Causal Discovery of Adverse Drug Events in Observational Data

Aubrey Barnard

PhD Dissertation
Computer Sciences
University of Wisconsin–Madison
2019

Version 2019-12-22.svn-4410.

Creative Commons Attribution 4.0 International License

Freedoms You are free to:

- **Share**: Copy and redistribute the material in any medium or format.
- **Adapt**: Remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Terms Under the following terms:

- **Attribution**: You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **No additional restrictions**: You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices

- You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.
- No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Abstract

Automatic causal discovery without experiments offers to accelerate scientific investigation and knowledge acquisition, for example, by searching databases of electronic health records to discover the unknown effects of drugs. However, effective causal discovery requires methods that control for confounders and that scale to large data sets which have the power to support or refute causal hypotheses. Accordingly, this dissertation first introduces a method for efficiently learning formal structural causal models of medical histories via parameter learning in log-linear temporal Markov networks. Such models work well when all of the effects of interest are already defined and measured, but it might not be the case that all possible effects are suspected beforehand, especially when considering the adverse effects of drugs. Therefore, this dissertation next develops machine learning methods for causal discovery, including differential classification and temporal inverse probability weighting, that hypothesize likely causal effects while analyzing controlled observational studies. Applying all of these methods to causal modeling and finding adverse drug effects in synthetic and real-world electronic health records demonstrates their ability to accurately discover causal effects despite the irregularity, noise, and sparsity of such data. This dissertation thus establishes (1) that scalable, causal methods discover causal effects more accurately than methods that ignore causality, do not scale to large databases, or are not robust to the messiness of medical data, and (2) that methods that hypothesize effects improve genuine causal discovery by avoiding the limitations of human bias. In summary, the methods herein distinguish themselves by bridging machine learning and epidemiology: they bring causal inference and observational studies to machine learning, and they apply learning techniques and formal causal models to tasks in epidemiology. By integrating multiple approaches to causality, these methods achieve a wider perspective that overcomes the limitations of the individual perspectives, and leads to new methods for automatic causal discovery from observational data.

Overview

Adverse drug events (ADEs) have a high impact on society, costing lives and an estimated $30 billion per year in the USA alone (Sultana et al., 2013). While ADEs are a serious and widespread health problem, their rarity makes them hard to detect. Big data sets are needed to detect such rare occurrences, but typically the only ones available that are large enough are databases of electronic health records (EHR). But EHR databases are noisy, contain only an incomplete, approximate view of a patient, and are observational, meaning any conclusions drawn from the data may be confounded. Consequently, robust yet

sensitive causal discovery methods are needed, ones that can control for confounding and are scalable enough to search large databases for evidence of unknown causal relationships. The following chapters describe several methods for causal discovery that aim to meet these criteria.

Chapter 1 introduces the problem, and Chapter 2 provides some background on causal inference, ADE discovery, and the ADE identification task set by the Observational Medical Outcomes Partnership (OMOP).

Chapter 3 introduces temporal Markov networks (TMNs), which are undirected, log-linear probabilistic graphical models. Through their use of flexible feature functions, TMNs are especially suited to modeling irregular, sparse, and noisy sequences of events such as patient histories in EHRs. By fitting the parameters of an undirected TMN to a data set of event sequences, one can learn the structure of a directed causal model, in this case a dynamic Bayesian network (DBN). One can then query the TMN or the derived structural causal model to estimate the likelihood of various potential ADEs. TMNs are scalable to large data sets because only one pass over the data is required to compute the necessary sufficient statistics. Fitting a TMN is reasonably quick and results in a globally optimal model because it formulates the structure learning problem as a smooth, convex optimization. This is in contrast to the combinatorial complexity and susceptibility to local optima of typical Bayesian network structure learning algorithms. However, like all observational studies and most machine learning methods, possible ADEs must already be suspected, defined, and measured.

Chapter 4 describes a method for identifying and hypothesizing ADEs by learning rules via inductive logic programming. The rules are learned directly from the relations (tables) in a database, without needing to transform the data into a single table as is required for most machine learning methods. In the context of a self-controlled study design based on a drug, the learned rules describe commonalities in patients that could be adverse effects. A causally-motivated scoring function then evaluates the ADE likelihood of each rule, thereby identifying known ADEs and hypothesizing new ones.

Chapter 5 explores a wider range of causally-motivated scores based on temporal dependence, comparing them to computational epidemiology methods on the task of identifying ADEs. The effects of confounding are reduced by adjusting the scores with a Markov network fit to the data.

Chapter 6 presents temporal inverse probability weighting (IPW), a method for causal discovery that extends regular IPW to the analysis of controlled before–after studies with off-the-shelf machine learning classifiers. The before–after studies compare two treatments to controls and reduce confounding through self-control and the matching inherent in similar treatments. This makes the method especially suited to discovering ADEs that are specific to the generic version of a drug. Because ADEs in generics are likely to be unanticipated, the method is able to hypothesize new effects as it learns to differentiate between brand and generic versions of a drug.

Preface

As a computer scientist and free / open software enthusiast, I am gladly obliged to have a version of this document publicly available on the world wide web at https://www.cs.wisc.edu/~barnard/dissertation.pdf. Compared to the required graduate school deposit format, it is nicely formatted as a book with appropriate chapter and section headings, and will be updated to incorporate corrections as needed. Accordingly, if you come across any errors, please send them to barnard@cs.wisc.edu.

Relationship to Published Material

Parts of this dissertation draw heavily on previously published work. Chapter 3 is based on Barnard and Page (2018), and Chapter 4 is based on Page et al. (2012). The other chapters are based on manuscripts that have been prepared for submission but not yet published.

Acknowledgments

This work has been supported by many people in many different ways over the years. Most significantly, I am indebted to my other, Erin, who has provided the missing cogs for this machine, but I also would not have gotten it up and running without the faith and encouragement of the rest of my family—Burton Barnard, Beatrice Bigony, and the Adlers. I especially thank Professor Bigony for all of her editing work, and accept full responsibility for any of her carefully plead suggestions that I have ignored.

Of others who have also provided emotional support and advice, I would particularly like to thank Tycho Andersen, Tristan Ravitch, Bess Jacques, and my other bike buddies and ski bums for helping keep the whole endeavor rolling and for helping get me back on the lift after everything went downhill. You inspire me in how you have followed your passions.

In the anointed arcades of academia, many have demonstrated what it means to be a researcher and scholar or how to earn a Ph.D., and I owe much of my success to following their example. I am especially grateful to: Scott Alfeld for his encouraging mentorship, theoretical bravery, and commitment to academic values; Finn Kuusisto for his ceaselessly positive attitude and willingness to bounce ideas; Eric Lantz for demonstrating such stoic dedication and determination; Kendrick Boyd for his candor and compassion; Ameet Soni for his generosity and cool routine; Taha Bahadori for showing me an alternative approach to research and how it can be done with an abundance of genuine enthusiasm; and, fondly, Vítor Santos Costa for exemplifying curiosity, work ethic, and collaboration.

Most importantly, I am deeply thankful to David Page for his guidance and generous support, without which none of this work would have happened. Likewise, I am thankful to my committee for their efforts and feedback, and I aspire to their standard of scholarship in my future academic endeavors.

Lastly and no lessly, I appreciate the large contributions of luck and coffee!

Committee

- C. David Page, Jr., advisor, Professor, Biostatistics and Medical Informatics, Computer Sciences
- Mark Craven, Professor, Biostatistics and Medical Informatics, Computer Sciences
- Xiaojin (Jerry) Zhu, Professor, Computer Sciences
- Theodoros Rekatsinas, Assistant Professor, Computer Sciences
- Po-Ling Loh, Associate Professor, Statistics

Contents

Abstract i

Preface iii

Contents v

1 Introduction 1
1.1 Thesis . . . 2

2 Background 5
2.1 Experiments and Observational Studies . . . 5
2.2 Causal Inference . . . 6
2.3 Inductive Logic Programming (ILP) . . . 9
2.4 The OMOP ADE Identification Task . . . 9

3 Causal Structure Learning via Temporal Markov Networks 13
3.1 Introduction . . . 13
3.2 Background . . . 16
3.3 Temporal Markov Networks . . . 17
3.4 Experiments . . . 21
3.5 Supplementary Experimental Details . . . 26
3.6 Discussion . . . 27
3.7 Conclusion . . . 31

4 Identifying ADEs using Relational Learning 33
4.1 Introduction . . . 33
4.2 Reverse Machine Learning for ADE Surveillance . . . 35
4.3 Experiments . . . 41
4.4 Conclusion . . . 44

5 Identifying ADEs using Markov Networks and Temporal Dependence 47
5.1 Empirical Causal Discovery . . . 47
5.2 Adverse Drug Event Detection . . . 49
5.3 Markov Network Model . . . 50
5.4 Temporal Scoring . . . 51
5.5 OMOP Task, Data, and Methods . . . 52

5.6 Experiments 53
5.7 Discussion 55
5.8 Conclusion 58

6 Temporal IPW for Discovering ADEs Especially in Generic Drugs **59**
6.1 Introduction 59
6.2 ADE Discovery 61
6.3 Methods for Finding Differential Effects of a Generic Drug 64
6.4 Results 70
6.5 Conclusion 78

7 Conclusion **79**

References **81**

Chapter 1

Introduction

For millenia, humans have wondered why things happen. Led by curiosity, they have observed and manipulated their environments in an effort to understand the causal mechanisms underlying reality. Yet, it is only over the last one and one-half centuries—well after the Renaissance, the Scientific Revolution, and the Age of Enlightenment—that the study of causality has gone beyond natural laws (Newton, 1687) and philosophy (Hume, 1740), and has begun to be formalized in fields such as statistics (Good, 1961; Suppes, 1970), epidemiology (Hill, 1965), and artificial intelligence (Pearl, 1988). Within AI, researchers have argued that incorporating causality into learning and reasoning algorithms is crucial for reliable progress to occur (Pearl, 2009; Peters et al., 2017; Bengio et al., 2019). Despite many foundational advances (e.g., Pearl, 1988; Spirtes et al., 2000; Pearl, 2009), accurate causal learning remains a challenging task, in part owing to the combinatorial nature of relationships between many variables, the difficulties of probabilistic reasoning and statistical inference, and the intrinsic limitations of observation as a source of information. These challenges define the core problems of computational causal discovery from observational data.

One application of computational causal discovery is discovering adverse drug events (ADEs) in electronic health records (EHR) data. The pharmaceutical industry, consumer protection groups, takers of medications, and government oversight agencies are all strongly interested in identifying adverse reactions to drugs. Adverse reactions may account for 10–30% of hospital admissions, with estimated costs of $30 billion or more annually in the U.S. alone (Lazarou et al., 1998; Sultana et al., 2013). Soberingly, half of the 180k annual life-threatening or fatal ADEs could have been prevented (Gurwitz et al., 2003). Although the U.S. Food and Drug Administration (FDA) and its counterparts elsewhere have rigorous approval processes for drugs that involve clinical trials, such processes cannot uncover everything about a drug. A clinical trial might enroll one thousand patients, but millions of patients may take a drug once it is released on the market. As a result, in many cases ADEs are observed in this larger, more diverse population that were not identified during the clinical trials. The scale of this problem and the unreliability of patient reporting create a need for continuous, automatic postmarketing surveillance of drugs to identify previously unanticipated ADEs.

However, the only reasonable data sets for such surveillance are observational, not experimental (see Chapter 2). It would be unethical to conduct a randomized controlled

trial to find ADEs if a drug were suspected of being substantially harmful. Moreover, the rarity of ADEs makes large sample sizes necessary for detection, but trials of the appropriate scale would be prohibitively expensive. Thus, EHR or health insurance claims databases are typically the only data sets available that are ethical and large enough. But the observational nature of the data means that any inferences drawn from the data may be confounded. So, for an algorithm to successfully detect ADEs, it must be able to adjust for confounding.

Observational data is normal for machine learning, but in EHR databases the data is also temporal and relational. Since most standard machine learning algorithms expect a single table of data, either the data needs to be transformed into a single table by joining or aggregating, which ruins the frequencies or details of the information, or specialized algorithms are required that are able to learn from the data in its native format. Furthermore, the data in EHR databases was collected for medical care and billing, not research, so the data presents a limited, sporadic view of patients, one that is often inaccurate. In order to successfully detect ADEs in such large, noisy data, an algorithm must be scalable and robust yet sensitive.

Most approaches to causal inference, including observational studies and structural causal models, assume that the possible effects are already defined, but in many cases the possible ADEs are not known nor even suspected. In such cases of genuine causal discovery, the possible effects must be hypothesized as well as estimated. But even algorithms that can hypothesize effects must be careful because any events that are associated with both the exposure and the outcome are likely to be hypothesized as part of an effect, but they might actually be confounders. For example, a machine learning algorithm might see more myocardial infarctions (MIs) in patients taking beta blockers compared to patients not taking beta blockers, and hence hypothesize that beta blockers are a cause of MI when in fact they help prevent MI-related deaths.

In total, to overcome these difficulties of ADE discovery, a method must be causally-aware and able to handle confounders, scalable and able to handle the messiness of EHR data, and have good statistical properties: sensitivity, specificity, and robustness. Additionally, when the goal is genuine discovery rather than detection of suspected ADEs, a method must be able to hypothesize effects.

Unsurprisingly given these requirements, previous work has only addressed parts of the task. Standard machine learning algorithms were never designed to infer causal relationships. While designed for causal inference and theoretically justified, structural causal models are not known for being scalable, in terms of data size or the number of variables. Observational studies are also theoretically justified and they scale well, but, like the rest, it is left to the investigator to hypothesize effects. Relational rule learning can hypothesize effects but it tends to be statistically unsophisticated and not very scalable. Clearly, all of these approaches have shortcomings. But they also have strengths that, when combined, may lead to new approaches with the potential to overcome many of the existing challenges of causal discovery.

1.1 Thesis

Motivated by the innateness of causal reasoning, the relevancy of discovering ADEs, and the natural occurrence and ubiquitousness of observational data, this work introduces methods

for causal discovery that support the following, broad argument:

> Methods that are causal, scalable, and that can learn from irregular, sparse, and noisy sequences of events will discover effects more accurately than methods that ignore one or more of those aspects. Furthermore, methods that can hypothesize effects will improve genuine causal discovery by overcoming the limitations of human imagination and bias.

In particular, causal relationships can be learned scalably from event sequences by learning temporal Markov networks (Chapter 3), but, like all structural causal models, they expect all effects to already be defined. To overcome this restriction, ADEs can be invented with relational rule learning (Chapter 4). Coupled with causally-aware and efficient-to-evaluate scores (Chapter 5), relational rule learning does well at recovering known ADEs and drug–drug interactions. Unfortunately, the learned rules often contain indications for the drug and other potential confounders. Analyzing before–after studies with temporal inverse probability weighting (Chapter 6) offers better control of confounding while still being able to invent effects and efficiently process messy EHR data. Overall, these methods bring together ideas from observational studies, formal causal modeling, and machine learning to improve on the tools available for accurate causal discovery from observational data.

Chapter 2

Background

Research into causal discovery from observational data has a rich history, of which Holland (1986) gives a good overview. Over time, two main schools of thought have developed: potential outcomes (Imbens and Rubin, 2015; Hernán and Robins, 2020) and structural causal models (Pearl, 2009; Peters et al., 2017). The framework of potential outcomes provides the formal grounds in statistics for drawing causal inferences from the results of experiments or nonexperimental (observational) studies. Structural causal models extend this framework to formal, statistical models of general causal systems, but were developed in other fields such as genetics, economics, and artificial intelligence.

2.1 Experiments and Observational Studies

The "gold standard" for causal evidence is an experiment where the conditions are carefully controlled to ensure comparability between experimental groups when the outcome is measured. Control can be accomplished by intervention, making the experimental conditions fixed and uniform, or by randomization, which ensures that assignment to treatment is not influenced by outside factors and which helps to avoid systematic differences between experimental units. Under both intervention and randomization, the goal is to make sure that the only differences between study groups are solely due to the treatments and not any other factors. Any such factor that can affect both the treatment and the outcome is a confounder. Confounders are a problem because they distort the measurement of the effect that the treatment has on the outcome if they are not properly accounted for. It is because experiments cannot be confounded that they are the best causal evidence.

However, there are situations where intervention and randomization are unethical or impossible. For example, it is unethical to assign (randomly or otherwise) a person to a study group where the treatment is known to be harmful, like smoking. Similarly, it is impossible to study the weather by intervening to control the global atmospheric conditions. In such situations, an investigator must use an observational study.

Observational studies, also known as nonexperimental studies or quasi-experimental designs, compare two or more study groups based on data collected by observation only. Because they are nonexperimental, observational studies must control for confounding after the fact. To adjust for potential confounders, they try to make the study groups as similar

as possible and avoid any systematic differences between groups. This can be done by matching, stratification, or propensity score methods.

However, without sufficient knowledge of the true underlying causal system, the investigator's analysis risks going astray. This can happen by adjusting for variables that are not actually confounders, which introduces bias into the measurement of the effect, or this can happen by not observing variables that are confounders, which means they cannot be adjusted for. Without omniscience, the possibility of unobserved confounders always exists, meaning that the investigator can never be sure that estimates are unbiased and true, although sensitivity analyses can bound the potential effects of confounders. This unescapable doubt about the true causality is what makes observational studies lesser causal evidence.

2.2 Causal Inference

The complexities of causal inference described above can best be illustrated in the context of a concrete example. Imagine an investigator wants to measure the effect of a treatment X, administering antipsychotics to a patient, on an outcome Y, whether that patient has a heart attack within a week. Both X and Y are binary random variables taking values in $\{0, 1\}$, where $X = 1$ means antipsychotics were administered, $Y = 1$ means the patient had a heart attack, and zero means the opposite in each case.

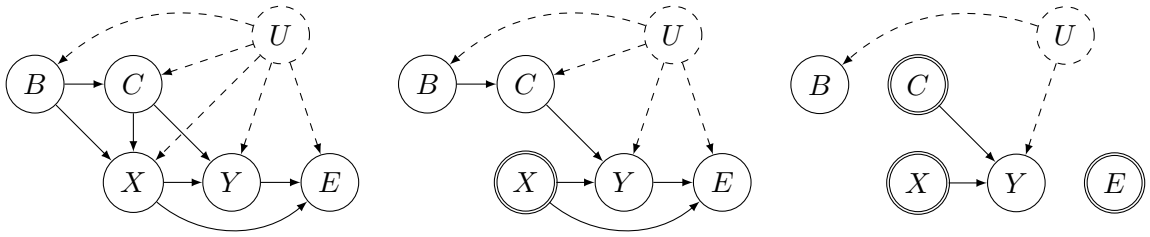
2.2.1 Potential Outcomes

In this scenario, the investigator wants to know whether X causes Y, or, equivalently, whether Y depends on X, expressed as $Y := f(X)$ or sometimes as $Y(X)$. Applying experience from the real world, the investigator might try to determine f, the relationship between X and Y, by comparing the outcomes Y under different settings of X, perhaps by (unethically) convincing a patient to take the drugs. The patient takes the antipsychotics and has a heart attack, but would not have had a heart attack had they not taken antipsychotics. For this patient $Y(X = 1) = 1$ and $Y(X = 0) = 0$. Now imagine a second patient who takes antipsychotics and has a heart attack, but that would have had a heart attack regardless of the antipsychotics. For the second patient $Y(X = 1) = 1$ but $Y(X = 0) = 1$. Based on this knowledge, the individual causal treatment effect, $Y(1) - Y(0)$, for the first patient is 1, but the individual effect for the second patient is 0. Were the investigator to have this knowledge, they could answer their question about X and Y, at least for these two patients.

However, a patient can have only one of the treatments at a time, and can only experience one of the outcomes given that treatment.[1] Thus, the investigator cannot measure the individual treatment effect because it depends on the result of something that, in fact, did not happen, a counterfactual. Nevertheless, by assuming that different patients will respond in the same way to the same treatment,[2] the investigator can allow the outcomes of other patients to stand in for the unobservable outcomes, and thereby determine the causal effect of X on Y with respect to the average in that group. Because of the uncertain nature of the outcomes, this framework for causal inference is known as "potential outcomes," or sometimes "counterfactual outcomes."

[1] The same patient at different times is considered to be separate experimental units.

[2] This is known as the stable-unit-treatment-value assumption or SUTVA (Rubin, 1980; Hernán and Robins, 2020), and is typically unrealistic.



(a) Observational study (b) Randomized controlled trial (c) Experiment

Figure 2.1: Causal models illustrating the variables and relationships in observational studies, randomized controlled trials, and experiments.

2.2.2 Structural Causal Models

While potential outcomes provide the framework for causal inference in statistics, it does not immediately have much to say about the context of X and Y, and that is where structural causal models come in. Returning to the antipsychotics and heart attack scenario, imagine the true underlying causal system has effects according to the arrows in Figure 2.1(a). Variables besides X and Y are also in the environment: B is a set of background variables that cannot be manipulated (like genetics), C is a set of observed confounders, U is a set of unobserved confounders, and E is a set of shared, downstream effects (but that the investigator suspects might affect X and/or Y). This causal system has an equivalent functional representation known as a structural equation model, even though the "equations" are actually formulas that represent one-way causal relationships, not algebraic identities. The formulas for Figure 2.1(a) follow, where the εs are noise variables that incorporate unobserved influences and intrinsic random variation.

$$B := f_B(\varepsilon_B) \quad (2.1) \qquad C := f_C(B, \varepsilon_C) \quad (2.2)$$
$$X := f_X(B, C, \varepsilon_X) \quad (2.3) \qquad Y := f_Y(C, X, \varepsilon_Y) \quad (2.4)$$
$$E := f_E(X, Y, \varepsilon_E) \quad (2.5)$$

In a randomized controlled trial, units are randomly assigned to the various treatments, and X is fixed at those treatment values ("levels"). This manipulation of X is indicated by the doubled circle in Figure 2.1(b). Randomization removes the possibility of any outside influences on the value of X, which, once the value x of X has been fixed, is expressed by replacing Equation 2.3 with $X := x$. Naturally, manipulating X does not stop it from influencing other variables, and those effects are what one wants to measure.

In an experiment with a controlled environment, steps are taken to fix all the suspected influences on Y and thereby isolate it from the environment as shown in Figure 2.1(c). After substituting fixed values, the updated formulas for the causal system are:

$$B := f_B(\varepsilon_B) \quad (2.6) \qquad C := c \quad (2.7)$$
$$X := x \quad (2.8) \qquad Y := f_Y(c, x, \varepsilon_Y) \quad (2.9)$$
$$E := e \quad (2.10)$$

Y is now a function of specific values, which allows the investigator to focus on estimating the form of f_Y. Furthermore, the only sources of noise are now ε_Y and whatever variation is intrinsic to f_Y. While Y can still be influenced by unknown outside factors (through ε_Y) in a study with interventions, those factors cannot confound the relationship between X and Y because they cannot affect X.

On the other hand, no such interventions can be done in an observational study, leaving an investigator or algorithm to handle the complications of all the relationships in Figure 2.1(a). While the effects of B and C can be controlled for, since X cannot be fixed, it is always possible that some U may affect both X and Y, leaving their true relationship in doubt. Note that E can safely be ignored because it is downstream from both X and Y. Indeed, adjusting for E *introduces* bias because conditioning X and Y on a particular value of E introduces dependence between X and Y. This consequence can be seen by letting the functional form of each formula be a conditional probability distribution and then comparing $P(Y \mid X)$ to $P(Y \mid X, E = e)$.

Letting each function be a conditional probability distribution results in a Bayesian network (BN). For example, the BN for the graph in Figure 2.1(a) has a distribution for each variable that is conditioned on its parents in the graph, and the joint distribution is the product of the individual conditional distributions:

$$P(B, C, X, Y, E) = P(B)\, P(C \mid B)\, P(X \mid B, C)\, P(Y \mid C, X)\, P(E \mid X, Y) \tag{2.11}$$

Because the structure of this BN is the same as that of the causal system it represents, it becomes a structural causal model rather than merely a probabilistic graphical model. Thus, with causal BNs, probabilistic inference can be used to answer causal queries, including queries about the effects of interventions. However, queries about interventions must be asked of a BN that represents the intervention, that is, a BN where the conditional probability distribution of each manipulated variable has been replaced with its fixed value as was done previously for the formulas. Under this perspective, Figure 2.1 shows the graphs for causal BNs that correspond to studies with the given manipulations.

Probabilistic inference also answers the question of how to properly adjust for observed confounders: just answer the query $P(Y \mid X, Z)$ with a causal BN and the appropriate conditioning set Z. The structure of the BN shows what variables need to be included in Z and which of them need to be observed to make an unbiased calculation of $P(Y \mid X)$: those that separate, and thereby make independent, X and Y from all unobserved variables. Of course, this is only accurate if the structure of the BN accurately reflects the structure of the causal system.

Unobserved confounders still pose a problem that is hidden by the algebraic notation (but not the graphs). The crucial assumption of BNs is that the conditional distributions are independent from one another (that the factorization of the joint distribution is correct). The analogous assumption in structural equation models is that all the noise variables are independent. Unobserved confounders introduce dependence among the noise variables or conditional distributions, thus ruining the intended independence of the mechanisms represented by the formulas or conditional distributions.

2.2.3 Machine Learning

One might think that machine learning (ML) is immune to the considerations of causal inference, but the data for most ML tasks is observational. Consequently, researchers implicitly conduct observational studies when running ML algorithms, so their conclusions are subject to bias and confounding in the same ways as any other observational study. Thus, incorporating knowledge and techniques from causal inference into ML can only help to improve ML practice. For example, selecting features based on the underlying causal system yields sets of features that are more accurate and parsimonious (Guyon et al., 2007; Aliferis et al., 2010).

2.3 Inductive Logic Programming (ILP)

Inductive logic programming is an approach to inducing rules in the form of logical clauses. The rules are learned from multiple tables (relations) as are typically found in relational databases. Each relation contains tuples (rows) which are the facts from which rules are assembled. To form a rule, the tuples are connected through variables that represent shared identifiers or values. At least one of the variables represents the objects being classified and connects back to a set of identifiers that define the positive and negative examples for the classification task. Introductions to logical rule learning can be found in Russell and Norvig (2003), Mitchell (1997), Getoor and Taskar (2007), and De Raedt (2008).

