Frequent Causal Pattern Mining: A Computationally Efficient Framework for Estimating Bias-Corrected Effects

Pranjul Yadav
Dept. of Computer Science
University of Minnesota
Minneapolis, MN
yadav023@umn.edu

Michael Steinbach
Dept. of Computer Science
University of Minnesota
Minneapolis, MN
stei0062@umn.edu

M. Regina Castro
Div. of Endocrinology
Mayo Clinic, Rochester, MN
castro.regina@mayo.edu

Pedro J. Caraballo
Dept. of Hospital Internal
Medicine
Mayo Clinic, Rochester, MN
caraballo.pedro@mayo.edu

Vipin Kumar
Dept. of Computer Science
University of Minnesota
Minneapolis, USA
kumar001@umn.edu

Gyorgy Simon
Dept. of Health Sciences
Research
Mayo Clinic, Rochester, MN
simon.gyorgy@mayo.edu

ABSTRACT

Our aging population increasingly suffers from multiple chronic diseases simultaneously, necessitating the comprehensive treatment of these conditions. Finding the optimal set of drugs for a combinatorial set of diseases is a combinatorial pattern exploration problem. Association rule mining is a popular tool for such problems, but the requirement of health care for finding causal, rather than associative, patterns renders association rule mining unsuitable. To address this issue, we propose a novel framework based for extracting causal rules from observational data, correcting for a number of common biases. Specifically, given a set of interventions (e.g. medications) and a set of items (e.g. diseases) that define subpopulations (e.g. people diagnosed with a set of diseases), we wish to find all subpopulations in which effective intervention combinations exist and in each such subpopulation, we wish to find all intervention combinations such that dropping any intervention from this combination will reduce the efficacy of the treatment. A key aspect of our framework is the concept of closed intervention sets which extend the concept of quantifying the effect of a single intervention to a set of concurrent interventions. Closed intervention sets also allow for a pruning strategy that is strictly more efficient than the traditional pruning strategy used by the Apriori algorithm for identifying frequent patterns. To implement our ideas, we introduce and compare five methods of estimating causal effect from observational data and rigorously evaluate them on synthetic data to mathematically prove (when possible) why they work. We also evaluated our causal rule mining framework on the Electronic Health Records data of a large cohort of patients from Mayo Clinic and showed that the patterns we extracted are sufficiently rich to explain the controversial findings in the medical literature regarding

the effect of a class of cholesterol drugs on Type-II Diabetes Mellitus.

Keywords

Causal Inference, Confounding, Counterfactual Estimation.

1 Introduction

Effective management of human health remains a major societal challenge as evidenced by the rapid growth in the number of patients with multiple chronic conditions. Type-II Diabetes Mellitus (T2DM), one of those conditions, affects 25.6 million (11.3%) Americans of age 20 or older and is the seventh leading cause of death in the United States [1]. Effective treatment of T2DM is frequently complicated by diseases comorbid to T2DM, such as high blood pressure, high cholesterol, and abdominal obesity. Currently, these diseases are treated in isolation, which leads to wasteful duplicate treatments and suboptimal outcomes. The recent rise in the number of patients with multiple chronic conditions necessitates comprehensive treatment of these conditions to reduce medical waste and improve outcomes.

Finding optimal treatment for patients who suffer from multiple associated diseases, each of which can have multiple available treatments is a complex problem. We could simply use techniques based on association, but a reasonable algorithm would likely find that the use of a drug is associated with some unfavorable outcome. This does not mean that the drug is harmful; in fact in many cases, it simply means that patients who take the drug are sicker than those who do not and thus they have a higher chance of the unfavorable outcome. What we really wish to know is whether a treatment *causes* an unfavorable outcome, as opposed to being merely associated with it.

The difficulty in quantifying the effect of interventions on outcomes stems from subtle biases. Suppose we wish to quantify the effect of a cholesterol-lowering agent, statin, on diabetes. We could simply compare the proportion of diabetic patients in the subpopulation that takes statin and the subpopulation that does not and estimate the effect of statin as the difference between the two proportions. This method would give the correct answer only if the statin-taking and non-statin-taking patients are identical in all respects that

influence the diabetes outcome. We refer to this situation as treated and untreated patients being *comparable*. Unfortunately, statin taking patients are not comparable to non-statin-taking patients, because they take statin to treat high cholesterol, which by and in itself increases the risk of diabetes. High cholesterol *confounds* the effect of statin. Many difference sources of bias exist, confounding is just one of the many. In this manuscript, we are going to address several different sources of bias, including confounding.

Techniques to address such biases in causal effect estimation exist. However, these techniques have been designed to quantify the effect of a single intervention. In trying to apply these techniques to our problem of finding optimal treatment for patients suffering from varying sets of diseases, we face two challenges.

First, patients with multiple conditions will likely need a combination of drugs. Quantifying the effect of multiple concurrent interventions is semantically different from considering only a single intervention. The key concept in estimating the effect of an intervention is *comparability*: to estimate the effect of intervention, we need two groups of patients who are identical in all relevant aspects except that one group receives the intervention and the other group does not. For a single intervention, the first group is typically the sickest patients who still do not get treated and the second group consists of the healthiest patient who get treatment. They are reasonably in the same state of health. However, when we go from a single intervention to multiple intervention and try to estimate their *joint* effect, comparability no longer exists. A patient requiring multiple simultaneous interventions is so fundamentally different from a patient who does not need any intervention that they are not comparable.

The other key challenge in finding optimal intervention sets for patients with combinatorial sets of diseases is the combinatorial search space. Even if we could trivially extend the methods for quantifying the effect of a single intervention to a set of concurrent interventions, we would have to systematically explore a combinatorially large search space. The association rule mining framework [2] provides an efficient solution for exploring combinatorial search spaces, however, it only detects associative relationships. Our interest is in causal relationships.

In this manuscript, we propose causal rule mining, a framework for transitioning from association rule mining towards causal inference in subpopulations. Specifically, given a set of interventions and a set of items to define subpopulations, we wish to find all subpopulations in which effective intervention combinations exist and in each such subpopulation, we wish to find all intervention combinations such that dropping any intervention from this combination will reduce the efficacy of the treatment. We call these *closed intervention sets*, which are not be confused with closed item sets. As a concrete example, interventions can be drugs, subpopulations can be defined in terms of their diseases and for each subpopulations (set of diseases), our algorithm would return effective drug cocktails of increasing number of constituent drugs. Leaving out any drug from the cocktail will reduce the efficacy of the treatment. Closed intervention sets allow us to go from estimating a single intervention to multiple interventions.

