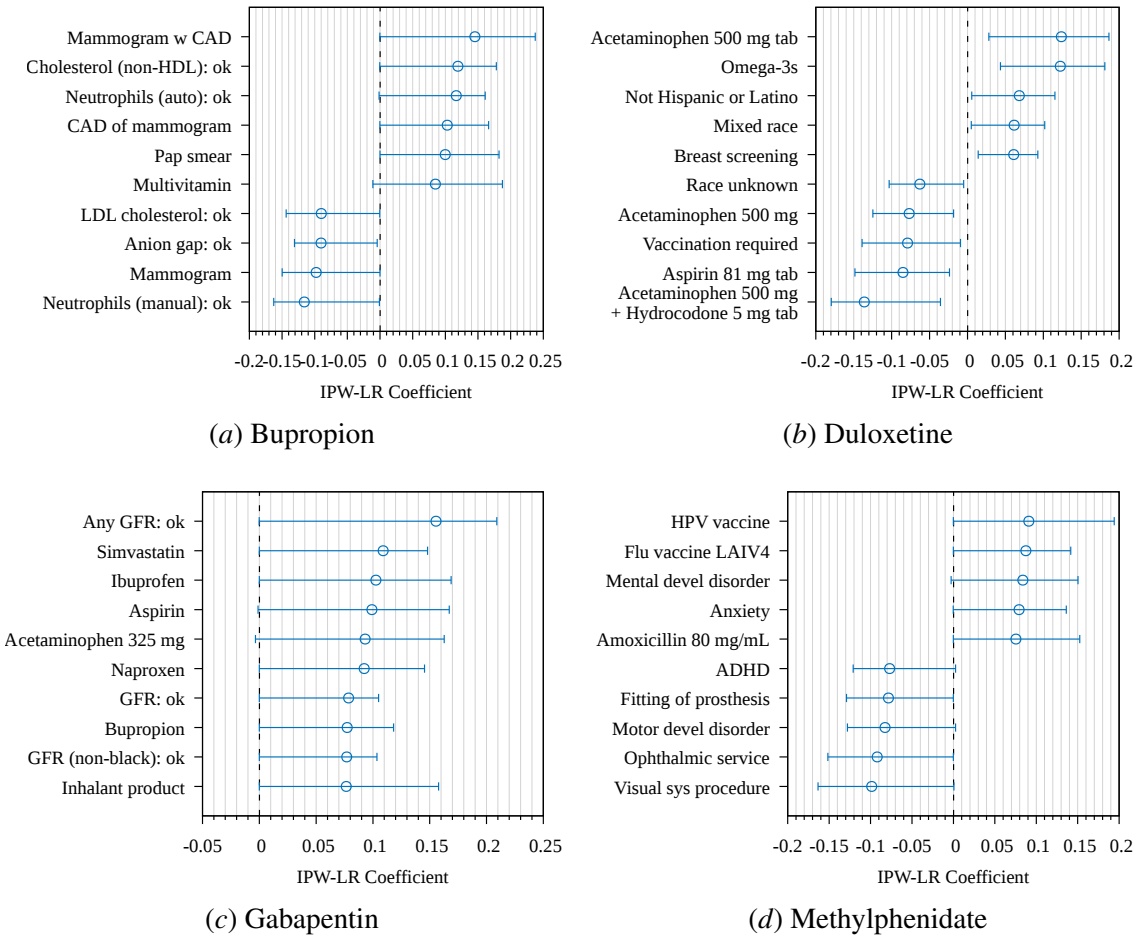
P_2 is the biggest difference between brand and generic, but that the correct event A_2 is the next one with a confidently nonzero coefficient. The two risk factors are confounders by indication, and P_1 is probably just an effect in common.