2.4 The OMOP ADE Identification Task

2.4.1 Drug Safety Surveillance

When developing a drug, the last stage before submission for approval by the U.S. Food and Drug Administration (FDA) is a clinical trial that demonstrates the drug's efficacy and safety. Such a trial is usually conducted on a study population of about 1000 people. While 1000 people provides good statistical power for detecting common or large effects, many more people, possibly millions, will take the drug once it is on the market. Inevitably, certain people will respond to the drug in ways not discovered during the trial, possibly in adverse ways. A well-known example of such an adverse drug event (ADE) is the pain management drug rofecoxib (Vioxx) causing increased risk of myocardial infarction (heart attack).

The FDA has a voluntary reporting system, the Adverse Event Reporting System (AERS), for these types of ADEs. The AERS collects "spontaneous" reports: anyone who believes they have experienced or witnessed an ADE makes a descriptive report of the events and submits it to the FDA. The FDA monitors these reports for systematic trends and may decide to conduct a safety investigation given accumulating evidence of harm. This reporting system has many flaws centering around the subjective and sporadic nature of its reports. A better system would be more automatic, accurate, timely, and reliable.

The need for such a better system and its value in improving drug safety has been recognized for some time. In 2005, the U.S. Department of Health and Human Services asked the FDA to expand and improve its ability to monitor the performance of medical products. In

Table 2.1: OMOP task ground truth drug–condition pairs. DoI: drug of interest, HoI: health outcome of interest, GI: gastrointestinal, MI: myocardial infarction, A: ADE, n: non-ADE, B: benefit.

DoI	HoI	angioedema	acute renal failure	acute liver failure	aplastic anemia	hip fracture	upper GI ulcer	acute MI	bleeding	death after MI	hospitalization
ACE inhibitors		A			n	n	n				B
amphotericin B		n	A	n	n	n				n	
antibiotics			n	A	n	n		n	n		
antiepileptics		n	n		A		n			n	
benzodiazepines		n	n	n	n	A		n	n		
bisphosphonates			n	n	n		A	n			
tricyclic antidepressants			n	n	n			A	n		
typical antipsychotics			n				n	A			
warfarin		n	n		n	n			A	n	
beta blockers		n	n	n	n	n	n			B	

2007, Congress passed legislation[3] mandating that the FDA plan and implement an active surveillance system, which is an automatic reporting system that constantly monitors incoming data for signs of ADEs. As a result, in 2008 the FDA created the Sentinel Initiative (U.S. Food and Drug Administration, 2008) to manage this project.

One of the organizations involved in the Sentinel Initiative was the Observational Medical Outcomes Partnership (OMOP) (Stang et al., 2010). OMOP was a public–private partnership tasked with developing the framework and methods necessary for an active surveillance system for ADEs. OMOP developed a common data model for integrating various sources of medical data, collected data from several organizations, and worked on methods for analyzing the data to search for ADEs. OMOP made their data available to research partners to use in developing ADE identification methods. OMOP also made their research methods available to other organizations who wished to search their own data for ADEs. These activities now continue under the Observational Health Data Sciences and Informatics (OHDSI) program (Hripcsak et al., 2015).

2.4.2 The OMOP Task

OMOP was interested in characterizing associations between drugs and conditions to determine if they were possible ADEs. They defined ten drugs of interest (DOIs) and ten health outcomes of interest (HOIs), which are listed in Table 2.1. The Cartesian product of these drugs and conditions forms a set of pairs whose associations are of interest. Based on known ADEs and the medical literature, OMOP defined 9 of these pairs as true ADEs,

[3]The Food and Drug Administration Amendments Act of 2007 (FDAAA).

42 as non-ADEs, and 2 as known benefits. The remaining pairs were considered unlabeled. There is evidence that some of the unlabeled pairs may actually be ADEs as well, such as a possible association between ACE inhibitors and renal failure, so treating them as additional negative examples would introduce noise into the labels. Estimating the causal relationship of each drug–condition pair was the OMOP task, which was often simplified to classifying each drug–condition pair as an ADE (after ignoring the two known beneficial effects).

While the primary OMOP task was assessing the ADE likelihood of the specified drug–condition pairs, the intention was to apply successful methods to the actual problem of active ADE surveillance. Methods were applied to all pairs of drugs and conditions in a database to see what associations they could find. The most confident of these associations were then evaluated by clinicians and possibly selected for further investigation.

2.4.3 The Common Data Model

OMOP's ADE surveillance methods were designed to operate on data in the form of the common data model (CDM). The CDM consists of the types of data typically included in databases of electronic medical records (EMRs) or electronic health records (EHRs), as illustrated in Figure 2.2. All the relations (tables) except patient demographics have associated timestamps or time intervals. OMOP curated several databases in this format drawn from various sources such as administrative claims, insurance, and computerized medical records systems.

Patient Demographics

Patient	Name	DOB	DOD	Sex	Eth
012298	John Snow	1813-03-15	1858-06-16	M	W
229049	Rachel Carson	1907-05-27	1964-04-14	F	W
628136	Alan Turing	1912-06-23	1954-06-07	M	W

Vital Signs

Patient	Time	Ht	Wt	°C	BP	HR	Resp
304525	2004-05-24 19:57:50	1.42	61.2	36.7	104/64	63	13
829533	2006-02-09 05:01:56	1.64	89.6	36.8	111/74	84	16
035194	2011-09-30 20:20:52	1.96	77.9	39.3	108/66	75	12

Genetics

Patient	1	2	3	...
185169	aa	aa	aa	.
292825	aa	bb	ab	.
496246	ab	aa	bb	.

Enrollment

Patient	Start	End	Ins
438802	2002-08-21	2007-12-18	GH
621929	2008-04-20	2011-07-17	UW
896880	2010-06-26	-	KP

Conditions

Patient	Condition	Date
281715	liver failure	2001-02-11
148098	GI ulcer	2004-08-26
181791	hip fracture	2007-03-02
367732	MI	2010-12-14

Procedures

Patient	Procedure	Date
750811	ecg	2004-04-18
234100	urine	2006-07-02
774789	cbc	2009-04-16
601764	lipids	2011-05-15

Drugs

Patient	Drug	Start	End
238336	beta blockers	2005-03-06	2005-09-02
181791	benzodiazepines	2006-06-20	2007-03-17
323172	beta blockers	2008-11-18	2009-05-17
281715	warfarin	2009-12-20	2010-12-19

Laboratory Results

Patient	Lab	Date	Value	Range	Code
990513	tsh	2001-12-28	-	-	lo
684403	rbc	2003-09-17	5.7	4.2–6.1	ok
133352	wbc	2004-07-25	12.0	4.5–10.0	hi

Figure 2.2: Example of data in an electronic medical records (EMR) database.

Chapter 3

Causal Structure Learning via Temporal Markov Networks

Having introduced the typical approaches to causal discovery and the ADE identification problem, this chapter introduces causal structure learning based on dynamic Bayesian networks. Using observational data, this method learns structures scalably, by making only a single pass over the data and by formulating the structure learning problem as a convex optimization problem, which can then be solved efficiently with common optimization algorithms. The method is implemented with a flexible, log-linear modeling approach that works well with the vagaries of electronic health records data.

3.1 Introduction

To understand how events unfold, scientists often analyze time series data with dynamic Bayesian networks (DBNs) (Dean and Kanazawa, 1989). Even when conditions are ideal—the data are available with the right time intervals and the right Markov order is selected for the DBN—learning the structure of the relationships between variables is a combinatorial problem. The enormous search space (Robinson, 1973, 1977) prevents a complete search, and the non-convexity of the likelihood prevents guarantees about the quality of the solution. Thus, a heuristic or greedy search is frequently employed.

The setting above is the classic search-and-score Bayesian network (BN) structure learning setting as introduced by Cooper and Herskovits (1992). Some algorithms, such as sparse candidate (Friedman et al., 1999), choose to manage the complexity of structure learning[1] by limiting the number of candidate parents to k for each of the n nodes. The result is a subset selection problem that has *combinatorial complexity*, $\sum_{i=1}^{k} \binom{n}{i}$, which is polynomial complexity of order k but tends to exponential complexity (2^n) as $k \to n$. Searching for subsets of size at most k is certainly better than searching over all of the 2^n subsets of the nodes, but it can still be extremely limiting in domains with networks of high degree, such as in biology. Other algorithms, such as K2 (Cooper and Herskovits, 1992) and that of Shojaie and Michailidis (2010), choose to presuppose an ordering of the variables. This requires strong assumptions or background knowledge, or it just exchanges the search

[1] Throughout this chapter, "complexity" refers to the complexity of structure learning only, excluding probabilistic inference.

over directed acyclic graphs (DAGs) for a search over permutations of variables (Teyssier and Koller, 2005).

Constraint-based methods also suffer from combinatorial complexity. Markov blanket induction (Aliferis et al., 2010), the PC algorithm (Spirtes et al., 2000), and the polynomial min-max skeleton algorithm (Brown et al., 2005) all search for possible separating sets. This is the subset selection problem rederived. In these cases, greedy search over subset members can be used to address the complexity but at the loss of accuracy. Constraint-based methods have additional problems with testing multiple statistical hypotheses and with the possible cascade of errors inherent in their greedy or sequential decision-making processes.

Greedy equivalence search (GES) (Chickering, 2002) may appear to avoid all of these difficulties by guaranteeing to find the optimal equivalence class after only a forward and backward pass (assuming faithfulness and infinite data), but it is still combinatorial. The number of neighboring search states encountered at each step can be exponential because adding or deleting edges involved in V-structures again faces a subset selection problem.

I propose to avoid the combinatorial nature of these search algorithms by reformulating the structure learning problem as a smooth, convex, non-combinatorial optimization problem in a log-linear model: first use a temporal Markov network (TMN) to learn the undirected skeleton and then direct the edges with time. TMNs provide a way to handle sequences of irregular events (which is important for analyzing medical data), and the optimization jointly estimates all the edges, avoiding issues of multiple testing and sequential decisions. Further, the quality of the model can be assessed by measuring the difference between the current model and the unique global optimum. Lee et al. (2006) harness the same optimization benefits, but they only learn an undirected skeleton, one that may be biased by using a L_1-regularizer. The proposed method learns the correct structure and is unbiased in the limit of the data (Theorem 3.5, p. 20).

3.1.1 ADE Discovery

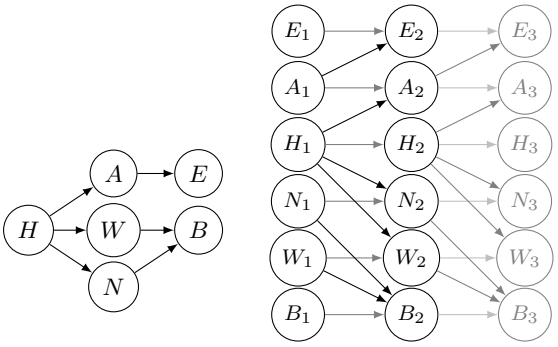
Identifying adverse drug events (ADEs) is the causal task that motivates this work. In the USA, ADEs are estimated to be the fourth leading cause of mortality, affecting more than 2 million people each year and incurring $136 billion in additional medical care (U.S. Food and Drug Administration, 2009). To combat this problem and improve patient safety, the Observational Medical Outcomes Partnership (OMOP) led research into drug safety surveillance methods by developing an ADE identification task (Figure 3.1, p. 15) and making electronic medical records (EMR) data sets (Figure 3.3, p. 25) available to researchers in a laboratory.[2]

One of the challenges of identifying ADEs is that it is an inherently causal task, and so requires appropriate methods. Causal methods fall into two broad categories: observational studies (e.g., cohort and case–control studies) and structural causal models (SCMs) (Pearl, 2009; Spirtes et al., 2000), such as causal BNs (Figure 3.1(a)). With SCMs, causal discovery becomes a structure learning problem. While most of the work on the OMOP ADE task has focused on observational studies (e.g., Ryan et al., 2012), a contribution of this work is the application of machine learning to the task: learning the structure of causal DBNs (Figure 3.1(b)).

[2] This work now continues under the Innovation in Medical Evidence Development and Surveillance (IMEDS) and Observational Health Data Sciences and Informatics (OHDSI) programs.

Table 3.1: Events (random variables) that make up the causal pairs of the OMOP ADE task. Additional negatives are non-ADE drug–condition pairs among the same events.

Drug	Condition	Label
A ACE inhibitors	E angioedema	+
T amphotericin B	R acute renal failure	+
I antibiotics	L acute liver failure	+
P antiepileptics	S aplastic anemia	+
Z benzodiazepines	F hip fracture	+
Φ bisphosphonates	U upper GI ulcer	+
D tricyclic antidepressants	M acute MI	+
Y typical antipsychotics	M acute MI	+
W warfarin	B bleeding	+
β beta blockers	X mortality after MI	–
N NSAIDs	H hypertension	



(a) Causal BN (b) Equivalent DBN

Figure 3.1: A real-world causal network and its equivalent unrolled DBN. The variables are those of the OMOP ADE task (Table 3.1).

The challenges of causal discovery are amplified in the EMR realm where the data is a messy collection of events. Patients interact with the medical system sporadically, on their own initiative, and usually only when they are ill, not when they are well. While EMR data contains thousands of variables describing the state of a patient's health, only a few are recorded at any visit. Thus, observations of a patient are irregular, subject to large time gaps, and very sparse. Furthermore, they are noisy and biased by patient health, by hospital procedure, or by convenience. Of course, EMR data is observational and so also susceptible to confounding.

3.1.2 Contributions

Learning the structure of causal DBNs is difficult due to the combinatorics of deciding which edges to include. This difficulty is worse when learning from EMR data, which is irregular, noisy, and sparse, and thus lacks the regular, full observations needed for DBN learning. Causal structure learning via TMNs addresses these problems by (1) learning the

directed structure using an undirected model, wherein the parameters indicate the edges and learning the parameters is a convex optimization problem, and (2) using features to model the irregularity, sparsity, and temporality of EMR data. As far as we are aware, combining structure learning via parameter learning and coarse temporal modeling is novel, and the results show that it is effective for causal structure learning.

3.2 Background

3.2.1 Related Work

While many methods could be used to identify ADEs in longitudinal data—ranging from graphical Granger methods (Arnold et al., 2007) to computational epidemiology (Simpson et al., 2013)—BN structure learning will be the focus here because of its potential to yield a SCM. Algorithms such as PC and fast causal inference (Spirtes et al., 2000) measure conditional independence to detect provably causal structures, but noise can affect independence tests and lead to a cascade of errors. Score-based BN structure learners (e.g., Heckerman et al., 1995) avoid these problems but are not guaranteed to learn a causal structure (although they may do so under certain conditions (Meek, 1997)). Local learners determine the neighborhood or the Markov blanket of each node before stitching them together (Margaritis and Thrun, 1999; Tsamardinos et al., 2003; Niinimäki and Parviainen, 2012). Aliferis et al. (2010) show that these "grow-shrink" algorithms can be sound and complete and therefore causal. A related algorithm learns an undirected skeleton with a local search method and then directs the edges in a greedy hill-climbing search (Brown et al., 2005).

Other, non-causal BN structure learning methods directly address the combinatorial optimization by using dynamic programming (Koivisto and Sood, 2004) or any-time, branch-and-bound search (de Campos et al., 2009). Similar linear programming approaches (Jaakkola et al., 2010; Cussens, 2011) operate in a continuous optimization space, but finding an integral solution to the relaxation may require combinatorial search.

Learning undirected structures over temporal variables, as is possible in the DBN setting where the temporal order of variables is given, opens the door to non-combinatorial structure learning algorithms. The classic example of such an algorithm is selection of Gaussian graphical models, where the zeros in the inverse covariance matrix indicate the absence of edges (Lauritzen, 1996). The same ideas have been developed for discrete variables, including methods for nodewise structure learning using L_1-regularized regression (Loh and Wainwright, 2013). In contrast, this method directly uses the zero parameters to indicate conditional independence, as in Liu and Page (2013) and Lee et al. (2006), but it is unbiased and also addresses the recovery of directed models.

3.2.2 Probabilistic Graphical Models

A probabilistic graphical model (PGM) is a model of a probability distribution over a set of random variables $X = \{X_1, \ldots, X_n\}$ that uses a fixed graph G to represent the conditional independence relationships of the distribution. In a PGM, each variable corresponds to a vertex in G.

The structure of a distribution refers to its factorization and conditional independence properties, which are related as follows (Lauritzen, 1996; Koller and Friedman, 2009; Loh

and Wainwright, 2013). These statements lay the foundation for Theorem 3.5 (p. 20), one of the main contributions of this research.

Definition 3.1 (Factorization property)**.** A distribution $P(X)$ *factorizes* according to an undirected graph G if its density can be expressed as a product of non-negative potential functions on the cliques C of G.

$$P(X) \propto \prod_{c \in C} \psi_c(X_c) \tag{3.1}$$

Definition 3.2 (Global Markov property)**.** $X_A \perp\!\!\!\perp X_B \mid X_S$ if and only if S separates A from B in G.

Theorem 3.3 (Proposition 3.8 (Lauritzen, 1996))**.** *For any undirected graph G and any probability distribution P on X, it holds that the factorization property implies the global Markov property.*

3.3 Temporal Markov Networks

This chapter introduces temporal Markov networks (TMNs), a type of log-linear PGM with feature functions for modeling timelines. TMNs are motivated by the need for a probabilistic causal model of EMR data, which does not have (1) synchronized timing of a consistently-observed set of events, as assumed by DBNs and other time series methods (e.g., Granger, 1969; Arnold et al., 2007), nor (2) detailed patient state and reliable timing of events, as needed by continuous time Bayesian networks (Nodelman et al., 2002) and piecewise-constant conditional intensity models (Gunawardana et al., 2011).

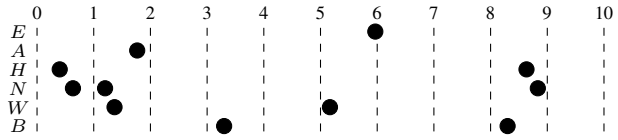
3.3.1 Timelines

A timeline (sequence) $\mathcal{S}$ is a set of random variables $X = \{X_1, \ldots, X_n\}$ that occur over a set of times $\mathcal{T}$: $\mathcal{S} = \{X_{i,t} : (X_i, t) \in X \times \mathcal{T}\}$, as in Figure 3.2(a) (p. 18). Each $X_{i,t}$ is a point event, so a timeline is equivalently a sequence of event tuples (t, X_i, x), where t is the time of occurrence, X_i is the event type, and x is its observed value. This is the form of typical EMR data, with such a sequence for each patient. This work considers only discrete times $\mathcal{T} \subseteq \mathbb{Z}_{0+}$ and binary variables[3] (event occurrences) as in Figure 3.2(b). A condensed timeline (Figure 3.2(c)) includes only observed events as a sequence of timesteps. It is constructed by ignoring empty timesteps, discarding durations between events, and treating the remaining timesteps in sequence.

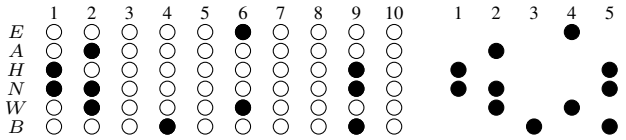
3.3.2 Log-Linear Model

Definition 3.4 (Temporal Markov network (TMN))**.** A TMN is a tuple (X, F, θ), where X is a set of event types (random variables), F is a set of binary feature functions $f_i(X_i \subseteq X) : X_i \mapsto \{0, 1\}$, and $\theta \in \mathbb{R}^{|F|}$ is set of weights corresponding to the features. A TMN defines a probability distribution over timelines S through the log-linear model in Equations 3.2

[3] While this may seem limiting, any discrete variable can always be encoded with binary indicator variables (Loh and Wainwright, 2013).



(a) Point events in continuous time (EMR)



(b) Fully-observed discrete time (DBN) (c) Condensed

Figure 3.2: Various forms of a sequence of events (timeline) as might be observed from a process like Figure 3.1(a) (p. 15).

and 3.3 (e.g., Koller and Friedman, 2009), where $f_i \in F$ and $\theta_i \in \theta$. The features must be (1) *hierarchical*: the variables in each feature define a clique and the cliques induce a graph G; in order to be hierarchical, F must contain a feature for each (sub-)clique in G (Lauritzen, 1996); and (2) *temporal*: F must include at least some features for temporal order or succession (see §3.3.3).

$$P(S = s) = \frac{1}{Z} \exp\left(\sum_i \theta_i f_i(s)\right) \tag{3.2}$$

$$Z = \sum_{s \in S} \exp\left(\sum_i \theta_i f_i(s)\right) \tag{3.3}$$

While being a log-linear model ensures that a TMN always represents a well-defined probability distribution, the additional semantics of a TMN depend on its features, as explained below.

3.3.3 Feature Functions

The following temporal indicator features model the most salient aspects of timelines as logical predicates. They are designed to capture the main effects of the events and their interactions, both temporal and atemporal. In the notation, S is a timeline, T is a timestep, X, Y, Z are events, uppercase indicates variables, and lowercase indicates instantiated values. When used in a TMN, the features are instantiated ($f_S(\cdot) \mapsto f_i(S)$) for each non-redundant combination of events and times. For example, co-occurrence ignores order, so $f_S(w, b)$ and $f_S(b, w)$ are redundant, but $f_S(w \to b)$ and $f_S(b \to w)$ are ordered, so they are not redundant. Note that both $f_S(w \to b)$ and $f_S(b \to w)$ can be true of the same sequence as shown in Figure 3.2.

- *event*, $f_S(x)$: true if event x occurs in S (atemporal)

- *event@*, $f_S(x_t)$: true if event x occurs at t in S (atemporal)
- *co-occur*, $f_S(x, y)$: true if events x and y occur in S (atemporal)
- *co-occur@*, $f_S(x_{t_1}, y_{t_2})$: true if x occurs at t_1 and y occurs at t_2 in S (temporal)
- *before*, $f_S(x \to y)$: true if x and y occur in S and x occurs before y (temporal)
- *before-δ*, $f_S(x_T \to y_{T+\delta})$: true if x and y occur in S and x occurs δ timesteps before y (temporal)
- *before3*, $f_S(\{x, y\} \to z)$: true if x, y, and z occur in S and both x and y occur before z (temporal)

The "@" features are anchored to specific timesteps, but the other features float. Floating, being less specific than anchoring, ties parameters across timesteps and makes an assumption of stationarity. All the floating features except *before-δ* span any number of timesteps, allowing them to capture short- and long-range effects. The *before3* feature exists to model temporal V-structures.

Depending on the choice of features and the parameter tying they induce, TMNs can represent undirected analogs of BNs, DBNs, and event networks (Arroyo-Figueroa and Sucar, 1999; Galán and Díez, 2002), and the semantics of a TMN follow those of the analogous model. Examples of TMNs that imitate BNs and DBNs are in §3.4 (p. 21).

3.3.4 Parameter Learning

The parameters are learned using standard maximum likelihood estimation. Finding the maximum of the log-likelihood is a continuous, convex optimization problem, which can be solved by gradient ascent. Because the maximum of the log-likelihood is global, it is reached when the gradient (Equation 3.4) is zero (e.g., Koller and Friedman, 2009).

$$\frac{\partial}{\partial \theta_i} \frac{1}{|D|} \log \mathcal{L}(\theta; D) = \mathbb{E}_D(f_i(s)) - \mathbb{E}_\theta(f_i(s)) \tag{3.4}$$

To compute the gradient, the expected statistics of the data ($\mathbb{E}_D$) must first be computed, but this needs to be done only once. Then, the expected statistics given the TMN ($\mathbb{E}_\theta$) must be computed, and this must be done every time the parameters change. Doing so requires inference, but inference is difficult because the graph structure defined by the features is a single clique, and hence not amenable to inference algorithms for factor graphs. This limits the inference options to sampling or, for small problems, exact inference. Exact inference was chosen for the sake of precision, and implemented the TMNs in Julia using L-BFGS optimization.

3.3.5 Causal Structure Learning via Parameter Learning

TMNs are used to learn the directed structure of a distribution of timelines by (1) detecting conditional independence between variables, (2) including only those edges that correspond to direct dependences, and (3) directing edges with time. Detecting conditional independence is done by constructing a TMN, learning the weights of its features, and comparing those weights to zero. A weight that is zero indicates the absence of the relationship modeled by that feature, and if all the weights of all the features involving a pair of variables are zero, then those variables are conditionally independent. This property allows weight learning in

TMNs to recover the conditional independence structure of the generating DBN as shown in the following theorem.

Theorem 3.5 (TMN Structure Learning)**.** *Given a DBN $\mathcal{M}$ that generates a true distribution $P(\mathcal{S})$ over timelines, the forward edges of the DAG G of $\mathcal{M}$ can be deduced from the weights of a TMN fit to $P(\mathcal{S})$ using maximum likelihood. Specifically, if the weights of $f_i(X \to Y)$ and all the other features containing X and Y are zero, then $X \to Y$ is not an edge in G:*

$$(\forall (i : f_i \supseteq \{X, Y\})\ \theta_i = 0) \implies X \to Y \notin G \tag{3.5}$$

Proof. If all of the weights of features involving X and Y are zero, then those weights contribute nothing to the sum in Equation 3.2 and hence contribute nothing to the product in Equation 3.1. Since the factorization of P does not include X and Y, they must be independent by Theorem 3.3, and there cannot be an edge between X and Y in any graphical model consistent with P. □

A TMN can only capture the undirected version of the generating DBN, but if it is a first-order, non-isochronal DBN (it has no edges within a timestep that represent instantaneous relationships) (Plis et al., 2015), then moralizing it adds no edges between timesteps, and the forward edges indicated by the TMN weights are exactly the edges of the DBN.

In summary, Theorem 3.5 shows how weight learning in a TMN can recover the DBN structure given the true distribution of timelines: include only those edges $x \to y$ that correspond to features $f_i(x \to y)$ (or $f_i(\{x, z\} \to y)$, etc.) with nonzero weights θ_i. The edges are already directed with time.

Is such a structure a causal model? If one assumes the causal Markov and causal faithfulness conditions (Spirtes et al., 2000), as is commonly done, then a DBN that has the correct independence structure is a causal DBN. Theorem 3.5 shows that, given the true distribution of a first-order, non-isochronal DBN, the weights of the learned TMN will indicate the correct independences and therefore describe a causal structure.