To address the exploration of the combinatorial search space, we propose a novel frequency-based anti monotonic pruning strategy enable by the closed intervention set concept. The essence of antimonotonic property is that if a set I of interventions does not satisfy a criterion, none of its supersets will. The proposed pruning strategy based on the closed intervention is strictly more efficient than the traditional pruning strategy used by the Apriori algorithm [2] to identify frequent patterns.

Underneath our combinatorial exploration algorithm, we utilize the Rubin-Neyman model of causation [3]. This model sets two conditions for causation: a set X of interventions causes a change in Y iff X happens before Y and Y would be different had X not occurred. The unobservable outcome of what would happen had a treated patient not received treatment is a *potential outcome* and needs to be estimated. We present and compare five methods for estimating these potential outcomes and describe the biases these methods can correct.

Typically the ground truth for the effect of drugs is not known. In order to assess the quality of the estimates, we conduct a simulation study utilizing five different synthetic data set that introduce a new source of bias. We also evaluate the effect of the bias on the five proposed methods underscoring the statements with rigorous proofs when possible.

We also evaluate our work on a real clinical data set from Mayo Clinic. We have data for over 52,000 patients with 13 years of follow-up time. Our outcome of interest is 5-year incident T2DM and we wish to extract patterns of interventions for patients suffering from combinations of common comorbidities of T2DM. First, we evaluate our methodology in terms of the computational cost, demonstrating the effectiveness of the pruning methodologies. Next, we evaluate the patterns qualitatively, using patterns involving statins. We show that our methodology extracted patterns that allow us to explain the controversial patterns surrounding statin [4].

Contributions. (1) We propose a novel framework for extracting causal rules from observational data correcting for a number of common biases. (2) We introduce the concept of closed intervention sets to extend the concept of quantifying the effect of a single intervention to a set of concurrent interventions sidestepping the patient comparability problem. Closed intervention sets also allow for a pruning strategy that is strictly more efficient than the traditional pruning strategy used by the Apriori algorithm [2] for finding frequent sets. (3) We compare five methods of estimating causal effect from observational data that are applicable to our problem and rigorously evaluate them on a real clinical data. We also evaluate them on synthetic data to mathematically prove (when possible) why they work.

2   Related Work

Causation has received substantial research interest in many areas. In computer science, Pearl [5] and Rosenbaum[6] laid the foundation for causal inference, upon which several fields, cognitive science, econometrics, epidemiology, philosophy and statistics have built their respective methodologies [7, 8, 9].

At the center of causation is a causal model. Arguably, one of the earliest and popular models is the Rubin-Neyman causal model [3]. Under this model X causes Y, if X occurs before Y; and without X, Y would be different. Beside the Rubin-Neyman model, there are several other causal models, including the Granger causality [10] for time se-

ries, Bayes Networks [11], Structural Equation Modeling [8], causal graphical models [12], and more generally, probabilistic graphical models [13]. In our work, we use the potential outcome framework from the Rubin-Neyman model and we use causal graphical models to identify and correct for biases.

Causal graphical models are tools to visualize causal relationships among variables. Nodes of the causal graph are variables and edges are causal relationships. Most methods assume that the causal graph structure is a priori given, however, methods have been proposed for discovering the structure of the causal graph [14, 15]. In our work, the structure is partially given: we know the relationships among groups of variables, however we have to assign each variable to the correct group based on data.

Knowing the correct graph structure is important, because substructures in the graph are suggestive of sources of bias. To correct for biases, we are looking for specific substructures. For example, causal chains can be sources of overcorrection bias and "V"-shaped structures can be indicative of confounding or endogenous selection bias [9]. Many other interesting substructures have been studied [16, 17, 18]. In our work, we consider three fundamental such structures: direct causal effect, indirect causal effect and confounding. Of these, confounding is the most severe and received the most research interest.

Numerous methods exist to handle confounding, which includes propensity score matching (PSM) [19], structural marginal models [9] and g-estimation [8]. The latter two extend PSM for various situations, for example, for time-varying interventions [9].

Propensity score matching is used to estimate the effect of an intervention on an outcome. The propensity score is the propensity (probability) of a patient receiving the intervention given his baseline characteristics and the propensity score is used to create a new population that is free of confounding. Many PSM techniques exist and they typically differ in how they use the propensity score to create this new population [20, 21, 22, 23].

Applications of causal modeling is not exclusive to social and life sciences. In data mining, Lambert et al. [24] investigated the causal effect of new features on click through rates and Chan et al. [25] used doubly robust estimation techniques to determine the efficacy of display advertisements.

Even extending association rules mining to causal rule mining has been attempted before [26, 27, 28]. Li et al. [26] used odds ratio to identify causal patterns and later extended their technique [28] to handle large data set. Their technique, however, is not rooted in a causal model and hence offers no protection against computing systematically biased estimates. In their proposed causal decision trees [29], they used the potential outcome framework, but still have not addressed correction for various biases, including confounding.

3 Background: Association Rule Mining

We first briefly review the fundamental concepts of association rule mining and extend these concepts to causal rule mining in the next section. Consider a set $\mathcal{I}$ of **items**, which are single-term predicates evaluating to 'true' or 'false'. For example, $\{age > 55\}$ can be an item. A k-**itemset** is a set of k items, evaluated as the conjunction (logical 'and') of its constituent items. Consider a dataset D = { $d_1, d_2.....d_n$ }, which consists of n **observations**. Each observation, denoted by d_j is a set of items. An itemset $I = i_1, i_2, \ldots, i_k$ $(I \subset \mathcal{I})$ **supports** an observation d_j if all items in I evaluate to 'true' in the observation. The **support** of I is the fraction of the observations in D that support I. An itemset is **frequent** if its support exceeds a pre-defined minimum support threshold.

A association rule is a logical implication of form $X \Rightarrow Y$, where X and Y are disjoint itemsets. The support of a rule is support(XY) and the **confidence** of the rule is

$$\text{conf}(X \Rightarrow Y) = \frac{\text{support}(XY)}{\text{support}(X)} = \text{P}(Y|X).$$

4 Causal Rule Mining

Given an **intervention** itemset X and an **outcome** item Y, such that X and Y are disjoint, a causal rule is an implication of form $X \rightarrow Y$, suggesting that X *causes* a change in Y. Let the itemset S define a **subpopulation**, consisting of all observations that support S. This subpopulation consists of all observations for which all items in S evaluate to 'true'. The **causal rule** $X \rightarrow Y|_S$ implies that the intervention X has causal effect on Y in the subpopulation defined by S. The quantity of interest is the **causal effect**, which is the change in Y in the subpopulation S caused by X. We will formally define the metric used to quantify the causal effect shortly.