4.3. Experiments on Real EHR Data

To explore what differences the methods could discover between brand and generic drugs in real EHR data, four drugs were studied that were available in a generic version and had widespread use: (1) bupropion, an NDRI antidepressant that also helps with smoking cessation, (2) duloxetine, an SSRI antidepressant that also treats anxiety, fibromyalgia, and neuropathic pain, (3) gabapentin, an anticonvulsant that also treats neuropathic pain, and (4) methylphenidate, a stimulant that treats attention deficit hyperactivity disorder (ADHD). Both generic bupropion and methylphenidate had issues with specific manufacturing runs (US FDA, 2012, 2016) during the period of study. For each of these four drugs, a controlled before–after study was constructed with data from a deidentified EHR database from Marshfield Clinic, which resulted in the following numbers of cases: bupropion: 8,343, duloxetine: 5,105, gabapentin: 26,022, methylphenidate: 7,801. As a whole, the database included tables for demographics, diagnoses, drugs, measurements (labs and vitals), procedures, observations, visits, and deaths. The data spanned years 1978–2018, and included 1.7M patients and 1.5G events, with patient histories having 872 events on average.

As in the experiments on the synthetic data, the features were counts of event occurrences plus demographic facts. Events for lab results were included and discretized as low, ok, or high according to the recorded interpretation in the result. Each event was also represented at various levels of generality as found in the hierarchies for conditions, drugs, labs, observations, and procedures in the OHDSI vocabulary. However, the most general parts of the hierarchies were pruned by manual curation to remove concepts that were too general to be meaningful on their own (e.g., "disease," "pill," "substance," "surgical procedure"). In order to focus the results for human interpretation, the best method from the synthetic experiments was used: IPW-LR using the the best hyperparameters from the synthetic experiments (perhaps a sort of transfer tuning). 1000 bootstrap samples were done to produce confidence intervals for the parameter estimates.

Figure 4 shows the top differences between brand and generic discovered by IPW-LR for the four drugs. The confidence intervals for three of the drugs all include zero, suggesting there are no real differences between brand and generic in this data. Despite the tenuousness of these associations, some point to real-world reasons. For example, with bupropion (*a*), mammograms, tests for neutrophils, and cholesterol tests appear to be associated with both brand and generic, but are distinguished by technology or test type. This suggests changes in testing policies over time (e.g., from insurance coverage), yet temporal IPW prevents such temporal confounders from being false discoveries. With gabapentin (*c*), surprisingly, all of the top features are more associated with generic. There appear to be two themes: pain management and testing for kidney function. One could look at the association of generic gabapentin with ibuprofen, aspirin, acetaminophen, and naproxen and speculate that generic gabapentin is not as effective at managing pain as the brand version, thus needing more supplemental pain medications. However, these associations more likely reflect the expansion of pain-related indications for gabapentin, both approved and off-label, to conditions where supplemental pain medications could be expected, such as fibromyalgia, neuralgia, and chronic pruritis (Peckham et al., 2018). The tests for glomerular filtration rate (GFR) likely reflect the monitoring necessary for using gabapentin as an analgesic in patients with kidney dis-



(*a*) Bupropion (*b*) Duloxetine

(*c*) Gabapentin (*d*) Methylphenidate

Figure 4: Top 10 features from IPW-LR by LR coefficient magnitude with bootstrapped 99% confidence intervals. Positive coefficients favor generic; negative coefficients favor brand.

ease because they are at risk of gabapentin toxicity (Zand et al., 2010). With methylphenidate (*d*), the features illustrate ADHD in children and adolescents: getting glasses (which are prostheses), prevention and treatment of diseases, dealing with developmental conditions. However, "Mental devel disorder," OHDSI concept 4043545,[6] is a direct parent of ADHD in the diagnosis concept hierarchy. ADHD codes occurring more frequently in generic-takers could be a hint of the issue that manufacturers of the generic had with inefficacy (US FDA, 2016).

In contrast to the other three drugs, duloxetine (*b*) is the only drug where the confidence intervals exclude zero, but the features do not suggest any differences between brand and generic nonetheless. At best, all the acetaminophen features might suggest changes in prescribing over time, and the features for race and ethnicity might suggest a shift in patient demographics or just a change in how demographics are recorded.

