2018

Using the Literature to Identify Confounders

Scott Malec
University of Texas Health Science Center at Houston, Scott.Malec@uth.tmc.edu

Using the Literature to Identify Confounders

A
Dissertation

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

By

Scott Alexander Malec, MLIS, MSIT

University of Texas Health Science Center at Houston

2018

Dissertation Committee:

Advisor
Trevor Cohen, MBChB, PhD[1], Advisor

Committee
Elmer Bernstam, MD, MSE[1], FACP, FACMI
Peng Wei, PhD[2]

[1]The School of Biomedical Informatics, University of Texas Health Science Center at Houston
[2]MD Anderson, University of Texas Health Science Center at Houston

Dedication

This dissertation is dedicated to the victims of hurricanes Harvey, Irma, and **Maria**[1].

I cannot imagine a universe without my wife, Jimena.

Our two children, Ada and Simon, remind me why life is worth living.

The sound and thoughtful lives of Hector and Claudia have been a revelation to me.

Thanks, to you mom and dad, who, as part of a long if unlikely chain of coincidences, as unlikely as any, yet were the necessary and sufficient preconditions to "cause" my existence!

Finally, this dissertation is also dedicated to my canine, Clyde, who has been a trusty boon companion.

Thank you all for being there.

[1] Harvard estimated approximately 5,000 deaths in Puerto Rico in the wake of Hurricane Harvey ("Hurricane Maria Contributed to Nearly 5,000 Deaths, Researchers Say - Scientific American," n.d.).

Acknowledgments

This work was supported by the **Brown Foundation**, **NIH NCATS** grants **UL1 TR000371** and **UL1 TR001105**, **NLM R01-LM011563**, **NIH/NIGMS R01 GM103859**, **NSF grant III 0964613**, and by a training fellowship from the Gulf Coast Consortia, on the **NLM Training Program in Biomedical Informatics and Data Science T15 LM007093**. I would like to thank (in no specific order) Frank Manion, Lex Frieden, Ram Dixit, Sándor Darányi, Swaroop Gantela, Richard Boyce, Harry Hochheiser, Amy Franklin, Greg Cooper, Marco Scutari, and anonymous reviewers for their helpful suggestions, professional or technical advice, and feedback.

I would like to thank Clark Glymour of Carnegie Mellon University for his efforts and contribution as serving as an external reader of this dissertation.

I would like to thank the members of my committee for their invaluable feedback, mentoring, support, expertise, and for setting such a fine example as both scientists and as human beings.

Abstract

Prior work in causal modeling has focused primarily on learning graph structures and parameters to model data generating processes from observational or experimental data, while the focus of the literature-based discovery paradigm was to identify novel therapeutic hypotheses in publicly available knowledge. The critical contribution of this dissertation is to refashion the literature-based discovery paradigm as a means to populate causal models with relevant covariates to abet causal inference. In particular, this dissertation describes a generalizable framework for mapping from causal propositions in the literature to subgraphs populated by instantiated variables that reflect observational data. The observational data are those derived from electronic health records. The purpose of causal inference is to detect adverse drug event signals. The *Principle of the Common Cause* is exploited as a heuristic for a defeasible practical logic. The fundamental intuition is that improbable co-occurrences can be "explained away" with reference to a common cause, or confounder. Semantic constraints in literature-based discovery can be leveraged to identify such covariates. Further, the asymmetric semantic constraints of causal propositions map directly to the topology of causal graphs as directed edges. The hypothesis is that causal models conditioned on sets of such covariates will improve upon the performance of purely statistical techniques for detecting adverse drug event signals. By improving upon previous work in purely EHR-based pharmacovigilance, these results establish the utility of this scalable approach to automated causal inference.

Vita

1996 ..... B.A., Foreign Languages, EUP

2003 ..... MLIS, Library and Information Science, University of Pittsburgh

2010 ..... MSIT, Information Systems Management, Carnegie Mellon University

2018 ..... PhD, Biomedical Informatics, School of Biomedical Informatics, University of Texas

Publications

Field of Study

Biomedical Informatics

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

> *Long after the solstice, near the equinox, wintry weather returned, and at the actual equinoctial period there were southerly winds with snow, but not for long. The spring southerly again, with no winds; many rains throughout until the Dog Star. The summer was clear and warm, with waves of stifling heat. The Etesian winds were faint and intermittent. But, on the other hand, near the rising of Arcturus there were heavy rains with northerly winds.*
> *The year having proved southerly, wet and mild, in the winter the general health was good except for the consumptives, who will be described in due course.*
>
> *Hippocrates, Epidemics III, pg. 241*

Scientific understanding progresses when evidence is marshaled to explain some hitherto misunderstood aspect of our world. As findings are shared, a debate within the scientific community ensues over their meaning and validity. The validity of any study hinges in part upon the integrity of the data and the suitability of the methods used to analyze them. Techniques that are well-suited for analyzing experimental data may not be applied without modification to assay observational data. In other words, novel approaches must be developed that account for the unique characteristics of any new stream of empirical evidence.

This dissertation explores the accommodations necessary for using observational clinical data derived from electronic health records[2] (EHRs) to detect putative drug/adverse event relationships. An adverse event is defined as:

> "An appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product" (Edwards & Aronson, 2000).

Randomized controlled trials (RCTs) were developed to determine the efficacy and safety of novel treatments for disease and are considered a gold standard in this regard. However, the capacity of RCTs is limited concerning what they can tell about either effectiveness or safety under conditions of everyday use (Cartwright, 2007). Spontaneous Reporting Systems such as the Federal Adverse Event Reporting System (FAERS) in the United States have been developed to gather data for pharmacovigilance, or the post-marketing surveillance of drugs and other treatments (*Federal Drug Administration Adverse Event Reporting System*, 2017). However, researchers have established that adverse events are underreported (Alvarez-Requejo et al., 1998; Gahr, Eller, Connemann, & Schonfeldt-Lecuona, 2016; Hasford, Goettler, Munter, & Muller-Oerlinghausen, 2002; Perez Garcia & Figueras, 2011).

Electronic Health Records have been proposed as a source of data to complement spontaneous reporting systems. Indeed, EHR represents a rich but imperfect record of

---

[2] According to International Standards Organization, an electronic health record is defined as "repository of information regarding the health status of a subject of care, in computer processable form" ("ISO/TR 20514:2005(en), Health informatics — Electronic health record — Definition, scope and context," n.d.).

routine clinical practice and can be used to complement other sources of data (Hersh et al., 2013). The HITECH Act of 2009 mandates the "meaningful use" or secondary reuse of electronic health records (EHR) in research to improve public health outcomes (Henricks, 2011). As the bulk of data of interest are embedded in the unstructured text, it is necessary that these data undergo extensive text processing to make them amenable to computation and downstream analysis (Haerian et al., 2012; X. Wang, Hripcsak, Markatou, & Friedman, 2009). Another issue, one that is the focus of this dissertation, is that of confounding, or the presence of variables that may introduce bias if left uncontrolled (Brookhart, Stürmer, Glynn, Rassen, & Schneeweiss, 2010; S. Greenland & Morgenstern, 2001).

Unrecognized confounding factors are a significant analytic concern that can lead to erroneous conclusions (S. Greenland & Morgenstern, 2001). However, if these covariates are identified, they may help to "explain away" spurious associations and facilitate detection of significant relationships. For example, a strong initial correlation between the medication rosiglitazone may be observed with the adverse event acute myocardial infarction (Dore, Trivedi, Mor, & Lapane, 2009; Florez et al., 2015). However, if diabetes mellitus is included as a variable, the strength of this association diminishes. The inclusion of the comorbidity of diabetes makes sense since rosiglitazone is used to treat diabetes mellitus and diabetes mellitus is a known to cause heart attacks (Hanssen et al., 2018). Moreover, diabetes mellitus is a mutual cause of both the

treatment[3] (rosiglitazone exposure) and the adverse event (acute myocardial infarction). Note, however, that the "causal" interpretation of confounding outlined above has only slowly gained traction only in the last thirty-five or so years[4] (J. Pearl & Mackenzie, 2018). This interpretation, referred to as "the principle of the common cause," was first noted in 1956 in the philosophy of science literature (Reichenbach, 2012). The common cause principle is illustrated in **Figure 1**.

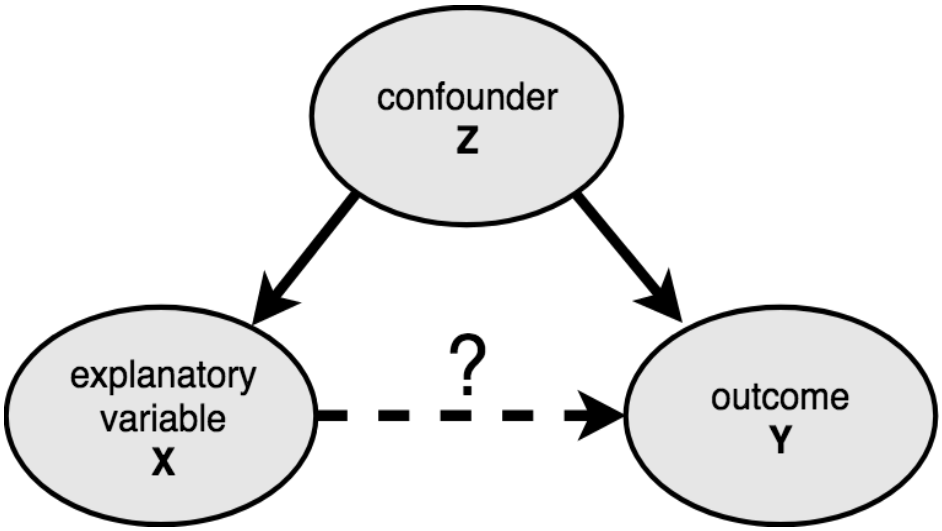


Figure 1. Illustration of a "true confounder" (S. Greenland, Pearl, & Robins, 1999). This illustration demonstrates the "forking pattern" exhibited by a confounder[5] on a putative predictor/explanatory/exposure variable (drug) and an outcome variable by an extraneous variable (a variable besides the exposure).

---

[3] Were it not for diabetes, the patient would not likely have been exposed to rosiglitazone. "*To treat*" is usually thought of as having the subject (the treatment, e.g., medication or device) exert its action (as a *force dynamic predicate* (Levin & Hovav, 2005)) upon an object (some pathological phenotype), but here it may be thought of in terms causation with the object exerting its influence, i.e., *causing* the exposure.

[4] More technically exacting definitions exist, but this explanation conveys the core of what is meant by "confounder" (T. J. VanderWeele & Shpitser, 2013).

[5] Greenland et al. offer this definition of confounder: "the variable is an ancestor (cause) of the outcome, and (2), the variable is associated with the exposure, but (3) the variable is not a descendant (effect) of the exposure or outcome" (S. Greenland, Pearl, & Robins, 1999; Weinberg, 1992).