Rubin-Neyman Causal Model. X has a causal effect on Y if (i) X happens earlier than Y and (ii) if X had not happened, Y would be different [3].

Our study design ensures that the intervention X precedes the outcome Y, but fulfilling the second conditions requires that we estimate the outcome for the same patient both under intervention and without intervention.

Potential Outcomes. Every patient in the dataset has two potential outcomes: Y_0 denotes their outcome had they not had the intervention X; and Y_1 denotes the outcome had they had the intervention. Typically, only one of the two potential outcomes can be observed. The observable outcome is the **actual** outcome (denoted by Y) and the unobservable potential outcome is called the **counterfactual** outcome.

Using the definition of counterfactual outcome, we can now define the metric for estimating the change in Y caused by X. **Average Treatment response on the Treated** (ATT) is a widely known metric in the causal literature and is computed as follows:

$$\text{ATT}(X \rightarrow Y|_S) = \mathbb{E}[Y_1 - Y_0]_{X=1} = \mathbb{E}[Y_1]_{X=1} - \mathbb{E}[Y_0]_{X=1},$$

where $\mathbb{E}$ denotes the expectation and the $X = 1$ in the subscript signals that we only evaluate the expectation in the treated patients ($X = 1$).

ATT aims to compute an average per-patient change caused by the intervention. $Y_0 = Y_1$, indicates that the intervention resulted in no change in outcome for the patient.

Biases. Beside X, numerous other variables can also exert influence over Y, leading to biases in the estimates. To correct for these biases, we have correctly account for these other effects. The quintessential tool for this purpose is the causal graph, depicted in Figure 1. The nodes of this graph are sets of variables that play a causal role and edges are causal effects. This is not a correlation graph (or dependence graph), because for example, U and Z are dependent given X, yet there is no edge between them.

Variables (items in $\mathcal{I}$) can exert influence on the effect of X on Y in three way: they may only influence X, they may only influence Y or them may influence both X and Y. Accordingly, variables can be categorized into four categories:

V are variables that directly influence Y and thus have *direct effect* on Y;

U are variables that only influence Y through X and thus have *indirect effect* on Y;

Z are variables that influence both X and Y and are called *confounders*; and finally

O are variables that do not influence either X or Y and hence can be safely ignored.
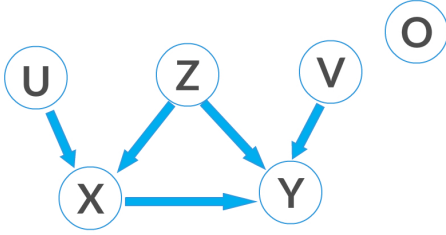

Figure 1: Rubin-Neyman Causal Model

Most of the causal inference literature assumes that the causal graph is known and true. In other words, we know apriori which variables fall into each of the categories, U, Z, V and O. In our case, only X and Y are specified and we have to infer which category each other variable (item) belongs to. Since this inference relies on association (dependence) rather than causation, the discovered graph may have errors, misclassifications of variables into the wrong category. For example, because of the marginal dependence between U and Y, variables in U can easily get misclassified as Z. Such misclassifications do not necessarily lead to biases, but they can cause loss of efficiency.

Problem Formulation. Given a data set D, a set $\mathcal{S}$ of **subpopulation-defining** items, a set $\mathcal{X}$ of **intervention** items, a minimal support threshold θ and a minimum effect threshold η, we wish to find all subpopulations S ($S \subset \mathcal{S}$) and all interventions X ($X \subset \mathcal{X}$), X and S are disjoint, such that the causal rule $X \rightarrow Y|_S$ is frequent and its intervention set X is closed w.r.t. our metric of causal effect, ATT.

Note that the meaning of θ, the minimum support threshold, is different than in association rule mining literature. Typically, rules with support less than θ are considered uninteresting, in other cases, it is simply a computational convenience, but in our case, we set θ to a minimum value such that ATT is estimable for the discovered patterns.

We call a causal rule **frequent** iff its support exceeds the user-specified minimum threshold θ

$$\text{support}(X \rightarrow Y|_S) = \text{support}(XYS) = \text{P}(XYS) > \theta$$

and we call an intervention set X **closed** w.r.t. to ATT iff

$$\forall x \in X, \quad |ATT(x \rightarrow Y|_{S, X \setminus x})| > \eta,$$

where η is the user-specified minimum causal effect threshold. In other words, a causal rule is closed in a subpopulation, if its (absolute) effect is greater than any of its subrules.

Example. In a medical setting, $\mathcal{X}$ may be drugs, $\mathcal{S}$ could be comorbid diseases. Then X is a drug-combination that hopefully treats set of diseases S. This set of drugs being *closed* w.r.t. ATT means that dropping any drug from X will reduce the overall efficacy of the treatment; the patient is not taking unnecessary drugs.

An itemset is closed if its support is strictly higher than all of its subitemsets'. Analogously, an intervention set is closed if its absolute causal effect is strictly higher than all of its subitemsets'.

5 Frequent Causal Pattern Mining Algorithm

We can now present our algorithm for causal pattern mining. At a very high level, the algorithm comprises of two nested frequent pattern enumeration [30] loops. The outer loop enumerates subpopulation-defining itemsets S using items in $\mathcal{S}$, while the inner loop enumerates intervention combinations using items in $\mathcal{X} \setminus \mathcal{S}$. More generally, $\mathcal{X}$ and $\mathcal{S}$ can overlap but we do not consider that in this paper. Effective algorithms to this end exists [31, 32], we simply use Apriori [2].

Once the patterns are discovered, the ATT of the interventions are computed, using one of the methods from Section 6 and the frequent, effective patterns are returned.

On the surface, this approach appears very expensive, however several novel, extremely effective pruning strategies are possible and we describe them below.

Potential Outcome Support Pruning. Let X be an intervention k-itemset, S be a subpopulation-defining itemset, and let X and S be disjoint. Further, X_{-i} be an itemset that evaluates to 'true' iff all items except the ith are 'true' but the ith item is 'false'. Using association rule mining terminology, all items in X except the ith are present in the transaction.