6. https://athena.ohdsi.org/search-terms/terms/4043545

Overall, the features that differ the most between the brand and generic versions of the drugs appear to result from differences in the way these drugs were delivered in the health care system, perhaps differences over time, in health plans, or in practices between locations. In particular, there are no features that obviously point to adverse events of either the generic or brand versions of the drugs. This accords with the differences between these brands and generics known to the FDA, which have been related to specific manufacturing runs (US FDA, 2012, 2016), and which would likely not be evident in this data because this data does not include drug manufacturers. While temporal IPW's lack of findings may be interpreted as inconclusive, we regard its control of false discovery as a success given that falsely discovering temporal confounders is the main challenge faced by methods on this task.

5. Conclusion

Temporal IPW is a new method for causal discovery from observational data that is effective at discovering generic-specific ADEs in combination with controlled before–after studies. Such studies address the unique challenges and opportunities of ADE discovery in generic drugs by supporting causal discovery ML for hypothesizing ADEs and, especially with self-control, offering better control of confounders, including temporal and unobserved confounders. The benefits of this study design are also available to general causal discovery with other methods, such as differential classification using off-the-shelf ML classifiers. Together, these contributions to causal discovery promote the study of drug safety and thereby help to mitigate the high impact of ADEs on society.

Acknowledgments

The authors would like to gratefully acknowledge support for this work from the US FDA (FDA BAA13-00119) (AB, DP), especially Meng Hu, from the Computation and Informatics in Biology and Medicine Training Program (NLM training grant 5T15LM007359) (AB), and from Irene Ong (AB). The views are the authors' own, and they do not speak on behalf of the FDA or NLM.

References

Aubrey Barnard and David Page. Causal structure learning via temporal Markov networks. In *International Conference on Probabilistic Graphical Models 9*, 2018.

Rishi J. Desai, Ameet Sarpatwari, Sara Dejene, Nazleen F. Khan, Joyce Lii, James R. Rogers, Sarah K. Dutcher, Saeid Raofi, Justin Bohn, John G. Connolly, Michael A. Fischer, Aaron S. Kesselheim, and Joshua J. Gagne. Comparative effectiveness of generic and brand-name medication use: A database study of US health insurance claims. *PLoS Medicine*, 16(3), 2019. doi: 10.1371/journal.pmed.1002763.

Katherine Eban. *Bottle of Lies: The Inside Story of the Generic Drug Boom*. HarperCollins, 2019.

Sinong Geng, Zhaobin Kuang, Peggy Peissig, and David Page. Temporal Poisson square root graphical models. In *International Conference on Machine Learning 35*, 2018.

I. J. Good. A causal calculus (I). *The British Journal for the Philosophy of Science*, 11(44), 1961.

Jerry H. Gurwitz, Terry S. Field, Leslie R. Harrold, Jeffrey Rothschild, Kristin Debellis, Andrew C. Seger, Cynthia Cadoret, Leslie S. Fish, Lawrence Garber, Michael Kelleher, and David W. Bates. Incidence and preventability of adverse drug events among older persons in the ambulatory setting. *Journal of the American Medical Association*, 289(9), 2003. doi: 10.1001/jama.289.9.1107.

Pierre Gutierrez and Jean-Yves Gérardy. Causal inference and uplift modelling: A review of the literature. In *International Conference on Predictive Applications 3*, 2017.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. CRC Press, 2020.

Sir Austin Bradford Hill. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58, 1965.

George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R. Rijnbeek, Johan van der Lei, Nicole Pratt, G. Niklas Norén, Yu-Chuan Li, Paul E. Stang, David Madigan, and Patrick B. Ryan. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. In *World Congress on Health and Biomedical Informatics 15*, 2015. doi: 10.3233/978-1-61499-564-7-574.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Maciej Jaśkowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *ICML 2012 Workshop on Machine Learning for Clinical Data Analysis*, 2012.