The exploitation of observational EHR data for research purposes necessitates a means to identify such confounding variables, variables that are known to affect both the predictor (e.g., drug or pathogen exposures, genetic variants) and the health outcome of interest. Such exogenous factors are unavoidable in EHR data as clinical data are not collected under controlled experimental conditions (Brookhart, Stürmer, et al., 2010; S. Greenland & Morgenstern, 2001). Unless confounding bias is mitigated, the quality of analysis will be suboptimal (S. Greenland & Morgenstern, 2001). One approach to confounding control is by identifying confounders and including them into one's statistical or causal model through experience or domain knowledge. The problem of specifying a causal model consistent with background knowledge is referred to as the "identification problem" (Freedman, 2004; Han, Xie, Wu, Li, & Zhu, 2015; W. Li, Jiang, Geng, & Zhou, 2018; J. Pearl & Mackenzie, 2018). If more powerful methods are not developed to address confounding, the EHR is likely to remain an underutilized resource (Brookhart, Stürmer, et al., 2010; Hersh et al., 2013; Samuels, McGrath, Fetzer, Mittal, & Bourgoine, 2015).

This dissertation directly addresses the identification problem of causal modeling within a meta-statistical causal modeling framework. Here, I emphasize "meta-statistical" because descriptive statistics derived from data alone are insufficient to identify confounders or other types of causal relationships (Amirkhani, Rahmati, Lucas, & Hommersom, 2016; Cooper, Gregory F., 1984; Heckerman, Geiger, & Chickering, 1995; Meek, 2013; J. Pearl & Mackenzie, 2018; Judea Pearl, 2009, 2010).

Causal modeling may be thought of as a fusion method wherein assumptions from background knowledge are combined with empirical data (experimental, observational, or both) (Cooper & Yoo, 2013) to provide insight into the mechanisms responsible for generating inter-variable covariance patterns (Heckerman et al., 1995; Meek, 2013, 2013; Judea Pearl, 2009; P. Spirtes, Glymour, & Scheines, 2012).

A universally satisfactory definition of causality eludes scholars and remains a topic of active research. This dissertation assumes a classic counterfactual definition of causality[6]: "were it not for X, Y would not have happened" (Beebee, Hitchcock, & Menzies, 2009; Hume, 1748; Lewis, 1979; Mackie & Press, 1980; Judea Pearl, 2009; Salmon, 1984). Relationships due to causation and probabilistic association often share common attributes: temporal precedence and spatial contiguity. To use an example from Cheng et al., a rooster may crow before dawn, but sound does not and cannot levitate objects[7] (Holyoak & Cheng, 2011). A key distinction is that a causal relationship is by definition not merely necessary[8], but *stable under varying conditions and sufficient to produce an effect* (Mackie & Press, 1980; Judea Pearl, 2009; Woodward, 2016).

Given background knowledge of confounders identified in the literature[9], the present work describes a causal modeling framework to use observational clinical data

---

[6] Is causality a stochastic process with deterministic surface manifestations or deterministic process with stochastic surface manifestations? David Hume offered not one but at least three distinct definitions of causality in his works (Beebee, Hitchcock, & Menzies, 2009).

[7] Per Glymour in personal note: "the rooster's crow does levitate some objects, e.g., sleepers who hear it, just like a trumpet at dawn causes soldiers to rise."

[8] Oxygen is necessary for arson to occur, but it is not sufficient. Oxygen is, however, an "enabling condition."

[9] By contrast, in "causal discovery," a closely related area of research, one may not necessarily entertain strong assumptions about the causal relationships, if there are any, between the entities under study.

derived from EHR to determine the likelihood that a particular medication might cause an adverse event. This domain knowledge provides a "*causal story*" to help filter out more likely cause-and-effect drug/adverse event relationships from less likely drug/adverse event pairs that tend to frequently co-occur due to their having a common cause (J. Pearl, Glymour, & Jewell, 2016). The associational methods of traditional statistics can only indicate statistical correlation, and not determine causal relationships, yet such relationships are what is desired in the sciences (C. Glymour, Scheines, & Spirtes, 2014; Mackie & Press, 1980; Salmon, 1984).

Before proceeding further, a necessary clarification should be made to emphasize the distinction between the ends of causal modeling and the evidentiary establishment of causal claims. Confusion remains as the selfsame word is used for both (both hypothesis and its establishment) and that both rest on "external validation" (Reeves et al., 2014).

Sir Bradford Hill defined a set of criteria for establishing causal relationships in medicine: these criteria include prevalence, exposure, incidence, consistency, temporality, biological gradient, and so forth[10] (Hill, 1965). The ***establishment*** of a causal claim is a product of consensus mediated by the scientific community and a body of evidence[11] (Thagard, 2018). A causal claim requires *explanatory coherence* between the hypothesis and the evidence to be accepted. The elements of explanatory coherence

[10] Robert Koch in 1882 described another related set of criteria required to demonstrate causal relationships between microorganisms and disease, called the "Koch postulates" (Brock, 1961; Thagard, 2018).

[11] Thagard assumes that the discussion is about social establishment rather than the rational justification of causal claims.

include voluminous evidence of consistent biological plausible empirical data and if possible randomized experimentation to rule out confounding factors. By contrast, causal modeling cannot by itself ***establish*** a causal relationship, as this is a social process mitigated over time within the scientific community, and as such beyond the purview of the present discussion (Thagard, 2018).

Rather, *causal modeling is a tool for data-driven exploration of causal hypotheses*. Its inputs are data and causal assumptions (including domain knowledge), and the output of these artifacts is a causal model, i.e., a configurable blueprint of the mechanisms that underlie the variables so subsumed[12] (Cartwright, 2004; J. Pearl et al., 2016; J. Pearl & Mackenzie, 2018).

Causal modeling incurs an extra cost beyond raw data, as domain knowledge of confounders (or "a causal story") facilitates its practical application and hence, the discernment of causal relationships (J. Pearl et al., 2016). Since it is not feasible to perform causal analysis without such knowledge (although the data can in fact indicate confounders or their absence), causal modeling at scale is not tractable without some means to automate the identification of contextually relevant covariates.

One promising source of domain knowledge to populate causal models is the biomedical literature. If it were possible to identify confounding variables[13] in the

---

[12] Cartwright writes: "Old causal knowledge must be supplied for new causal knowledge to be had" (Cartwright, 1994)

[13] Other types or "roles" of causal variables exist: mediators which lie along the causal path from an explanatory variable to the outcome (Richiardi, Bellocco, & Zugna, 2013; T. J. VanderWeele, 2012, 2012), instrumental variables which are correlated with the explanatory variable but not the outcome, to name a few (Bowden & Turkington, 1985).

literature, this would obviate the need for manual construction of causal models by domain experts and hence permit the application of causal modeling and the evaluation of causal hypotheses at scale.

The primary hypothesis of this dissertation is that causal models informed by domain knowledge will outperform purely statistical methods for detecting drug/adverse event relationships in clinical data derived from EHR. To this end, this dissertation develops and describes a concrete application of this approach within the problem domain *clinical research informatics* (as it pertains to the secondary re-use of electronic health records), and more specifically to the field of pharmacovigilance. This approach leverages the literature to predict and identify confounders in EHR-derived data and uses these covariates within a probabilistic generative causal modeling framework to estimate the magnitude of drug/adverse event relationships. To access domain knowledge in the literature, this work describes techniques derived from Literature-Based Discovery (LBD) methods. LBD is a program of research pioneered by librarian-turned-researcher Don Swanson in the 1980s that focuses on identifying implicit relationships embedded in publicly available knowledge as a means to generate novel hypotheses[14] (Bruza & Weeber, 2008). The intuition behind LBD, which Swanson referred to as the "**A-B-C**" model is as follows: **A** is associated with **B** and **B** is associated with **C** in the literature (Smalheiser, 2012, 2017). The task of LBD is to identify what "**B**" is, referred to as a "bridging term" (Hristovski, Friedman, Rindflesch, & Peterlin, 2006). In the case of

[14] More recently, LBD work has expanded into detecting of pharmacovigilance signals in the biomedical literature (Shang, Xu, Rindflesch, & Cohen, 2014).

causal modeling for pharmacovigilance, the confounders identified in the literature derive from these bridging terms.

This dissertation tests the hypothesis above with three specific aims: Aim 1 describes the task of creating a database of clinical data derived from EHR and establishing a baseline using unadjusted traditional statistical methods (logistic regression) using a curated publicly available reference data set. Aim 2 defines a method to integrate the literature-identified variables into the causal graphs with a focus on the graph structure. Aim 3 presents a strategy to estimate average treatment effects using conditional probability queries on data simulated from "modified" or "mutilated" causal graphs.

These aims are described in detail as follows:

1.1 SPECIFIC AIM I: Extract clinical data from EHR and evaluate baselines scores using traditional statistical methods with and without confounding adjustment (from incorporating literature derived covariates).

Objective: Extract clinical data from a clinical data repository after having attained approval from the [UTHealth] IRB, and construct a statistical baseline using logistic regression for comparison with causal models.

Rationale: Create clinical data database and obtain a performance baseline for EHR-based pharmacovigilance for comparison with more sophisticated models that will incorporate literature-derived confounders. Demonstrate the feasibility of using LBD to identify contextually relevant confounders given a drug and an adverse event as "cue terms."

Research Question(s): How difficult is it to detect pharmacovigilance signals in EHR with traditional statistical methods (logistic regression)? Do literature-derived confounders help to reduce confounding bias in statistical models of EHR data? Which discovery patterns are most effective at reducing confounding? Can the literature be used to identify confounders? Can the identified confounders be used to "explain away" spurious correlations in observational clinical data, thereby improving the accuracy of predictions based on statistical associations?

1.1.1 Extract clinical data from EHR that will be utilized in evaluating the literature-identified covariates.

1.1.2 Evaluate baseline performance both with and without confounding adjustments. Use a publicly available reference dataset[15] with EHR data by calculating the Area Under the Curve of a Receiver Operator Characteristic (AUROC) from the aggregated ranked-order logistic regression coefficients for each drug/adverse event pair in the reference dataset (Ryan, Schuemie, et al., 2013; *UTHealth BIG*., 2017).

1.2 ***SPECIFIC AIM II: Develop and test the utility of the literature-identified confounders in causal models and compare the results with those from statistical models*****.**

Objective: Incorporate literature-identified confounders into causal models.

[15] Reference datasets in pharmacovigilance are used for methodological evaluation of novel methods for detecting drug/adverse event signals that may be worthy of further investigation.

Rationale: Demonstrate the utility of graph topology learned from EHR data within a pharmacovigilance use case, a core subfield of biomedicine.