DEFINITION 1 (POTENTIAL OUTCOME SUPPORT PRUNING). *We only need to consider itemsets X such that*

$$\min\{\text{support}(S, X), \quad \text{support}(\{S, X_{-1}), \ldots,$$
$$\text{support}(S, X_{-k})\} \quad > \theta.$$

In order to be able to estimate the effect of $x \in X$ in the subpopulation S, we need to have observations with x 'true' and also with x 'false' in S.

LEMMA 1. *Potential Outcome Support Pruning is anti-monotonic.*

PROOF: Consider a causal rule $X \rightarrow Y|_S$. If the causal rule $X \rightarrow Y|_S$ is infrequent, then

$$\text{support}(XS) < \theta \quad \lor \quad \exists i, \text{support}(X_{-i}S) < \theta.$$

If support$(X_{-i}S)$ had insufficient support, then any extension of it with an intervention item x will continue to have insufficient support, thus the $Xx \rightarrow Y|_S$ rule will have insufficient support. Likewise, if support(XS) had insufficient support, then any extension of it with an intervention item x will also have insufficient support.

Pruning based on Causal Effect.

PROPOSITION 1. *Effective causal rule pruning condition is anti-monotonic.*

RATIONALE: To explain the rational, let us return to the medical example, where X is a combination of drugs forming a treatment. Assuming that the effects of drugs are additive, if a casual rule $X \rightarrow Y|_S$ is ineffective because

$$\exists x_i \in X, \qquad |\text{ATT}(x_i \rightarrow Y|_{S, X \setminus x_i})| < \eta,$$

then forming a new rule $X x_j \rightarrow Y|_S$ will also be ineffective because

$$|\text{ATT}(x_i \rightarrow Y|_{S, x_j, X \setminus x_i})|$$

will be ineffective. In the presence of positive interactions (that reinforce each other's effect) among the drugs, this statement may not hold true. Beside statistical reasoning, one can question why a patient should receive a drug that has no effect in a combination.

6 Causal Estimation Methods

ATT, our metric of interest, with respect to a single intervention x in a subpopulation S is defined as

$$\text{ATT}(x \rightarrow Y|_S) = \mathbb{E}\left[Y_1 - Y_0\right]_{S, X=1},$$

which is the expected difference between the potential outcome under treatment Y_1 and the potential outcome without treatment Y_0 in patients with S who actually received treatment. Since we consider treated patients, the potential outcome Y_1 can be observed, the potential outcome Y_0 cannot. Thus at least one of the two must be estimated. The methods we present below differ in which potential outcome they estimate and how they estimate it.

For the discussion below, we consider the variables X, Z, U and V from the causal graph in Figure 1. X is a single intervention, U, V and Z can be sets of items. For regression models, we will denote the matrix defined by U, V and Z in the subpopulation S as U, V and Z (same letter as the variable sets).

Counterfactual Confidence (CC). This is the simplest method. We simply assume that the patients who receive intervention $X = 1$ and those who do not $X = 0$, do not differ in any important respect that would influence Y. Under this assumption, Y_1 in the treated is simply the actual outcome in the treated and the potential outcome Y_0 is simply the actual outcome in the non-treated ($X = 0$). Thus

$$\begin{aligned}
\text{ATT} &= \text{conf}((X=1) \rightarrow Y|_S) - \text{conf}((X=0) \rightarrow Y|_S), \\
&= \text{P}(Y|S, X=1) - \text{P}(Y|S, X=0)
\end{aligned}$$

In the followings, to improve readability, we drop the S subscript. All evaluations take place in the S subpopulations.

Direct Adjustment (DA). We cannot estimate Y_0 in the treated ($X = 1$) as the actual outcome Y in the untreated, because the treated and untreated populations can significantly differ in variables such as Z and V that influence Y. In Direct Adjustment, we attempt to directly remove the effect of V and Z by including them into a regression model. Since a regression model relates the means of the predictors with the mean of the outcome, we can remove the effect of V and Z by making their means 0.

Let R be a generalized linear regression model, predicting Y via a link function g

$$g(Y|V, Z, X) = \beta_0 + \beta_V V + \beta_Z Z + \beta_X X.$$

Then the (link-transformed) potential outcome under treatment is $g(Y_1) = \beta0 + \beta_V V + \beta_Z Z + \beta_X$ and the potential outcome without treatment is $g(Y_0) = \beta0 + \beta_V V + \beta_Z Z$. The ATT is then

$$\begin{aligned}
\text{ATT} &= \mathbb{E}\left[g^{-1}(Y_1|V, Z, X=1)\right]_{X=1} - \\
&\quad \mathbb{E}\left[g^{-1}(Y_0|V, Z, X=0)\right]_{X=1}.
\end{aligned}$$

where $g^{-1}(Y_1|V, Z, X=1)$ is prediction for an observation with the observed V and Z but with X set to 1. The $\mathbb{E}(\cdot)_{X=1}$ notation signifies that these expectation of the predictions are taken only over patients who actually received the treatment.

The advantage of DA (over CC) is manyfold. First, it can adjust for Z and V as long the model specification is correct, namely the interaction terms that may exist among Z and V are specified correctly. Second, we get correct estimates even if we ignore U, because U is conditionally independent of Y given X. This unfortunately only is a theoretical advantage, because we have to infer from the data whether a variable is a predictor of Y and U is marginally dependent on Y, so we will likely adjust for U, even if we don't need to.

Counterfactual Model (CM). In this technique, we build an explicit model for the potential outcome without treatment Y_0 using patients with $X = 0$. Specifically, we build a model

$$g(Y|V, Z, X=0) = \beta_0 + \beta_V V + \beta_Z Z.$$

and estimate the potential outcome as

$$g(Y_0|V, Z) = g(Y|V, Z, X=0).$$

The ATT is then

$$\text{ATT} = \text{P}(Y|X=1) - \mathbb{E}\left[g^{-1}(Y_0|V, Z)\right]_{X=1}.$$

Similarly to Direct Adjustment, the Counterfactual Model does not depend on U. However, in case of the Counterfactual Model, we are only considering the population with $X = 0$. In this population, U and Y are independent, thus we will not include U variables into the model.

Propensity Score Matching (PSM). The central idea of Propensity Score Matching is to create a new population, such that patients in this new population are comparable in all relevant respects and thus the expectation of the potential outcome in the untreated equals the expectation of the actual outcome in the untreated.

Patients are matched based on their propensity of receiving treatment. This propensity is computed as a logistic regression model with treatment as the dependent variable

$$\log \frac{\text{P}(X)}{1 - \text{P}(X)} = \beta_0 + \beta_V V + \beta_Z Z.$$

Patient pairs are formed, such that in each pair, one patient received treatment and the other did not and their propensities for treatment differ by no more than a user-defined caliper difference ρ.