Of course, in practice the distribution is not the population one but an empirical one. The noise in such a sample alters the learned weights and obscures the independences. Thus, it becomes necessary to employ regularization or a threshold to determine the zeros. Regularization introduces bias, so thresholding the weight magnitudes was chosen. This approach provides an unbiased estimator that has a straightforward interpretation: as soon as the magnitude of the noise gets larger than the magnitude of the signal, the thresholding will start to get edges wrong. This can be mitigated by choosing features that are expressive enough to accurately model the distribution, that can diffuse or absorb the noise, and that isolate relationships of interest. For example, one could use only $f(x \to y)$ and $f(y \to x)$ to model a temporal relationship, but also including $f(x, y)$ isolates their atemporal co-occurrence from their temporal precedence and splits the noise accordingly. Choosing features that are expressive enough to accurately model the distribution means choosing features that match the level of interactions (cliques) in the underlying process. At one extreme, the saturated model (e.g., Wasserman, 2004, §19.4) makes no assumptions about independence or the level of interactions, but is intractable due to the number of features involved. (The number of features in the saturated model is $2^n - 1$ for n binary variables, and there are $|X||\mathcal{T}|$ binary variables.) At the other extreme, one can assume only pairwise interactions, but this will almost certainly lead to inaccurate weights and an incorrect ranking

of edges. Note that using only pairwise features is similar to making an assumption that the conditional probability distributions of the underlying process are noisy-ORs.

3.4 Experiments

To evaluate TMNs, experiments were conducted to compare them to other methods on DBN structure learning tasks using synthetic and real-world data. The experiments were designed to measure how accurately the designated methods could recover the structure of dynamic causal networks in a variety of scenarios. With the synthetic data, the methods sought to recover known, complete causal networks having observed all the relevant variables. With the real-world EMR data, they sought to recover the causal structure among the variables in Figure 3.1 (p. 15), having observed only those same variables. This is the OMOP ADE task, which involves only a small, known subset of the causal structure in EMR data.

For comparison methods, the PC algorithm and BNFinder were chosen to represent the two major BN structure learning paradigms. The causal, constraint-based paradigm was represented by the PC algorithm (Spirtes et al., 2000). Even though it only works for static data, it was also applied to timelines by using separate variables for each timestep, unrolling the model as in Figure 3.1(b), and by reversing edges that went backwards in time. The comparison TMN, TMN-PC, equivalently used anchored features ($f(x_t)$, $f(x_{t_1}, y_{t_2})$). The score-based paradigm was represented by BNFinder (Wilczyński and Dojer, 2009; Dojer, 2006), which finds the optimal-scoring BN structure in polynomial time given a partial order of the variables and a maximum number of parents. Being optimal, BNFinder subsumes GES (Meek, 1997; Chickering, 2002) and other score-based structure learners on the task of learning DBNs. The comparison TMN, TMN-DBN, used features to represent the initial and transition distributions of a first-order DBN ($f(x_{t=0})$, $f(x_{t=0}, y_{t=0})$, $f(x)$, $f(x_T, y_T)$, $f(x_T \to y_{T+1})$). A third TMN, TMN-Bf3, extended the TMN-DBN approach with long-range temporal features and three-way interactions ($f(x_{t=0})$, $f(x_{t=0}, y_{t=0})$, $f(x)$, $f(x, y)$, $f(x \to y)$, $f(\{x, z\} \to y)$). While higher-order interactions would be necessary to represent distributions in general, tuning indicated three-way features were sufficiently rich.

Both the synthetic and real-world experiments shared the same setup and analysis. The data was timelines of events (§3.3.1, p. 17). Based on the timelines, the utilized methods scored each possible forward edge in a first-order DBN to produce a weighted, bipartite graph. The edge score was the weight magnitude of the corresponding temporal feature for TMNs, aggregate posterior edge probability for BNF-DBN, and edge existence $\{0, 1\}$ for PC. The weighted graphs were evaluated as (soft) binary classification tasks: which of the edges belong to the true DBN graph. To do this, the edges were ranked by their score, and then classification accuracy was assessed with precision-recall (PR) analysis because class skew (edge density of the true graph) varied widely. The methods were developed and tuned using a separate set of hand-crafted and randomly-generated test cases prior to running any experiments. The specific parameters are in §3.5.1 (p. 26).

3.4.1 Synthetic DBN Experiments

In the synthetic data experiments, the goal was to recover the structure of random DBNs given datasets of timelines sampled from those DBNs. Each dataset received four data treatments ("regimes") designed to test the methods in the face of noise, missing timesteps,

and confounding. The plain treatment left the data unaltered. The noisy treatment selected $X_{i,t}$ IID if $\text{Bernoulli}(\varepsilon)$ and replaced selected values with $x_{i,t} \sim \text{Bernoulli}(1/2)$. Each dataset had its own noise level $\varepsilon \sim \text{Uniform}(0.1, 0.9)$. The missing treatment selected timesteps IID if $\text{Bernoulli}(\eta)$ and hid their values. This was meant to imitate how patients are unobserved in real EMR data and to measure the influence of assuming unobserved values to be false. Each dataset had its own missingness level $\eta \sim \text{Uniform}(0.1, 0.9)$. The confounding treatment randomly selected a subset of confounders (variables with at least two children) and removed them from the data and the ground truth graph. Specifically, the DBN graph was compressed (rolled up) (Plis et al., 2015), confounders were randomly selected so that no more than $2/5$ of the variables would be hidden and so that the class proportion remained in $[0.1, 0.9]$, the confounding variables were removed from the graph by summing them out, and the graph was uncompressed to become the new ground truth graph. For each of the four data treatments the data was represented in two ways: fully-observed and condensed, as illustrated in Figures 3.2(b) and 3.2(c) (p. 18). The condensed data imitates real EMR data where negatives are typically not recorded and absolute times are not reliable, but it also simplifies the problem of modeling events that occur over widely-varying time scales.

Each DBN was generated by (1) drawing a number of variables $n \in 2{:}10$ from a distribution that favors numbers in proportion to their size, (2) drawing an edge probability $p_e \sim \text{Uniform}(0, 1)$ and drawing each of the n^2 possible forward DBN edges IID as $\text{Bernoulli}(p_e)$ to create a bipartite graph representing two timesteps, (3) creating conditional probability tables by sampling a probability $p \sim \text{Uniform}(0, 1)$ for each setting of a node's parents, and (4) rejecting any DBN with edge density ($|E|/n^2$) not in $[0.1, 0.9]$ (which kept the class skew less than $9:1$).

The synthetic data consisted of datasets sampled from 1k random DBNs. Each dataset had 10k timelines and each timeline had 10 timesteps. Experiments were performed on the first 100, 1k, and 10k timelines of each dataset to assess statistical efficiency. The number of DBNs was determined by a power calculation for a 0.9 probability of detecting a PR area difference of 0.01 at $\alpha = 0.01$ with a two-tailed paired t-test. In total, there were 120k experiments: 1k random DBNs, 4 data treatments, 2 data representations, 3 data sizes, and 5 methods (PC, TMN-PC, BNF-DBN, TMN-DBN, TMN-Bf3).

3.4.2 Synthetic DBN Results

To assess how well the methods recovered DBNs from the synthetic data, the PR areas of their structure recovery were compared. Figure 3.3 (p. 23) shows the average PR area achieved by each method on each data regime. Overall, BNF-DBN scored the best on average, followed by TMN-PC, PC, TMN-Bf3, and TMN-DBN. The researcher believes that BNF-DBN did so well because its assumptions exactly match the data generating model.

Behind the averages in Figure 3.3, the performance of the methods varied substantially by data regime and other characteristics of the DBN structure learning problems. To assess the influence of these characteristics on the achieved PR areas, a linear regression was performed using PR area as the dependent variable and method, data regime, data size, etc. as the independent variables. The results are in Table 3.2 (p. 24). Each coefficient is interpreted as the change in PR area attributable to a unit change in X, everything else held constant. The covariates are able to explain a reasonable amount of variance ($R^2 = 0.699$),

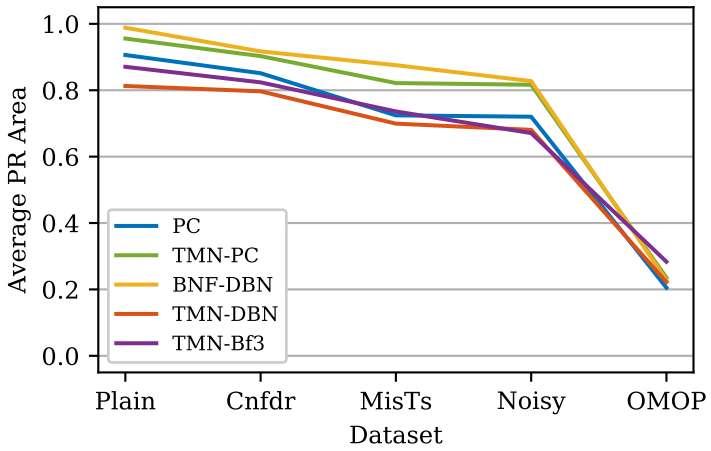


Figure 3.3: Summary of results. Average PR areas of the methods in each of the synthetic data regimes and OMOP, using the largest data size (synthetic: 10k, OMOP: 100k timelines). MisTs: missing timesteps, Cnfdr: confounders.

and most of the coefficients have intuitive interpretations. Those of the method indicators show that each method achieves significant, positive contributions to PR area compared to random. BNF-DBN improves over random by the largest amount (0.400), probably because it is the only method whose assumptions exactly match the data. However, the coeffecients also show that BNF-DBN suffered the most in the face of confounding, noise, and missing timesteps. TMN-DBN was the most robust to noise while TMN-Bf3 was the most robust to confounding and missing timesteps. Complex networks are harder to recover as indicated by the negative coefficients on maximum in-degree, number of nodes, and number of V-structures. Data size was important but condensing the data had almost no effect. Of the data treatments, missingness was the least detrimental of the three in terms of its interactions with the methods. These results suggest that treating missing data as false and condensing it is reasonable to do with EMR data (where the majority of data is not observed).

Perhaps counterintuitively, increasing the network density or increasing the number of confounders helps performance. In the case of confounders, hiding variables removes them from the problem, leaving a smaller, easier problem. In the case of density, having more of the possible edges be true reduces the chance that misranking a single edge will affect the PR area.

3.4.3 OMOP Experiments

In the OMOP experiments, the goal was to discover ADEs in real-world EMR and claims databases. This was formulated as a DBN structure learning task rather than a causal effect size estimation task as is the case with many other methods for causal discovery. The DBN structure learning task was based on the OMOP ADE task, which defines 9 true ADEs and 44 non-ADEs among the same events (Figure 3.1, p. 15). OMOP selected these positives and negatives based on drug labeling and evidence in the literature. For the experiments, the positives defined the edges of the ground truth graph (Figure 3.5(a), p. 27). The methods learned DBNs over all of the drugs and conditions, but only edges corresponding to pairs in

Table 3.2: Linear regression of PR areas on attributes of synthetic DBN experiments, including interactions between method and data regime, ranked by $\hat{\beta}$ magnitude. R^2 = 0.699. The method indicators contrast with random guessing.

Rank	X	$\hat{\beta}$	se($\hat{\beta}$)	TStat	P-Value
1	BNF-DBN? * cnfdr	-0.713	0.0139	-51.2	0
2	density, e/n^2	0.663	0.0117	56.9	0
3	TMN-PC? * cnfdr	-0.553	0.0139	-39.7	0
4	PC? * cnfdr	-0.499	0.0139	-35.9	1.69e-280
5	TMN-DBN? * cnfdr	-0.475	0.0139	-34.1	7.15e-254
6	TMN-Bf3? * cnfdr	-0.454	0.0139	-32.6	3.16e-232
7	BNF-DBN? * noise	-0.414	0.00547	-75.7	0
8	BNF-DBN?	0.400	0.00188	213	0
9	TMN-PC?	0.309	0.00188	165	0
10	BNF-DBN? * mists	-0.305	0.00537	-56.8	0
11	PC? * noise	-0.287	0.00547	-52.5	0
12	TMN-PC? * noise	-0.281	0.00547	-51.4	0
13	TMN-Bf3? * noise	-0.251	0.00547	-45.8	0
14	PC? * mists	-0.248	0.00537	-46.1	0
15	PC?	0.245	0.00188	131	0
16	TMN-DBN? * noise	-0.225	0.00547	-41.1	0
17	TMN-DBN?	0.217	0.00188	115	0
18	TMN-PC? * mists	-0.216	0.00537	-40.2	0
19	TMN-Bf3?	0.209	0.00188	111	0
20	TMN-DBN? * mists	-0.169	0.00537	-31.5	2.11e-216
21	TMN-Bf3? * mists	-0.157	0.00537	-29.2	1.26e-186
22	# cnfdr $/n$	0.130	0.00996	13.1	3.99e-39
23	log # data	0.0747	0.000430	174	0
24	missingness	-0.0227	0.00380	-5.99	2.16e-09
25	noise level	-0.0213	0.00387	-5.50	3.79e-08
26	intercept	-0.0121	0.00698	-1.73	0.0833
27	avg in-deg	-0.0109	0.00820	-1.33	0.184
28	max in-deg	-0.00795	0.000607	-13.1	3.39e-39
29	# edges, e	0.00533	0.000421	12.7	1.07e-36
30	# nodes, $2n$	-0.00167	0.00117	-1.42	0.156
31	# V-structures	-0.00138	6.06e-05	-22.7	1.06e-113
32	condensed?	0.000733	0.000702	1.04	0.296
33	max edges, n^2	-0.000375	0.000184	-2.04	0.0417
34	max out-deg	-0.000243	0.000638	-0.381	0.703
35	# CPT θs	3.11e-05	1.86e-06	16.7	1.75e-62

the OMOP task were used in the evaluation.

The five methods learned DBNs from data sets of timelines extracted from the five OMOP databases (Figure 3.3, p. 25). The OMOP databases contained dated event tuples (§3.3.1, p. 17), which can be viewed as timelines discretized by day (Figure 3.2(a), p. 18). To create a data set from each database, a timeline for each patient was extracted and then condensed as in Figure 3.2(c) (p. 18). Variables not observed were assumed to be false. Twenty samples of 100k timelines were drawn without replacement from each data set. These replicates were drawn because PC and BNF-DBN could not scale to the full data size. (TMNs, needing only sufficient statistics, have no direct data size limitations.) The number of replicates was determined by a power calculation for a 0.9 probability of detecting a PR area difference of 0.05 at $\alpha = 0.01$ with a two-tailed paired t-test.

Table 3.3: Summary statistics of the five OMOP datasets: GE Centricity, MarketScan Commercial Claims and Encounters, MarketScan Medicaid, MarketScan Medicare, MarketScan Lab Results. EoI: events of interest, PoI: people with EoI.

Name	Type	People	PoI	EoI	Years
GE	EMR	11.2M	4.1M	7.1M	1995–2009
CCAE	claims	46.5M	25.6M	47.7M	2003–2009
MDCD	claims	10.8M	7.3M	14.0M	2002–2007
MDCR	claims	4.6M	3.9M	12.7M	2003–2009
MSLR	claims	1.2M	1.1M	2.1M	2003–2008

3.4.4 OMOP Results

Figure 3.4 (p. 26) shows the results of the experiments on the OMOP data sets in terms of PR area distributions of replicates.[4] The TMNs do especially well on the GE EMR data, but the performance on the claims data is mixed. Looking at the medians, TMN-PC beats PC on 3 data sets, TMN-DBN beats BNF-DBN on 2 data sets, and TMN-Bf3 is the best on all 5 data sets. The significance results in Table 3.4 (p. 28) lead to a similar ranking of the methods by wins in a pairwise tournament: TMN-Bf3, TMN-PC, BNF-DBN, TMN-DBN, PC. I hypothesize that the success of TMN-Bf3 on the OMOP task is due to its ability to effectively model higher-order interactions and detect independence in the presence of noise.

The (min, avg, max) run times, in hours, on the OMOP task were BNF-DBN (0.6, 0.8, 0.9), TMNs (0.5, 1.3, 2.8), and PC (0.1, 2.9, 9.9). While this makes BNF-DBN look fast, the experiments had to be limited to 100k timelines to make BNF-DBN and PC tractable[5] (whereas the TMNs were able to run on the millions of timelines in the full-size OMOP data sets (Table 3.3)). Furthermore, TMN weight learning could run faster by stopping as soon as the ranking of weights is settled (because PR area only depends on the ranking of edges), but this was not implemented. Perhaps this explains why TMNs were successful

[4] A linear regression was not appropriate because only the one BN structure of the OMOP ADE task was involved, and so there was no variation in most of the experimental attributes in Table 3.2 (p. 24).

[5] BNF-DBN has $|E|^{\log|D|}$ in its complexity polynomial. The complexity of PC is $O(n^q|D|)$ in the worst case, where q bounds the degree of the graph.

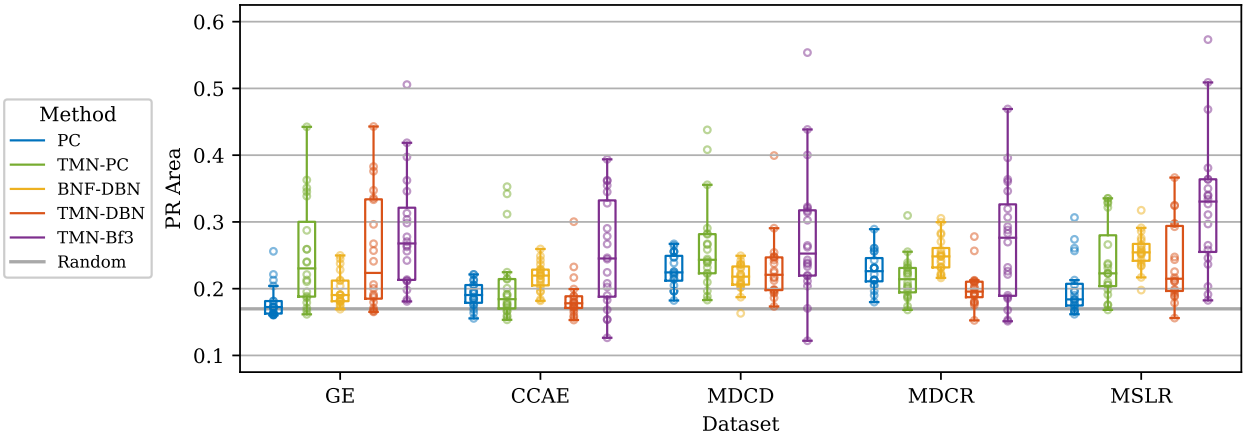


Figure 3.4: Distributions of PR areas from the 20 replicates drawn from each OMOP data set by method. The whiskers are 1.5 interquartile range.

even though in many cases they did not converge within their allotted 1000 iterations. On the other hand, the lack of convergence is likely a large factor in the variation of the TMN results.

The success of BNF-DBN on both the synthetic and OMOP tasks demonstrates that DBNs may be applicable to modeling EMR data more than previously thought (Hyttinen et al., 2016) despite its sparseness and irregularity, and suggests that condensed data may also work for other discrete-time models that assume fully-observed, regularly-sampled data.

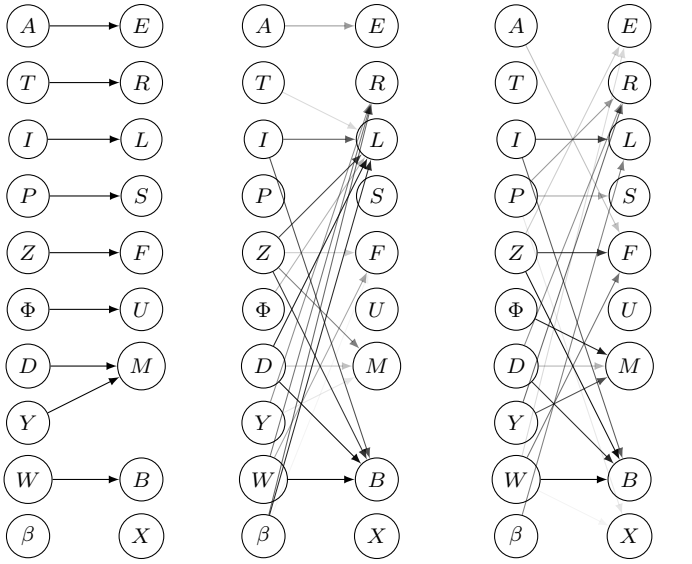
In a qualitative view of performance, additive ensembles of the networks learned by PC and TMN-Bf3 on MSLR are shown in Figure 3.5 (p. 27) along with ground truth. Both methods have six correct edges among the top 20, but TMN-Bf3's six are higher in its ranking. (The other methods have four or fewer correct edges in the top 20.) PC and TMN-Bf3 agree on four correct edges. PC concentrates many relationships on renal and liver failure, while TMN-Bf3 spreads out its edges more evenly. Both methods concentrate on bleeding, apparently one of the more confounded relationships. These results demonstrate that causal structure learning methods are applicable and relevant to problems in epidemiology despite not estimating effect sizes.

3.5 Supplementary Experimental Details

3.5.1 Method Parameters

Here are the specific parameters of the methods and their rationales. We used the Center for Causal Discovery's Java implementation of the PC algorithm which is based on the Tetrad implementation from Carnegie Mellon University.

- PC
  - α: 0.01. Decided to ensure approximately one Type 1 error per 10-node graph.
  - depth: 10. Decided to correspond with the maximum number of nodes in the synthetic data and the maximum number of parents (drugs) in the OMOP data.



(a) Ground truth (b) PC, MSLR (c) TMN-Bf3, MSLR

Figure 3.5: Ground truth and selected learned networks for the OMOP task showing the top 20 edges from an ensemble of that method's MSLR runs. Figure 3.1 (p. 15) lists the variables.

- BNFinder
  - score: BDe. Tuned, but the Bayesian Dirichlet equivalence score performs no differently than the minimum description length score.
  - maximum parents: 10. Decided to correspond with the maximum number of nodes in the synthetic data and the maximum number of parents (drugs) in the OMOP data.
- Temporal Markov networks using Optim.jl
  - maximum optimization iterations: 1000. Software default.
  - gradient infinity-norm bound: 1e-8. Software default.
  - L-BFGS approximation vectors: 10. Software default.

3.5.2 Pairwise Comparisons

Table 3.4 (p. 28) contains the detailed results of the pairwise comparisons and their statistical significance.

3.6 Discussion

There are many advantages to treating structure learning as a smooth, convex optimization problem rather than a combinatorial one. Convexity guarantees that there is a global optimum and that there are no impediments to getting there, like plateaus or local optima.

Table 3.4: Pairwise comparisons between methods within the five data regimes using two-tailed paired t-tests, ranked by p-value. The row with the lines indicates the significance cutoff of a paper-wise false discovery rate controlled at 0.01 with the Benjamini-Hochberg procedure.

Rank	Better	Worse	DiffMeans	TStatistic	P-Value
1	BNF-DBN-Plain	Random-Plain	0.480	92.2	0
2	TMN-PC-Plain	Random-Plain	0.447	89.5	0
3	TMN-Bf3-Plain	Random-Plain	0.362	83.4	0
4	TMN-DBN-Plain	Random-Plain	0.304	81.0	0
5	PC-Plain	Random-Plain	0.398	77.4	0
6	BNF-DBN-MisTs	Random-MisTs	0.367	68.5	0
7	TMN-PC-MisTs	Random-MisTs	0.313	65.5	0
8	BNF-DBN-Noisy	Random-Noisy	0.319	60.2	0
9	BNF-DBN-Cnfdr	Random-Cnfdr	0.260	59.2	0
10	TMN-PC-Noisy	Random-Noisy	0.308	59.1	0
11	TMN-PC-Cnfdr	Random-Cnfdr	0.246	56.8	0
12	TMN-DBN-MisTs	Random-MisTs	0.191	53.1	0
13	TMN-Bf3-MisTs	Random-MisTs	0.228	52.9	0
14	BNF-DBN-Plain	TMN-DBN-Plain	0.176	50.0	0
15	TMN-Bf3-Cnfdr	Random-Cnfdr	0.167	49.9	0
16	BNF-DBN-Noisy	PC-Noisy	0.107	49.8	0
17	TMN-DBN-Noisy	Random-Noisy	0.172	48.6	0
18	TMN-PC-Noisy	PC-Noisy	0.0963	47.1	0
19	BNF-DBN-MisTs	PC-MisTs	0.151	47.0	0
20	PC-Cnfdr	Random-Cnfdr	0.194	46.6	7.41e-322
21	PC-MisTs	Random-MisTs	0.216	46.6	7.91e-322
22	BNF-DBN-Plain	PC-Plain	0.0821	45.2	8.61e-308
23	PC-Noisy	Random-Noisy	0.212	44.1	1.15e-297
24	TMN-PC-Plain	TMN-DBN-Plain	0.143	43.6	5.15e-292
25	TMN-DBN-Cnfdr	Random-Cnfdr	0.140	41.9	2.36e-276
26	BNF-DBN-MisTs	TMN-DBN-MisTs	0.176	41.6	5.19e-273
27	TMN-Bf3-Noisy	Random-Noisy	0.162	40.9	6.11e-266
28	TMN-PC-MisTs	PC-MisTs	0.0973	40.0	1.39e-257
29	BNF-DBN-Noisy	TMN-DBN-Noisy	0.147	37.5	9.95e-234
30	BNF-DBN-Noisy	TMN-Bf3-Noisy	0.157	36.5	1.20e-223
31	TMN-PC-Noisy	TMN-DBN-Noisy	0.136	35.2	2.64e-211
32	BNF-DBN-Cnfdr	PC-Cnfdr	0.0657	34.3	5.39e-203
33	BNF-DBN-Cnfdr	TMN-DBN-Cnfdr	0.120	33.5	1.93e-195
34	TMN-PC-Noisy	TMN-Bf3-Noisy	0.146	33.2	6.38e-193
35	TMN-PC-Plain	PC-Plain	0.0495	33.2	1.45e-192
36	TMN-PC-MisTs	TMN-DBN-MisTs	0.122	32.3	2.95e-184
37	BNF-DBN-MisTs	TMN-Bf3-MisTs	0.139	31.8	9.87e-180
38	TMN-PC-Cnfdr	TMN-DBN-Cnfdr	0.106	30.9	3.77e-172
39	BNF-DBN-Plain	TMN-Bf3-Plain	0.118	30.4	1.65e-167
40	TMN-PC-Cnfdr	PC-Cnfdr	0.0515	29.5	3.56e-159

Table 3.5: Pairwise comparisons, part 2 of Table 3.4 (p. 28).