Research Question(s): How useful is the structure of causal graphs for disentangling pharmacovigilance signal from noise? What is the optimal number of confounding variable candidates to incorporate into a model? Is automated causal inference feasible using literature-based discovery as a feature selection technique for potential confounders?

1.2.1 Define a method to integrate not only the literature-identified variables, but a causal subgraph[16] of these covariates into causal models. Evaluate the utility of the literature-derived confounders on the task of disentangling adverse event signals from noise in logistic regression models of EHR data using a publicly available reference dataset (Ryan, Schuemie, et al., 2013; *UTHealth BIG.*, 2017).

1.2.2 Evaluate the utility of causal models with literature-identified confounders and graph structure for improving signals of causal drug/adverse drug event relationships.

1.2.3 Evaluate the utility of causal models with literature-identified confounders and graph structure for improving signals of causal drug/adverse event relationships. Use a publicly available reference set with EHR data by calculating the AUROC from the aggregated ranked-order logistic regression coefficients for each

[16] "Causal subgraph" refers to the directed acyclic graph (which consists of nodes and directed edges, or arrows) wherein domain knowledge facilitates the orientation of the edges of the literature-derived confounders, i.e., the edges point fork-like from the confounder to the medication and putative adverse event.

drug/adverse event pair in the reference dataset (Ryan, Schuemie, et al., 2013; *UTHealth BIG.*, 2017).

1.3 *SPECIFIC AIM III: Develop and test methods to estimate Average Treatment Effect (ATE) from simulations using causal models informed by literature-derived confounders.*

Objective: Incorporate literature-identified confounders into causal models and estimate treatment effects using simulated data from the topology and structures learned from the EHR data.

Rationale: Demonstrate the generalizability of causal models with the automatic selection of literature-derived confounders using another set of clinical data derived from EHR and another reference dataset (Rave Harpaz, 2014).

Research Question(s): How useful is parameter estimation[17] for reducing error? Is it possible to calculate average treatment effect?

1.3.1 Develop the data generating process capabilities of causal models and exploit these simulations to estimate Average Treatment Effect (ATE). Evaluate methods using a publicly available reference set with EHR data by calculating the AUROC from the aggregated ranked-order logistic regression coefficients for each drug/adverse event pair in the reference dataset (Ryan, Schuemie, et al., 2013; *UTHealth BIG.*, 2017).

[17] "Parameter estimation" refers to estimating the strength of the causal relationship that is inferred between, for example, a particular medication and an alleged adverse event.

The research domains addressed by this thesis include literature-based discovery, causal modeling and discovery, and pharmacovigilance methodology. Natural language processing (NLP) facilitates this work but is not the theoretical focus. The literature-based discovery research paradigm is used as a means of accessing salient domain knowledge given a problem space defined by cue terms (a drug and an adverse event). The use of LBD methods to identify confounding variables is a novel application of these methods. Literature-based discovery in our configuration harnesses semantic constraints in the form of predicates to identify salient covariates for our models. In the approach, literature-based discovery is used to automate feature selection combined with constraints on the directed acyclic graphs in our causal models.

The contributions of this thesis to the discourse of biomedical informatics are as follows:

1. Demonstration of a novel domain of application for LBD methods.
2. Demonstration of LBD methods to automatically identify confounding variables and improve the accuracy of predictive models (both classical statistical and contemporary causal modeling methods).
3. Demonstration of how domain knowledge may be used as a means to constrain hypothesis space and automatically devise hypothetical explanations of empirical observational data. The automated generation of hypotheses to explain observations using domain knowledge is tantamount to implementing what is referred to as a "defeasible practical logic" *in*

silico (Dziurosz-Serafinowicz, 2012; Gabbay & Woods, 2005; Reichenbach, 2012).

4. The methods proposed may be generalized to other areas of biomedicine. For example, these methods may be adapted to identify control groups for RCTs (Fokkema, Smits, Zeileis, Hothorn, & Kelderman, 2017).

As a desirable bonus feature, the graphs that result from the approach should be clinically insightful. Interpretability being a notable feature of our approach, clinicians may use the graphs to enhance their comprehension of the underlying pathophysiology, inspire unexpected questions, or discern potential risk factors when considering a course of treatment, thereby improving health outcomes.

This dissertation is organized as follows. *Chapter 2* reviews previous work in EHR-based pharmacovigilance, identify methodological gaps and provide the background and motivation for understanding the impetus for refashioning the literature-based discovery framework to inform causal models using EHR-based observational clinical data as a primary data source. *Chapter 3* presents an overview of literature-based discovery and causal modeling components and how they complement each other. *Chapter 4* tests Specific Aim 1 by incorporating the confounding variable candidates identified in the literature into logistic regression models. *Chapter 5* and *Chapter 6* implement and evaluate variants of the causal modeling, and discuss their implications for the secondary analysis of observational data (addressing Specific Aim 2 and 3). Finally, *Chapter 7* summarizes the contributions that this work offers to the practice and

theory of informatics, discusses the limitations of the present study, and presents promising future directions for this program of research.

Chapter 2: Literature Review

There are also a number of things for which it is not enough to name one cause, but many, one of which is nevertheless the true cause: just as if you should yourself see some man's body lying lifeless at a distance, you may perhaps think proper to name all the causes of death in order that the one true cause of the man's death be named. For you could not prove that steel or cold had been the death of him, or disease, or it may be poison, but we know that what has happened to him is something of this sort.

Lucretius, De Rerum Natura, 6.703 – 709

This chapter provides an overview of related work in pharmacovigilance, causal inference, and literature-based discovery. To circumscribe the scope of the current review, unresolved challenges of using clinical text derived from electronic health record (EHR) systems for pharmacovigilance will be presented first. Noting the unresolved challenges will help to specify the problem, and thereby isolate the gaps in the current literature, that this dissertation seeks to address.

2.1 Pharmacovigilance

Some 770,000 adverse events occur annually in the United States alone, resulting in morbidity, mortality, and the increased cost is a considerable onus on healthcare

systems worldwide[18] (Diaz-Garelli, Bernstam, Mse, Rahbar, & Johnson, 2015; Hersh et al., 2013). Pharmacovigilance aims to address the set of challenges posed by adverse events, including those detected after drugs are released to market after regulatory approval. It seeks to ascertain the risks and benefits of exposure, identify contraindications, and in extreme cases withdraw products altogether in the event of severe adverse events[19] (J. K. Aronson, 2017). Recognizing the need to monitor adverse effects of drugs systematically, since surveillance cannot and does not end after approval, regulatory agencies such as the US Food and Drug Administration (FDA) have implemented spontaneous reporting systems, through which clinicians and administrators of clinical trials can report potential adverse events as they are observed. However, spontaneous reporting systems such as the Federal Adverse Event Report System (FAERS) have limitations, including incomplete clinical information, under-reporting of side-effects, and unacknowledged sources of bias (Alvarez-Requejo et al., 1998; *Federal Drug Administration Adverse Event Reporting System*, 2017; Hasford et al., 2002).

Researchers in pharmacovigilance have sought to evaluate the utility of other sources of data as input for pharmacovigilance methods. These have included both structured and unstructured data and from social media, claims data, and clinical data derived from electronic health record (EHR) systems (Edlinger et al., 2014; Eshleman & Singh, 2016; Haerian et al., 2012; Rave Harpaz et al., 2017; Rave Harpaz, DuMouchel, &

[18] In one UK study, adverse events resulted in ~ 6.5% of all hospital admissions (Pirmohamed et al., 2004).

[19] For a historical introduction to the history of pharmacovigilance from the case of thalidomide which led to the passage of the Kefauver-Harris Act of 1962, and the Vioxx case, see (Nesi, 2008; "News & Events > 50 Years," 2013; Stephens & Brynner, 2009).

Shah, 2015; Hersh et al., 2013; Y. Li et al., 2014; Lin & Schneeweiss, 2016; Liu, Zhao, & Zhang, 2016; Pierce et al., 2017; Samuels et al., 2015; X. Wang et al., 2009). The focus of this dissertation is on the use of unstructured clinical text recorded in EHR systems.

Clinical text (notes) recorded in EHR systems are a potentially useful source of data for pharmacovigilance, yet drawing reliable conclusions from routinely collected clinical data is notoriously challenging (Diaz-Garelli et al., 2015; Hersh et al., 2013). In the pharmacovigilance literature, the term "signal" refers to anything that warrants further investigation. Signals may be easier to detect in clinical trials where subjects are deliberately monitored for side-effects, by contrast data embedded clinical narratives were not collected for pharmacovigilance, and are often beset with redundancy or missing data, use of non-standard abbreviations, misspellings, and so on. In addition, clinical data contain confounding variables (Hersh et al., 2013; Y. Li et al., 2014).

2.1.2 Confounding in Electronic Health Records

An association between two variables, let's denote them **X** and **Y**, cannot be explained with reference only to themselves (Wasserman, 2013). One needs to introduce at least one other variable, and usually many other such covariates, to rule out non-causal reasons for an observed statistical correlation. If one were to possess domain knowledge of **X** and **Y**, one might include a third variable **Z** that experience or domain knowledge has determined may be responsible for producing them both and increasing the likelihood

of their co-occurrence (referred to as the principle of the common cause)[20] (Reichenbach, 2012). For example, there may be a robust initial correlation between a drug, e.g., rosiglitazone, and myocardial infarction. However, upon the introduction of the third variable of diabetes, which rosiglitazone is used to treat and which also is known to cause heart attacks, the robust initial correlation is diminished. The identification of such covariates may be acquired either through analysis of the data itself or by experience or domain knowledge (Y. Li et al., 2014a). Given such knowledge, the determination of the existence and/or magnitude of the effect of **X** on **Y** may be measured in one of two ways:

1.) By a randomized experiment, with samples chosen for similar risk characteristics, e.g., those who are carriers of an allele associated with risk of developing a disease or family history.

2.) From non-experimental observational data that infer the influence of latent confounding or with the inclusion of such covariates into statistical and/or causal models as shall subsequently be described.

Experimental conditions are by all means preferred when possible since the researcher may carefully control the introduction of exogenous or independent variables to measure the outcome and determine the existence and magnitude of any causal relationship. However, the opportunity and the advantages of such conditions may not always be feasible: randomized control trials may be unethical to obtain or intractably

[20] The Principle of the Common Cause does not always hold: correlation may be present in the absence of a common cause, e.g., the number of pirates and atmospheric carbon.

expensive, e.g., studying the relationship between tobacco use and lung cancer (Kovesdy & Kalantar-Zadeh, 2012). This brings us to the focus of this dissertation: the determination and estimation of the risk of adverse events arising from the exposure to medications using non-experimental data, specifically observational clinical data derived from unstructured clinical text (notes) in electronic health records [21].

The following section will discuss study designs of related work in pharmacovigilance using clinical data derived from electronic health records within the context of confounding.