The matched population has an equal number of treated and untreated patients, is balanced on V and Z, thus the patients are comparable in terms of their baseline risk of Y. Hopefully, the only factor causing a difference in outcome is the treatment.

For estimating ATT, the potential outcome without treatment is estimated from the actual outcomes of the patients

in the matched population who did not receive treatment:

$$ATT = \mathbb{E}[Y_1 - Y_0]$$
$$- P(Y|X=1, M) - P(Y|X=0, M),$$

where M denotes the matched population.

Among the methods we consider, propensity score matching most strictly enforces the patient comparability criterion, however, it is susceptible to misspecification of the propensity regression model, which can erode the quality of the matching.

Stratified Non-Parametric (SN). In the stratified estimation, we directly compute the expectation via stratification. The assumption is that the patient in each stratum are comparable in all relevant respects and only differ in the presence or absence of intervention. In each stratum, we can estimate the potential outcome Y_0 in the treated as the actual outcome Y in the untreated.

$$ATT = \mathbb{E}[Y_1 - Y_0]_{X=1}$$
$$= \sum_l P(l|X=1) [P(Y_1|l, X=1) - P(Y_0|l, X=1)]$$
$$= \sum_l P(l|X=1) [P(Y|X=1) - P(Y|X=0)],$$

where l iterates over the combined levels of V and Z. If we can identify the items that fall into U, then we can ignore them, otherwise, we should include them as well into the stratification.

The stratified method makes very few assumptions and should arrive at the correct estimate as long as each of the strata are sufficiently large. The key disadvantage of the stratified method lies in stratification itself: when the number of items across which we need to stratify is too large, we may end up dividing the population into excessively many small subpopulations (strata) and become unable to estimate the causal effect in many of them thus introducing bias into the estimate.

7 Results

After describing our data and study design, we present three evaluations of the proposed methodology. The first evaluation demonstrates the computational efficiency of our pruning methodologies, isolating the effect of each pruning methods: (i) Apriori support-based pruning, (ii) Potential Outcome Support Pruning, and (iii) Potential Outcome Support Pruning in conjunction with Effective Causal Rule Pruning. In the second section, we provide a qualitative evaluation, looking at patterns involving statin. We attempt to use the extracted patterns to explain the controversial findings that exist in the literature regarding the effect of statin on diabetes. Finally, in order to compare the treatment effect estimates to a ground truth, which does not exists for real drugs, we simulate a data set using proportions we derived from the Mayo Clinic data set.

Data and Study Design. In this study we utilized a large cohort of Mayo Clinic patients with data between 1999 and 2013. We included all adult patients (69,747) with research consent. The baseline of our study was set at Jan. 1, 2005. We collected lab results, medications, vital signs and status, and medication orders during a 6-year *retrospective period* between 1999 and the baseline to ascertain the patient's baseline comorbidities. From this cohort, we excluded all patients with a diagnosis of diabetes before the baseline (478 patients), missing fasting plasma glucose measurements (14,559 patients), patients whose lipid health could not be determined (1,023 patients) and patients with unknown hypertension status (498 patients). Our final study cohort consists of 52,139 patients who were followed until the summer of 2013.

Patients were phenotyped during the retrospective period. Comorbidities of interest include Impaired Fasting Glucose (IFG), abdominal obesity, Hypertension (HTN; high blood pressure) and hyperlipidemia (HLP; high cholesterol). For each comorbidity, the phenotyping algorithm classified patients into three broad levels of severity: normal, mild and severe. Normal patients show no sign of disease; mild patients are either untreated and out of control or are controlled using first-line therapy; severe patients require more aggressive therapy. IFG is categorized into normal and prediabetic, the latter indicating impaired fasting plasma glucose levels but not meeting the diabetes criteria yet. For this study, progression to T2DM within 5 years from baseline (i.e. Jan 1, 2005) was chosen as our outcome of interest. Out of 52,139 patients 3627 patients progressed to T2DM, 41028 patients did not progressed to T2DM and the remaining patients (7484) dropped out of the study. In Table 1 we present statistics about our patient population.

	T2DM	
	Present	Absent
Total Number of Patients	3627	41028
Average Age	44.73	35.58
Male(%)	51	41
Female(%)	49	59
Patient Diagnosis Status (%)		
NormFG	42	84
PreDM	58	16
Normal Obesity	29	59
Mild Obesity	25	30
Severe Obesity	46	11
Normal Hypertension	48	69
Mild Hypertension	33	23
Severe Hypertension	19	08
Normal Hyperlipidemia	12	29
Mild Hyperlipidemia	72	64
Severe Hyperlipidemia	16	07
Patient Medication Status(%)		
Statin	26	11
Fibrates	03	01
Cholesterol.Other	02	01
Acerab	17	07
Diuret	18	07
CCB	08	04
BetaBlockers	22	10
HTN.Others	01	01

Table 1: Demographics statistics of patient population

7.1 Pruning Efficiency

In our work, we proposed two new pruning methods. First, we have the Potential Outcome Support Pruning, which aims to eliminate patterns for which the ATT is not estimable. Second, we have the Effective Causal Rule Pruning, where we eliminate patterns that do not improve treatment effectiveness relative to the subitemsets.

In Figure 2 we present the number of patterns discovered using (i) the traditional Apriori support based pruning, (ii) our proposed Potential Outcome Support Pruning (POSP), and (iii) POSP in conjunction with Effective Causal Rule Pruning (ECRP).
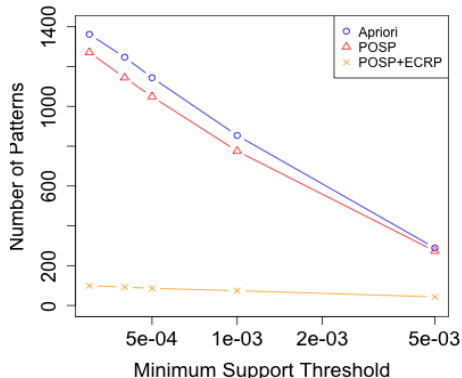

Figure 2: Comparison of Pruning Techniques

The number of patterns discovered by POSP is strictly less than the number of patterns discovered by the Apriori pruning. POSP in conjunction with ECRP is very effective.

7.2 Statin

In this section, we demonstrate that the proposed causal rule mining methodology can be used to discover non-trivial patterns from the above diabetes data set.