Evan D. Kharasch, Alicia Neiner, Kristin Kraus, Jane Blood, Angela Stevens, Julia Schweiger, J. Philip Miller, and Eric J. Lenze. Bioequivalence and therapeutic equivalence of generic and brand bupropion in adults with major depression: A randomized clinical trial. *Clinical Pharmacology & Therapeutics*, 105(5), 2019. doi: 10.1002/cpt.1309.

Zhaobin Kuang, Peggy Peissig, Vítor Santos Costa, Richard Maclin, and David Page. Pharmacovigilance via baseline regularization with large-scale longitudinal observational data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 23*, 2017. doi: 10.1145/3097983.3097998.

Finn Kuusisto, Vítor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page, and Jude Shavlik. Support vector machines for differential prediction. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, 2014. doi: 10.1007/978-3-662-44851-9_4.

Jacinthe Leclerc, Claudia Blais, Louis Rochette, Denis Hamel, Line Guénette, and Paul Poirier. Impact of the commercialization of three generic angiotensin II receptor blockers on adverse events in Quebec, Canada. *Circulation: Cardiovascular Quality and Outcomes*, 10(10), 2017. doi: 10.1161/CIRCOUTCOMES.117.003891.

Pawel Lewek and Przemyslaw Kardas. Generic drugs: The benefits and risks of making the switch. *Journal of Family Practice*, 59(11), 2010.

Robert L. Linn. Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology*, 63(4), 1978. doi: 10.1037/0021-9010.63.4.507.

Po-Ling Loh and Martin J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Annals of Statistics*, 41(6), 2013. doi: 10.1214/13-AOS1162.

David Madigan and Patrick Ryan. What can we really learn from observational studies?: The need for empirical assessment of methodology for active drug safety surveillance and comparative effectiveness research. *Epidemiology*, 22(5), 2011. doi: 10.1097/EDE.0b013e318228ca1d.

Houssam Nassif, Vítor Santos Costa, Elizabeth S. Burnside, and David Page. Relational differential prediction. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012. doi: 10.1007/978-3-642-33460-3_45.

Uri Nodelman, Christian R. Shelton, and Daphne Koller. Continuous time Bayesian networks. In *Uncertainty in Artificial Intelligence 18*, 2002. arXiv:1301.0591.

David Page, Vítor Santos Costa, Sriraam Natarajan, Aubrey Barnard, Peggy Peissig, and Michael Caldwell. Identifying adverse drug events by relational learning. In *AAAI Conference on Artificial Intelligence 26*, 2012.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

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

Alyssa M. Peckham, Kirk E. Evoy, Leslie Ochs, and Jordan R. Covvey. Gabapentin for off-label use: Evidence-based or cause for concern? *Substance Use: Research and Treatment*, 12, 2018. doi: 10.1177/1178221818801311.

N. J. Radcliffe and P. D. Surry. Differential response analysis: Modeling true response by isolating the effect of a single action. In *Credit Scoring and Credit Control 6*, 1999.

Md. Motiur Rahman, Yasser Alatawi, Ning Cheng, Jingjing Qian, Annya V. Plotkina, Peggy L. Peissig, Richard L. Berg, David Page, and Richard A. Hansen. Comparison of brand versus generic antiepileptic drug adverse event reporting rates in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). *Epilepsy Research*, 135, 2017. doi: 10.1016/j.eplepsyres.2017.06.007.

James M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics – Theory and Methods*, 23(8), 1994. doi: 10.1080/03610929408831393.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 1994. doi: 10.1080/01621459.1994.10476818.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 1983. doi: 10.1093/biomet/70.1.41.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 65(5), 1974.

William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company, 2002.