2.1.2.1 Confounding control

If individual variables exogenous to but influencing an **X** and a **Y** of interest may be identified and have been measured, it may be possible to determine the existence of a relationship and/or estimated measure of the risk of an adverse event given exposure from observational data under certain conditions (C. Wang, Dominici, Parmigiani, & Zigler, 2015; G. Wang, Jung, Winnenburg, & Shah, 2015; X. Wang et al., 2009). Researchers have developed a convenient taxonomy for the various types of confounder that may be present in electronic health records for pharmacovigilance. These include confounding by *co-morbidity*, confounding by *co-medication* (where exposure to another drug is responsible for producing the adverse event), and confounding by *indication*

[21] As noted in Chapter 1, other forms of non-experimental observational data, e.g., claims and social media data, exist and are available for research; however, these lie beyond the scope of the present dissertation.

(wherein complications from the disease being treated may be culpable for producing the adverse event) (Y. Li et al., 2014a).

2.1.2.2 Confounding and Confounders

Confounding is present when the influence of common causes, alternate etiologies, or uncontrolled latent variables introduce bias in one's model (Judea Pearl, 2009; T. J. VanderWeele & Shpitser, 2011, 2013). The term "confounding" is the more generally accepted word for the overarching phenomenon that denotes bias endemic to observational data, while the term "confounder," which refers to an individual variable that introduces confounding, has only recently gained acceptance and traction among researchers (T. J. VanderWeele & Shpitser, 2013). Etymologically, confounding refers to a state of being mixed up or confused ("confound | Definition of confound in English by Oxford Dictionaries," n.d.).

If a confounder can be identified, then it may be incorporated into one's model to de-confound that model and reduce confounding and hence potentially improve the quality of one's analysis. Insofar as it reduces bias due to confounding, it is a *proper confounder* (S. Greenland et al., 1999; Judea Pearl, 2009; T. J. VanderWeele & Shpitser, 2013). However, if such a covariate does not serve to improve the model, then it should be ignored, a condition that is referred to as "ignorability" (S. Greenland & Morgenstern, 2001; S. Greenland & Robins, 1986; Sander Greenland & Robins, 2009).

2.1.2.3 Selection Bias, Measurement Error, and Confounding[22]

Since observational clinical data were not produced under experimental conditions, other types of bias may be present. These may include *selection bias* and *measurement error* (Haneuse, 2016; Talbot & Aronson, 2012). Selection bias occurs when a sample of individuals for a study is not properly randomized[23] and is therefore not representative (Haneuse, 2016; Miguel A. Hernan, Hernandez-Diaz, & Robins, 2004). For example, particular practices may have self-selected for patients that could be particularly vulnerable to an adverse event. Additionally, errors may be introduced in measurement in data collection as a result of typographical errors or other factors, such as clinician fatigue or burnout (Collier, 2017; EHRIntelligence, 2018; Wachter & Goldsmith, 2018). However, however methods to control these other sources of bias are beyond the scope of this dissertation and have been addressed elsewhere (Cartwright, 2004; Elwert & Winship, 2014; Haneuse, 2016; Miguel A. Hernan et al., 2004; P. H. Lee & Burstyn, 2016; Suzuki, Tsuda, Mitsuhashi, Mansournia, & Yamamoto, 2016). Instead, this dissertation focuses on the prospect of controlling exogenous variables that may be identified in the clinical text. The next section describes methods to control for confounding either through analysis of the data or by other means, such as the application

[22] I acknowledge that I am excluding missing data bias, reporting bias, design bias, protophatic bias, clinician facility bias, clinical NLP bias, and so on.

[23] Glymour in personal note: "randomization does not guarantee representativeness, nor does absence of randomization entail that a sample is not representative."

of background knowledge extraneous to the raw data inputs to inform structural causal models (the focus of this dissertation).

2.2 Related Work: confounding control in EHR-based pharmacovigilance

Confounding control in observational data study designs may be categorized by two general classes: control by design (e.g., cohort studies, case-control) and control by analysis (Talbot & Aronson, 2012).

2.2.1 Control by Design

Cohort and case-control models are two methods that have been proposed to address confounding in EHR data that can be characterized by "control by design" (Talbot & Aronson, 2012). Similar to RCTs, subgroups are identified and included in these studies for their respective populations having similar risk factors and other characteristics "with the purpose of mitigating the effects of confounders" (Lewallen & Courtright, 1998). A primary difference between case-control and cohort studies is that case-control studies are retrospective and cohort studies are prospective (Lanza, Ravaud, Riveros, & Dechartres, 2016; Pugh, Bronsvoort, Handel, Summers, & Clements, 2014; Talbot & Aronson, 2012, p. 376). Such study designs are notable for producing accurate and useful results at detecting and verifying adverse events given exposures, particularly rare drug/adverse event relationships, but may be susceptible to selection bias (Backenroth, Chase, Friedman, & Wei, 2016; de Bie et al., 2015; Lanza et al., 2016; Norén et al., 2013; Talbot & Aronson, 2012; Thygesen et al., 2017, p. 376). Cohort

studies excel at detecting rare events (Pugh et al., 2014). However, it is important to point out an important limitation: such study designs are not informative about risk factors, enabling conditions, contraindications, and alternate etiologies that can help to explain the nature of an alleged drug/adverse event relationship. Furthermore, in observational data there is always the risk of unobserved and unmeasured covariates.

2.2.2 Control by Analysis

Control by analysis implies the application of statistical techniques to identify covariates to control for confounding from data or background knowledge. One research tack in confounding control in EHR-based pharmacovigilance has focused on using the data to identify individual confounders (Backenroth et al., 2016; Y. Li et al., 2014). Li et al. recently developed a method to identify confounders in data. Li used a propensity score method (PSM) (Tatonetti, Ye, Daneshjou, & Altman, 2012) to identify factors associated with the treatment and another technique to identify risk factors associated with the outcome (R. Harpaz et al., 2012). An overlap between these two sets (of the "comorbidity" subtype in pharmacovigilance) were collected. Li processed subsets of these data-derived confounders using penalized regularization methods, specifically lasso regression (Y. Li et al., 2014). Lasso regression is a regularization and variable selection technique that shrinks multivariate predictor coefficients that fall beneath a threshold down to zero (Tibshirani, 1996). However, Least Angle Regression (LAR), the algorithm that is most commonly used to perform lasso regression, can be computationally expensive, depending on input, being either quadratic $\boldsymbol{O}(\boldsymbol{n}^{\boldsymbol{2}})$ or cubic $\boldsymbol{O}(\boldsymbol{n}^{\boldsymbol{3}})$ in its computational complexity. Recent innovations such as cyclic coordinate descent, have

produced improvements in this regard (Efron, Hastie, Johnstone, & Tibshirani, 2004; J. Friedman, Hastie, & Tibshirani, 2010; N. Simon, Friedman, Hastie, & Tibshirani, 2011).

Using these methods, Li reported a precision of 83.3% (95% CI: 62.2% to 100%) for rhabdomyolysis and 60.8% (95% CI: 47.4% to 74.2%) for pancreatitis. These scores improved upon scores and confidence intervals using either risk factors or PSM scores alone. Li noted that having a sufficient number of samples is critical to decreasing the number of false negatives, as many adverse events are rare.

2.2.2.1 Disproportionality Metrics

Disproportionality metrics are a traditional method of detecting drug/adverse event relationships in SRSs. These include Odds Ratio, Reporting Odds Ratio (ROR), and Reporting Risk Ratio (RRR). As these metrics are derived from the occurrence statistics of pairs of entities (without addressing other [potentially confounding] entities), these techniques were found to have little utility in terms of either sensitivity or specificity for EHR data (DuMouchel, Ryan, Schuemie, & Madigan, 2013; Ryan, Stang, et al., 2013).To account for sampling bias and differences between sample size, to increase sensitivity to rarer adverse events, the Multi-item Gamma Poisson Shrinker (MGPS) was developed and is currently in use by the United States Food and Drug Administration (FDA) (Commissioner, n.d.; Rave Harpaz et al., 2013). Longitudinal Gamma Poisson Shrinker (LGPS), a variation of MGPS, has been applied to claims and EHR data (Schuemie, 2011; Zorych, Madigan, Ryan, & Bate, 2013). The inability of

these methods to account for confounding has been cited as a limitation (Shrier & Pang, 2015).

2.2.2.2 Meta-analytic methods

Other innovative signal detection approaches have involved combining multiple data sources (including at times "omics" data) via meta-analysis (Evans, Chaix, Lobbedez, Verger, & Flahault, 2012; Y. Li, 2015; Oliveira et al., 2013; G. Trifiro et al., 2014). Li was able to achieve 0.73 overall AUROC by combining EHR with FAERS data (improving AUROC of the EHR by 0.22 from the overall baseline of 0.51) (Y. Li, 2015). However, these techniques insofar as they apply to the detection of drug safety signals in EHR belong arguably to a higher ontological order of pharmacovigilance – that of substantiation and validation (Bauer-Mehren et al., 2012; Talbot & Aronson, 2012; Gianluca Trifiro, Sultana, & Bate, 2018). Meta-analysis has the potential to improve the performance of any individual method to the extent that it can be applied to multiple data sources.

2.2.2.3 Other common methods

Other regression and regularization based methods have been applied to EHR systems, including propensity scores – wherein the characteristics of a large set of covariates is collated into a single score that measures the likelihood that a patient

received a treatment (Rave Harpaz et al., 2015; Madigan, Ryan, & Schuemie, 2013; Rosenbaum & Rubin, Donald, 1983). Ideally, the treated and untreated subgroups should have similar characteristics. However, Ali et al. and Jackson report that the selection of covariates that are used in studies that that use the PSM rarely report their covariates or how they came to chose them (Ali et al., 2015; Jackson, Schmid, & Stuart, 2017).

There exist hybrid methods as well, e.g., recent research incorporating confounders from enriched data (Backenroth et al., 2016). Data enrichment refers to the utility of introducing relevant evidence to a problem of interest to contextualize and understand it (Boyce et al., 2014).

The question remains: what is the best way to integrate these data, this information, this knowledge? Associational studies using traditional statistics offer only descriptions of data, yet a critical component of what science desires is a means to derive insight into mechanisms and interrelationships. The next section describes work that has the means to peer beyond the associational approach.

2.2.2.4 Causal inference methods

The objective of causal inference is to estimate the likelihood of one variable producing a change in another under varying conditions and may be thought of as "the counterpart to experiment" for observational data (Ranganath & Perotte, 2018).To date, most work that has been done with causal inference methods in the field of pharmacovigilance have been in application of instrumental variables (Brookhart, Rassen, & Schneeweiss, 2010; Brookhart, Wang, Solomon, & Schneeweiss, 2006; Davies, Smith, Windmeijer, & Martin, 2013). An instrumental variable is a variable that

is chosen for fulfilling two criteria: 1.) it should not share a mutual cause with the outcome (in this case, an adverse event), and hence is unconfounded with it; and 2.) it should be correlated strongly with the explanatory variable, or estimator (in this case, a drug) (S. Greenland, 2000).