In recent years, the use of statins, a class of cholesterol-lowering agents, have been prescribed increasingly. High cholesterol (hyperlipidemia) is linked to cardio-vascular mortality and the efficacy of statins in reducing cardio-vascular mortality is well documented. However, as evidenced by a 2013 BMJ editorial [4] devoted to this topic, statins are surrounded in controversy. In patients with normal blood sugar levels (labeled as NormalFG), statins have a detrimental effect, they increase the risk of diabetes; yet in pre-diabetic patients (PreDM), it appears to have no effect. What we demonstrate below is that this phenomenon is simply disease heterogeneity.

First, we describe how this problem maps to the causal rule mining problem. Our set of interventions ($\mathcal{X}$) consists of statin and our subpopulation defining variables consist of the various levels of HTN, HLP and IFG ($\mathcal{S}$). Our interest is the effect of statin (x) on T2DM (Y) in all possible subpopulations S, $S \subset \mathcal{S}$.

In this setup, HTN, which is associated with both hyperlipidemia (and statin use), as well as with T2DM, is a confounder (Z). A cholesterol drug, other than statin, (say) fibrates, are in the U category: they are predictive of statin (patients on monotherapy who take fibrates do not take statins), but have no effect on Y, because its effect is already incorporated into the hyperlipidemia severity variables that defined the subpopulation. Variables that only influence diabetes but not statin use (say a diabetes drug) would fall into the V category. All subpopulations have variables that fall into Z and U and some subpopulation may also have V.

The HLP variable in Table 1 uses statin as part of its definition, thus we constructed two new variables. The first one is HLP1, a variable at the borderline between HLP-Normal

and HLP-Mild, consisting of untreated patients with mildly abnormal lab results (these fall into HLP-Normal) and patients who are diagnosed and receive a first-line treatment (they fall into HLP-Mild). Comparability is the central concept of estimating causal effects and these patients are comparable at baseline. Similarly, we also created another variable, HLP2, which is at the border of HLP-Mild and HLP-Severe, again consisting of patients who are comparable in relevant aspects of their health at baseline.

S	CC	DA	CM	PSM	SN
PreDM	0.145	0.022	0.010	0.022	0.017
NormFG	0.060	0.023	0.034	0.017	0.029
HLP1	0.078	0.019	0.014	0.010	0.010
HLP2	0.021	-0.013	-0.010	-0.021	-0.015
PreDM,HLP1	0.067	0.018	0.021	0.004	0.002
PreDM,HLP2	0.001	-0.038	-0.031	-0.048	-0.043
NormFG,HLP1	0.043	0.020	0.015	0.014	0.013
NormFG,HLP2	0.017	-0.002	-0.002	-0.005	-0.004

Table 2: ATT due to statin in various subpopulations S as estimated by the 5 proposed methods.

Table 2 presents the ATT estimates obtained by the various methods proposed in Section 3.4 for some of the most relevant subpopulations. Negative ATT indicates beneficial effect and positive ATT indicates detrimental effect.

Counterfactual confidence (CC) estimates statin to be detrimental in all subpopulations. While statins are known to have detrimental effect in patients with normal glucose levels [4], it is unlikely that statins are universally detrimental, even in patients with severe hyperlipidemia, the very disease it is supposed to treat.

The results between DA, CM, PSM and SN are similar, with PSM and SN having larger effect sizes in general. The picture that emerges from these results is that patients with severe hyperlipidemia appear to benefit from statin treatment even in terms of their diabetes outcomes, while statin treatment is moderately detrimental for patients with mild hyperlipidemia.

Bootstrap estimation was used to compute the statistical significance of these results. For brevity, we report the results only for PSM. The estimates are significant in the following subpopulations: NormFG, PreDM+HLP2 (p-values are $<.001$) and NormFG+HLP1 (p-value .05).

The true ATT in these subpopulations is not know. To investigate the accuracy that the various methods achieve, we use simulated that is largely based on this example [4, 33].

7.3 Synthetic Data

In this section, we describe five experiments utilizing synthetic data sets, each of which introduces a new potential source of bias. Our objective is to illustrate the ability of the five methods from Section 3.4 for adjusting for these biases. We compare their ATT estimates to the true ATT we used to generate the data set and discuss reasons for their success or failure.

The rows of Table 3 correspond to the synthetic data sets in increasing order of the biases we introduced and the columns corresponds to the methods: Conf (confidence), CC (Counterfactual Confidence), DA (Direct Adjustment), CM (Counterfactual Model), PSM (Propensity Score Matching) and

SNP (Stratified Non-Parametric).

Some of these methods, DA, CM, PSM and SNP take the causal graph structure into account while estimating ATT. Specifically, they require the information whether a variable is a confounder (Z), has a direct effect (V), an indirect effect (V), or no effect (O). PSM and SNP are not sensitive to the inclusion of superfluous variables, they simply decrease the method's efficiency.

In all of the data sets, we use a notation consistent with Figure 1: Z is the central disease with outcome Y; X is the intervention of interest that treats Z; V is another disease with direct causal effect on Y, but V is independent of X; and U is a third disease, which can be treated with X, but has no impact on Y. All data sets contain 5000 observations.

I. Direct Causal Effect from V. We assume that every patient in the cohort has disease Z at the same severity. They are all comparable w.r.t. Z. 30% of the patients are subject to the intervention X aimed at treating Z, while others are not. Untreated patients face a 25% chance of having Y, while treated patients only have 10% chance. Some patients, 20% of the population, also have disease V, which directly affects Y: it increases the probability of Y by 5%.

In this example the true ATT is -.15, as X reduces the chance of Y by 15%. Our causal graph dictates that X and V be marginally independent, hence this this effect is homogeneous across the levels of V. (Otherwise V would become predictive of X and it would become a confounder. Confounding is discussed in experiments III-V.) All methods estimated the ATT correctly, because ATT does not depend on V. We can demonstrate this by stratifying on V and using the marginal independence of X and V.

$$
\begin{aligned}
ATT &= \mathbb{E}\left[P(Y|X=1) - P(Y|X=0)\right] \\
&= \sum_{v \in V} P(V=v)\left[P(Y|V=v, X=1) - P(Y|V=v, X=0)\right] \\
&= \sum_{v \in V}\left[P(Y, V=v|X=1) - P(Y, V=v|X=0)\right] \\
&= P(Y|X=1) - P(Y|X=0)
\end{aligned}
$$

where v denotes the levels of V. The marginal independence of X and V is used in step three:

$$
P(Y|V,X) = \frac{P(Y,V,X)}{P(V,X)} = \frac{P(Y,V|X)P(X)}{P(X,V)} = \frac{P(Y,V|X)}{P(V)}.
$$

II. Indirect Causal Effect. The setup for this experiment is the same as for the 'Direct Causal Effect' experiment, except we have disease U instead of V. Just like Z, disease U is also treated by X, but U has no direct effect on Y; its effect is indirect through X. U is thus independent of Y given X. The true ATT continues to be -.15.