Rank	Better	Worse	DiffMeans	TStatistic	P-Value
41	PC-Plain	TMN-DBN-Plain	0.0934	26.5	1.30e-132
42	BNF-DBN-Cnfdr	TMN-Bf3-Cnfdr	0.0933	24.4	2.49e-115
43	TMN-PC-Plain	TMN-Bf3-Plain	0.0851	22.8	2.17e-102
44	BNF-DBN-Plain	TMN-PC-Plain	0.0327	21.3	8.01e-91
45	TMN-PC-Cnfdr	TMN-Bf3-Cnfdr	0.0790	21.1	1.05e-89
46	TMN-Bf3-Plain	TMN-DBN-Plain	0.0578	20.2	4.54e-83
47	TMN-PC-MisTs	TMN-Bf3-MisTs	0.0855	20.1	5.41e-82
48	BNF-DBN-MisTs	TMN-PC-MisTs	0.0538	17.8	4.57e-66
49	PC-Cnfdr	TMN-DBN-Cnfdr	0.0546	15.8	3.29e-53
50	PC-Noisy	TMN-Bf3-Noisy	0.0493	12.6	7.45e-35
51	BNF-DBN-OMOP	Random-OMOP	0.0580	18.1	3.05e-33
52	TMN-Bf3-MisTs	TMN-DBN-MisTs	0.0368	10.9	5.90e-27
53	PC-Noisy	TMN-DBN-Noisy	0.0396	10.8	2.61e-26
54	TMN-Bf3-Cnfdr	TMN-DBN-Cnfdr	0.0270	10.4	1.19e-24
55	TMN-Bf3-OMOP	Random-OMOP	0.114	12.2	1.70e-21
56	BNF-DBN-Noisy	TMN-PC-Noisy	0.0110	9.11	2.01e-19
57	PC-Plain	TMN-Bf3-Plain	0.0356	8.79	3.29e-18
58	PC-OMOP	Random-OMOP	0.0353	10.4	1.68e-17
59	TMN-PC-OMOP	Random-OMOP	0.0655	10.3	2.73e-17
60	BNF-DBN-Cnfdr	TMN-PC-Cnfdr	0.0142	7.91	4.18e-15
61	TMN-DBN-OMOP	Random-OMOP	0.0524	8.77	5.14e-14
62	PC-Cnfdr	TMN-Bf3-Cnfdr	0.0275	7.45	1.37e-13
63	TMN-Bf3-OMOP	PC-OMOP	0.0783	8.11	1.40e-12
64	BNF-DBN-OMOP	PC-OMOP	0.0227	7.86	4.70e-12
65	PC-MisTs	TMN-DBN-MisTs	0.0249	6.80	1.39e-11
66	TMN-Bf3-OMOP	BNF-DBN-OMOP	0.0556	5.93	4.47e-08
67	TMN-Bf3-OMOP	TMN-DBN-OMOP	0.0612	5.45	3.76e-07
68	TMN-Bf3-OMOP	TMN-PC-OMOP	0.0481	4.40	2.78e-05
69	TMN-DBN-Noisy	TMN-Bf3-Noisy	0.00977	4.06	5.00e-05
70	TMN-PC-OMOP	PC-OMOP	0.0303	4.21	5.59e-05
	—	—	—	—	— BH 0.01
71	TMN-Bf3-MisTs	PC-MisTs	0.0119	2.81	0.00494
72	TMN-DBN-OMOP	PC-OMOP	0.0171	2.52	0.0132
73	TMN-PC-OMOP	TMN-DBN-OMOP	0.0131	1.97	0.0513
74	TMN-PC-OMOP	BNF-DBN-OMOP	0.00755	1.04	0.300
75	BNF-DBN-OMOP	TMN-DBN-OMOP	0.00559	0.822	0.413

This guarantees progress with every iteration, and the optimization can be stopped at any time to yield an approximate solution with the gradient giving a sense of how close the current model is to the optimum. Furthermore, the optimization focuses first on the most important features, which are those with the largest gradients. Framing the problem as an optimization means that all the edges are estimated jointly, avoiding sequential decisions and multiple testing. This framing also removes the need for greedy or heuristic search as the optimization space is tractable and amenable to well-understood approximation (e.g. stochastic gradient descent). All these advantages combine to make this approach faster, more robust, and better able to handle noise than approaches based on combinatorial search.

The formulation as a log-linear model also comes with advantages and disadvantages. In terms of advantages, it allows arbitrary features, which can be used to handle irregular events and model short- and long-range dependencies. The data can be completely summarized by the sufficient statistics of the features, which separates the data processing from the optimization and enables the straightforward application of large-scale data processing techniques such as partitioning and parallelism. By comparison, updating a BN structure score requires a pass over the data even if it only involves a few of the variables. The sufficient statistics, being aggregates, are also robust to noise. In terms of disadvantages, there is now a modeling problem as one must choose the right features. Part of this relates to choosing the level of interactions that the features can express. Depending on how many features are chosen, their complexity, and how many combinations of events for which they are instantiated, there can be a very large number of features and a correspondingly large optimization space, which may be challenging for optimization algorithms. The optimization challenges are amplified by the inference difficulties of an extremely large, unfactorable PGM.

One limitation of this approach is that it relies on time for direction. But this is true for all DBN learners and for other algorithms (e.g., Shojaie and Michailidis, 2010) that operate on temporal data and assume edges go forward in time. However, because our approach estimates both $x \to y$ and $y \to x$, it can still pick the more important between the two. This can be useful if one desires to roll the DBN up into a BN (roll Figure 3.1(b) up into 3.1(a) (p. 15)): the edge with smaller magnitude can be discarded to break cycles.

Unfortunately, due to the inability of undirected PGMs to express the independence in a V-structure, exact recovery of DBN edges by TMNs is limited to first-order, non-isochronal DBNs. However, assuming a first-order DBN is relatively innocuous because any higher-order DBN can be converted to an equivalent first-order DBN. Assuming a non-isochronal DBN is reasonable in cases where the timescale of the DBN is smaller than that of the system (Plis et al., 2015). This is the case for EMR data where data is available on the same scale as disease progression in both inpatient and outpatient settings. Anything that happens more immediately is comparatively easy to notice and is probably already well-understood, making it unlikely to be the subject of a causal modeling task.

This work represents an alternative approach to the OMOP task, one that uses structure learning instead of causal effect estimation (as would be done in epidemiology). Structure learning and effect estimation are not directly comparable because they handle direct and indirect effects differently. Effect estimation does not care about the path, only its overall effect, whereas structure learning cares only about the direct effects that make up the path, not its overall effect. Unfortunately, this mismatch means that the OMOP task is not necessarily a suitable evaluation for structure learning methods; it depends on how many of

the OMOP pairs are direct effects in terms of the observable variables in the data. Perhaps this explains why the results achieved by the DBN learners in this paper are lower than published results from epidemiological methods on the OMOP task (e.g., Ryan et al., 2012). Naturally, the answer is to use the learned structure to estimate a causal DBN and then query it to determine effect sizes. Investigating this and determining how to better apply structure learning to epidemiological tasks is future research.

3.7 Conclusion

In learning the relationships among events, TMNs avoid the combinatorial nature of classical BN structure learning algorithms by reformulating structure learning as a smooth, convex optimization problem in a log-linear model. As shown in Theorem 3.5 (p. 20), TMNs learn the correct structure given enough data and sufficiently expressive features, and the learned structure corresponds to a causal DBN. This enables TMNs to do causal discovery, and their flexible, expressive features enable them to handle the irregularity, sparsity, and noise of EMR data. Therefore, TMNs have the characteristics necessary to address the challenges of the OMOP ADE task. In practice, they demonstrate their effectiveness by performing as well or better than representative methods for DBN structure learning. Thus, with characteristics and performance that complement existing methods, TMNs establish an alternative to DBNs for causal discovery from observational time series data.

Attribution

This chapter was previously published as Barnard and Page (2018).

Chapter 4

Identifying ADEs using Relational Learning

Learning the structure of a causal model is a technique for causal discovery that is theoretically justified and works well when all the effects of interest are observed in the data and included as random variables in the model. However, this limits discovery to associations between events that have already been hypothesized, defined, and measured. It is often the case that open-ended discovery is the real goal, in which case both learning structural causal models and conducting observational studies fall short: they cannot hypothesize new effects. The following chapters describe methods for causal discovery that can both hypothesize and then score causal effects. These methods are not based on structural causal models but rather are based on novel applications of machine learning to analyzing observational studies. This chapter covers one such method which uses inductive logic programming (ILP) to hypothesize ADEs and a causally-motivated score to rank them.

4.1 Introduction

The pharmaceutical industry, consumer protection groups, users of medications, and government oversight agencies are all strongly interested in identifying adverse reactions to drugs. Adverse drug events (ADEs) are estimated to account for 10–30% of hospital admissions, with costs in the United States alone between 30 and 150 billion dollars annually (Lazarou et al., 1998), and with more than 180k life threatening or fatal ADEs annually, of which 50% could have been prevented (Gurwitz et al., 2003). Although the U.S. Food and Drug Administration (FDA) and its counterparts elsewhere have preapproval processes for drugs that are rigorous and involve controlled clinical trials, such processes cannot possibly uncover everything about a drug. While a clinical trial of a drug may use only a thousand patients, once a drug is released on the market it may be taken by millions of patients. As a result, in many cases adverse drug events (ADEs) are observed in the broader population that were not identified during clinical trials. Therefore, there is a need for continued, postmarketing surveillance of drugs to identify previously unanticipated ADEs.

This chapter proposes *reverse machine learning* as a postmarketing surveillance tool in order to predict or detect adverse reactions to drugs from EHR data. This approach is applied to actual EHR data sets, including data sets provided by the Observational Medical

Outcomes Partnership (OMOP). This task poses several novel *challenges* to the machine learning (ML) community:

1. One cannot assume advance knowledge as to an ADE that a particular drug might cause. In some cases, one may suspect a specific ADE, such as increased risk of myocardial infarction (MI, heart attack); in such a case, supervised learning can be employed with MI as the class variable. But if one does not know the ADE in advance, what class variable can one use? This work proposes using the *drug* itself as the class variable and claims that, while one already knows who is taking the drug, examination of a model that accurately predicts drug use can give insight into ADEs. Because this work seeks to discover the ADE by building a model to "predict" drug use (who has been on the drug), rather than to predict the actual entity of interest (the ADE), one can refer to this approach as *reverse machine learning*.

2. The data are *multi-relational*. Several objects such as doctors, patients, drugs, diseases, and labs are connected through relations such as visits, prescriptions, diagnoses, etc. If traditional ML techniques are to be employed, they require flattening the data into a single table. All known flattening techniques, such as computing a join or summary features, result in either (1) changes in frequencies on which machine learning algorithms critically depend or (2) loss of information.

3. There are *arbitrary* numbers of patient visits, diagnoses and prescriptions for different patients, i.e., there is no fixed pattern in the diagnoses and prescriptions of the patients. It is incorrect to assume that there are fixed number of diagnoses or that only the last diagnosis is relevant. To predict ADEs for a drug, it is important to consider the other drugs prescribed for the patient, as well as past diagnoses, procedures, and laboratory results.

4. Since all the preceding events and their interactions are *temporal*, it is important to explicitly model time. For example, some drugs taken at the same time can lead to side effects, while in other cases one drug taken after another can cause a side effect. As is demonstrated in the experiments, it is important to capture such interactions to be able to make useful predictions.

5. ML researchers need to learn lessons from *epidemiology*, especially *pharmacoepidemiology*, about how to construct cases and controls (positive and negative examples) as well as how to address *confounders*. Otherwise ML methods will simply identify disease conditions associated with the drug for other reasons, such as drug indications or conditions correlated with use of the drug for other reasons.

4.1.1 Contributions to Machine Learning

This chapter presents a machine learning approach to studying an important, real-world, high-impact task—identifying ADEs—for which data sets are available through the Observational Medical Outcomes Partnership. The chapter shows how relational learning (Lavrač and Džeroski, 1994; De Raedt, 2008) is especially well-suited to the task, because of the multi-relational nature of EHR data. In addition, this chapter provides technical lessons for ML that should be applicable to a number of other domains as well. These lessons are listed

here, discussed as they arise in the presentation of the empirical analysis of this approach, and then reviewed again at the end of this chapter.

1. In some ML applications, one may not have observations for the class variable. For example, one might hypothesize an unknown genetic factor in a disease or an unknown subtype of a disease. In such situations, one typically resorts to unsupervised learning. The task of identifying previously unanticipated ADEs is such a situation: without a hypothesized ADE, how can one run a supervised learning algorithm to model it? Without knowing in advance that MI is an ADE for COX-2 inhibitors, how can one provide supervision such that the algorithm will predict that MI risk is raised by these drugs? This work shows that the problem can be addressed by running supervised learning "in reverse," to learn a model to predict who is on a COX-2 inhibitor. If an algorithm can identify some subgroup of COX-2 inhibitor patients based on the events occurring after they start COX-2 inhibitors, this can provide evidence that the subgroup might be sharing some common effects of COX-2 inhibitors. The researcher anticipates that this same approach can also be applied to other situations where the class variable of interest is not observed. This lesson is referred to as *reverse ML*.

2. This work introduces to ML some useful ideas from epidemiology, including treating each patient as their own control, by drawing as positive examples patients and their data *after* they begin use of a drug and as negative examples the same patients but *before* they begin use of the drug. Another idea one can employ from epidemiology is to use a domain-specific scoring function that includes normalization based on other drugs and other conditions. This work introduces to epidemiology the idea of learning rules to characterize ADEs, rather than simply scoring drug–condition pairs which require the ADE to correspond to an already-defined condition.

3. Finally, this work reinforces the need for iteration between human and computer in order to obtain the models that provide the most insight for the task. In ADE identification, rules that are predictive of drug use can be taken as *candidate* ADEs, but these candidate ADEs must then be vetted by a human expert. If some of the rules are found to still capture other factors besides drug effects such as indications, then these rules should be discarded. This lesson is referred to as *iterative interaction*. Note that the prediction is in reverse not only in terms of causality, but more importantly in terms of the label of interest.

4.2 Reverse Machine Learning for ADE Surveillance

Learning adverse drug events can be defined as follows:

> **Given:** Patient data (from claims databases and/or EHRs) and a drug D,
>
> **Do:** Determine if evidence exists that associates D with some previously unanticipated adverse event.

Note that no specific associated ADE has been hypothesized, and there is a need to identify the event to be predicted.

How might one hypothesize ADEs in this setting? One may not be able to define an unanticipated ADE, but one knows what drugs each patient has taken. So, one might have the model predict drugs instead of predicting conditions. If a model can predict which patients are taking a drug D, then there must be some combination of clinical experiences more common among patients on the drug. If the data is limited to those events occurring after the drug started, then the model will focus not on common causes but on common effects. The contents of the model can then be treated as the definition of an ADE and used for subsequent association estimation, prediction, or inspection by an expert. This is the idea of *reverse learning*.

4.2.1 Formalizing Learning in Reverse

Given a (large) EHR database and a drug, the task is to find a condition that is related to the drug. To better understand the complexity of the problem, consider a model of a patient where the patient's attributes and medical history influence current disease status and laboratory results. The states in the model are a set of partially observed variables $\langle \mathbf{A}, \mathbf{C}, \mathbf{L}, \mathbf{D} \rangle$ at various points in time. $\mathbf{A}$ is a vector of attributes of the patient, such as gender, age, family history, and genetic information, $\mathbf{C}$ are conditions (diagnoses), $\mathbf{L}$ are lab tests, and $\mathbf{D}$ are drugs. Each vector contains a large number of variables; for example, an EHR typically includes over 10k reported conditions, and 4k to 5k different drugs. Given the dimensionality of the task, latent variables were not included in this model (Saria et al., 2010).

An ADE was defined as an unexpected dependency between an observed variable in $\mathbf{C}$ and an observed variable in $\mathbf{D}$, in the simplest case, or even some combination of variables in $\mathbf{D}$. To the best of this researcher's knowledge, this work is the first to consider the more complex case of combinations, but the simpler case of a single drug is considered first.

4.2.1.1 Related Work

A standard approach to this problem is to assume two timesteps: events that happened before (step 0) and after taking a drug D_j (step 2). Techniques such as disproportionality analysis (Wilson et al., 2003; Zorych et al., 2011) then search for a condition C_i such that its probability increases after taking drug D_j, i.e.,

$$P(C_{i,t_0} \mid D_{j,t_1}) < P(C_{i,t_2} \mid D_{j,t_1}) \text{ s.t. } t_0 < t_1 < t_2, \tag{4.1}$$

where $C_{i,t}$ denotes the condition C_i at time t. To do so, one must obtain estimators $\hat{P}(C_{i,t_0} \mid D_{j,t_1})$ and $\hat{P}(C_{i,t_2} \mid D_{j,t_1})$ and test against the null hypothesis. In practice, estimates can be confounded by other parameters. Typically, one will consider $\mathbf{A}$ and stratify at least over age and gender, and then weight the estimates. One can also go a step further and count time of exposure, as in observational screening (Ryan et al., 2013), where the condition C_i is considered the result of a non-homogeneous Poisson process with two rates, during and after usage of drug D_j. A different method is to take into account confounding between different drugs. For example, a Bayesian logistic regression method (Caster et al., 2010) takes into account all drugs, plus gender and age information, to estimate $\hat{P}(C_i)$.

Essentially, these different methods search conditions $C_{i,t}$ such that their posterior probabilities of occurrence are greater than some threshold δ

$$P(C_{i,t} \mid \mathbf{A}_{1:t}, \mathbf{C}_{1:t}, \mathbf{L}_{1:t}, \mathbf{D}_{1:t}) > \delta, \tag{4.2}$$

i.e., they search through the entire EHR for some conditions occurring with a non-trivial probability given the drug history. Given the size of the problem, they focus on different combinations of $\langle \mathbf{A}, \mathbf{C}, \mathbf{L}, \mathbf{D} \rangle$. The previous approaches to the problem can be described as an enumeration of

$$P(C_{i,t} \mid \mathbf{A}_{1:t}, \mathbf{C}_{1:t}, \mathbf{L}_{1:t}, \mathbf{D}_{1:t}), \tag{4.3}$$

given some fixed

$$\langle \mathbf{A}_{1:t}, \mathbf{C}_{1:t}, \mathbf{L}_{1:t}, \mathbf{D}_{1:t} \rangle. \tag{4.4}$$

4.2.1.2 Reverse Machine Learning

This chapter proposes *reverse learning*. Instead of a direct search for C_i, it proposes to enumerate over Equation 4.4 and compute

$$P(D_{j,k} \mid \mathbf{A}_{1:t}, \mathbf{C}_{1:t}, \mathbf{L}_{1:t}, \mathbf{D}_{1:t}) \tag{4.5}$$

for some time k, as one knows that, if C_i is an ADE for D_j, then $C_{i,l}$ will be in a learned model for $D_{j,k}$ where $k \leq l$. If one were to go forward in time and predict effects based on causes, one would have to explore an increasing number of effects. By going backward in time, one traces effects back to their causes. Since the interest is only in a limited number of causes in reverse learning, the set of all possible causes is much smaller than the set of all possible effects. Thus, the problem of learning models for every condition C_i is reduced to the problem of finding out whether C_i is in a model for D_j. As a result, standard learning technology can perform the search.

Note that this approach is akin to Bayesian inference, where one computes $P(C \mid E)$ by estimating $P(E \mid C)$. Indeed it reduces to this in the case where one just searches over fixed subsets. On the other hand, the advantage is not in the Bayesian approach itself, as Equation 4.5 is not necessarily always easier to estimate than Equation 4.3: both are estimated from counts. The advantage is in transforming the learning process and making the problem supervised.

The strong relation between this work and Bayesian learning suggests a connection between reverse learning and abduction (Sato and Kameya, 2002; Kakas and Flach, 2009). Notice that in this setting the goal is not as much to learn a set of abducibles for an existing procedure, as to learn a new concept. The problem is thus closer to the problem of predicate invention (Muggleton, 1994; Richards and Mooney, 1995; Davis et al., 2007; Muggleton et al., 2009). Such insights may help guide further progress in reverse learning.

4.2.2 Study Design

Reverse learning can be seen as a case–control study, where cases (positive examples) are the patients on drug D_j, and controls (negative examples) have not taken D_j. Choosing controls is fundamental to obtaining good study quality, which relies on making the case and control groups as similar as possible (Rosenbaum, 2004; Rothman et al., 2008). One

way to do that is treat each case patient as its own control, accomplished by splitting the patient timeline into one or more control and case periods based on when they received treatment. This is exactly what is done in this ADE identification approach (Figure 4.1) and is known as a before–after design (Cochran, 1983; Shadish et al., 2002). Alternatively, one could search for age- and gender-matched controls and use them as negative examples. In this case, for each positive example, a control is a patient of the same age and gender who is not on drug D_j. (Controls could be selected to be similar to the cases in other ways, for example, by sharing smoking status; age and gender are just the most common such features in clinical studies.)

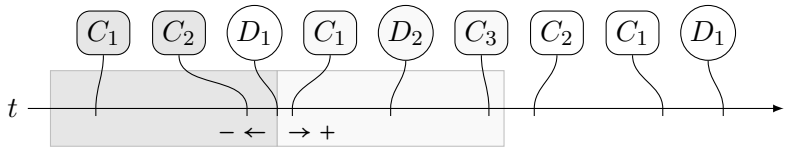


Figure 4.1: An example patient timeline with time windows for before and after a drug occurrence

Specifically, reverse learning works by comparing two intervals, one before the cause and one after, as illustrated by the darker and lighter boxes in Figure 4.1. The interval after the cause contains the events of interest, the possible effects, while the interval before the cause serves as a comparison baseline. For example, learning would exclude C_1 from consideration as a possible effect because it also occurs in the before interval; this leaves C_3 as the effect, possibly interacting with D_2.

4.2.3 Implementing Reverse Learning with Inductive Logic Programming

To apply this reverse learning algorithm, it needs to be considered in greater detail along with the following three factors:

1. EHR data are multi-relational and temporal, necessitating relational learning for this task (De Raedt, 2008).

2. The output of the learning process should be easy to interpret by the domain expert (Page and Srinivasan, 2003).

3. Generally, only a few patients on a drug D will experience novel ADEs (ADEs not already found during clinical trials). The learned model need not, and indeed most often should not, correctly identify everyone on the drug, but rather merely a subgroup of those on the drug while not generating many false positives (individuals not on the drug). This argues that this reverse learning problem actually can be viewed as "subgroup discovery" (Wrobel, 1997; Klösgen, 2002; Železný and Lavrač, 2006), in this case finding a subgroup of patients on drug D who share some subsequent clinical events.

This suggests using a relational rule-based classifier, since relational rules naturally induce subgroups on the data, are discriminant, and are often easy to understand. The experiments

utilize the ILP system, Aleph (Srinivasan, 2007). In the remainder of this section, for concreteness, the discussion is presented in terms of Aleph.

Another advantage of the multi-relational approach is that the body (precondition) of the rule does not have to be a single condition, but it can be a combination of conditions and lab results, possibly in a temporal order. Hence, ADEs that do not neatly correspond to an exact pre-existing diagnosis code can be discovered. Furthermore, the body of the rule can involve other drugs. So, ADEs caused by drug interactions can be captured. For example, it has recently been observed that patients on clopidogrel (Plavix) may have an increased risk of stroke (ordinarily prevented by clopidogrel) if they are also on omeprazole (Prilosec). This can be represented by the following rule:

$$\text{clopidogrel}(X) \leftarrow \text{omeprazole}(X) \wedge \text{stroke}(X). \tag{4.6}$$

Just because the rule is representable does not mean it will be learned. This depends on its support in the data, and the support of other rules that could score better, specifically as the support impacts the scoring function employed.

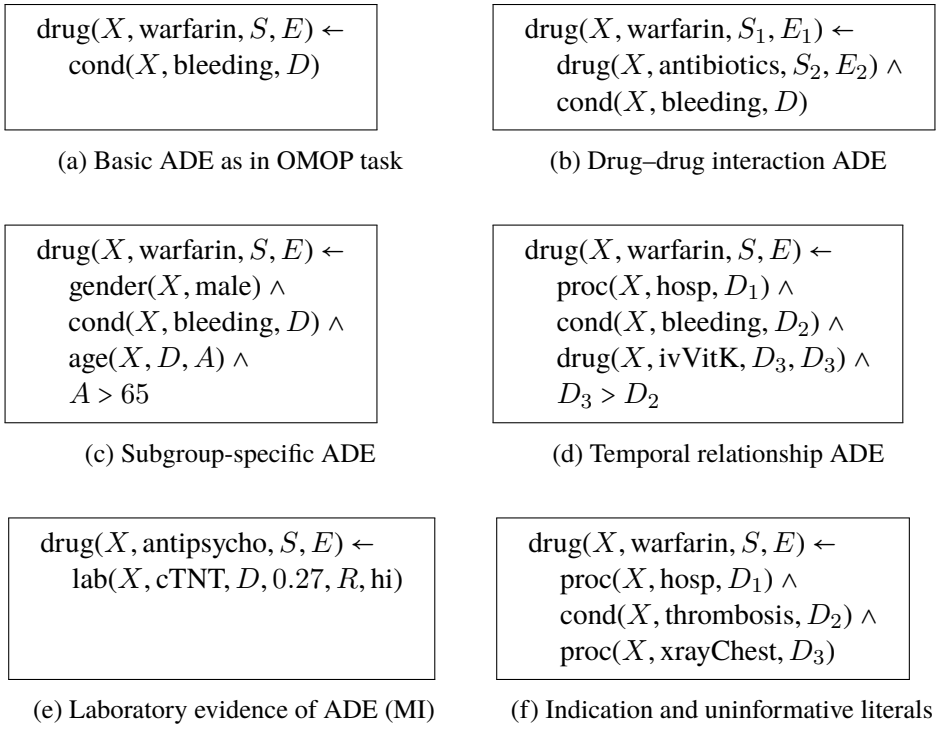
Aleph learns rules that predict a certain classification, in this case whether patients are on a particular drug or not. The data indicates who has taken a drug and who has not. These patients become the positive (on drug) and negative (non-using) examples for the supervised learning task. To learn a rule that distinguishes between positive and negative examples, Aleph picks a positive example, collects all the literals (facts in the data and background knowledge) pertaining to that example into a "bottom clause," and then searches for subsets of those literals that form a rule that maximizes a scoring function (typically coverage; see §4.2.4). Once Aleph has found the best rule for an example, it repeats the process on an unexplored positive example until there are enough rules to classify all positive examples as positive.