An example of an instrumental variable that has been used in a pharmacovigilance study would be the inclusion of a variable that represents the attending clinician, since physicians may tend to be partial to the medications that they prescribe. However, controversy exists over the validity of such variables (Brookhart et al., 2006). One meta-analysis of instrumental variable studies found that only 16/28 offered empirical evidence that their choice of instrument fulfilled both criteria (Chen & Briesacher, 2011). Other examples of instrumental variable types included patient history, financial status, and calendar time. The intuition that underlies instrumental variables is that the instrument, e.g., patient income, may be thought of as a "cue ball" in a game of billiards, where that ball is struck by the player and is used strategically to propel other balls into the pockets of the gaming table. If the instrumental cue ball is correlated with the outcome (the adverse event), the causal effect of the candidate cause may be estimated with a simple ordinary least squares regression (OLS) or two state least squares[24] (Bowden & Turkington, 1985; S. Greenland, 2000). Instrumental variables may be thought of as side-stepping the issue of *identifying individual confounders*, since instrumental variables may be used to estimate effects "even in the presence of unmeasured common causes"

[24] OLS will work assuming that the dependency is linear.

(confounders)[25] (Chu, Scheines, & Spirtes, 2013). One recent exception to the focus on instrumental variables in pharmacovigilance explicitly addresses the issue of confounding, but the method described sidesteps the issue of identifying confounders by inferring the magnitude of shared confounders between independent treatment regimes and estimating upper and lower bounds for the influence of the latent (and unidentified) confounders (Ranganath & Perotte, 2018).

The causal modeling techniques presented above excels in the case when there is hidden latent and/or *unmeasured* confounding as no knowledge of covariates is required and so the covariates it follows do not have to be measured[26] (Ali et al., 2015). Indeed, many of the methods that have been reviewed thus far are agnostic of covariates. What happens if we have rich though flawed EHR data extracted from narrative text: a case in which many individual confounders may have been measured? One promising source of knowledge that may be exploited to identify potentially relevant covariates is the literature.

2.4 Literature-Based Discovery

A fundamental limitation of the approaches mentioned above is that etiological insights into confounders, risk factors, and enabling conditions that are present in the literature may be amiss in the analyses of researchers. Absent the identification of exogenous

[25] Semi-instrumental variables wherein the non-confounding assumptions concerning the outcome of interest are weakened may be used as estimators given that common causes between the instrument and the outcome are controlled for and measured (Chu, Scheines, & Spirtes, 2013).

[26] Koller likens such methods to modelling a process of, to paraphrase, all latent background stochastic processes (Koller, Friedman, & Bach, 2009).

variables in observational data, researchers may not be able to take full advantage of EHR data. Since EHR is so rich, might it be possible to identify confounders and control for them in observational data so as to gain additional insight?

Traditionally, statisticians have depended upon the domain knowledge elicited painstakingly from domain experts to identify relevant confounding factors (Y. Li et al., 2014). While this approach to gathering background knowledge is likely to result in the identification of confounding variable candidates pertinent to an individual study, it would be financially intractable to hire the quantity and diversity of experts needed to conduct pharmacovigilance across large numbers of marketed drugs and each of their potential adverse events.

The focus of literature-based discovery has historically been on the identification of novel therapies. Literature-based discovery methods have recently been utilized to assess the plausibility of drug/adverse event associations (Cohen & Widdows, 2017; Hristovski et al., 2006; Mower, Subramanian, Shang, & Cohen, 2016; Shang, Xu, Rindflesch, & Cohen, 2014). As literature-based discovery has not traditionally been a paradigm to interoperate with observational data, the current work represents a novel problem domain for literature-based discovery methods, where such methods are repurposed to identify confounders in observational data for adjustment.

2.5 Summary

The gap the current work addresses is the lack of methods tools to LBD-informed causal modeling. The purpose of this work is to assess the extent to which feasibly sized

sets of confounding variable candidates that have been observed in the literature can be used to identify concepts that reflect this "causal story" for the task of confounding adjustment of clinical data derived from EHR. The data-driven, knowledge-based method described aims to find a middle ground between human-intensive expert-guided confounding variable discovery and computationally intensive selection of such variables based on empirical data alone. The following chapter provides an overview of literature-based discovery methods of identifying such a causal story with which to populate statistical and causal inference models.

Chapter 3: Methodological Framework

This chapter explains the core components of the approach developed and tested in this dissertation (causal modeling and literature-based discovery), and then provides theoretical context to understand how these components complement each other.

Causal Modeling and Literature-based Discovery: a Synthesis. The method developed uses the literature to identify contextually relevant variables to include in statistical and/or causal models as a means to control for confounding bias in observational clinical data derived from EHR. This chapter leaves out some specifics, e.g., version of SemMedDB used, which Discovery Patterns, the advanced methods for estimate the Average Treatment Effect (ATE) (*described in Chapter 6*). Such specific details of the experiments will be described and clarified in the chapters to follow.

3.1 Overview of Causal Modeling

In recent decades, causal inference methods have been developed by such seminal figures as Judea Pearl, Gregory Cooper, and Clark Glymour (Cooper, Gregory F., 1984; Judea Pearl, 2009; P. Spirtes et al., 2012; Peter Spirtes & Glymour, 1991). The patrimony of the field of causal modeling is remarkably diverse and interdisciplinary. The fields that have contributed to causal inference include the following: its epistemological basis (philosophy of science), potential outcomes framework and path diagrams (agronomy and genetics) ("Corn and hog correlations / by Sewall Wright. v.1300(1925). - Full View | HathiTrust Digital Library | HathiTrust Digital Library," n.d.; Neyman, 1937; Neyman,

Iwaszkiewicz, & Kolodziejczyk, 1935; Rubin, 1990; S, 1921), structural equation models (economics and the social sciences) (A, 1990; "CFM 14 | Cowles Foundation for Research in Economics," n.d.; Duncan, 1975; Goldberger & Duncan, 1973; Haavelmo, 1944; J, 2008), counterfactual analysis (philosophy) (Cartwright, 2007; C. N. Glymour, 2001, 2001; Hume, 1740, 1748; Lewis, 1979; Mill, 1843; Scheines, Spirtes, Glymour, Meek, & Richardson, 1998), computational tractability (computer science and cognitive science) (Gabbay & Woods, 2005; Gopnik et al., 2004; Griffiths, Sobel, Tenenbaum, & Gopnik, 2011; Griffiths & Tenenbaum, 2005; Holyoak & Cheng, 2011, 2011; Judea Pearl, 2009). While an exhaustive review of the field is beyond the scope of the present discourse, I will provide historical details when it may prove insightful or for clarification.

The objectives of causal inference are manifold. One researcher may wish to specify the data generating process responsible for a set of observations to understand the underlying mechanisms. Another objective might be to provide insight into how to weigh policy options or risk/benefit factors when considering an intervention, e.g., the effects of raising the rate of interest on employment, or choosing between chemotherapy or surgery (J. Pearl & Mackenzie, 2018; Judea Pearl, 2009). Exact causal inference has been shown to be NP-hard[27]; however, causal inference from data may be approximated (Gavril, 1977; Heckerman et al., 1995; Koller, Friedman, & Bach, 2009; Marco Scutari, Vitolo, & Tucker, 2018).

[27] The search space is super-exponential with the number N of nodes (N. Friedman, 2013).

The causal graphs (also referred to as "causal Bayesian networks") and the potential outcomes (also referred to as the counterfactual intervention) framework have been synthesized into a coherent discourse only in the recent past several decades[28] (M.A. Hernan & Robins, 2017; Morgan & Winship, 2015; J. Pearl & Mackenzie, 2018; Judea Pearl, 2009). Researchers realized that causal graphs and counterfactual approaches are essentially different languages for emphasizing and expressing aspects of the same core underlying ideas, with each suited for the sub-tasks for which they were initially devised (Morgan & Winship, 2015; J. Pearl & Mackenzie, 2018; Judea Pearl, 2009). These core tasks of causal modeling are the following: inferring causal graph structures (qualitatively determining which variables influence each other), quantitatively estimating relative strength of such relationships encoded in the graph, and predicting how the variables subsumed in such models might behave under perturbation.

Causal inference requires additional assumptions (although these can be tempered): faithfulness, the causal Markov condition, and absence of latent confounders (the latter will be defined and explained in the next section). *Faithfulness* assumes that the causal graph represents valid dependency or causal relationships (J. Ramsey, Zhang, & Spirtes, 2012; Uhler, Raskutti, Bühlmann, & Yu, 2013; Zhang & Spirtes, 2012). The causal Markov condition describes how parent nodes in a DAG define child nodes. The intuition behind this is that sans quantum entanglement or "spooky action at a distance," causal relations tend to be spatially and temporally contiguous. Finally, there is the

[28] Contention remains about whether or not these two frameworks have been reconciled, however. At there remain two camps: those who use graph search vs. the potential outcomes (where the predictor and an outcome variable are pre-specified). In this dissertation, both approaches are used.

assumption of the *absence of latent confounding*[29]. That is to say that it is assumed that the model includes all relevant mutual causes and influences. Other perils exist for the would-be causal modeler: for example, selecting confounders that are themselves confounded, and "over-controlling" for confounding (Elwert & Winship, 2014; M.A. Hernan & Robins, 2017; Morgan & Winship, 2015; Judea Pearl, 2009).

Structure Learning. There are two main algorithm classes for identifying graphical structures that are compatible with input data. These are constraint-based and score-based learning algorithms. Constraint-based learning entails a search for dependency relationships between input data representing nodes in a graph. If two nodes are independent (their correlation falls beneath a threshold), then no edge is detected between them (correlation is a necessary but insufficient pre-condition for a causal edge). The most famous constraint-based causal structure learning algorithm is the "PC" algorithm, first developed by Peter Spirtes and Clark Glymour (Peter Spirtes & Glymour, 1991). Another is the Fast Causal Inference (FCI) algorithm (Scheines et al., 1998; Peter Spirtes & Zhang, 2016; Zhang, 2008).