Again, the ATT does not depend on U, hence all methods estimated it correctly. To demonstrate that ATT does not depend on U, we use stratification and the conditional inde-

pendence of Y and U.

$$
\begin{aligned}
ATT &= \mathbb{E}\left[P(Y|X=1) - P(Y|X=0)\right] \\
&= \sum_{u \in U}\left[P(Y|U=u, X=1)P(U=u|X=1)\right. \\
&\qquad \left. -P(Y|U=u, X=0)P(U=u|X=0)\right] \\
&= \sum_{u \in U}\left[P(Y|X=1)P(U=u|X=1)\right. \\
&\qquad \left. -P(Y|X=0)P(U=u|X=0)\right] \\
&= P(Y|X=1)\sum_{u} P(U=u|X=1) - \\
&\qquad P(Y|X=0)\sum_{u} P(U=u|X=0) \\
&= P(Y|X=1) - P(Y|X=0)
\end{aligned}
$$

III. Confounding. In this experiment, we consider the simplest case of confounding, involving a single disease Z, a single treatment X and outcome Y. 20% of the patients have disease Z and 95% of the diseased patients are treated with X, while 5% are not. All treated patients have Z. 25% of the untreated patients ($Z=1$ and $X=0$) have outcome Y; 10% of the treated patients ($Z=1$ and $X=1$) have the outcome; and only 5% of the healthy patients ($Z=0$) have it. The true ATT is -.15.

In the presence of confounding, the counterfactual confidence and ATT do not coincide. With z denoting the levels of Z and $P(z)$ being a shorthand for $P(Z=z)$,

$$
\begin{aligned}
ATT &= \mathbb{E}\left[P(Y|X=1) - P(Y|X=0)\right] \\
&= \sum_{z} P(z)\left[P(Y|X=1, z) - P(Y|X=0, z)\right],
\end{aligned}
$$

while the counterfactual confidence (CC) is

$$
\begin{aligned}
CC &= P(Y|X=1) - P(Y|X=0) \\
&= \sum_{z}\left[P(Y|X=1, z)P(z|X=1)\right. \\
&\qquad \left. -P(Y|X=0, z)P(z|X=0)\right].
\end{aligned}
$$

When $P(z|X) \neq P(z)$, these quantities do not coincide. However, any method that can estimate $P(Y|X, Z)$ for all levels of Z and X will arrive at the correct ATT estimate. We used logistic regression in our implementation of the Direct Adjustment method, which can estimate $P(Y|X, Z)$ when X and Z have no interactions. Note that the causal graph admits interaction between X and Z, thus model misspecification can cause biases in the estimate.

IV. Confounding with Indirect Effect. In addition to the Confounding experiment, we also have an indirect causal effect from U. We now have two diseases, Z and U, each of which can be treated with X. 20% of the population has Z and independently, 20% has U. 25% of the patients who have Z and have no treatment ($X=0$) have Y, while only 10% of the treated ($X=1$) patients have it, regardless of whether the patient has U. (If the probability of Y was affected by U, it would be another confounder, rather than have an indirect effect.)

X has a beneficial ATT of -.15 in patients with $Z == 1$ (and $X == 1$) and has no effect in patients with $Z = 0$ (who get X because of U). Thus the true ATT=-.0833.

In this experiment, the counterfactual model was the best-performing model. The counterfactual model estimates the

ATT through the definition

$$\mathrm{ATT} = \mathbb{E}\left[\mathrm{P}(Y_1|X=1) - \mathrm{P}(Y_0|X=1)\right],$$

where Y_0 is the potential outcome the patient would have without treatment $X = 0$ and $\mathrm{P}(Y_0|X=1)$ is the counterfactual probability of Y (the probability of Y had they not received X) in the population who actually got $X = 1$. Note that the potential outcome $Y_1|X=1$ in the patients who actually got $X = 1$ is the observed outcome $Y|X = 1$. With u and z denoting the levels of U and Z, respectively and $\mathrm{P}(u)$ being a shorthand for $\mathrm{P}(U = u)$,

$$
\begin{aligned}
ATT &= \mathbb{E}\left[\mathrm{P}(Y|X=1) - \mathrm{P}(Y_0|X=1)\right] \\
&= \sum_u \sum_z \mathrm{P}(u,z)\left[\mathrm{P}(Y|X=1,u,z) - \mathrm{P}(Y_0|X=1,u,z)\right] \\
&= \sum_z \mathrm{P}(z)\sum\left[\mathrm{P}(Y|X=1,z) - \mathrm{P}(Y_0|X=1,z)\right] \\
&= \sum_z \mathrm{P}(z)\sum\left[\mathrm{P}(Y|X=1,z) - \mathrm{P}(Y|X=0,z)\right],
\end{aligned}
$$

which coincides with the data generation mechanism, hence the estimate is correct.

In the derivation, step 2 holds because U and Z are independent given X and step 3 uses the fact that the counterfactual model estimates $P_0(Y|X=1,z,u)$ from the untreated patients, thus

$$\mathrm{P}(Y_0|X=1,z,u) = \mathrm{P}(Y|X=0,z,u) = \mathrm{P}(Y|X=0,z).$$

V. Confounding with Direct and Indirect Effects. In this experiment, we have three diseases: our index disease Z, which is a confounder; U having an indirect effect on Y via X; and V having a direct effect on Y. 20% of the population has each of Z, V and U independently. 95% of patients with Z or U get the intervention X. 25% of the untreated patients with Z get Y, while only 10% of the treated patients do, regardless of whether they have U. Patients with V face a 5% in their chance of experiencing outcome Y. X has a beneficial ATT of -.15 in patients with $Z = 1$ and have no effect in patients with $Z = 0$ (who get X because of U). Whether a patient has V does not influence the effect of X. The true ATT is thus -.0833. None of the methods estimated the effect correctly, but Propensity Score Matching came closest. Analytic derivation of why it performed well is outside the scope of this paper, but in essence, its success is driven by its ability to maximally exploit the independence relationships encoded in the causal graph. It can ignore V when it constructs the propensity score model, because X and V are independent (when Y not given); and it can ignore U and V when it computes the ATT in the propensity matched population. On the other hand, the causal graph admits interaction among U, Z and X, thus a logistic regression model as the propensity score model can be subject to model misspecification. The Stratified Non-Parametric method, which is essentially just a direct implementation of the definition of ATT, underestimated the ATT by almost 25%. The reason lies in the excessive stratification across all combinations of the levels of U, V, and Z. Even with just three variables, most strata did not have sufficiently many patients (either treated or untreated) to estimate $\mathrm{P}(Y|X,u,v,z)$. In the discussion, we will describe remedies to overcome this problem.