Shawn E. Simpson, David Madigan, Ivan Zorych, Martijn J. Schuemie, Patrick B. Ryan, and Marc A. Suchard. Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics*, 69(4), 2013. doi: 10.1111/biom.12078.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

Paul E. Stang, Patrick B. Ryan, Judith A. Racoosin, J. Marc Overhage, Abraham G. Hartzema, Christian Reich, Emily Welebob, Thomas Scarnecchia, and Janet Woodcock. Advancing the science for active surveillance: Rationale and design for the Observational Medical Outcomes Partnership. *Annals of Internal Medicine*, 153(9), 2010. doi: 10.7326/0003-4819-153-9-201011020-00010.

Janet Sultana, Paola Cutroneo, and Gianluca Trifirò. Clinical and economic burden of adverse drug reactions. *Journal of Pharmacology and Pharmacotherapeutics*, 4(5), 2013. doi: 10.4103/0976-500X.120957.

US FDA. The Sentinel Initiative: National strategy for monitoring medical product safety. https://www.fda.gov/media/75240/download, 2008. Accessed November 2, 2024.

US FDA. Budeprion XL 300 mg not therapeutically equivalent to Wellbutrin XL 300 mg. https://www.fda.gov/drugs/postmarket-drug-safety-information-patients-and-providers/update-bupropion-hydrochloride-extended-release-300-mg-bioequivalence-studies, 2012. Accessed November 2, 2024.

US FDA. Methylphenidate hydrochloride extended release tablets (generic Concerta) made by Mallinckrodt and Kudco. https://www.fda.gov/drugs/drug-safety-and-availability/methylphenidate-hydrochloride-extended-release-tablets-generic-concerta-made-mallinckrodt-and-kudco, 2016. Accessed November 2, 2024.

US FDA. What is the approval process for generic drugs? https://www.fda.gov/drugs/generic-drugs/what-approval-process-generic-drugs, 2017. Accessed November 2, 2024.

S. Vansteelandt and E. Goetghebeur. Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4), 2003. doi: 10.1046/j.1369-7412.2003.00417.x.

Ladan Zand, Kevin P. McKian, and Qi Qian. Gabapentin toxicity in patients with chronic kidney disease: A preventable cause of morbidity. *The American Journal of Medicine*, 123(4), 2010. doi: 10.1016/j.amjmed.2009.09.030.

Ivan Zorych, David Madigan, Patrick Ryan, and Andrew Bate. Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Statistical Methods in Medical Research*, 22(1), 2011. doi: 10.1177/0962280211403602.

Appendix A. Proof of Proposition 1

Proof Let the sizes of the positive groups be $x = |\mathcal{T}_{2A}|$ and $y = |\mathcal{C}_B|$. Then the respective TPRs are

$$\text{TPR}_{\text{TIPW}} = \alpha_{\mathcal{T}} x / x \tag{6}$$

$$\text{TPR}_{\text{DFC}} = (\alpha_{\mathcal{T}} x + \beta_{\mathcal{C}} y)/(x + y). \tag{7}$$

Expressing y in terms of x, $\gamma x = y$, noting $x > 0 \wedge y > 0 \implies \gamma > 0$, the TPR for DFC becomes

$$= (\alpha_{\mathcal{T}} x + \beta_{\mathcal{C}} \gamma x)/(x + \gamma x) \tag{8}$$

$$= (\alpha_{\mathcal{T}} + \beta_{\mathcal{C}} \gamma)/(1 + \gamma). \tag{9}$$

Substituting this into the right-hand side of (5) yields the proposed result,

$$\alpha_{\mathcal{T}} \geq (\alpha_{\mathcal{T}} + \beta_{\mathcal{C}} \gamma)/(1 + \gamma) \tag{10}$$

$$\alpha_{\mathcal{T}} + \alpha_{\mathcal{T}} \gamma \geq \alpha_{\mathcal{T}} + \beta_{\mathcal{C}} \gamma \tag{11}$$

$$\alpha_{\mathcal{T}} \geq \beta_{\mathcal{C}}. \tag{12}$$

■

Appendix B. Supplemental Information on Methods

The following example illustrates comparing the groups in a controlled before–after study. Table 2 contains the counts of events of the study periods in Figure 2(*a*) in the form of feature vectors. Consider applying the method of difference in differences, Equation 1, reproduced here,

$$(f(\mathcal{T}_{2A}) - f(\mathcal{T}_{1B})) - (f(\mathcal{C}_A) - f(\mathcal{C}_B)).$$