Score-based algorithms on the other hand stochastically add, delete, orient edges, and optimize for a fitness criterion, e.g., Bayesian Information Criterion, wherein the loss function is minimized between the model and the data. Examples of these include the Fast Greedy Equivalency Search (FGeS) algorithm (J. D. Ramsey, 2015), derivative of Chickering and Meek (Chickering, 2015). Score-based algorithms are typically more

[29] This assumption is often weakened or qualified as algorithms exist specifically to detect hidden latent confounding in the data, e.g. Greedy Fast Causal Inference (GFCI) and Fast Causal Inference (FCI) .

efficient than those which are constraint-based since they do not perform an exhaustive search. While score-based methods excel in computational efficiency and are easily parallelized, they are often poor at detecting latent variables (J. D. Ramsey & Malinsky, 2016; Zhang & Spirtes, 2012). Hybrid structure learning algorithms exist as well. Most notable among these would be Greedy Fast Causal Inference (GFCI), which first runs FGeS and then prunes edges with a constraint-based structure learning algorithm (Ogarrio, Spirtes, & Ramsey, 2016; J. D. Ramsey & Malinsky, 2016; *TETRAD*, 2017). Parameter estimation and the quantification of estimated treatment or causal effects are addressed in Chapter 6.

The output of most score-based causal structure learning algorithms is a set of directed graphs with the observed variables as vertices and is known as a completed partially directed acyclic graph (also called a "pattern"), or CPDAG. A CPDAG describes a family of graphs which are score-equivalent (Geiger, 1990; Ogarrio et al., 2016; Judea Pearl, 2009; P. L. Spirtes, 2013). That is, based on the input data, the structure learning algorithm cannot specify a unique form of a graph as there may be latent variables (Chu et al., 2013; Geiger, 1990; P. L. Spirtes, 2013). In the case of a CPDAG, this means that some edges may be bi-directed. However, given domain knowledge, many of these edges can be oriented. If a known confounder is introduced, for example, directed edges should be oriented toward both the predictor or explanatory variable and the outcome. The next subsection discusses how confounder variable candidates are identified in the literature.

3.2 Literature-Based Discovery

Literature-based Discovery (LBD) is an idea first developed by Don Swanson as a means of using the biomedical literature to investigate therapeutically useful associations (Smalheiser, 2017; Swanson, 1988, 1989, 1989). Swanson's approach involves examining latent relationships between concepts that may suggest undiscovered therapeutic relationships. More recently, LBD researchers have explored the idea of using semantic constraints to reduce the search space of relevant associations (Bruza & Weeber, 2008; Hristovski et al., 2006). For example, the following discovery pattern describes a set of semantic constraints with which to identify a set of concepts: "drug **INHIBITS** *x*; *x* **CAUSES** disease." The variable "*x,*" referred to as a "bridging term," may indicate a confounding concept that is associated with a drug and an adverse event. An example of a bridging term would be a term that appeared in Don Swanson's original research in LBD: Eicosapentaenoic acid, or EPA, an ingredient in fish oil. Swanson noticed that the *mechanisms* that underlie the pathology of those who suffer from Raynaud disease are the opposite of the changes that are produced by consuming fish oil, e.g., *decreased* vs. *increased* blood viscosity (Swanson, 1986). In this way, Swanson was able to identify a novel therapy for Raynaud's disease and proceeded to identify other potential therapies (Swanson, 1988, 1989). These patterns of relationships are known as "***discovery patterns***" (Hristovski et al., 2006).

While discovery patterns have been used to identify therapeutic and harmful relationships previously, their application to identify confounding variables for causal and statistical models is unprecedented. Some means to traverse discovery patterns is

required to make LBD scalable. The next section will address such computational considerations.

3.2.1 Predication-Based Semantic Indexing (PSI)

Recent work in the area of discovery pattern-based LBD has leveraged high-dimensional vector space representations derived from large (tens of millions) repositories of concept-by-predicate-by-concept triples[30] (known as semantic predications – e.g., x **INHIBITS** y) to facilitate efficient search and accurate inference, using a technique called Predication-based Semantic Indexing (PSI) (Cohen, Schvaneveldt, & Rindflesch, 2009; Cohen, Widdows, Schvaneveldt, Davies, & Rindflesch, 2012). Vector Space Architecture (VSA) theory provides the infrastructure for PSI (Gayler, 2004; Levy & Gayler, 2008). After the imposition of a reversible vector transformation to encode the nature of a relationship connecting two concepts, these elemental vectors can be superposed upon each other to generate composite semantic structures called semantic vectors (Cohen et al., 2009; Cohen, Schvaneveldt, & Widdows, 2010; Cohen, Whitfield, Schvaneveldt, Mukund, & Rindflesch, 2010; Cohen, Widdows, Schvaneveldt, & Rindflesch, n.d.; Kanerva, Kristoferson, & Holst, 2000; Widdows & Cohen, 2015).

3.2.2 PSI and for identifying confounding variable candidates

PSI uses random vectors. These random vectors are an effective way to represent elemental components, since there is a high probability of mutual near-orthogonality in

[30] Predications are relationships in which pairs of concepts are connected by predicates (or "verbs", e.g., "TREATS", "CAUSES", "INHIBITS", "STIMULATES") in biomedical literature.

higher dimensions. Semantic vectors, on the other hand, are composed as superpositions of the "bound products" of the aforementioned elemental vectors of predicate-argument pairs (as extracted from the literature using SemRep) (Kanerva, 1994; Kanerva et al., 2000). The binding operator, which varies in implementation across VSAs, is a multiplication-like operator that provides the means to encode additional information, such as the nature and context of a relationship, into the resulting vector space. Since the same predication can be encountered in multiple documents, PSI can be thought of as a distributional model of predications. Critical to PSI, semantic representations of concepts are built up as vectors from relations found in the literature. PSI facilitates the rapid search for, and retrieval of, concepts that are related to one another in particular ways (i.e., through particular predicates). As such a space is distributional, concepts in which a relationship of interest occur more frequently will be retrieved first (analogous to the way in which other information retrieval systems facilitate ranked results). In the current work, PSI is used to facilitate rapid retrieval of concepts related to other concepts along discovery patterns that suggest potential confounders. PSI can be used to retrieve the most strongly associated concepts (called "bridging terms") across any particular predicate discovery pattern of interest and even identify novel discovery patterns themselves (Cohen et al., 2012). A discussion of predicate discovery patterns that indicate discovery patterns will follow. PSI and its other applications are discussed in detail elsewhere (Cohen, Schvaneveldt, et al., 2010, 2010; Cohen et al., 2012; Widdows & Cohen, 2015).

The operative assumption is that if plausible therapeutic relationships in a knowledge base of biomedical literature (which has arguably been the primary focus of work in LBD to date), then semantic constraints may be used to identify common causes and exploit the common cause principle. Since LBD uses distributional semantics, such concepts that are identified in the literature may be predictive of entities that tend to co-occur with drug/adverse event pairs of interest in observational clinical data. Confounding variable candidates may be identified by matching the pre-defined relationship-types encoded by discovery patterns. In the next section the question of how the structure suggested by discovery patterns can inform the construction of causal graphs instantiated with EHR data.

3.3 Framework for LBD-informed Statistical and Causal Modeling

This section discusses how literature-derived confounding variable candidates may be incorporated into predictive (statistical or causal) models.

3.3.1 LBD-informed Statistical Modeling

As there is no structure in statistical modeling, the integration of literature-derived covariates is simple. Once a set of covariates has been chosen, a model and a regression coefficient may be extracted for further processing.

Domain knowledge extracted from the literature assists in the automated construction of statistical models in at least two ways:

1. By identifying relevant covariates informed by semantic constraints, the input may be supplied for processing by statistical methods;

2. Given semantic constraints of the literature-derived confounders, this reduces the computational requirements when using a regularization method (ridge or lasso regression) to reduce the set of covariates.

3.3.2 LBD-informed Causal Modeling

Causal inference requires some means to automate the process of identifying contextually relevant covariates (feature selection) to scale. Domain knowledge extracted from the literature assists in the automated construction of causal models in three ways:

1. By identifying relevant covariates informed by semantic constraints, the causal graphs are populated with nodes/covariates;
2. Given semantic constraints of the literature-derived confounders, the edges of the graph can be oriented in the event of "data equivalence" of the learned graph structure.
3. Prior information concerning the orientation of graph edges helps computationally to reduce the search space for learning the graph structure.

Once the structure has been learned, the relationship between nodes may be estimated. So, LBD can provide contextually relevant domain knowledge to facilitate the automation of causal modeling. This section will describe the methods that have been developed to facilitate predictive modeling in pharmacovigilance.

Causal modeling for pharmacovigilance. Moving from these theoretical considerations, I will describe how sets of such confounders can be identified for inclusion into causal models for de-confounding observational clinical data derived from EHR data.

3.3.3 Methodological Overview and Evaluation Framework

The next subsections describe the process of constructing a causal knowledge base using PSI and the steps for using this knowledge base to identify, validate, and include literature-derived covariates in statistical and causal models.

3.3.3.1 Materials

This subsection introduces the data that will be used in a series of experiments for evaluating the method. The data that one first receives are seldom in the shape that they need to be for practical use. The next section describes how the raw data were collected and processed into a representation amenable to downstream analysis.

EHR data. Clinical text derived from electronic health records is the primary source of raw data for this dissertation. Data from structured fields were not included in the analysis, only unstructured text. Unstructured text data is known to contain valuable information that is not contained in structured data – this may include indications of "temporal relations, severity and degree modifiers, causal connections, clinical explanations, and rationale" (Johnson et al., 2008). The primary unit of analysis is the individual electronic health record which represents a single, unique visit by a patient to a clinician. EHR data was obtained after obtaining permission from the [UTHealth] Internal Review Board (IRB). A corpus of approximately 2.2 million electronic health records (EHR) was extracted from the UTHealth's clinical data warehouse concerning outpatient encounters for ~ 364,000 patients in the Houston metropolitan area between 2004-2012 (*UTHealth BIG.*, 2017). *MedLEE*, a clinical Natural Language Processing

system, was used to normalize concepts in the EHR corpus (C. Friedman, Shagina, Lussier, & Hripcsak, 2004). MedLEE has been shown to perform accurately on clinical notes, for example with a recall of 0.77, and precision of ~ 0.89 for the task of extracting clinical concepts (C. Friedman et al., 2004). In addition to identifying concept types ("health outcomes of interest," "medications"), MedLEE also encodes each extracted concept with a concept unique identifier (CUI) from the Unified Medical Language System (UMLS) (A. R. Aronson, 2001; *The Unified Medical Language System (UMLS)*., 2017).

Next, for convenience, I indexed the MedLEE-process corpus of EHR data using Apache Lucene[31] to facilitate the extraction of document-level co-occurrence statistics. From this index, document-by-concept arrays were obtained. A list of document ids specific to the Lucene index was extracted for each UMLS CUI in the index. Next, each concept (drug, adverse event, or adjustment set of confounders) was persisted as a large sparse binary array and compressed. In these binary arrays (input for causal algorithms), a value of 1 or 0 represents presence or absence of that concept within a document in the corpus index. These arrays are later used as input for statistical and causal algorithms. These matrices can be constructed quickly, and can represent the observational data for each of the drugs, adverse events, and their confounders, and can provide input matrices for the causal methods to be described.