	Conf	CC	DA	CM	PSM	SN
I.	+.110	-.150	-.150	-.150	NA	-.150
II.	+.099	-.150	-.150	-.150	-.151	-.149
III.	+.099	+.047	-.136	-.136	-.136	-.136
IV.	+.077	+.024	-.019	-.083	-.068	-.064
V.	+.072	+.038	-.037	-.105	-.074	-.067

Table 3: The ATT estimates by the 6 methods in the five experiments. The experimental conditions, the full names of the methods and the true ATT value are specified in the text.

8   Discussion And Conclusion

We proposed the causal rule mining framework, which transitions pattern mining from finding patterns that are associated with an outcome towards patterns that cause changes in the outcome. Finding causal relationships instead of associations is absolutely critical in health care, but also has appeal beyond health care.

The numerous biases that arise in establishing causation make quantifying causal effects difficult. We use the Neyman-Rubin causal model to define causation and use the potential outcome framework to estimate the causal effects. We correct for three kinds of potential biases: those stemming from direct causal effect, indirect causal effect and confounding. We compared five different methods for estimating the causal effect, evaluated them on real and synthetic data and found that three of these methods gave very similar results.

We have demonstrated on real clinical data that our proposed method can effectively enumerate causal patterns in a large combinatorial search space due to the two new pruning methods we developed for this work. We also demonstrated that the patterns discovered from the data were very rich and we managed to illustrate how the effect of statin is different in various subpopulations. The results we found are consistent with the literature but go beyond what is already known about statin's effect on the risk of diabetes.

The discussions and experimental results provided in this paper provide some general guidance on when to use the different methods we described. We recommend counterfactual confidence if no confounding is suspected as counterfactual confidence is computationally efficient and can arrive at the correct solution even when direct effects and indirect effects are present. In the presence of confounding, propensity score matching gave the most accurate results, but due to the need to create a matched population, it has built-in randomness, increasing its variance. Moreover, the counterfactual model as well as the propensity score model are susceptible to model misspecification. If unknown interactions among variables are suspected, we recommend the stratified non-parametric method. With this technique, model misspecification is virtually impossible, however, its sample size requirement is high. The stratified model is suboptimal if we need to stratify across many variables. Stratifying across many variables can fragment the population into many strata too small to afford us with the ability to estimate the effects correctly. If the estimates use some strata but not others, they may be biased.

9 References

[1] National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the united states, 2011. *Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention*, 201, 2011.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[3] J. Sekhon. The neyman-rubin model of causal inference and estimation via matching methods. *The Oxford handbook of political methodology*, pages 271–299, 2008.

[4] R. Huupponen and J. Viikari. Statins and the risk of developing diabetes. *BMJ*, 346, 2013.

[5] V. Didelez and I. Pigeot. Judea pearl: Causality: Models, reasoning, and inference. *Politische Vierteljahresschrift*, 42(2):313–315, 2001.

[6] P. Rosenbaum. *Observational studies*. Springer, 2002.

[7] D. Freedman. From association to causation via regression. *Advances in applied mathematics*, 18(1):59–110, 1997.

[8] W. Bielby and R. Hauser. Structural equation models. *Annual review of sociology*, pages 137–161, 1977.

[9] J. Robins, M. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pages 550–560, 2000.

[10] C. Granger. Some recent development in a concept of causality. *Journal of econometrics*, 39(1):199–211, 1988.

[11] F. Jensen. *An introduction to Bayesian networks*, volume 210. UCL press London, 1996.

[12] F. Elwert. Graphical causal models. In *Handbook of causal analysis for social research*, pages 245–273. Springer, 2013.

[13] K. Murphy. Dynamic bayesian networks. *Probabilistic Graphical Models, M. Jordan*, 2002.

[14] D. Heckerman. A bayesian approach to learning causal networks. In *UAI*, pages 285–295. Morgan Kaufmann Publishers Inc., 1995.

[15] D. Heckerman. Bayesian networks for data mining. *Data mining and knowledge discovery*, 1(1):79–119, 1997.

[16] G. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224, 1997.

[17] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2-3):163–192, 2000.

[18] S. Mani, P. Spirtes, and G. Cooper. A theoretical study of y structures for causal discovery. *arXiv preprint arXiv:1206.6853*, 2012.

[19] P. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[20] J. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

[21] P. Austin and E. Stuart. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical methods in medical research*, page 0962280215601134, 2015.

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

[23] P. Austin and E. Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

[24] D. Lambert and D. Pregibon. More bang for their bucks: Assessing new features for online advertisers. In *Workshop on Data mining and audience intelligence for advertising*, pages 7–15. ACM, 2007.

[25] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. Lambert. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–16. ACM, 2010.

[26] J. Li, T. Le, L. Liu, J. Liu, Z. Jin, and B. Sun. Mining causal association rules. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pages 114–123. IEEE, 2013.

[27] P. Holland and D. Thayer. Differential item performance and the mantel-haenszel procedure. *Test validity*, pages 129–145, 1988.

[28] J. Li, T. Le, L. Liu, J. Liu, Z. Jin, B. Sun, and S. Ma. From observational studies to causal rule mining. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):14, 2015.

[29] J. Li, S. Ma, T. Le, L. Liu, and J. Liu. Causal decision trees. *arXiv preprint arXiv:1508.03812*, 2015.

[30] B. Goethals. Survey on frequent pattern mining. *Univ. of Helsinki*, 2003.

[31] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu. Freespan: frequent pattern-projected sequential pattern mining. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 355–359. ACM, 2000.

[32] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[33] M Regina Castro, Gyorgy Simon, Stephen S Cha, Barbara P Yawn, L Joseph Melton III, and Pedro J Caraballo. Statin use, diabetes incidence and overall mortality in normoglycemic and impaired fasting glucose patients. *Journal of General Internal Medicine*, pages 1–7.