The literals in a rule are attributes shared by a subset of patients and can be considered together as a common effect of the drug. Such literals can be conditions (diagnoses) or condition classes (suggesting an ADE), drugs or drug classes (suggesting a drug–drug interaction), demographics (selecting subsets of the population), vitals or lab results (suggesting an ADE or a common criterion), procedures (suggesting a "corrected" ADE or a common criterion), or other background knowledge such as (temporal) relationships between the above events (which tend to serve as additional criteria). Depending on the collection of literals, a rule may indicate an ADE or other medical relationship, or it may be uninformative. Figure 4.2 (p. 40) shows examples of these various types of rules.

Aleph tries to find a predictor of the drug for every patient that discriminates well, according to the scoring function, between those on the drug and those not on the drug. Because Aleph looks at rules based on each individual patient while scoring those rules on all patients, Aleph can find specific or rare relationships that only apply to a few patients, or find general relationships that apply to many patients.

4.2.4 Evaluating Rules as Potential ADEs

Aleph learns rules in the form of Prolog (Horn) clauses and scores candidate rules by coverage. Coverage refers to the difference in the number of positive and negative examples that satisfy a rule. Specifically, if p positive examples and n negative examples satisfy a rule, then the rule is given the coverage score $p - n$. Consider COX-2 inhibitors, for example.



(a) Basic ADE as in OMOP task

(b) Drug–drug interaction ADE

(c) Subgroup-specific ADE

(d) Temporal relationship ADE

(e) Laboratory evidence of ADE (MI)

(f) Indication and uninformative literals

Figure 4.2: Example rules for the ADE task. The schema is in Figure 2.2 (p. 12).

Because COX-2 inhibitors double the risk of MI, one can expect the distribution of selected patients to have twice as many MIs among the positive patients (those on COX-2 inhibitors). For example, in a sample of 200 positive patients who suffer a MI, one can expect about 100 negative patients to have a MI. The following rule says that a patient is likely on a COX-2 inhibitor if they suffered a MI. It would have a strong score of $p - n = 100$ and hence would be selected by Aleph unless some other rule scores even better.

$$\mathrm{cox2ib}(X) \leftarrow \mathrm{mi}(X) \tag{4.7}$$

Coverage is only the default score and can easily be replaced by any user-defined scoring function. Coverage does well at learning rules that capture interesting relationships and discriminate well, but does not do well at identifying ADEs. Thus, coverage is used to learn rules that hypothesize interesting ADEs, but then those rules are passed onto a second stage that evaluates the ADE potential of each rule according to a different, causally-motivated temporal score.

4.2.4.1 Temporal Score

The most critical part of identifying ADEs is evaluating proposed ADEs based on their causal plausibility. Epidemiologists often estimate causal plausibility by estimating relative risk, which compares the likelihood of some adverse event, given that a patient took a drug, to the likelihood of that same adverse event, given that a patient did not take the drug. The temporal scoring function accomplishes similar goals. It estimates the probability of an

adverse event given a drug and uses it to rank many possible ADEs, thereby comparing ADEs with non-ADEs.

Specifically, the temporal score (Equation 4.8) estimates the probability that a specific condition, c, occurs after a specific drug, d, in patients who experience both. The score then adjusts for the relative frequencies of other associated drugs and conditions by dividing by their probabilities: the probability that *any drug* occurs before *this condition* and the probability that *any condition* occurs after *this drug*. Here, D and C are the sets of drugs and conditions under investigation, so $d \in D$ and $c \in C$, and t is the time of the first occurrence in a patient.

$$\frac{P(t_d < t_c \mid d, c)}{P(t_D < t_c \mid D, c)P(t_d < t_C \mid d, C)} \tag{4.8}$$

This score is how the ADE likelihood of a particular drug–condition pair is estimated.

4.3 Experiments

These experiments consider two cases. In the first case, drugs were associated with with specific conditions or candidate ADEs. In terms of relational learning, an association is represented by a rule, or definite clause, whose head is an atomic formula built from a predicate naming the drug and a variable standing for the patient, and whose tail is an atomic formula built from a predicate naming the condition and the same patient variable; this form is illustrated by the COX-2 inhibitor and MI rule above. In this case the reverse learning approach is another way to carry out a standard association study, differing only in the scoring function employed. In the second case, a list of candidate ADEs or conditions was not assumed; instead an ADE was represented by any conjunction of atomic formulas with predicates naming entities from the EMR such as conditions, observations (labs or vitals), or other drugs, or possibly predicates defined in a background theory such as before. In this case reverse learning extended beyond the standard association study methodology.

4.3.1 OMOP Experiments and Results

The first set of experiments was with a large, real-world health insurance claims database available through OMOP and containing over 1.2 million subjects, 17M drug reports, and 29M condition reports, for a total of 1300 drugs and over 10k conditions. This was one of several databases available for evaluation of methods for ADE discovery (Ryan et al., 2010). Rule learning methods were evaluated on the OMOP ADE identification task, a set of drug–condition pairs including 9 known ADEs and 44 non-ADEs (see §2.4.2, p. 10).

As a first study, because all the other methods tested by OMOP ranked only drug–condition pairs, Aleph was limited to rules consisting of only a single condition in the body of the rule, that is, rules of the form of the following example:

$$\text{warfarin}(X) \leftarrow \text{bleeding}(X). \tag{4.9}$$

With its default coverage scoring function and this rule length constraint, the rules learned by Aleph scored no better than chance. Consequently, later experiments learned rules by coverage and then scored their ADE likelihood with the temporal score.

The temporal score estimated ADE likelihood competitively with other methods on the OMOP task (Madigan and Ryan, 2011; Ryan et al., 2012). Those results reported that the best area under the ROC curve achieved by any method on any OMOP database is 0.78, while the best AUC ROC achieved for the MSLR database is 0.73. This is quite high considering that many approaches did no better than chance (AUC ROC of 0.5). As seen in Figure 4.3, this method achieves 0.76 AUC ROC for MSLR, which is the best result on MSLR and competitive with other results in general.

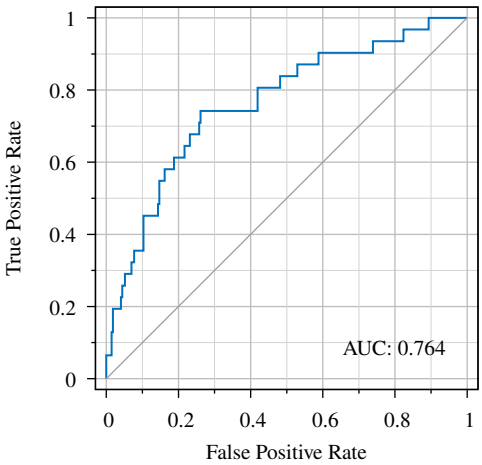


Figure 4.3: OMOP task ROC curve for MSLR.

This researcher believes the temporal score is effective because it captures aspects of causality. The score uses temporal ordering to help it decide whether a drug causes a condition. Further, it tries to focus on the main effect by adjusting for the influences of other drugs and conditions. The structure of the score is similar to pointwise mutual information which is another possible reason the score is effective.

The main benefit of using reverse machine learning with Aleph comes only with extending the possible lengths of the rule bodies. The next experiment was to do so with the same data set. Runs of this type take substantially longer, varying from twenty minutes to almost seven hours depending on the drug. There was no longer a ground truth against which to score these more complex rules, but their potential value was able to be evaluated, especially their ability to pick up on drug–drug interactions. One of the top-scoring rules was:

$$\text{warfarin}(X) \leftarrow \text{bleeding}(X) \wedge \text{antibiotics}(X). \tag{4.10}$$

This rule represents a rediscovery that antibiotics elevate the risk of bleeding in patients on warfarin (Baillargeon et al., 2012), and the rule scores significantly better than a rule with bleeding alone.

4.3.2 Marshfield Clinic EHR Experiments and Results

The second set of experiments was with a very different EHR. Marshfield Clinic has one of the oldest internally developed EHRs in the U.S., with coded diagnoses dating back to the early 1960s. It has over 13,000 users throughout central and northern Wisconsin. Data

Table 4.1: Bodies of top 10 rules learned for COX-2 inhibitors, using only single conditions.

Rule	Pos	Neg	Tot	P-Value
condition(A,_,'790.29','Abnormal glucose').	333	137	470	6.80e-20
condition(A,_,'V54.89','Aftercare, other orthopedic').	403	189	592	8.59e-19
condition(A,_,'V58.76','Aftercare, surgery genitourinary sys').	287	129	416	6.58e-15
condition(A,_,'V06.1','Diphtheria, tetanus, pertussis').	211	82	293	2.88e-14
condition(A,_,'959.19','Other injury of other sites of trunk').	212	89	301	9.86e-13
condition(A,_,'959.11','Other injury of chest wall').	195	81	276	5.17e-12
condition(A,_,'V58.75','Aftercare, surgery teeth, oral cavity, digestive sys').	236	115	351	9.88e-11
condition(A,_,'V58.72','Aftercare, surgery nervous sys, NEC').	222	106	328	1.40e-10
condition(A,_,'410','Myocardial infarction').	212	100	312	2.13e-10
condition(A,_,'790.21','Impaired fasting glucose').	182	80	262	2.62e-10

collected for clinical care is transferred daily into a data warehouse where it is integrated. The data warehouse is the source of data for this study. Programs were developed to select, de-identify by removing direct identifiers, and then transfer the data to a collaboration server. For this work, the specific tables used were: ICD-9 diagnoses, observations (lab results and other observations such as weight, blood pressure, and height), three sources of medication information, and patient demographics (gender and birth date). Also associated with every entry was a date, so Aleph was with background knowledge predicates to compare dates.

Two drugs were studied, warfarin and rofecoxib (Vioxx). For warfarin, the approach easily rediscovered the known ADE of bleeding, together with the common treatment for warfarin-induced bleeding (phytonadione, or vitamin K1).

$$\text{warfarin}(X) \leftarrow \text{bleeding}(X, T_1) \wedge \text{phytonadione}(X, T_2) \wedge \text{before}(T_1, T_2) \qquad (4.11)$$

Rofecoxib is a drug that was pulled from the market because it was found to double the risk of MI. The next experiment tested whether Aleph would uncover this link with MI if the link were unknown. Rofecoxib belongs to a larger class of drugs called COX-2 inhibitors. The overall goal was to identify possible ADEs caused by COX-2 inhibitors. In the reverse ML approach, the specific goal of running Aleph was to learn rules that accurately predict which patients used COX-2 inhibitors. These rules would then be vetted by a human expert to distinguish which were merely associated with indications of the drug (diseases or conditions for which the drug is prescribed) and which constituted possible ADEs (or other interesting associations, such as off-label uses for the drug).

First, the methodology was validated with a run in which only diagnoses were used and rules were kept as short as possible—one body literal (precondition) per rule. The run tested if the method would automatically uncover MI, a known ADE of COX-2 inhibitors. Table 4.1 shows the ten most significant rules identified by Aleph for a single run. Note that the penultimate rule (highlighted) identifies diagnosis 410 (MI) as a possible ADE of COX-2. The fact that this ADE can be learned from data demonstrates that this method is capable of identifying important drug interactions and side effects.

In some cases, a drug may cause an ADE that does not neatly correspond to an existing diagnosis code (e.g., ICD-9 code), or that only occurs in the presence of another drug or other preconditions. In such a case, simple one-literal rules will not suffice to capture the ADE.

Table 4.2: Bodies of top 10 rules learned for COX-2 inhibitors, using all relations.

Rule	Pos	Neg	Tot	P-Value
gender(A,'F'), drug(A,_,'Ibuprofen'), condition(A,_,'305.1','Tobacco use disorder').	509	177	686	4.25E-38
condition(A,B,'462','Acute pharyngitis'), drug(A,B,'Ibuprofen').	457	148	605	1.27E-37
drug(A,_,'Norgestimate-ethinyl estradiol'), gender(A,'F').	339	88	427	8.12E-36
condition(A,_,'V70.0','Routine medical exam'), drug(A,_,'Ibuprofen').	531	199	730	1.00E-35
condition(A,_,'724.2','Lumbago').	433	144	577	1.44E-34
condition(A,_,'462','Acute pharyngitis'), gender(A,'M').	502	186	688	2.02E-34
condition(A,_,'89.39','Nonoperative exams NEC'), condition(A,_,'305.1','Tobacco use disorder').	415	135	550	4.12E-34
drug(A,_,'Cyclobenzaprine HCl'), gender(A,'M'), drug(A,_,'Fluoxetine HCl').	493	189	682	3.60E-32
gender(A,'F'), lab(A,B,'Calcium',9.8), condition(A,B,'724.5','Backache NOS').	487	189	676	3.28E-31
condition(A,_,'V71.89','Observation for other suspected condition'), gender(A,'M').	492	193	685	5.35E-31

The next run used all of the background knowledge, including labs, vitals, demographics and other drugs. Table 4.2 (p. 44) shows the top ten most significant rules that were learned in the run, and demonstrates that rules learned by ILP are easy to interpret. According to Fisher's exact test, the rules identified positive cases significantly better than by chance.

The sobering aspect of this result is that Aleph learns over a hundred rules, and while some are potential ADEs, most appear to simply describe combinations of features associated with indications for the drug, as illustrated in Figure 4.2(f) (p. 40). At present a clinician must then sort through this large set of rules in order to find any evidence for possible ADEs, a laborious and imprecise process. Research is required to find ways to reduce the burden on the clinician, including automatically focusing the set of rules toward possible ADEs and presenting the remaining rules in a manner most likely to ease human effort.

4.4 Conclusion

This paper presents an initial study of machine learning for the discovery of unanticipated adverse drug events (ADEs). The key contributions and lessons learned for ML are:

- ML can be used "in reverse" when the real class value of interest—in this case, some unanticipated ADE—is not known at learning time. The experiments showed that this approach is able to successfully uncover ADEs.
- This research demonstrates the importance of learning from years of epidemiology research in selecting positive and negative examples for machine learning, as well as in setting scoring functions. The goal is not to find patterns in the patients who get prescribed a particular drug, because such patterns are already known—they are the indications of the drug. Hence, it is important to control by using data about patients before the drug, as well as by total amounts of data on various conditions following various drugs.

- Another lesson is that, despite censoring the data, a subgroup discovered with high accuracy does not automatically mean one or more ADEs have been uncovered. Instead, all rules must be vetted by a human expert to determine if they are representative of an ADE or of some other phenomenon, such as that patients on arthritis medication such as COX-2 inhibitors also suffer from other correlated ailments. Once these associated conditions are also censored, learning ideally should be re-run in case ADEs were masked by other rules that scored better.

- Another lesson is that data are multi-relational, including longitudinal (temporal), and hence may be best analyzed by methods that can directly handle such data. It would be desirable to take into account time from drug exposure to events, but this is a challenging direction because different drugs can cause ADEs over different ranges of time. Some drugs may cause an ADE within hours after they are taken, whereas others may have permanent effects that only manifest themselves as an ADE years later.

4.4.1 Applications for Machine Learning in Active Surveillance

In addition to the task of ADE identification that has been presented, machine learning approaches could support many drug safety needs, including:

1. *Identify and characterize temporal relationships between drugs and conditions across the population.* Is there an association between exposure to rofecoxib and cardiovascular events such as MI? If so, what is the likely time-to-onset of the event, relative to exposure? Does the risk increase over time and vary by dose?

2. *Identify drug–condition relationships within patient subpopulations.* Among elderly, what are the observed effects of a particular medicine? Among patients with renal impairment, what is rate of adverse events?

3. *Identify drug–drug interactions that produce harmful effects.* Which concomitant drug combinations produce elevated risks, relative to exposure to individual products?

4. *Identify risk factors and define patient subgroups with differential effects of a drug-related adverse event.* Which patients are more likely to experience adverse events? Which patients less likely to experience adverse events?

5. *Create models for predicting event onset.* Which patients are likely to have experienced a MI, based on available information about diagnoses (AMI and other CV terms), diagnostic procedures (EKG), treatments (PCI), lab tests (troponin, CK-MB), and other observations.

Identifying previously unanticipated ADEs, predicting who is most at risk for an ADE, and predicting safe and efficacious doses of drugs for particular patients all are important needs for society. With the recent advent of "paperless" medical record systems, the pieces are in place for machine learning to help meet these important needs.

Attribution

A version of this chapter was previously published as Page et al. (2012).

Chapter 5

Identifying ADEs using Markov Networks and Temporal Dependence

Logical, relational rules excel at modeling structured, complex, or relational data, while probabilities excel at modeling uncertainty and noise. Consequently, their combination, statistical relational learning (e.g., Getoor and Taskar, 2007), naturally suits machine learning from EHR data. This perspective anticipates that modeling patient histories with logical predicates combined into a probabilistic model will improve ADE discovery by representing the data without alteration or approximation, and by teasing apart the noisy relationships between all of the medical events. And these relationships, once adjusted for other influences and isolated by the model, will show evidence of ADEs and support additional, causally-motivated scoring functions. As explained in the following chapter, this idea is implemented by using feature functions (equivalent to logical predicates) to represent patient event sequences, a log-linear Markov network to model their joint probability distribution in the EHR database, and various scoring functions based on temporal dependence to evaluate the likelihood of ADEs.

5.1 Empirical Causal Discovery

The detection of adverse drug events (ADEs) in observational data is a challenging task that presents significant opportunities for improving public health. ADE detection typically proceeds by designing and conducting an observational study, a difficult, resource-intensive, and error-prone endeavor, even without considering that there is no guaranteed way to remove all confounding. The scale and difficulty of the problem requires automatic, computational approaches to augment the existing spontaneous reporting, clinical research, and epidemiological studies. This chapter develops such an algorithmic technique for causal discovery in observational medical data that is free from the errors of human study design while accurately identifying ADEs.

In particular, this chapter presents a method for *empirical causal discovery*, where the goal is to distinguish causal relationships by ranking them from causal to non-causal, given only observational data. The method employs an undirected probabilistic graphical model (a Markov network, MN) to model the temporal relationships between random variables in context, and then scores the relationships by their adjusted probabilistic temporal precedence.

This method will be called Markov-network-assisted temporal scoring, or MNATS. By incorporating features of other causal discovery methods into standard probabilistic models, MNATS combines benefits of purely causal approaches and purely predictive ones.

MNATS differs from other approaches in that it is intentionally approximate and free of assumptions. It only makes the arguably fundamental assumption that causes precede their effects in time. This is in contrast to many techniques which make strong assumptions or are restricted to idealized settings. Such techniques fall into three main categories: directed probabilistic graphical models, time series, and computational epidemiology.

Directed probabilistic graphical models (Bayesian networks, BNs) are structural causal models (SCMs) when their structure matches that of the underlying causal system. SCMs are a class of formal causal models that include BNs, structural equation models, and the classic framework of potential outcomes (counterfactuals) (Pearl, 2009). In the perspective of these models, causal discovery is a network structure learning problem. (The measurement of causal effects and adjustment for confounders is straightforward given a network.) The structure learning algorithms developed for SCMs, such as PC (Spirtes et al., 2000), rely on measuring conditional independence to detect provably causal structures. MNATS differs from these constraint-based structure learners because it is not concerned with sound inferences about causal relationships, only their empirical assessment. Further, MNATS is temporal whereas SCMs often operate atemporally.

On the other hand, methods for time series are explicitly temporal, but they often assume samples at regular time steps, or require a way to interpolate or approximate such complete histories of variable values. This is this case for Granger causality (Granger, 1969) and dynamic Bayesian networks (Dean and Kanazawa, 1989). Continuous time Bayesian networks (CTBNs) (Nodelman et al., 2002) model the value of every variable at every instant and therefore require complete continuous trajectories of variables for learning. Likely trajectories can be inferred, but CTBN inference is very costly and often intractable. In contrast to continuous trajectories or regular samples, MNATS can handle point events at arbitrary times.

The most similar approaches to MNATS are computational versions of epidemiological study designs (such as retrospective cohort, case–control, and self-controlled studies), which can handle events at arbitrary points in time and adjust for bias and confounding. These studies are designed to estimate the effect size of a hypothesized cause–effect pair while controlling for known factors. They fit into the potential outcomes framework (Rubin, 1974), in which outcomes are compared between groups of exposed and unexposed, assuming exchangeability of subjects. Propensity scores (Rosenbaum, 2002) and inverse probability of treatment weighting (Robins et al., 2000) ensure the comparability of treatment groups by matching/weighting subjects so that they are conditioned on the same measured (and unmeasured) factors.

These types of epidemiological study designs are usually repeated for each such cause–effect pair and therefore must be well-calibrated to produce a good ranking of many pairs. MNATS has the benefit of considering all desired pairs at once so as to estimate the best ranking, but does not attempt to estimate effect sizes. In this sense, it addresses the harder problem of collective estimation but does so in a less precise way, seeking only the correct ranking of effect sizes rather than their precise magnitudes.

Some methods do not fit into the above categories, particularly methods that model aspects of the distribution of events and then draw inferences based on properties of the

distribution. Some examples of this type of method are Shimizu et al. (2006), Peters et al. (2009), and Mooij et al. (2011). MNATS shares this overall approach in its unsupervised modeling of temporal relationships between events, and its evaluation by properties of the probability distribution according to the learned model.

5.2 Adverse Drug Event Detection

Detecting ADEs earlier in the process of bringing drugs to market could significantly improve public health. While side effects are usually discovered during the clinical trials for a medication, it is possible that rare side effects will remain undetected until the drug is on the market and taken by many more people. Therefore, automatic, algorithmic techniques for detecting ADEs "in the wild" are crucial for patient safety. In response, organizations such as OMOP and OHDSI, in combination with many universities, have researched such pharmacosurveillance techniques, but there are still problems to be solved.

Even though analysis methods are still being developed, the data needed for observational study of drug safety is conveniently being collected by computerized medical care systems, including electronic medical records (EMR) and administrative claims (coding and billing). EMR systems can track vital signs, symptoms, diagnoses, notes and reports, prescriptions, tests, imaging, procedures, etc.: all the events and records of medical care. Administrative systems usually track a subset of this information. In both cases, the data is naturally in relational database form, which may not simply be "flattened" into a single table for use with typical statistical and machine learning methods without losing information or altering frequencies in the data. Thus, special techniques for relational data must be used (such as multi-relational machine learning) or care must be taken in designing a transformation. Thankfully, for ADE detection, everything centers around the patient, and so a natural and equivalent representation is a timeline of clinical events.

A patient's clinical history is a continuous timeline of dynamic factors, but it only ever appears as a sequence of samples. Each sample, or event, can be thought of as a tuple, $(t, e, v)_i$, containing the time t, event type e, and value v of the event. The events are ordered by time and indexed by i. The values can be complex, like notes and images, or simple, like lab and vitals measurements, or nonexistent, like for occurrences of diagnoses and prescriptions. In the following sections, "event" will usually refer to an event type rather than an event tuple.

Medical data comes with complications, mainly because it is very noisy, but also because it was most likely recorded for a purpose other than analysis. There can be temporal uncertainty (events can be shifted in time or reordered), event uncertainty (events can be repeatedly recorded without having reoccurred in the patient or be recorded differently from reality, usually for billing purposes), and value uncertainty. The wrong level of abstraction is another source of problems, like insufficient time precision or overly specific medications. All these issues can be hard to address as they often depend on aspects of institutional processes and on other (possibly unrecorded) events and information.

5.3 Markov Network Model

In this work, relational medical data is represented as sequences S of clinical events, one sequence s per patient. This solves the problem of trying to flatten the data for each patient into a feature vector, but it introduces the problem of modeling arbitrary sequences of events. One could choose to have random variables represent time steps and range over events, or represent events and range over time steps. The latter does not handle multiple occurrences of the same event, but limiting events to their first occurrences solves this problem (and helps address temporal uncertainty). Thus, each event is represented by a random variable which can take on the allowed time step values or "none," meaning that event did not occur.

A log-linear Markov network (MN, Equations 5.1 and 5.2, (e.g., Koller and Friedman, 2009)) is used to model the distribution of patient event sequences $P(S)$.

$$P(S = s) = \frac{1}{Z} \exp\left(\sum_i \theta_i f_i(s)\right) \tag{5.1}$$

$$Z = \sum_{s \in S} \exp\left(\sum_i \theta_i f_i(s)\right) \tag{5.2}$$

The MN incorporates binary features for the occurrences of each event and for the temporal ordering of each pair of events. These features are the feature functions f_i in Equation 5.1, and can be thought of as the following logical predicates applied to a sequence S:

- *event*(S, X): true if event X occurs in sequence S
- *co-occur*(S, X, Y): true if both events X and Y occur in sequence S
- *before*(S, X, Y): true if events X and Y occur in sequence S and X occurs earlier than Y

The *before* and *co-occur* features span any number of time steps, which allows them to pick up on long-range effects.

Given a data set of patient event sequences, the MN is fitted to the data using maximum likelihood. Then the weights of the features and the corresponding inferred probabilities are fed into various scoring functions.

These features induce a MN graph structure that completely connects all the random variables. Since this structure is not decomposable, the options for inference are limited to sampling or exact inference. For the sake of precision, exact inference was chosen along with limiting the size of the problem to ensure feasibility. The size of the space of sequences with at most one occurrence of each event is a function of the number of events e and the maximum sequence length n:

$$\sum_{l=0}^{n} \frac{e!}{(e-l)!}. \tag{5.3}$$

The purpose of modeling patient event sequences with these features is to capture the marginal temporal relationships between every pair of events and to adjust those relationships in the context of all the other events. The temporal relationships can then be used to infer causal relationships. Of course, the main problem in causal inference is identifying and adjusting for the effects of confounders. By connecting each pair of events, each pairwise relationship can take on its individual responsibility for the observed data. This allows the

MN to automatically adjust effects in the presence of all other effects and therefore address the issues of confounding.

5.4 Temporal Scoring

Traditionally, investigators have used relative risk (in cohort studies) and the odds ratio (in case–control studies) to assess causality in observational studies. In the case of a MN modeling temporal relationships, it is not clear what score would best assess causal relationships among pairs of events. Several possibilities are considered here, focused on the idea of quantifying the strength of temporal relationships in context.