Suppose f is the simplest, most naïve estimator, the identity function, and directly and literally apply Equation 1 elementwise to the count vectors. The first column of Table 2 represents this sum, and the last row shows the result. Finding the features with the maximum value picks out R_2 and A_2. Ruling out R_2 using temporality or background knowledge leaves A_2 as the generic-specific ADE, which is correct. Of course, involving actual estimators or classifiers and interpreting the differences conceptually rather than literally is what leads to differential prediction and classification.

		R_1	R_2	R_3	I_1	I_2	I_3	D_2	A_1	A_2	A_3	P_1	P_2
−	T_{1B}	1	0	1	0	1	1	0	1	0	0	1	0
+	T_{2A}	1	0	0	0	1	0	1	1	1	0	0	1
+	C_B	1	1	0	0	1	0	1	0	0	0	0	0
−	C_A	1	0	0	0	1	1	2	0	0	0	1	1
=		0	1	-1	0	0	-2	0	0	1	0	-2	0

Table 2: Count vectors for the timelines in Figure 2(*a*) and an example comparison according to Equation 1. D_1 is omitted because it is used for the label ($b \veebar g$).

Appendix C. Supplemental Information on Experiments

Table 3 shows the sizes of the synthetic data sets. There are four examples for each case because there are four study periods in a before–after study: each case and control in a matched pair contributes both a before and after period.

Patients	Cases	Examples
10	2	8
100	18	72
1 000	174	696
10 000	1 739	6 956
100 000	17 684	70 736
1 000 000	176 683	706 732
10 000 000	1 770 382	7 081 528

Table 3: Sizes of synthetic data sets.

Appendix D. Supplemental Information on Results

Table 4 shows the OHDSI concept IDs of the top features for the four generic drugs.

Feature Name	Concept ID
Mammogram w CAD	2617289
Cholesterol (non-HDL): ok	3044491
Neutrophils (auto): ok	3013650
CAD of mammogram	2211809
Pap smear	2720580
Multivitamin	19135832
LDL cholesterol: ok	40795800
Anion gap: ok	40789527
Mammogram	42737560
Neutrophils (manual): ok	3017501

(*a*) Bupropion

Feature Name	Concept ID
Acetaminophen 500 mg tab	19020053
Omega-3s	19106973
Not Hispanic or Latino	38003564
Mixed race	4212311
Breast screening	44823895
Race unknown	8552
Acetaminophen 500 mg	1127527
Vaccination required	37109774
Aspirin 81 mg tab	19059056
Acetaminophen 500 mg + Hydrocodone 5 mg tab	40162494

(*b*) Duloxetine

Feature Name	Concept ID
Any GFR: ok	1029770
Simvastatin	1539403
Ibuprofen	1177480
Aspirin	1112807
Acetaminophen 325 mg	1127524
Naproxen	1115008
GFR: ok	40771922
Bupropion	750982
GFR (non-black): ok	3049187
Inhalant product	36217207

(*c*) Gabapentin

Feature Name	Concept ID
HPV vaccine	2213435
Flu vaccine LAIV4	43527981
Mental devel disorder	4043545
Anxiety	441542
Amoxicillin 80 mg/mL	19083578
ADHD	438409
Fitting of prosthesis	4287998
Motor devel disorder	4148091
Ophthalmic service	4195164
Visual sys procedure	4155790

(*d*) Methylphenidate

Table 4: OHDSI concept IDs of top features.