[31] https://lucene.apache.org

Reference Dataset. Reference data sets are used for methodological evaluation in pharmacovigilance. These datasets customarily consist of a number of medication/adverse event pairs wherein there are cases (a causal relationship between exposure and adverse event has been established) and controls (no known relationship). A reference set of curated drug/adverse event associations that was developed by Ryan and his colleagues as a standard for evaluating pharmacovigilance method was used for methodological evaluation (Ryan, Schuemie, et al., 2013). This reference set includes 399 drug/adverse event pairs for four clinically important adverse events with both positive (drug/adverse event relationships supported by the literature and other sources, including package labeling events) and negative (drug/adverse event relationships without support) control groups per adverse event. The four adverse events are as follows: acute kidney injury (**AKI**), acute liver injury (**ALI**), gastrointestinal bleeding (**GIB**), and acute myocardial infarction (**MI**). These adverse events were chosen for their importance to pharmacovigilance and for their impact on financial and personal cost. I mapped and expanded drug/adverse event synonyms to make the EHR data amenable to additional processing. I used RxNorm for drug synonyms at the clinical drug level, and I assigned the reference set's Observational Medical Outcomes Partnership (OMOP)adverse event to UMLS CUIs (Nelson, Zeng, Kilbourne, Powell, & Moore, 2011; *Observational Medical Outcomes Partnership (OMOP)*, 2017; *RxNorm*., 2017). For example, the generic concept "Ibuprofen" is encoded with a CUI string of "C0020740," while the specific concept that refers to a brand-name instance of "Advil Ibuprofen Caplets" is

"C0305170." To gain statistical power, these CUIs are reconciled into a single representation. Table 1 presents an overview of the reference dataset:

Table 1

Ryan et al. (2013) Reference Dataset

Adverse Events	**Case**	**Control**	**Total**
AKI (Acute Kidney Injury)	24	64	88
ALI (Acute Liver Injury)	81	37	118
AMI (Acute Myocardial Infarction)	36	66	102
GIB (Gastrointestinal Bleeding)	24	67	91
Total	164	164	399

Note. Categories in far left denote health outcomes of interest phenotypes in the OMOP reference data set.

The next subsections discuss the confounding variable candidate discovery and validation process.

3.3.3 Confounding Variable Discovery Process

To construct a knowledge base for the present experiment, Predication-based Semantic Indexing was supplied to SemMedDB, a knowledge repository developed by the National Library of Medicine using SemRep (Cohen et al., 2009; Kilicoglu, Shin, Fiszman, Rosemblat, & Rindflesch, 2012; *SemMedDB*, 2017). PSI uses random projections and reversible vector transformations to derive distributed concept vector representations from SemMedDB, mediating efficient but approximate search, retrieval, and inference. The higher the dimensionality that is used, the better the recall and precision of the model (with a trade-off of computational efficiency). When searching for the missing argument of a predicate-argument pair, concepts that fill this role most frequently will be retrieved first, analogous to the ranking of results in search engines. In the current work, PSI is used to facilitate rapid retrieval and rank ordering of concepts related to other concepts through particular predicates.

3.3.3.1 Querying PSI Spaces with Discovery Patterns to Identify Potential Confounders

The Semantic Vectors package (*Semantic Vectors*, n.d.) provides an interface that permits searching PSI spaces for concepts that populate particular predicate discovery patterns, which I used to identify the most strongly associated confounders for each drug/adverse event pair. If the following discovery pattern is used to identify confounding variable candidates: **"**drug **TREATS** confounder; confounder **CAUSES-INV adverse_event**,**"** given acarbose (used to treat diabetes mellitus) and myocardial

infarction as cue terms, “metabolic disorder” will be obtained as one of the results. The order in which these covariates are retrieved reflects their ranked relevance given the distributional semantics of the query terms in the index. Sample confounders for abacavir, an antiretroviral used to treat AIDS in the negative control group for gastrointestinal bleeding, include (by ranked order of relevance): Dieulafoy’s vascular malformation, HIV infections, lipoatrophy, HIV encephalopathy, and angiodysplasia. Further down the list, confounders become less specific: peptic ulcers and diabetes. In the evaluation, I excluded spurious associations (as all vectors in the space are a measurable distance apart) from confounders, making use of a frequency threshold, such that only bridging terms with association strengths 2.5 standard deviations were included. More confounders from the discovery patterns that were used are included below in **Table 2**:

Table 2

Discovery Patterns ("DPs") for Statistical and Causal Modeling

DPs	Confounding Type	Example Confounding Variable Candidates[32] **drug:** allopurinol **adverse event:** liver failure
X CAUSES adverse_event	co-medication, comorbidity	transplantation, embolism
X PREDISPOSES adverse_event	co-medication, comorbidity	transplantation, embolism
drug TREATS X; **X COEXISTS_WITH adverse_event**	comorbidity	pericarditis, gout, kidney failure
drug TREATS X; **X CAUSES adverse_event**	comorbidity	hyperuricemia, gout

Note. Items are in far left and denote the discovery patterns used in the experiments. The first three discovery patterns were used only in the first experiment (Chapter 5), whereas the fourth discovery pattern was used in Chapter 5 ("graph structure") and Chapter 6 ("using simulations to quantify average treatment/causal estimates").

A browser interface called EpiphaNet has been developed for querying PSI-based representation of SemMedDB using a query language with metavariables that specify predicate vectors, elemental vectors, semantic vectors along with binding and

[32] In previous research in causal modeling, Cheng refers to such covariates as "focal sets" (Cheng, n.d.). Greenland et al. refer to confounding variable candidates as "potential confounders" (S. Greenland et al., 1999). Pearl et al. refer to the "focal set" of "potential confounders" as the "causal story" (J. Pearl, Glymour, & Jewell, 2016; J. Pearl & Mackenzie, 2018).

superposition operators (Cohen et al., 2009; Cohen, Schvaneveldt, et al., 2010; Cohen, Whitfield, et al., 2010; Cohen et al., n.d.; Kanerva et al., 2000; Widdows & Cohen, 2015). PSI vector space model of SemMedDB may be queried to identify confounding variable candidates for each drug/adverse event pair with PSI queries that represent discovery patterns. The results of a query through EpiphaNet's (http://www.epiphanet.uth.tmc.edu) web interface are shown below in **Figure 2**:

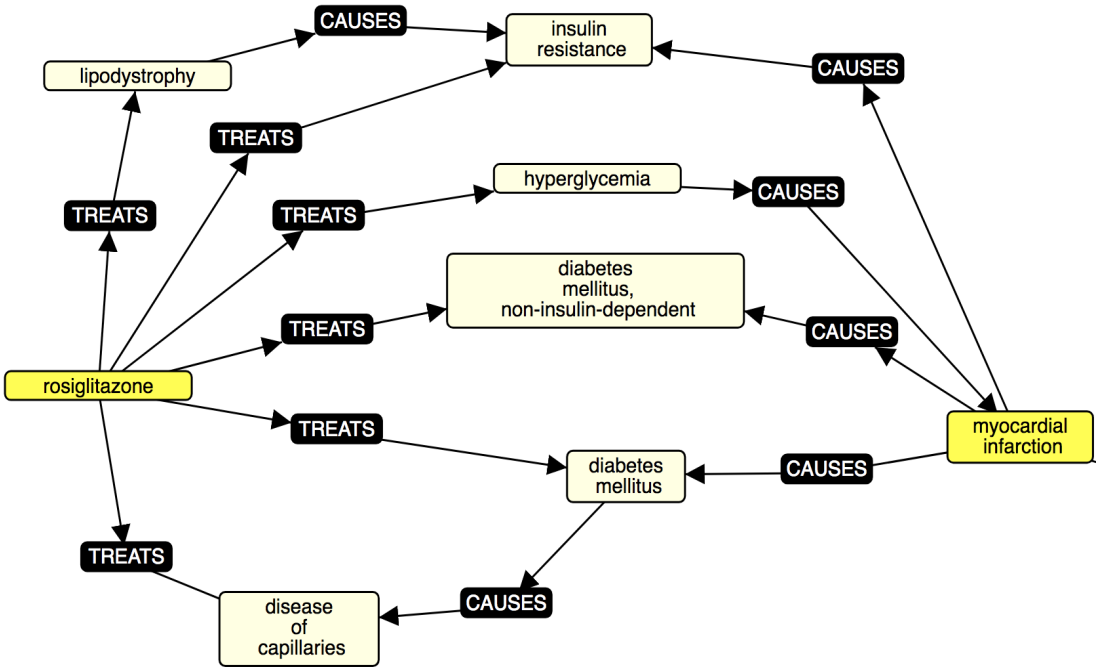


Figure 2. Illustration of EpiphaNet query results. EpiphaNet query results for rosiglitazone and myocardial infarction are depicted. Note that the direction of the arrows for **TREATS** reflects the semantics of that predicate. In causal terms, the edge is oriented in the opposite direction in the sense that the disease **CAUSES** the exposure to the drug.

3.3.3.2 Individual validation of confounder candidates with EHR data

When applying machine learning algorithms, it is vital to perform feature selection as it will reduce the number of variables that are used to construct the model (current reference). Having variables that are irrelevant can hinder analysis by increasing the processing time, reducing available resources, and often lowering the accuracy and precision of models through the introduction of multicollinear covariates. The implication of this is that in either statistical or causal models, each literature-derived confounder must meet a series of constraints for inclusion into downstream predictive models. The first constraint entails first determining that it has been measured in the EHR data, as this is a prerequisite to both further validation and predictive modeling. Secondly, each confounding variable candidate must be correlated with at least the outcome (Chapter 4 – initial experiments), if not both the predictor and the outcome (Chapter 5 and Chapter 6). Validation entails first determining that it has been measured in the EHR data. Secondly, each confounding variable candidate must be correlated with at least the outcome (Chapter 4), if not both the predictor and the outcome (Chapters 5 and Chapter 6). The justification for excluding variables that are not correlated with both predictor and outcome is that such covariates fail to meet the criteria as confounders and so to be of potential use in de-confounding observational data.

3.4 Summary and Roadmap for following chapters

In this chapter, I have sketched an outline of the framework that I have developed for identifying confounders using the literature and integrating these into statistical and

causal models. I will further develop and refine these ideas given the course of iterations and refinements in the chapters that follow. Additional details concerning the data and/or methods are specified in those respective sections of the forthcoming chapters. The items enumerated below provide an outline of the general steps that hold across the experiments to be described in detail in the forthcoming chapters.

1. ***Extract, process, and persist observational clinical data*** from the clinical data warehouse (after obtaining approval from the [UTHealth] IRB.

2. ***Construct knowledge base*** of domain knowledge extracted from the literature.