Treating the MN like a regression model, the weights of the *before* features show the marginal relative importances (marginal multiplicative probability effect) of those ordered event pairs in the distribution of sequences. Therefore, one can rank drug–condition pairs directly by the weight of the corresponding *before* feature,

$$\mathrm{BF} \coloneqq w(d \to c). \tag{5.4}$$

Considering that the strength of the temporal precedence is more important, one can rank by the smoothed ratio of the exponentiated weights (REW),

$$\mathrm{REW} \coloneqq \frac{e^{w(d\to c)} + \alpha_1}{e^{w(c\to d)} + \alpha_2}. \tag{5.5}$$

The alphas shrink the estimate towards the null hypothesis and help avoid extreme ratios when the size of either weight is small.

Another measure of the strength of temporal precedence is the prevalence of that event order compared to the prevalence if the events were independent. One way to formulate plain dependence as a score is

$$\frac{P(d,c)}{P(d)\,P(c)}. \tag{5.6}$$

To adapt this to the temporal realm recognize that $P(d,c) = P(d \to c) + P(c \to d)$, and that, if d and c are independent, $P(d \to c) = P(c \to d)$ and $P(d,c) = 2P(d \to c) = 2P(c \to d)$. Thus, a version of temporal dependence for $d \to c$ is

$$\mathrm{TD} \coloneqq \frac{2P(d \to c)}{P(d)\,P(c)}. \tag{5.7}$$

A very similar score replaces the denominator terms with their transition probability equivalents to form temporal transition dependence (TTD),

$$\mathrm{TTD} \coloneqq \frac{2P(d \to c)}{P(d \to C)\,P(D \to c)}. \tag{5.8}$$

Here D and C are the sets of drugs and conditions of interest, $f(x \to Y) = \sum_{y' \in Y} f(x \to y')$, and $f(X \to y) = \sum_{x' \in X} f(x' \to y)$, in terms of some function f (probability, counts, etc.).

These denominator terms are very similar to the marginal probabilities $P(d)$ and $P(c)$, but focus on the space of drugs occurring before conditions, and therefore can be considered better normalizing terms.

The above score was inspired by its connection to temporal dependence and also by the temporal score in Page et al. (2012),

$$\mathrm{TS} := \frac{P(d \to c \mid d, c)}{P(d \to C \mid d, C)\, P(D \to c \mid D, c)}, \tag{5.9}$$

which is estimated in terms of counts on event sequences with

$$\frac{\frac{\#(d\to c)+\alpha}{\#(d\to c)+\#(c\to d)+2\alpha}}{\frac{\#(d\to C)+\alpha}{\#(d\to C)+\#(C\to d)+2\alpha}\,\frac{\#(D\to c)+\alpha}{\#(D\to c)+\#(c\to D)+2\alpha}} \tag{5.10}$$

using an additive smoothing value of $\alpha = 1$.

Adjusted relative risk is the standard score for assessing causality in cohort studies. Similarly, the probabilities as modeled and adjusted by the MN can be used to calculate the relative risk. However, the conditional probabilities in the definition of relative risk (RR) must be interpreted as temporally ordered, as is done in the context of a cohort study. This means the partition of patients having the condition before the drug ($c \to d$) is considered unexposed to the drug at the time of the condition. The result is the following expression in terms of MN probabilities.

$$\mathrm{RR} := \frac{P(c \mid d)}{P(c \mid \neg d)} = \frac{\frac{P(c,d)}{P(d)}}{\frac{P(c,\neg d)}{P(\neg d)}} \overset{\text{temporally}}{=} \frac{\frac{P(d\to c)}{P(d)-P(c\to d)}}{\frac{P(c)-P(d\to c)}{1-P(d)+P(c\to d)}} \tag{5.11}$$

This MN-adjusted relative risk is compared to the unadjusted (crude) relative risk which is computed using the following expression in terms of counts on the event sequences. This expression retains the temporal interpretation. "$\#(\cdot)$" is the total count.

$$\mathrm{cRR} := \frac{\frac{\#(d\to c)}{\#(d)-\#(c\to d)}}{\frac{\#(c)-\#(d\to c)}{\#(\cdot)-\#(d)+\#(c\to d)}} \tag{5.12}$$

Finally, ranking sum ensembles are used to combine the predictive benefits of the above scores. A ranking sum ensemble takes as input the ranks of two or more scores and re-ranks by the negative of the sum of corresponding elements. (The negative keeps small sums at the top of the ranking.) Concretely, the ranking sum ensemble score for a given example x (a drug–condition pair in this case) and two scoring functions f_1 and f_2 is

$$\mathrm{RSES}(x) := -\left(\mathrm{rank}(f_1(x)) + \mathrm{rank}(f_2(x))\right). \tag{5.13}$$

5.5 OMOP Task, Data, and Methods

Causal discovery performance was tested on an ADE detection task developed by the Observational Medical Outcomes Partnership (OMOP) (see §2.4.2, p. 10). OMOP was an organization that created standards for interoperable observational healthcare databases and researched methods to evaluate the safety of medical products. The general task they developed is to rank drug–condition pairs from causal to non-causal using data from an EMR database. OMOP developed the task by selecting 10 drug classes of interest and 10

health outcomes of interest. They then selected 53 pairs from the Cartesian product of those sets, 9 of which are true, known ADEs. The selections were made according to research objectives, drug labeling, and extensive literature review (Ryan et al., 2012). The specific pairs are listed in the same article in Table III, or in Table 2.1 (p. 10), and were used exactly as defined to ensure fair comparison.

The task was evaluated on data from five databases (Table 3.3, p. 25), also provided by OMOP, although they were commercially available. In the OMOP research laboratory environment, these databases were converted to a common data model with a single schema and unified coding systems.

MNATS was compared on the OMOP task against eight methods from the OMOP methods library whose performance was studied by Ryan et al. (2012). The OMOP methods are described in that paper and in the following summaries.

CCO Case–crossover implements a crossover study design that compares the outcome risk during an interval of exposure to the risk in a non-exposure interval from the same person (Schneeweiss et al., 1997).

CCS Case–control surveillance implements a standard case–control study design that estimates risk odds ratios after matching by demographic factors (Rosenberg et al., 2012).

DP Disproportionality analysis compares the occurrence frequency of exposure–outcome pairs to their expected frequency if the events were independent (Zorych et al., 2011).

HDPS High-dimensional propensity score implements a new-user cohort design that adjusts risk estimates by matching stratified propensity scores estimated with logistic regression (Schneeweiss et al., 2009).

ICTPD Information component temporal pattern discovery compares patterns of outcomes before and after exposure in the same person using information component disproportionality analysis (Norén et al., 2010).

IUD Incident user design implements an inception cohort design with a Cox proportional hazards model and propensity score matching for risk adjustment (Ray, 2003).

OS Observational screening implements a traditional cohort study design that compares outcome risk between groups of exposed and unexposed people (Ryan et al., 2009).

USCCS Univariate self-controlled case series compares outcome risk under exposure and non-exposure within each person and only among people who have had the outcome (Whitaker et al., 2006).

5.6 Experiments

Experiments were run to compare MNATS to the OMOP methods described above. The experiments used the OMOP task and the five OMOP databases. MSLR served as the development and tuning data. The scoring functions to be evaluated on all five databases

were picked based on their performance on MSLR.[1] This contrasts with the results from the OMOP methods which were not tuned. For each database, Ryan et al. (2012) reported the best performance of each method across all its parameter settings.

Patient event sequences were extracted from each of the five databases in temporal order. When multiple events occurred on the same day in the same patient, their order was randomized. These events were then limited by selecting only the first occurrences of events of interest in the task. Events within each patient were sampled to limit the sequences to a target length of six. This preserved longer-term effects compared to picking consecutive events. Because of this sampling, five data sets of sequences were extracted from each EMR database and their results were averaged. Note that the only differences between these extracted data sets are the sequences from patients who had more than six events of interest and therefore underwent sampling.

A log-linear MN with *event*, *co-occur*, and *before* features was then fitted to the data using exact inference and L-BFGS optimization. Probabilities for all the events, co-occurrences, and temporal relationships (all the feature predicates) were computed. These and the fitted parameters were then fed into the various scoring functions. The resulting scores were used to rank the drug–condition pairs and compute corresponding ROC areas. The crude scoring functions (cTTD, cTS, cRR) were computed from the same patient event sequences used to fit the MNs, rather than computed from all the data in the databases, to ensure fair comparison. The ensembles are indicated with hyphens (e.g., TTD–cTS means the ranking sum ensemble of TTD and cTS). The ROC areas for all the scoring functions and methods are in Table 5.1. The OMOP methods have the advantage of using all the available data in the databases rather than only the extracted patient event sequences, and also have their best results reported.

Each pair of methods was tested for better performance using a two-sided, paired t-test across the five databases. The top 20 of these comparisons are listed in Table 5.2, along with a few other notable comparisons. Comparisons with cRR have been omitted from the listing, but not the ranking, because almost all methods outperformed it. The pairwise performance analysis ignored BF and REW for the same reasons. Due to the large number of pairwise comparisons among the 16 included methods, no comparison was found to be statistically significant after controlling the false discovery rate at the 0.01 level with the Benjamini–Hochberg procedure. Nevertheless, some of the pairs had notable AUC ROC differences.

The ensemble TTD–cTS achieved the best average score overall (0.78). Among the OMOP methods, HDPS had the best average (0.76). Among the crude scoring functions, cTS had the best average (0.69), surprisingly outperforming the adjusted scoring functions. Both TTD and cTS were comparable to the averages of the OMOP scores. TTD–cTS performed quite a bit better than the average of the OMOP scores (by 0.10), but less so (by 0.08) after omitting the worst performer, IUD. TTD–cTS also outperformed its constituent scores, TTD (by 0.11) and cTS (by 0.09).

All methods performed consistently better than random (AUC ROC 0.5) except IUD. The weakest methods were IUD, cRR, TD, and REW on average. Indeed, even the OMOP average performed better than these methods.

[1] A database was used for tuning in this way because splitting up the 53 positive and negative pairs of the task into training, tuning, and testing sets would leave only three positives per split and be impractical.

Table 5.1: The ROC areas for the various methods and databases. Results above the double line come from Table VI in Ryan et al. (2012). Methods below the double line were developed in this work (except cTS) and are MN-assisted unless prefixed with "c." Category maxima are in bold. Some CCS runs were computationally infeasible and do not have results.

Method	MSLR	GE	MDCD	MDCR	CCAE	Avg
CCO	0.64	0.65	0.71	0.65	0.67	0.66
CCS	0.69	**0.67**	—	—	—	0.68
DP	**0.74**	0.63	0.70	0.68	0.69	0.69
HDPS	**0.74**	0.66	**0.83**	**0.79**	0.76	**0.76**
ICTPD	0.72	0.66	0.77	0.77	**0.78**	0.74
IUD	0.55	0.41	0.61	0.60	0.58	0.55
OS	0.67	0.64	0.68	0.65	0.68	0.66
USCCS	0.68	0.62	0.70	0.73	0.73	0.69
Avg OMOP	0.68	0.62	0.71	0.70	0.70	0.68
Avg no IUD	0.70	0.65	0.73	0.71	0.72	0.70
TTD–cTS	**0.86**	**0.66**	**0.80**	**0.77**	**0.81**	**0.78**
RR–cTS	0.76	0.64	0.74	0.72	0.73	0.72
BF	0.66	0.56	0.66	0.62	0.57	0.62
REW	0.68	0.58	0.62	0.58	0.52	0.60
TD	0.65	0.52	0.61	0.57	0.59	0.59
TTD	**0.72**	0.57	**0.73**	0.68	**0.68**	0.67
RR	0.68	0.52	0.62	0.63	0.61	0.61
cTS	0.71	**0.66**	0.70	**0.74**	0.64	**0.69**
cTTD	0.64	0.57	0.71	0.67	0.63	0.64
cRR	0.59	0.50	0.60	0.61	0.56	0.57

Notably, TTD outperformed TD by 0.08, suggesting that marginal transition probabilities are more important to causal discovery than plain marginal probabilities. It also appears that the MN was successful in automatically adjusting for confounding effects, as RR outperformed cRR on every database by an average of 0.04, and TTD likewise outperformed cTTD on every database. However, despite adjustment, RR was still weaker than the average of the OMOP methods, but its ensemble, RR–cTS, had a better average than the OMOP methods.

5.7 Discussion

Scoring techniques based on temporal dependence performed as well as the OMOP methods despite their different perspectives. The methods developed in this work are simple and do not rely on epidemiological approaches or other expert knowledge. However, they do incorporate functionality important for causality: temporality and adjustment for other effects. Further, that ensembling outperforms the OMOP methods and improves performance

Table 5.2: The top 20 and other notable comparisons using a two-sided, paired t-test. Almost all methods outperform cRR; those rows are not listed.

Rank	Better	Worse	P-Value
05	TTD–cTS	TD	1.17e-03
06	ICTPD	TD	1.28e-03
07	TTD–cTS	IUD	1.53e-03
08	HDPS	TD	1.55e-03
09	HDPS	IUD	2.12e-03
10	RR–cTS	TD	2.12e-03
11	ICTPD	IUD	2.20e-03
16	TTD–cTS	RR	3.28e-03
17	RR–cTS	IUD	3.53e-03
19	HDPS	RR	4.88e-03
20	ICTPD	RR	4.90e-03
21	cTS	TD	7.21e-03
22	DP	TD	7.36e-03
23	USCCS	TD	8.19e-03
24	TTD–cTS	OS	8.55e-03
25	TTD–cTS	cTTD	8.56e-03
26	cTS	IUD	8.95e-03
27	RR–cTS	RR	9.16e-03
28	DP	IUD	9.31e-03
29	USCCS	IUD	9.36e-03
49	TTD–cTS	TTD	3.93e-02
50	TTD	TD	4.14e-02
51	TTD–cTS	cTS	4.50e-02

over its constituents, demonstrates that basic machine learning techniques are relevant and applicable to problems in epidemiology.

MNATS with the TTD scoring function works because it focuses on temporal relationships, by modeling the transitions between variables, and because it adjusts effects in the context of other effects, by modeling the joint distribution in detail. Evidence of adjustment can be seen in the performance of RR when compared to cRR and TTD when compared to cTTD. On the score side, the performance of TDD compared to TD suggests that it is much more important what events transition to what other events rather than merely if the events occur. MNATS thus incorporates temporality and adjustment, two fundamental aspects of causal discovery.

MNATS relies on no other assumptions and therefore is more widely applicable than methods built for idealistic scenarios or that require lots of background knowledge or expertise. It also only requires the minimum of temporal information: just the order of events and not their precise times or lags.

One of the main disadvantages of MNATS is that it does not estimate effect size. However, in many scenarios, it may be sufficient to collectively rank causal relationships, especially if a good ranking would be as effective at prioritizing ADEs for further investi-

gation.

One of the more serious limitations of MNATS compared to other approaches is its inability to handle the full database worth of data, both in terms of the length of event sequences and in terms of the number of events of interest. Including full patient histories would provide evidence of more relationships to the model. While administrative claims data is somewhat sparse and sequences of length six cover the large majority of patient histories without sampling, this is partly an effect of the small number of events, and, in any case, does not extend to the larger, denser histories found in EMR data. Expanding the number of events considered at once is likely crucial to improving the performance of MNATS. Considering more events allows the MN to learn more specific, truer, and, consequently, sparser relationships than what is possible when those relationships are represented in proxy by their "projection" down to fewer events. Including more events also increases the likelihood of observing variables that would otherwise be unmeasured confounders, or observing variables that screen off confounding effects.

Both the issues of number of events and sequence length are issues of scale. Thus, increasing the scale at which MNATS operates is an important next step. This will undoubtedly involve approximate inference. Other than this, it is not clear how to address scaling MNATS, but there may be ways to approximate and sparsify the graphical structure of the model to enable message passing inference algorithms.

Using a ranking sum ensemble is causally justified because it essentially averages its inputs, similar to forming a consensus, which does not change the causal validity of the inputs. Imagining the algorithms are experts, there is greater confidence when their opinions agree yet nothing is taken away from the causal knowledge in their individual opinions. In this way, the ensemble affords the agreements greater weight, as the sum of the inputs moves towards the extreme. Disagreements are afforded lesser weight, as their sum moves towards the middle of the range. It is these averaging effects that make ensembles more robust and less prone to overfitting.

However, ensembles only provide benefit if the inputs are sufficiently different; there must be disagreements to settle. The idea of using an ensemble with MNATS was born out of noticing differences in the rankings produced by various scores, but it was not clear what was responsible for the differences. In the case of TTD–cTS, the constituent scores differ in that the MN adjusts TTD and cTS uses conditional probabilities. It may be possible, then, that over-adjustment (TTD) combined with under-adjustment (cTS) is responsible for their joint performance. In any case, the ensemble combines their strengths and is more robust as a result.

The lack of success of the BF and REW scoring functions suggests that it is difficult to interpret the parameters of a MN on their own as if they were regression coefficients. This counterintuitive result makes the interpretation of MN parameters an attractive topic for further study. These results were also surprising in that the REW scoring function performed very well on synthetic data from the second generation OMOP simulator (OSIM2). However, it was later discovered that cRR performs just as well as REW on OSIM2 data, which suggests that the synthetic relationships in that data are not representative of real-world data.

5.8 Conclusion

The performance of MNATS on the OMOP task shows that incorporating causal features, especially time, into standard machine learning methods can succeed at causal discovery. Indeed, such causal features appear to be necessary, as ignoring them leads to poorer results. The performance of MNATS also shows that combinations of simple (even naïve) scoring functions can surpass the best performance of sophisticated, domain-specific methods. That these patterns hold over multiple databases demonstrates that MNATS generalizes to other EMR settings. Collectively, these results establish that ML is relevant and applicable to causal discovery problems in epidemiology, including identifying ADEs in observational data.

Chapter 6

Temporal Inverse Probability Weighting for Discovering ADEs Especially in Generic Drugs

Even though "reverse machine learning" and the various temporal dependence scoring functions may have done reasonably well at identifying ADEs in the OMOP task, they were not very successful at controlling for confounding. Despite applying reverse machine learning with a self-controlled study design, it was especially prone to associating a drug with the diseases it is used to treat (its indications), suffering "confounding by indication." This chapter addresses such confounding with a study design that controls for both treatment and changes over time, and with analysis methods that correct for treatment propensity and hypothesize effects. The methods employ off-the-shelf machine learning classifiers, making them easy to implement and widely applicable. This combination of study design and novel analysis successfully identifies ADEs in synthetic data and discovers causally-relevant differences between brand and generic versions of drugs.

6.1 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 (U.S. Food and Drug Administration, 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 patient choice and time-varying confounding, which motivate the development of a method that addresses 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 (U.S. Food and Drug Administration, 2017). This similarity actually means that any effect specific to a generic is unlikely to have been suspected based on evidence from the brand. 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, bioequivalence studies may surface similar hypotheses, 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 (see §5.2), 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 does not require 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. Furthermore, when the generic debuts 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 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. 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 an approach to causal discovery from observational data that analyzes an observational 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. This approach potentially generalizes to other tasks in similar settings, but evaluating its generalizability beyond generic drug pharmacosurveillance is 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 it hypothesizes plausible effects of generic

drugs when analyzing real EHR data.

6.2 ADE Discovery

6.2.1 Adverse Drug Events

ADEs are estimated to account for up to 30% of hospital admissions and at least $30 billion in annual healthcare costs (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, and studies generally show equivalence (e.g., Desai 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 especially 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 (U.S. Food and Drug Administration, 2016).

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. More recently, many of the original OMOP investigators continue working 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, some of which are described in §6.2.3.

6.2.2 Causal Discovery

The ADE detection task can be viewed as a special case of general causal discovery. The objective of causal discovery is to determine what direct causal relationships exist among a set of variables given measurements of those variables. For example, the variables may include many drugs and conditions, and one might want to know if Vioxx (when it was on the market) can cause myocardial infarction (MI). Or, one might want to know if switching patients from brand to generic methylphenidate can cause an increase in a particular ADHD symptom or code. The gold standard for testing for such causal relationships is a randomized controlled trial (e.g., a clinical trial). But generic drugs are not required to undergo clinical trials, so the main source of information about them is pharmacosurveillance, which is *observational*: one does not get to intervene and randomize patients to the brand or generic drug; one can only observe what drug they take and what happens to them subsequently, as recorded in EHR or claims data.

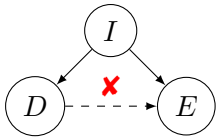


Figure 6.1: Structural causal model in which indication I (e.g., MI) has a causal effect on both a drug D (e.g., beta blocker) and an ADE E (e.g., death). I is a confounder of D and E: if I is not observed one might falsely conclude that D causes E.

In general, inferences from observational data are subject to *confounding* by other variables. For example, beta blockers are known to reduce the risk of MI, but from purely observational data one would be tempted to conclude the reverse, since patients on a beta blocker appear to have a higher probability of MI than those not on a beta blocker. The reason of course is that being at higher risk for MI leads to taking beta blockers, but it also leads to more MIs. Thus, MI confounds the effect of beta blockers on death. This is a common scenario called confounding by indication (MI "indicates" prescribing beta blockers) and is shown in Figure 6.1. In general, any variable could be a confounder, not just an indication. The antidote to confounding is to observe I and measure its influence on both D and E. Once measured, its influence can be removed, leaving just the effect of D on E (which could be nothing). This is called *adjusting* or *controlling* for confounding. However, it is impossible to observe and control for all variables that might influence a particular causal relationship of interest, so one cannot guarantee that inferences based on observational data are free from confounding.

6.2.3 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 method that OMOP evaluated was disproportionality analysis (Zorych et al., 2011). Disproportionality analysis constructs a 2×2 table of patient counts for a drug and a condition, and uses measures such as odds ratio or relative risk to ask if a higher association exists between the drug and condition than would be expected by chance. OMOP found disproportionality analysis methods perform relatively poorly at ADE detection in EHR data, most likely because of confounding variables.

One method that performed especially well in the OMOP evaluations was multiple self-controlled case series (MSCCS) (Simpson et al., 2013). It performs a regularized Poisson regression to predict the count of any event type, such as MI, based on a patient's exposure to drugs over a given time interval. Its success may lie in its use of a patient-specific baseline risk that partially adjusts for unobserved confounders. If multiple patients suffered an MI while on Vioxx, one might explain MI risk in part with a substantial positive coefficient on Vioxx in the Poisson regression, whereas a patient who took no drugs might have his MI explained instead by a high baseline risk. Other subsequent approaches have attempted to extend this idea by modeling risks that vary over time for a single patient (Kuang et al., 2017), though such a time-varying baseline must be heavily regularized, or by combining

patient-specific baselines with probabilistic graphical model learning (Geng et al., 2018).

The classic framework for causal inference is the Rubin causal model (Rubin, 1974), which provides the foundation for modern causal inference in both randomized and non-randomized studies. Units are divided into treated and control groups; then the response to treatment in each group is measured and compared. When the only difference between the groups is the treatment (as is the case in randomized studies), then the difference between the groups is the treatment effect. In nonrandomized (observational) studies, confounders may affect both the likelihood of treatment and response, thus obscuring the treatment effect. One way to control for this confounding is to first model a patient's likelihood for treatment, their propensity score (Rosenbaum and Rubin, 1983), and then select or reweight patients to balance the distributions of propensities in the treated and control groups, perhaps by inverse probability weighting (IPW) (Robins et al., 1994). Propensity scoring is sensitive to the type of propensity models constructed (often logistic regression models) and the validity of its assumptions, such as that there are no unobserved confounders.

Differential prediction[1] (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 in standardized testing, to search for systematic biases, and in marketing, to evaluate the effectiveness of targed advertising, 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 two variables at a time. Another body of work estimates all of the direct causal relationships among a set of variables at once: structural causal modeling (Spirtes et al., 2000; Pearl, 2009). In this framework, a structural equation model or causal Bayesian network represents the causal system (the "laws of nature"). Such a model does causal inference by answering queries about the effects of interventions or counterfactual situations. Causal discovery is done by learning the structure of the model, that is, by learning which variables directly affect which other variables. Alternatively, a structure can be presumed or assembled from other knowledge, such as individual relationships derived from controlled studies or experiments. Because structural causal models represent a larger causal system, they subsume the Rubin causal model (Pearl, 2009), but are more difficult to learn and are still subject to confounding, as described in Figure 6.1.

All of the work reviewed so far assumes that possible ADEs have been identified and precisely defined before the analysis. For example, OMOP's evaluation identified ten ADEs of interest and included precise, sometimes complex, definitions of ADE occurrence. One piece of prior work asserted that it may not be known what ADEs a drug might cause (Page et al., 2012). Therefore, one cannot put the unknown ADE into a graphical model as a variable, or use it as the target for supervised machine learning, such as for Poisson regression in MSCCS or differential prediction. Instead, Page et al. (2012) proposed using "reverse machine learning" to build a model to "predict" who takes a drug compared to controls, based on events that happened after starting the drug. In their approach, every case (drug taker) has a matching control (never taker) of the same age and gender (and

[1] Differential prediction is also known in various fields as uplift modeling, difference in differences, or structural mean models.

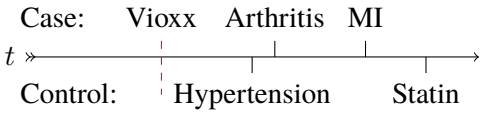


Figure 6.2: One case–control pair in reverse machine learning, which builds a model that distinguishes cases from controls given data starting after the first event of interest (Vioxx, red line).

ideally propensity, though they did not use that), and the only data used to discriminate between cases and controls is data *after* the case patient started the drug. For example, with the patient history in Figure 6.2 and what is known about Vioxx, reverse machine learning might build a model that predicts "case" if the patient has an MI and "control" otherwise. They used inductive logic programming to classify cases and controls, but in principle any classification method could be used.

6.3 Methods for Finding Differential Effects of a Generic Drug

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 appropriate study designs they become instruments of causal inference. This combination produces a general approach to causal discovery that applies equally well to any two treatments as it does to brand and generic versions of a drug.

6.3.1 Hypothesizing Effects Using Causal Discovery Machine Learning

Since effects are not predefined, the first challenge posed by generic ADE discovery is hypothesizing the causal effects. The proposed approach tackles this by searching for a function that maximimally distinguishes two treatments while minimizing confounding. The function is a ML model and the search is the training process. Training operates in terms of the features of the data, so building a model effectively selects a subset of informative features, features that in this case explain the differences between brand and generic. These principal features define the effects.

To make this work, the proposed approach constructs a supervised binary classification task. The typical way to do this would be to make a suspected ADE the class label and then learn to predict it. For example, if it was suspected that the generic version of a drug caused myocardial infarction (MI), then a model could be trained to predict who will have a MI. Then the model could be inspected to see if the generic drug proved to be a useful predictor of MI. But because it is *not known* in advance what differences exist between patients taking the brand and generic versions of a drug, one cannot perform such ordinary supervised learning. Therefore, the proposed approach instead sets up the classification task with the drug as the class label: generic-takers are positive examples and brand-takers are negative examples. This is learning in reverse in terms of time and causation: while normally one would predict an ADE given a drug of interest, the proposed approach "predicts" who *has been* on what version of the drug given the medical events they experience after starting that drug. After training the model, it can be inspected to see what differences between brand and generic have been found. Page et al. (2012) proposed a similar "reverse machine learning" approach for hypothesizing ADEs, but they did not control for confounding.

6.3.2 Reducing Confounding Using Self-Controlled Studies

Since any approach that attempts causal discovery from observational data faces the possibility of confounding, the second challenge posed by generic ADE discovery is reducing such possibilities, especially for temporal confounding. The proposed approach tackles this by setting up a self-controlled study, which takes advantage of the similarities between brand-takers and generic-takers to implicitly match on observed and unobserved variables, thereby reducing the effects of confounders, especially unobserved ones.

First, patients taking brand or generic versions of a drug are taking it for the same reasons: they share indications. Recall Figure 6.1 and consider a study of two unrelated treatments where units do not share predisposition towards treatment. In such a study, confounding by indication I might manifest itself by causing an analysis to propose effects E that are associated with I. For example, if drug D is a beta blocker, then an analysis might propose MI as an effect because MI is more common in patients on beta blockers. But MI is an indication for (cause of) taking beta blockers, not an effect. A study of brand versus generic avoids this problem because patients take either of the drug versions for the same reasons, the same indications I, thus controlling for confounding by indication.

Nevertheless, other variables could be confounders, including variables not observed nor even imagined. In fact, this risk is greater when hypothesizing effects because any variable that appears more associated with brand or generic could be part of the hypothesized effect, and such variables are likely to be confounders that will bias the analysis. This leads to the second reason that the brand versus generic setting has an advantage: many patients switch from brand to generic. Entering these patients into the study makes it self-controlled, which means the same patients are in both treatment groups. This matches the groups exactly on observed and unobserved variables, thereby controlling for even unobserved confounders.

However, patients that switch from brand to generic (or generic to brand) might change over time, meaning that self-control cannot help with temporal confounding. Temporal or time-varying confounding can occur when the relationship between two variables changes over time. For example, when generic gabapentin became available in early 2005, the healthcare system studied here quickly switched patients from brand (Neurontin) to generic.

Around the same time, the healthcare system also switched from paper to electronic prescriptions. As a result, the variable "prescription transmitted electronically" is the best discriminator between brand and generic when the study does not control for changes over time. This variable does not cause generic; it is a temporal confounder. Because things can change between when a patient takes different versions of a drug, temporal confounding is exacerbated in the setting of brand versus generic. The proposed approach tackles this by additionally employing temporally-matched control groups. While all of these similarities reduce confounding, they can never guarantee to eliminate confounding due to the nature of observational studies.

6.3.3 Study Design

To discover differences between brand and generic versions of drugs, observational studies were constructed and then analyzed with machine learning. A typical observational study compares a treated group with a control group, but that does not work in the case of brand and generic drugs because there are two treatments and there exists the possibility of other factors changing over time, leading to temporal confounding. Thus, a study design was used that paired each treatment group with a temporally-matched control group, as in Figure 6.3.

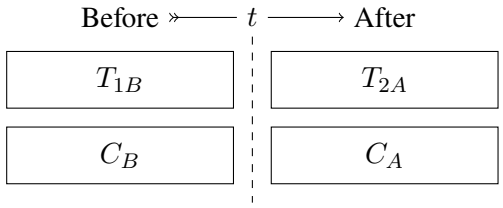


Figure 6.3: Controlled before–after study for two sequential treatments, T_1 and T_2.

The study design in Figure 6.3 is a type of controlled before–after study (Shadish et al., 2002). It contains two treatments, T_1 and T_2, 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 eliminate confounding.

In the case of brand versus generic, many patients switch from brand to generic and are therefore members of both T_{1B} and T_{2A}. Including these switchers in the study makes it *self-controlled*. Because the number of switchers tends to be relatively small, all of the patients that ever took brand or generic were entered in the study to increase its power and robustness. Specifically, there were three matching scenarios: (1) brand to generic switchers were matched with generic to brand switchers as in a crossover design, (2) leftover switchers were matched with never-takers, and (3) brand-only-takers were matched with generic-only-takers and served as each other's controls: the brand-taker served as T_{1B} and C_A and the generic-taker served as T_{2A} and C_B. Accordingly, the time threshold was chosen per matched pair. All matching included the time period of the drug as well as demographics and measures of interaction with the health system.

How can such a study be analyzed? Let T_{1B} be the group taking the brand version of a drug, T_{2A} be the group taking the generic version, and f be some outcome measure of each group. Then the difference between treated groups $f(T_{2A}) - f(T_{1B})$ is the effect, the

difference between control groups $f(C_A) - f(C_B)$ models the changes over time, and the difference in differences is the temporally-adjusted effect, Equation 6.1.

$$(f(T_{2A}) - f(T_{1B})) - (f(C_A) - f(C_B)) \tag{6.1}$$
$$= f(T_{2A}) - f(T_{1B}) - f(C_A) + f(C_B) \tag{6.2}$$
$$= (f(T_{2A}) - f(C_A)) - (f(T_{1B}) - f(C_B)) \tag{6.3}$$

Equation 6.1 is equivalent to first finding the effect of each treatment group compared to the baseline of its control group and then finding the difference between those effects, which is Equation 6.3. (Note that the same relationships hold after replacing differences with ratios.)

The typical analysis approach in fields such as statistics or epidemiology would be to estimate Equation 6.2—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: $+T_{2A}$, $-T_{1B}$, $-C_A$, $+C_B$. That is, the positive examples are T_{2A} and C_B together, and the negative examples are T_{1B} and C_A together, pitting the diagonals of Figure 6.3 against each other. This design controls for temporal and other differences because each classification group includes both before and after units, and both treated and control units. Also, 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.

6.3.4 Analyses

The observational studies were constructed according to the study design in Figure 6.3 and then analyzed with classification, differential prediction, and a method developed here, temporal inverse probability weighting (IPW). The classification method applied binary classifiers directly to the positives and negatives from the study design. Any binary classifier could work in this setup, but this work focused on those that produced interpretable feature weights so that humans could follow up on any potential ADE discoveries. Specifically, logistic regression (LR) was chosen because it is a commonly used model in causal inference, and support vector machines (SVMs) with linear kernels were chosen because they were also used for differential prediction. It turns out that applying binary classification to this study design is already 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 "proper" differential prediction using SVMs (with linear kernels) modified to maximize uplift (Kuusisto et al., 2014). (Uplift is a measure of differences between groups analogous to Equation 6.2.) Differential prediction seeks to predict whether units will respond to treatment by comparing treated and control groups. It works by building a model of response to treatment, building a separate model of response despite no treatment, and then comparing the two models, although some methods model both responses with a single, combined model that explains the differences between the treated and control groups. It was originally conceived as a way to target marketing at individuals who would respond by making a purchase (or not), but it turns out to also be applicable to analyzing controlled studies. The standard setting of predicting the response from the treatment must be adapted to the setting of discovering differences between brand

and generic because there are two treatments and the response is to be discovered. This can be done by noticing that there are still the four groups of Figure 6.3: brand–before, generic–after, control–before, and control–after. Thus, the groups remain the same, but instead of treatment and response, the experimental dimensions are drug and time.

6.3.4.1 Temporal Inverse Probability Weighting

In addition to differential classification and prediction, the studies were analyzed using inverse probability weighting (IPW) (Rosenbaum and Rubin, 1983; Imbens and Rubin, 2015), which this work adapts to the temporal setting of controlled before–after studies 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 or after for each of the brand (before) and generic (after) 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 taking brand or generic (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 brand and generic. Finally, a second classifier is trained on the reweighted brand- and generic-takers to discover the differences between them. LR and SVMs were the classifiers used.

More precisely, given pairs $(x, y)_i$ of feature vectors and labels for control units C and treated units T, temporal IPW is done by (1) modeling the control units $-C_B$ and $+C_A$ with one model $\mathcal{M}_1$, (2) using $\mathcal{M}_1$ to predict the labels of the units in the treated groups $-T_{1B}$ and $+T_{2A}$, (3) reweighting the treated units by the reciprocal of the predicted probability of their true label (Equation 6.4), and (4) modeling the reweighted treated units with a second model $\mathcal{M}_2$. Feature weights then come from inspecting $\mathcal{M}_2$.

$$w(x_i, y_i) := \frac{1}{\hat{P}_{\mathcal{M}_1}(Y = y_i \mid X = x_i)} \tag{6.4}$$

Temporal IPW can be understood as finding a feature that maximizes the ratio of time- and treatment-specific relative risks. The relative risk of a feature (outcome) f in the brand and generic groups compared to controls is, respectively,

$$\frac{P(f \mid T = b, W = \text{bef})}{P(f \mid T = 0, W = \text{bef})} \text{ and } \frac{P(f \mid T = g, W = \text{aft})}{P(f \mid T = 0, W = \text{aft})} \tag{6.5}$$

where T is the treatment, brand b, generic g, or neither 0, and W is when, before or after. The ratio, generic over brand, of these relative risks is

$$\frac{P(f \mid g, \text{aft})/P(f \mid 0, \text{aft})}{P(f \mid b, \text{bef})/P(f \mid 0, \text{bef})} = \frac{P(f \mid g, \text{aft})}{P(f \mid b, \text{bef})} \frac{P(f \mid 0, \text{bef})}{P(f \mid 0, \text{aft})}. \tag{6.6}$$

The second term in the right-hand side of Equation 6.6 reweights according to the reciprocal of changes over time in the control groups, corresponding to the inverse probability in Equation 6.4. Considering the right-hand side of Equation 6.6, temporal IPW works by

first finding a set of features that maximizes the second term and then finding a (possibly different) set of features that maximizes the first term. Thus, in actuality, the maximized quantity is a pseudo-relative risk where the relative outcomes may differ.

Compared to the other methods, temporal IPW has an optimization objective that cannot be gamed and has greater statistical efficiency. The risk with IPW in general is that a few patients get enormous weights. The strength of IPW based on *temporal* propensity scores is that such uneven weighting is unlikely if changes over time in population health and the health system are small and gradual. In theory then, if there exist differences in health events that make it possible to learn a model $\mathcal{M}$ in model class $\mathcal{C}$ (say, the class of linear SVMs) that can distinguish between patients on generic and brand with AUC ROC $A > 0.5$, then temporal IPW, together with an effective learning algorithm for $\mathcal{C}$ and enough data, should make it possible to approximate $\mathcal{M}$ and achieve AUC A. In contrast, differential prediction is unlikely to succeed in this same way because it can drive its objective up by learning a model that has an especially *low* AUC on the *control* patients rather than an especially high AUC on the treated patients. While differential classification does not suffer this malady, it is also unlikely to approximate $\mathcal{M}$ or achieve AUC A because fully half of its training examples (the controls) are not actually labeled according to $\mathcal{M}$, that is, according to the true class of brand or generic. Hence, if $\mathcal{M}$ has an AUC of $A = 0.5 + a$ when distinguishing brand $-T_{1B}$ and generic $+T_{2A}$, $\mathcal{M}$ will have an AUC only half as good, $A' = 0.5 + \frac{a}{2}$, when applied to the differential classification training data, $\{-T_{1B}, -C_A\}$ and $\{+T_{2A}, +C_B\}$. As a result, differential classification may find a model different from $\mathcal{M}$ that spuriously has a higher AUC.

6.3.4.2 Evaluation

To evaluate the analyses on the synthetic data, the experiments used the AUC ROC of identifying the true effects specific to the generic. That is, to do well, a model needed to score the events that are the true effects of the generic higher than all other events. The score of an event was the weight given to the corresponding feature X_j by the model. For logistic regression, this was the regression coefficient (log odds of the feature), and for SVMs (both differential and plain) this was the coefficient of the feature in the linear kernel.

6.3.5 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; otherwise it can be considered a fact. 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. All of the data was analyzed in the form of patient histories: the relevant period of time was extracted from the patient's history, the events during the period were counted, and the counts formed a feature vector along with demographic facts.

Experiments were conducted on both synthetic and real-world EHR data, the former to provide a ground truth for evaluating the methods, and the latter to apply those methods to finding actual differences between brand and generic drugs.

6.3.5.1 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 (ADEs). The temporal differences were the availability of the generic version of drug D_1 and a distractor, the introduction of procedure P_2. These were introduced midway through the samples and are indicated by dashed lines in Figure 6.4. 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 causes 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).

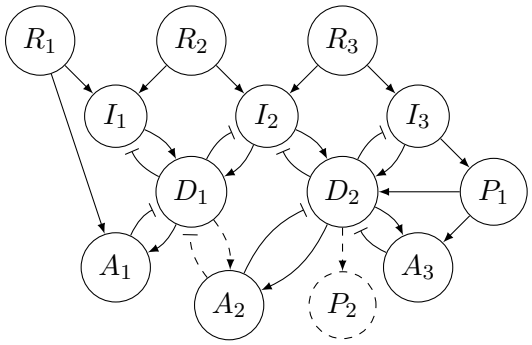


Figure 6.4: Network for CTBN for synthetic data. (R)isk factor, (I)ndication, (D)rug, (P)rocedure, (A)DE. P_2 is introduced at the same time as generic D_1, midway through the sampled patient histories. Generic D_1 causes A_2 whereas brand does not. The dashed lines indicate these temporal differences. Perpendicular arrowheads ⊣ mark inhibitors.

6.3.5.2 Real EHR Data

The real EHR data was deidentified medical records data from Marshfield Clinic. It 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. Patient histories had 872 events on average (quartiles: 0%: 1, 25%: 22, 50%: 140, 75%: 727, 100%: 134k).

6.4 Results

The methods from §6.3 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 been on the market long enough to have demonstrated their safety.

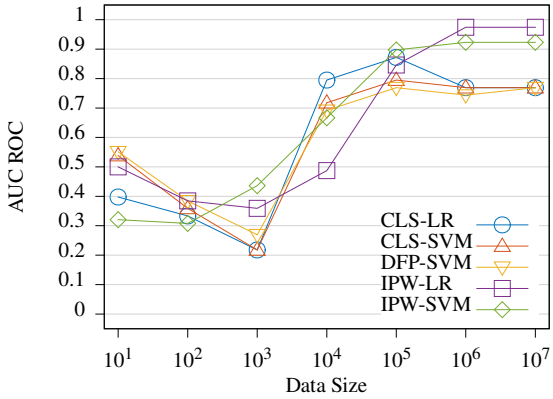
6.4.1 Synthetic Data

Samples were drawn from the CTBN in Figure 6.4 to produce three data sets each with 10M patients. Subsets of each large data set were then formed by taking the first n patients,

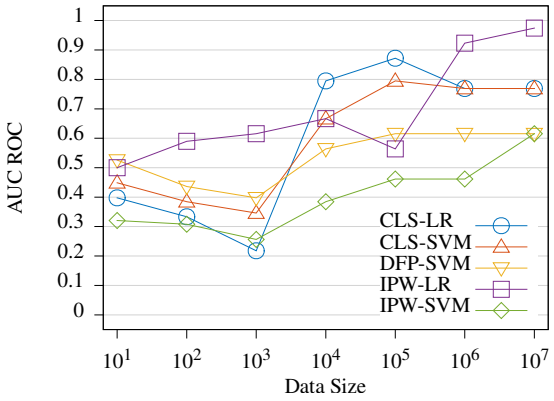
where n ranged from 10^1 to 10^7 (the full size). In a given subset of patients, only some took the brand or generic version of the drug of interest; these patients became the cases in a before–after study. Furthermore, each case and its matched control contributed an example for each of the before and after periods, leading to the classification task having four examples for each case–control pair. The sizes of the data sets in terms of these numbers are in Figure 6.5(a), averaged over the three data sets.

Patients	Cases	Examples
10	2	8
100	20	81
1 000	174	697
10 000	1 772	7 088
100 000	17 698	70 792
1 000 000	177 008	708 033
10 000 000	1 770 737	7 082 948

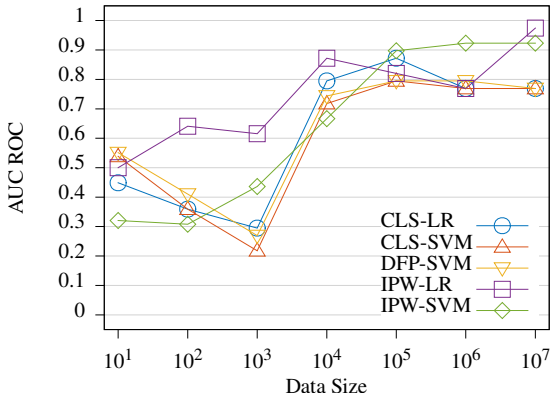
(a) Data sets



(b) Default



(c) Best classification ("Tuned")



(d) Best ADE discovery

Figure 6.5: Data sets and learning curves for discovering the true generic-specific ADEs in the experiments on the synthetic data. CLS: differential classification, DFP: differential prediction, IPW: temporal IPW, LR: logistic regression, SVM: support vector machine with a linear kernel, Cases: patients taking the brand or generic of interest, Examples: feature vectors for the classification task.

Each method was applied to each subset of each data set and then evaluated by how well it identified the correct generic-specific ADE (see §6.3.5.1). Figures 6.5(b)–6.5(d) show the results of this evaluation in the form of learning curves after averaging the results over the three data sets. 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), the standard notions of tuning

do not directly apply, which meant there was no standard way of picking hyperparameters.[2] Thus, learning curves from three different ways of picking hyperparameters are shown. In Figure 6.5(b), the methods with the default $C = 1$ were picked; in Figure 6.5(c), the methods with the best AUC ROC on the classification task were picked; and in Figure 6.5(d), the methods with the best AUC ROC on the ADE discovery task were picked, as if an oracle had provided the correct hyperparameters. The AUCs ROC were averaged over all the data sets and data sizes before picking the best. One can see that "tuning" the hyperparameters by picking the best on the classification task is not a good tactic; the classification accuracy is not very predictive of the causal discovery accuracy. Indeed, it appears that SVMs are the most sensitive to hyperparameters, with IPW-SVM having an inverse relationship between accuracy and ADE discovery. On the other hand, LR appears to be the most robust to hyperparameters, as it does consistently well for both differential classification and IPW. Overall, the IPW methods tend to do better than the differential classification and prediction methods, which perform similarly to each other.

6.4.2 Discussion of Synthetic Data Results

These results demonstrate the success of temporal IPW at causal discovery, which can be explained in part by temporal IPW's statistical efficiency and rigorous formulation. Unlike differential classification, temporal IPW does not dilute the ADE signal by combining treated and control groups, and so it is more sensitive to differences between the before (brand) and after (generic) groups. Unlike differential prediction, temporal IPW cannot game its objective by inadequately modeling the controls, and so it is more likely to discover actual differences between brand and generic.

These results also illustrate a remaining challenge: how to tune hyperparameters for causal discovery. Because causal discovery is not a supervised task, the standard techniques do not directly apply. However, they indirectly apply, and can be adapted by collecting known true / false causal relationships and treating them as a binary classification task. A set of such labeled cause–effect pairs can be divided into sets for training, tuning, and testing in the standard ways. One problem with this is that the cause–effect pairs are unlikely to be independent, even given the data. The more important problem is that the number of known causal effects in a given domain is usually fairly small, so dividing them up results in unreliable tuning estimates. This paucity of known causal effects is exacerbated when studying brand and generic versions of drugs because the generic should have few, if any, causal differences, much less generic-specific ADEs. Thus, causal discovery, especially for generic ADEs, would benefit from new ideas for tuning, perhaps related to how to transfer hyperparameters between tasks.

The strength of experiments on synthetic data is that the true causal relationships are known and can be used to directly evaluate the accuracy of causal discoveries. Of course, the weakness is that synthetic data may be unrepresentative of the real-world task in various ways. The next experiments study actual brand and generic drugs in real EHR data, which brings the challenge that not all differences and effects are known.

[2] The hyperparameters were just the regularization strength parameter for SVMs (C) and LR (λ).

Table 6.1: Top 10 / 50k features from IPW-LR by LR coefficient magnitude. Features that favor generic have positive coefficients.

(a) Bupropion

Score	Feature
-0.029	Ex-smoker
-0.014	Albuterol 0.09 mg inhaler
0.013	Chiropractic manipulative treatment
-0.012	Polyethylene glycol powder for oral solution
-0.012	Furosemide 20 mg tablet
-0.012	Ferrous sulfate 325 mg tablet
-0.011	Vitamin B12
-0.011	Omeprazole 20 mg capsule
-0.011	Acetaminophen 250 mg / aspirin 250 mg / caffeine 65 mg tablet
-0.011	Mometasone furoate 0.05 mg nasal spray

(b) Duloxetine

Score	Feature
-0.020	Ex-smoker
-0.016	Albuterol 0.83 mg/ml inhalant solution
-0.015	Cholecalciferol 2000 unt oral capsule
-0.014	Glucose in capillary blood
-0.014	Glucose in blood by test strip
-0.014	Glucose finger stick
-0.013	Fluticasone propionate 0.1 mg / salmeterol 0.05 mg dry powder inhaler
-0.013	Albuterol 0.09 mg inhaler
-0.013	Outpatient visit
0.012	Non-smoker

(c) Gabapentin

Score	Feature
0.269	Other measurements / exams
0.184	Glomerular filtration rate
0.180	eGFR with normals for non-black
0.113	Lidocaine 0.05 mg patch
0.113	Cyclobenzaprine hydrochloride 10 mg tablet
0.113	Naproxen 500 mg tablet
0.112	Citalopram 40 mg tablet
0.109	Obstructive sleep apnea syndrome
0.108	Essential hypertension
0.107	Albuterol 0.09 mg inhaler

(d) Methylphenidate

Score	Feature
-0.004	Year of birth
-0.002	Smoker
-0.001	BP systolic
-0.001	BP diastolic
-0.001	No matching concept
-0.001	Albuterol 0.09 mg inhaler
-0.000	Sertraline 100 mg tablet
-0.000	Insulin lispro pen injector
-0.000	Insulin glargine pen injector
-0.000	Budesonide 0.16 mg / formoterol fumarate 0.0045 mg inhaler

6.4.3 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, a NDRI antidepressant that also helps with smoking cessation, (2) duloxetine, a 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). For each of these drugs, a controlled before–after study was constructed by extracting data from a deidentified EHR database from Marshfield Clinic. As in the experiments on the synthetic data, the features were counts of event occurrences plus demographics, for a total of 49045 features in the EHR data. In order to focus the results for human interpretation, the best methods from the synthetic experiments were used. These were IPW-LR and

Table 6.2: Top 10 / 50k features from IPW-SVM by SVM coefficient magnitude. Features that favor generic have positive coefficients.

(a) Bupropion

Score	Feature
-1.791	Outpatient visit
-1.584	Tetrahydrocannabinol in urine
1.564	Adverse reaction to food
-1.513	Tobramycin 0.003 mg ophthalmic ointment
-1.507	Dexamethasone 6 mg tablet
-1.487	Injection of sacroiliac joint
-1.480	Quetiapine fumarate
1.469	Hysteroscopy, surgical
-1.446	H1N1 immunization
-1.384	Abnormal sexual function

(b) Duloxetine

Score	Feature
-2.231	Outpatient visit
-1.408	Pharmacologic management with minimal psychotherapy
1.271	Phentermine
-1.141	Individual psychotherapy
1.138	Drug screen
-1.095	Outpatient visit, established patient
-1.064	Linaclotide 0.29 mg capsule
1.044	Topiramate 50 mg capsule
1.006	Alopecia
-0.973	Mirtazapine

(c) Gabapentin

Score	Feature
1.898	Injections of muscle trigger points
-1.741	Outpatient visit
1.624	Fluoroscopic guidance for spinal injection procedures
-1.539	Gynecological exam
-1.533	Lipid metabolism disorder
-1.315	Speech / Language deficit from cerebrovascular accident
-1.314	Spinal anesthetic injection
1.306	Naproxen sodium 220 mg tablet
1.254	Other measurements / exams
-1.253	Reiter's disease

(d) Methylphenidate

Score	Feature
1.828	Epinephrine 0.5 mg/ml injector pen
-1.764	Outpatient visit
-1.679	Ciprofloxacin 3 mg / dexamethasone 1 mg otic suspension
-1.579	Asthma exacerbation
1.478	One-on-one cognitive skills training
1.444	Vertebral column disorder
1.443	Immunization for a minor
-1.438	Male
-1.428	Unknown race
-1.413	Hydroxyzine hydrochloride 10 mg tablet

IPW-SVM, using the the best hyperparameters from the synthetic experiments.

The top differences between brand and generic discovered by IPW-LR and IPW-SVM are shown in Tables 6.1 (p. 73) and 6.2 (p. 74), respectively. What patterns can be seen in these groups of features? For brand duloxetine (Table 6.1(b)), IPW-LR turns up several features having to do with diabetes. Diabetes also appears to be associated with brand methylphenidate (Table 6.1(d)), which is also associated with a higher year of birth (younger patients) and asthma. Asthma and related treatments are a pattern in these results, appearing 8 times with brand and once with generic when considering both IPW-LR and IPW-SVM together. Perhaps relatedly, steroids and antihistamines occur two times with brand and once with generic. Another pattern seen in both IPW-LR and IPW-SVM is the association of generic gabapentin with pain management events (Tables 6.1(c) and 6.2(c)), namely higher numbers of prescriptions for NSAID naproxen, lidocaine anesthetic patches, and muscle relaxant cyclobenzaprine. Another pattern that IPW-SVM finds across all drugs

Table 6.3: Top 10 / 50k features from CLS-LR by LR coefficient magnitude. Features that favor generic have positive coefficients.

(a) Bupropion

Score	Feature
11.044	Female
10.980	Male
9.050	White
6.524	Mixed racial group
6.450	Unknown race
5.891	Nepafenac 3 mg/ml ophthalmic suspension
5.783	Heat syncope
-5.261	Ammonia in plasma
-5.109	H1N1 immunization
4.982	Changes in skin texture

(b) Duloxetine

Score	Feature
4.609	Male
4.586	Female
-4.353	General exam
-4.270	Bipolar disorder
-4.203	MMRV vaccine
4.087	Methylsulfonylmethane 1000 mg capsule
4.059	Anodontia
-3.972	Psychotherapy service
-3.828	Periodic comprehensive exam
-3.792	Bulimia nervosa

(c) Gabapentin

Score	Feature
6.096	Mixed racial group
5.998	Female
5.873	Male
5.775	Unknown race
4.045	Nasal hemorrhage control
-3.889	Radiologic exam of wrist
-3.631	Spinal anesthetic injection
-3.496	Periodic comprehensive exam
3.360	Problem-focused oral exam
3.357	Hypertensive heart and chronic kidney disease

(d) Methylphenidate

Score	Feature
8.257	Female
7.926	Male
6.633	Throat pain
-6.499	Loratadine 10 mg tablet
5.762	White
5.712	Epinephrine 0.5 mg/ml injector pen
5.684	Vertebral column disorder
5.463	General well-being finding
5.403	Clinical finding procedure
-5.390	Benign neoplasm of choroid

(Table 6.2) is the association of brand with more visits, even psychotherapy visits, which can be expensive. None of the patterns in the features suggest obvious temporal confounders.

For reference and comparison, the results from the other methods are also included, using the best hyperparameters from the synthetic experiments as above. The differential classification methods, CLS-LR (Table 6.3, p. 75) and CLS-SVM (Table 6.4, p. 76), have a number of top features in common, and the top features from CLS-SVM and DFP-SVM (Table 6.5, p. 77) for each drug are quite similar. Any further discussion of patterns in the results from these methods is foregone here, however, because their lower causal discovery accuracy makes meaningful interpretation questionable.

6.4.4 Discussion of EHR Results

IPW-LR and IPW-SVM rediscovered some known relationships, and suggested some new, plausiably causal, relationships. The associations of diabetes with brand duloxetine and methylphenidate in the IPW-LR results can be readily explained: duloxetine treats diabetic

Table 6.4: Top 10 / 50k features from CLS-SVM by SVM coefficient magnitude. Features that favor generic have positive coefficients.

(a) Bupropion

Score	Feature
-2.770	Year of birth
1.918	Rhythm ECG
-1.872	Tetrahydrocannabinol in urine
-1.648	Methscopolamine bromide
-1.626	Morphine sulfate 30 mg capsule
-1.611	Health and behavior assessment
-1.564	Thyroid imaging
1.557	Scrotal varices
1.556	Radiologic exam of clavicle
-1.540	Toe contusion

(b) Duloxetine

Score	Feature
1.555	Blood test
1.534	Refraction disorder
1.497	Male
1.484	Methylsulfonylmethane 1000 mg capsule
-1.478	Eszopiclone 3 mg tablet
-1.473	Hemorrhoids
1.459	Female
-1.433	Bipolar disorder
1.347	Clorazepate 3.75 mg tablet
1.340	Positive pregnancy test

(c) Gabapentin

Score	Feature
-3.333	Year of birth
1.998	Benazepril 10 mg tablet
1.938	Hypertensive heart and chronic kidney disease
-1.770	Dental caries
1.760	Nasal hemorrhage control
1.740	Pityriasis versicolor
-1.685	Palindromic rheumatism
-1.655	Ketoprofen 200 mg capsule
-1.632	Dobutamine
1.626	Injections of muscle trigger points

(d) Methylphenidate

Score	Feature
2.104	Female
2.052	Desoximetasone 0.5 mg/ml cream
1.973	Male
-1.801	Crushing injury of hand
1.781	Epinephrine 0.5 mg/ml injector pen
1.669	Vertebral column disorder
1.632	Altered mental status
1.587	Senile hyperkeratosis
1.568	Throat pain
1.544	Hearing problem

neuropathy (Smith and Nicholson, 2007) and methylphenidate treats ADHD which is associated with higher risk of diabetes mellitus type 2 (Chen et al., 2018). However, it is not clear why these associations appear with only the brand drugs. Perhaps health plans that provide better diabetic care also pay for brand drugs, or perhaps patients who are more involved with their care prefer brand drugs. The association of brand methylphenidate with younger patients is curious. Is there a mistrust by parents or physicians of prescribing generic methylphenidate to younger patients, perhaps as a reaction to the news that some generics had reduced efficacy?

The association of generic gabapentin with increased pain management events is a consistent signal across both IPW-LR and IPW-SVM. While it is possible that generic-takers interact with the health system more to manage their pain, this association suggests a possible lack of efficacy of generic gabapentin. (There is evidence that taking brand improves outcomes through easier adherence (Sicras-Mainar et al., 2015; Candido et al., 2016).) Another consistent signal is the association of brand drugs with asthma. This is probably

Table 6.5: Top 10 / 50k features from DFP-SVM by SVM coefficient magnitude. Features that favor generic have positive coefficients.

(a) Bupropion

Score	Feature
-3.264	Year of birth
1.956	Rhythm ECG
-1.788	Human insulin
-1.768	Methscopolamine bromide
-1.743	Morphine sulfate 30 mg capsule
-1.737	Abnormal sexual function
-1.733	Tetrahydrocannabinol in urine
1.693	Nepafenac 3 mg/ml ophthalmic suspension
-1.636	Toe contusion
-1.629	Thyroid imaging

(b) Duloxetine

Score	Feature
-10.423	Year of birth
-10.246	Age
2.244	Methylsulfonylmethane 1000 mg capsule
-1.780	Individual psychotherapy service
1.779	Blood test
-1.776	Eszopiclone 3 mg tablet
-1.707	General exam
1.616	Refraction disorder
1.602	Methylphenidate 10 mg capsule
-1.585	Hemorrhoids

(c) Gabapentin

Score	Feature
-4.173	Year of birth
2.359	Benazepril 10 mg tablet
2.112	Nasal hemorrhage control
2.017	Mixed racial group
-1.952	Dental caries
-1.949	Palindromic rheumatism
-1.851	Spinal anesthetic injection
1.836	Female
-1.801	Retroperitoneal ultrasound
-1.798	Certification procedure

(d) Methylphenidate

Score	Feature
-6.225	Year of birth
-3.054	Age
2.613	Hearing problem
2.522	Desoximetasone 0.5 mg/ml cream
2.477	Clarithromycin 250 mg tablet
-2.412	Crushing injury of hand
-2.382	Albuterol 1 mg inhalation solution
2.367	Interpersonal relationship finding
2.341	Epinephrine 0.5 mg/ml injector pen
2.312	Vertebral column disorder

due to people with chronic conditions such as asthma preferring brand drugs, but treatments for asthma that reduce serotonin may worsen ADHD and other psychological conditions (Pretorius, 2004), leading to treatment with bupropion, duloxetine, or methylphenidate.

The pattern of brand drugs and more visits is also worth examining, as it relates to access to care and the quality thereof. The association can be explained by differences in health plans where a plan that pays for brand also pays for more services in general. Such a plan usually costs more, so this explanation reinforces the positive correlation between brand drugs, better healthcare, and higher socioeconomic status, which is just another way of saying that patients who take generics are at a disadvantage. Unfortunately, the plausibility of this inference cannot be checked because socioeconomic features are not included in the data, and thus our methods cannot control for them. One remedy would be to include more socioeconomic features as routine demographics in EHR data.

6.5 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.

Chapter 7

Conclusion

While accurate causal discovery will remain a challenging task due to the intrinsic limitations of observation as a source of information, the various methods for causal discovery described in this work contribute to the progress being made in taming the uncertainties of causal inference from observational data. Of these methods, TMNs focus on learning the structure of dynamic causal models of patient event sequences, thereby representing the structural causal modeling paradigm. Representing the observational studies paradigm, relational rule learning and temporal IPW focus on causal discovery ML for hypothesizing effects in the context of self-controlled and before–after studies. Filling in between the paradigms, the MNATS method models patient histories with temporal Markov networks, evaluating potential ADEs with model-adjusted scores of temporal dependence. Collectively, these methods offer ways to address a number of the challenges of causal discovery.

One challenge, often regarded as the most significant in causal inference, is confounding, which threatens to undermine any analysis of observational data. Partial solutions come from the two schools of thought regarding causal inference: properly-conducted observational studies and accurate structural causal models. Self-controlled studies reduce confounding from both observed and unobserved variables in the relational rule learning analyses. Temporal IPW employs a more sophisticated study design, a before–after study that is self-controlled and involves crossover and matching. This not only controls for observed and unobserved confounders, it also minimizes time-varying confounding. Bias is reduced by reweighting to adjust for propensity for treatment. Nested within a self-controlled study, the temporal score improves estimates of causal effects by adjusting for the influences of other variables. A more global approach to adjustment is taken by MNATS, which adjusts other scores of temporal dependence by estimating them with probabilities from a model of the distribution of patient histories. While an explicit study design is not involved, analyzing patient event sequences as they evolve over time achieves similar control of confounding through self comparison. These benefits of patient-specific modeling are also available to TMNs, but, more importantly, TMNs build on the formal foundations of structural causal models in order to learn causally-accurate structures.

Effective postmarketing pharmacosurveillance for ADEs requires large volumes of data involving many variables and therefore requires any applicable methods to be scalable. Reformulating structure learning in terms of parameter learning avoids the combinatorial nature of most algorithms for learning the structure of causal models, thereby making them

scalable to larger numbers of variables. That the sufficient statistics needed for learning can be computed in one pass over the data solves the data scalability problem. Similarly, temporal dependence scores are efficient to evaluate, even on large data sets. When setting up observational studies, as for causal discovery ML or temporal IPW, subsets of the data are selected corresponding to the exposures and outcomes of interest. Even with millions of patients, the selected data sets remain small enough to analyze quickly with most off-the-shelf ML classifiers, and those that are slower usually offer advantages worth waiting for, such as inventiveness in the case of ILP, or accuracy in the case of SVMs.

Other challenges stem from the relational format and messiness of EHR data, which is the kind of data typically used for pharmacosurveillance. The relational format is addressed with relational rule learning and by transforming the data into event sequences which can be analyzed with time series methods or observational studies. TMNs excel in this scenario through their feature functions, which not only model time but demonstrate an ability to handle the irregularity, missingness, and noise of EHR event sequences. Observational studies have similar abilities to handle the vagaries of such sequences through their flexibility in how events are defined, counted, and aggregated over the study periods. However, leaving researchers so much leeway to determine how to prepare data for modeling can lead to unprincipled analyses compared to a comprehensive modeling approach.

A final problem for genuine causal discovery is hypothesizing appropriate effects. This is where ML methods have a potentially large advantage over epidemiology methods, which are not designed for such exploratory analysis. Causal discovery ML develops a solution to this problem by inverting the temporal order of the problem: a model that accurately classifies known labels from the past, given data from the future, has probably discovered some consequence of the labels. Such effects, whether described by relational rules or by important features in a model, are not limited by human imagination or bias, and point towards causal relationships that could be medically relevant, such as ADEs.

Thesis The various causal discovery methods developed in this body of work have been tested on ADE detection tasks in both synthetic and real-world EHR data. Together with the evidence presented above, the results of these experiments demonstrate that:

> Methods that are causal, scalable, and applicable to irregular, sparse, and noisy event sequences discover causal effects more accurately than methods that ignore causality or cannot handle the scale and messiness of EHR data. Furthermore, methods that can hypothesize effects improve genuine causal discovery by avoiding the limitations of human bias.

The methods in this dissertation address ADE discovery in EHR data, contributing new techniques to the broader body of work on computational pharmacosurveillance. In contrast to the broader body, these methods focus on the situations in which the possible effects are unknown, and distinguish themselves by bridging ML and epidemiology: they bring aspects of causal inference and observational studies to ML, and apply learning techniques and formal causal models to tasks in epidemiology. The wider perspective achieved by combining multiple approaches to causality illuminates the dark corners and blindspots of the individual perspectives, and creates a path for advancing causal discovery. This path has already led to the new methods for discovering ADEs herein, and promises to lead to additional techniques for augmenting our ability to understand why things happen.

References

Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D. Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(Jan), 2010.

Andrew Arnold, Yan Liu, and Naoki Abe. Temporal causal modeling with graphical Granger methods. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13*, 2007.

Gustavo Arroyo-Figueroa and Luis Enrique Sucar. A temporal Bayesian network for diagnosis and prediction. In *Uncertainty in Artificial Intelligence 15*, 1999.

Jacques Baillargeon, Holly M. Holmes, Yu-Li Lin, Mukaila A. Raji, Gulshan Sharma, and Yong-Fang Kuo. Concurrent use of warfarin and antibiotics and the risk of bleeding in older adults. *The American Journal of Medicine*, 125(2), 2012. doi: 10.1016/j.amjmed.2011.08.014.

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

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. arXiv:1901.10912, 2019.

Laura E. Brown, Ioannis Tsamardinos, and Constantin F. Aliferis. A comparison of novel and state-of-the-art polynomial Bayesian network learning algorithms. In *AAAI Conference on Artificial Intelligence 20*, 2005.

Kenneth D. Candido, Joseph Chiweshe, Utchariya Anantamongkol, and Nebojsa Nick Knezevic. Can chronic pain patients be adequately treated using generic pain medications to the exclusion of brand-name ones? *American Journal of Therapeutics*, 23(2), 2016. doi: 10.1097/MJT.0000000000000098.

Ola Caster, G. Niklas Norén, David Madigan, and Andrew Bate. Large-scale regression-based pattern discovery: The example of screening the WHO global drug safety database. *Statistical Data Analysis and Data Mining*, 3(4), 2010. doi: 10.1002/sam.10078.

Mu-Hong Chen, Tai-Long Pan, Ju-Wei Hsu, Kai-Lin Huang, Tung-Ping Su, Cheng-Ta Li, Wei-Chen Lin, Shih-Jen Tsai, Wen-Han Chang, Tzeng-Ji Chen, and Ya-Mei Bai. Risk

of type 2 diabetes in adolescents and young adults with attention-deficit/hyperactivity disorder: A nationwide longitudinal study. *Journal of Clinical Psychiatry*, 79(3), 2018. doi: 10.4088/JCP.17m11607.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov), 2002.

William G. Cochran. *Planning and Analysis of Observational Studies*. John Wiley & Sons, 1983.

Gregory F. Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4), 1992. doi: 10.1023/A:1022649401552.

James Cussens. Bayesian network learning with cutting planes. In *Uncertainty in Artificial Intelligence 27*, 2011.

Jesse Davis, Irene Ong, Jan Struyf, Elizabeth Burnside, David Page, and Vítor Santos Costa. Change of representation for statistical relational learning. In *International Joint Conference on Artificial Intelligence 20*, 2007.

Cassio P. de Campos, Zhi Zeng, and Qiang Ji. Structure learning of Bayesian networks using constraints. In *International Conference on Machine Learning 26*, 2009.

Luc De Raedt. *Logical and Relational Learning*. Springer, 2008.

Thomas Dean and Keiji Kanazawa. A model for reasoning about persistence and causation. *Computational Intelligence*, 5(3), 1989.

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.

Norbert Dojer. Learning Bayesian networks does not have to be NP-hard. In *Mathematical Foundations of Computer Science 2006*, 2006. doi: 10.1007/11821069_27.

Nir Friedman, Iftach Nachman, and Dana Pe'er. Learning Bayesian network structure from massive datasets: The "sparse candidate" algorithm. In *Uncertainty in Artificial Intelligence 15*, 1999.

Severino F. Galán and Francisco J. Díez. Networks of probabilistic events in discrete time. *International Journal of Approximate Reasoning*, 30(3), 2002. doi: 10.1016/S0888-613X(02)00071-3.

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

Lise Getoor and Ben Taskar, editors. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

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

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3), 1969.

Asela Gunawardana, Christopher Meek, and Puyang Xu. A model for temporal dependencies in event streams. In *Advances in Neural Information Processing Systems 24*, 2011.

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.

Isabelle Guyon, Constantin Aliferis, and André Elisseeff. Causal feature selection. In Huan Liu and Hiroshi Motoda, editors, *Computational Methods of Feature Selection*. Chapman & Hall / CRC Press, 2007. doi: 10.1201/9781584888796.

David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3), 1995. doi: 10.1007/BF00994016.

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.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 1986.

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.

David Hume. *A Treatise of Human Nature*. John Noon, 1740.

Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. Causal discovery from subsampled time series data by constraint optimization. In *Probabilistic Graphical Models 8*, 2016.

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.

Tommi Jaakkola, David Sontag, Amir Globerson, and Marina Meila. Learning Bayesian network structure using LP relaxations. In *Artificial Intelligence and Statistics 13*, 2010.

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

Antonis Kakas and Peter Flach. Abduction and induction in artificial intelligence. *Journal of Applied Logic*, 7(3), 2009. doi: 10.1016/j.jal.2008.11.001.

Willi Klösgen. Types and forms of knowledge (patterns): Subgroup patterns. In Willi Klösgen and Jan M. Zytkow, editors, *Handbook of Data Mining and Knowledge Discovery*, chapter 5.2. Oxford University Press, 2002.

Mikko Koivisto and Kismat Sood. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5(May), 2004.

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

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.

Steffen L. Lauritzen. *Graphical Models*. Clarendon Press, 1996.

Nada Lavrač and Sašo Džeroski. *Inductive Logic Programming: Techniques and Applications*. Prentice Hall, 1994.

Jason Lazarou, Bruce H. Pomeranz, and Paul N. Corey. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *Journal of the American Medical Association*, 279(15), 1998. doi: 10.1001/jama.279.15.1200.

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.

Su-In Lee, Varun Ganapathi, and Daphne Koller. Efficient structure learning of Markov networks using l_1-regularization. In *Advances in Neural Information Processing Systems 19*, 2006.

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.

Jie Liu and David Page. Structure learning of undirected graphical models with contrastive divergence. In *ICML Workshop on Structured Learning: Inferring Graphs from Structured and Unstructured Inputs*, 2013.

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

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. In *Advances in Neural Information Processing Systems 12*, 1999.

Christopher Meek. *Graphical Models: Selecting Causal and Statistical Models*. PhD thesis, Carnegie Mellon University, 1997.

Tom M. Mitchell. Learning sets of rules. In *Machine Learning*, chapter 10. WCB / McGraw-Hill, 1997.

Joris M. Mooij, Dominik Janzing, Tom Heskes, and Bernhard Schölkopf. On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems 24*, 2011.

Stephen Muggleton. Predicate invention and utilization. *Journal of Experimental & Theoretical Artificial Intelligence*, 6(1), 1994. doi: 10.1080/09528139408953784.

Stephen Muggleton, Aline Paes, Vítor Santos Costa, and Gerson Zaverucha. Chess revision: Acquiring the rules of chess variants through FOL theory revision from examples. In *International Conference on Inductive Logic Programming 19*, 2009. doi: 10.1007/978-3-642-13840-9_12.

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.

Isaac Newton. *Philosopæ Naturalis Principia Mathematica*. Reg. Soc. Præses, 1687.

Teppo Niinimäki and Pekka Parviainen. Local structure discovery in Bayesian networks. In *Uncertainty in Artificial Intelligence 28*, 2012.

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

G. Niklas Norén, Johan Hopstadius, Andrew Bate, Kristina Star, and I. Ralph Edwards. Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery*, 20(3), 2010. doi: 10.1007/s10618-009-0152-3.

David Page and Ashwin Srinivasan. ILP: A short look back and a longer look forward. *Journal of Machine Learning Research*, 4(Aug), 2003.

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.

Jonas Peters, Dominik Janzing, Arthur Gretton, and Bernhard Schölkopf. Detecting the direction of causal time series. In *International Conference on Machine Learning 26*, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Sergey Plis, David Danks, Cynthia Freeman, and Vince Calhoun. Rate-agnostic (causal) structure learning. In *Advances in Neural Information Processing Systems 28*, 2015.

Etheresia Pretorius. Asthma medication may influence the psychological functioning of children. *Medical Hypotheses*, 63(3), 2004. doi: 10.1016/j.mehy.2003.12.049.

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.

Wayne A. Ray. Evaluating medication effects outside of clinical trials: New-user designs. *American Journal of Epidemiology*, 158(9), 2003. doi: 10.1093/aje/kwg231.

Bradley L. Richards and Raymond J. Mooney. Automated refinement of first-order Horn-clause domain theories. *Machine Learning*, 19(2), 1995. doi: 10.1007/BF01007461.

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.

James M. Robins, Miguel Ángel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 2000.

R. W. Robinson. Counting labeled acyclic digraphs. In Frank Harary, editor, *New Directions in the Theory of Graphs*. Academic Press, 1973.

R. W. Robinson. Counting unlabeled acyclic digraphs. In Charles H. C. Little, editor, *Combinatorial Mathematics V*, volume 622 of *Lecture Notes in Mathematics*. Springer, 1977. doi: 10.1007/BFb0069178.

Paul R. Rosenbaum. *Observational Studies*. Springer, 2nd edition, 2002.

Paul R. Rosenbaum. Matching in observational studies. In Andrew Gelman and Xiao-Li Meng, editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. John Wiley & Sons, 2004.

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.

Lynn Rosenberg, Patricia F. Coogan, and Julie R. Palmer. Case–control surveillance. In Brian L. Strom, Stephen E. Kimmel, and Sean Hennessy, editors, *Pharmacoepidemiology*. John Wiley & Sons, 5th edition, 2012. doi: 10.1002/9781119959946.ch19.

Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. *Modern Epidemiology*. Lippincott Williams & Wilkins, 3rd edition, 2008.

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

Donald B. Rubin. Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, 75(371), 1980. doi: 10.1080/01621459.1980.10477517.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2nd edition, 2003.

Patrick Ryan, Emily Welebob, Abraham G. Hartzema, Paul Stang, and J. Marc Overhage. Surveying US observational data sources and characteristics for drug safety needs. *Pharmaceutical Medicine*, 24(4), 2010. doi: 10.1007/BF03256821.

Patrick B. Ryan, Gregory E. Powell, Ed N. Pattishall, and Kathleen J. Beach. Performance of screening multiple observational databases for active drug safety surveillance. In *International Conference on Pharmacoepidemiology & Therapeutic Risk Management 25*, 2009. doi: 10.1002/pds.1806. Abstract.

Patrick B. Ryan, David Madigan, Paul E. Stang, J. Marc Overhage, Judith A. Racoosin, and Abraham G. Hartzema. Empirical assessment of methods for risk identification in healthcare data: Results from the experiments of the Observational Medical Outcomes Partnership. *Statistics in Medicine*, 31(30), 2012. doi: 10.1002/sim.5620.

Patrick B. Ryan, Martijn J. Schuemie, and David Madigan. Empirical performance of a self-controlled cohort method: Lessons for developing a risk identification and analysis system. *Drug Safety*, 36(S1), 2013. doi: 10.1007/s40264-013-0101-3.

Suchi Saria, Daphne Koller, and Anna Penn. Discovering shared and individual latent structure in multiple time series. arXiv:1008.2028, 2010.

Taisuke Sato and Yoshitaka Kameya. Statistical abduction with tabulation. In Antonis C. Kakas and Fariba Sadri, editors, *Computational Logic: Logic Programming and Beyond*. Springer, 2002. doi: 10.1007/3-540-45632-5_22.

Sebastian Schneeweiss, Til Stürmer, and Malcolm Maclure. Case–crossover and case–time–control designs as alternatives in pharmacoepidemiologic research. In *European Society of Pharmacovigilance Annual Meeting 4*, 1997.

Sebastian Schneeweiss, Jeremy A. Rassen, Robert J. Glynn, Jerry Avorn, Helen Mogun, and M. Alan Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4), 2009. doi: 10.1097/EDE.0b013e3181a663cc.

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

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7 (Oct), 2006.

Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high dimensional directed acyclic graphs. *Biometrika*, 97(3), 2010. doi: 10.1093/biomet/asq038.

Antoni Sicras-Mainar, Javier Rejas-Gutiérrez, and Ruth Navarro-Artieda. Comparative effectiveness and costs of generic and brand-name gabapentin and venlafaxine in patients with neuropathic pain or generalized anxiety disorder in Spain. *ClinicoEconomics and Outcomes Research*, 7, 2015. doi: 10.2147/CEOR.S85756.

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.

Timothy Smith and Robert A. Nicholson. Review of duloxetine in the management of diabetic peripheral neuropathic pain. *Vascular Health and Risk Management*, 3(6), 2007.

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

Ashwin Srinivasan. The Aleph manual. http://www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph.html, 2007. Accessed November 01, 2019.

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.

Patrick Suppes. *A Probabilistic Theory of Causality*. North Holland Publishing Company, 1970.

Marc Teyssier and Daphne Koller. Ordering-based search: A simple and effective algorithm for learning Bayesian networks. In *Uncertainty in Artificial Intelligence 21*, 2005.

Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander Statnikov. Algorithms for large scale Markov blanket discovery. In *International Florida Artificial Intelligence Research Society 16*, 2003.

U.S. Food and Drug Administration. The Sentinel Initiative: National strategy for monitoring medical product safety. https://www.fda.gov/media/75240/download, 2008. Accessed October 28, 2019.

U.S. Food and Drug Administration. Preventable adverse drug reactions: A focus on drug interactions. http://www.fda.gov/drugs/developmentapprovalprocess/developmentresources/druginteractionslabeling/ucm110632.htm, 2009. Accessed September 12, 2016.

U.S. Food and Drug Administration. 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 August 23, 2019.

U.S. Food and Drug Administration. What is the approval process for generic drugs? https://www.fda.gov/drugs/generic-drugs/what-approval-process-generic-drugs, 2017. Accessed October 28, 2019.

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.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

Heather J. Whitaker, C. Paddy Farrington, Bart Spiessens, and Patrick Musonda. Tutorial in biostatistics: The self-controlled case series method. *Statistics in Medicine*, 25(10), 2006. doi: 10.1002/sim.2302.

Bartek Wilczyński and Norbert Dojer. BNFinder: Exact and efficient method for learning Bayesian networks. *Bioinformatics*, 25(2), 2009. doi: 10.1093/bioinformatics/btn505.

Andrew M. Wilson, Lehana Thabane, and Anne Holbrook. Application of data mining techniques in pharmacovigilance. *British Journal of Clinical Pharmacology*, 57(2), 2003. doi: 10.1046/j.1365-2125.2003.01968.x.

Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In *Principles of Data Mining and Knowledge Discovery 1*, 1997. doi: 10.1007/3-540-63223-9_108.

Filip Železný and Nada Lavrač. Propositionalization-based relational subgroup discovery with RSD. *Machine Learning*, 62(1–2), 2006. doi: 10.1007/s10994-006-5834-0.

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.