3. ***Apply or research relevant synonym mappings*** for medications and adverse events from a publicly available reference dataset to be used for methodological evaluation.

4. ***Evaluate baseline performance of EHR data*** for pharmacovigilance signal detection using desired method (odds ratio, logistic regression) without confounding adjustment by calculating the area under the curve of the receiver operator characteristic curve using baseline scores for each drug/adverse event pair in the respective reference dataset.

5. ***Query the knowledge base for contextually relevant confounders*** using each drug/adverse event in the reference dataset as cue terms, given a discovery pattern, and determine if the literature-derived covariate behaves like a "confounder."

6. Construct statistical and/or causal models by incorporating the verified confounders.

7. ***Evaluate confounder adjusted performance of EHR data*** for pharmacovigilance signal detection using desired method (odds ratio, logistic regression) without confounding adjustment by calculating the area under the curve of the receiver operator characteristic curve using adjusted scores (parameter metrics, regression coefficients) for each drug/adverse event pair in the reference dataset.

If the pre-processing procedures may be abstracted out, since these have been completed by the time individual experiments begin, the steps above may be reduced to those illustrated in **Figure 3** below:

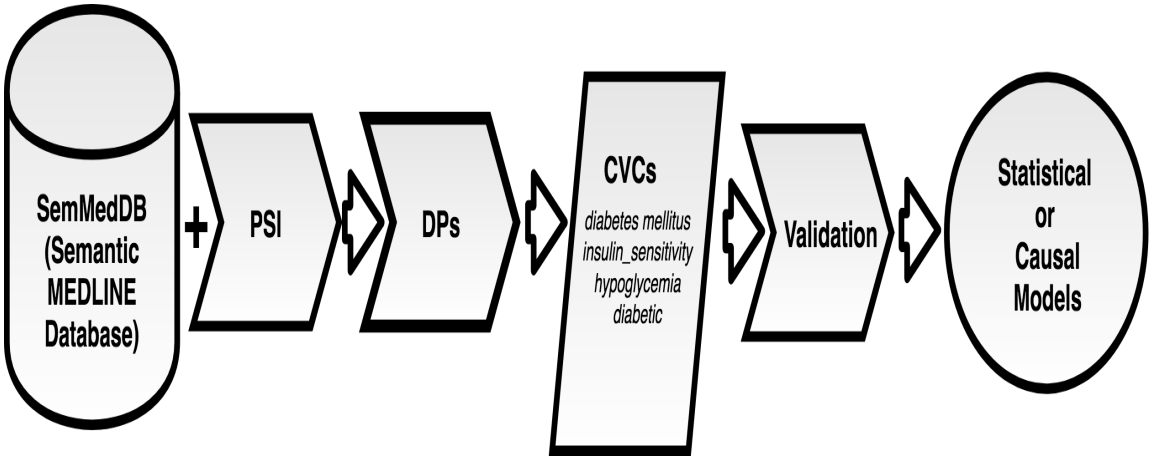


Figure 3. Illustration of LBD-informed causal modeling framework. This illustration represents the study design of the method and experiments. Each experiment can be thought of a "variation on a theme": the particular elements may vary (reference data set, the provenance of clinical data, modeling class [statistical or causal] or algorithm), but the overall workflow remains consistent across the experiments.

Variations on **Figure 3** will be revisited and adapted for the experiments presented in the chapters that follow. Individual elements concerning implementation details will vary (statistical vs. causal modeling, discovery pattern), but the overall workflow will be consistent.

Roadmap for the forthcoming chapters. Chapter 4 describes an experiment wherein I integrated literature-derived confounders into statistical models using step-forward logistic regression and three "discovery patterns". Chapter 5 presents an experiment where I used combinatorial permutation upon sets of literature-derived confounders as a scoring mechanism using a discovery pattern that identifies co-morbidity-type confounders with a "drug **TREATS** x; x **CAUSES** adverse_event" dual predicate discovery pattern. Chapter 6 describes and evaluates a method to obtain estimates of the average treatment effect of the likelihood of medications to cause adverse events, using the same discovery pattern from Chapter 5. Finally, Chapter 7 presents the accomplishments, contributions, limitations, and possible directions for forthcoming work in detail.

Chapter 4: Using the Literature to De-confound Statistical Models

The material in this chapter was presented at AMIA Symposium 2016, Chicago, IL:

Malec, S. A., Wei, P., Xu, H., Bernstam, E. V., Myneni, S., & Cohen, T. (2016). Literature-Based Discovery of Confounding in Observational Clinical Data. AMIA Annual Symposium Proceedings, 2016, 1920–1929.

This chapter constitutes the initial exploratory analysis of LBD-informed predictive modeling. As can be expected from an exploratory analysis, the methods and concomitant results presented are as yet in their infancy. LBD techniques are used to cast a wide net to identify intervening variables. Initially, any sort of intervening variable will do (confounder, mediators, alternate etiologies[33]). These intervening variables must meet at least two criteria in order to be considered for inclusion in statistical models: 1.) they must be mechanistically related to at least the health outcome (adverse event) in the literature through pre-defined relationships ("discovery patterns"); and 2.) they must be correlated with both (predictor and outcome) in the EHR data. The objective of this study was to see which discovery pattern or type of intervening variable would work the best (or at all) for improving the quality of pharmacovigilance signals in EHR data. Critical lessons were learned and

[33] The broadness of this initial motivation does not fit with the exclusive focus on confounders, but at the conclusion establishes this focus, preparing the way for the next two chapters/experiments, where the focus is exclusively on confounders from the outset.

ideas evolved quickly from this initial conception, as shall be seen in the next two chapters.

4.1 Introduction

The hypothesis evaluated in this chapter is that statistical models, adjusted with literature-derived covariates, will more accurately identify causative drug/adverse event relationships than baseline unadjusted models. In this experiment, intervening variable candidates (hereon to be referred to as confounding variable candidates, or confounding variable candidates) from three discovery patterns are integrated into statistical models using forward stepwise logistic regression, and their capacity to de-confound drug/adverse event signal in EHR data (from the UTHealth clinical data repository (*UTHealth BIG*., 2017)) is evaluated using the OMOP reference dataset for pharmacovigilance (Ryan, Schuemie, et al., 2013). EHR data and the reference dataset were pre-processed as described in section 3.3.2 of the previous chapter.

Definitions. A confounding variable influences or biases the magnitude of the correlation between a predictor variable (e.g., drug exposure/treatment) and a response variable (i.e., outcome/adverse event). In the context of creating models for pharmacovigilance, when a confounding relationship exists between a falsely associated drug/adverse event pair and adjustments are made to account for its influence, the association strength for that relationship should be diminished. For example, given a set of observational clinical notes, it is observed that fish oil intake is highly correlated with acute liver injury (ALI). However, after adjusting the model for the presence of known causal agents of ALI, (e.g., acetaminophen, liver cirrhosis, hepatitis c), that correlation

should approach zero. Let us consider another example of an acetaminophen exposure (predictor) and a hepatitis B infection (predictor) where the patient subsequently suffers ALI (response). In this case, each of these predictors, independently, are sufficient preconditions for ALI. As they occur together, these two predictors may confound each other. When the association of either predictor is adjusted in the absence of the other predictor, the association may diminish, but not as dramatically as in the first example. Li et al. introduced a taxonomy of confounding in pharmacovigilance with the following categories: confounding by indication (e.g., preexisting conditions), confounding by comorbidity (e.g., diabetes), and confounding by co-medication (e.g., aspirin) (Y. Li et al., 2014; Talbot & Aronson, 2012).

A mediating variable, by contrast, lies distinctly along the causal discovery pattern between the predictor variable and the response variable themselves and may be neither necessary nor sufficient to cause an adverse event by itself. Mediators may sometimes be thought of as "risk factors" or as aspects of the etiology of the adverse event itself (Richiardi, Bellocco, & Zugna, 2013; Valente, Pelham, Smyth, & MacKinnon, 2017). Examples of mediators include bile duct obstruction for (acute liver injury/failure) ALI or hypertension for myocardial infarction (MI). For a more detailed discussion of mediation, see (Judea Pearl, 2009; Richiardi et al., 2013; Tchetgen Tchetgen & Vanderweele, 2014; T. VanderWeele, 2015; T. J. VanderWeele, 2012).

4.2 Materials and Methods

Knowledge Base. A PSI vector space derived from version 24_32 of SemMedDB (processed with version 1.5 of SemRep), containing 23.9 million citations and 70.4 million semantic predications(*SemMedDB*, n.d.) was used to construct a knowledge base. A knowledge base consists of assertions of an etiological nature derived from the literature. In this case, this knowledge base will be used to identify sets of covariates that may be pertinent to an evaluation of putative drug/adverse event relationships of interest. A 48,000-dimensional binary vector PSI space was built using the Semantic Vectors package (version 5.9) (*Semantic Vectors*, n.d.). Predicates were excluded that indicate negation (e.g., **DOES_NOT_TREAT**), as well as terms ("stop words") with occurrence $\geq$ 500,000.

Derivation of confounding variable candidates from the literature. The discovery patterns used in this chapter's experiment are presented and summarized in Table 3 below. These discovery patterns were identified while studying how domain experts used of the EpiphaNet LBD interface to interpret results of drug/adverse event queries. EpiphaNet would at times generate reasoning discovery patterns that suggest confounding relationships. Note that "drug **TREATS** x; x **COEXISTS_WITH** adverse_event" is referred to as a "double predicate" discovery pattern in that it is composed of two predicates that yield confounding variable candidates that link to both the drug/adverse event cue terms.

Table 3

Discovery Patterns ("DPs") for Statistical Modeling

DPs	Confounding Type	Examples **drug:** allopurinol **adverse event:** liver failure
X CAUSES adverse_event	co-medication, comorbidity	transplantation, embolism
X PREDISPOSES adverse_event	co-medication, comorbidity	transplantation, embolism
drug TREATS X; **X COEXISTS_WITH adverse_event**	comorbidity	pericarditis, gout, kidney failure

Note. Categories in far left in **bold** denote the discovery patterns used in this experiment.

Methods

Preprocessing steps for the data and knowledge base were described previously in Chapter 3. Note that the input consists of document level concept occurrence statistics for drugs, adverse events, and literature-derived covariates for statistical models.

Establish baseline performance by calculating the area under the curve from the receiver operator characteristic[34] (AUROC) curves from ranked-order of coefficients from unadjusted logistic regression models.

[34] AUROC is a popular metric used to summarize the performance of binary classifiers in statistical machine learning, and calculated as area under the Receiver Operating Characteristic curve (from sensitivity/1-specificity).

1. **Query PSI vector** space for confounding variable candidates given each drug/adverse event pair for confounding variable candidates and extract data from the index. Fifty confounding variable candidates were extracted for each pattern per drug/adverse event pair. The constraint for the inclusion of a confounding variable candidate into a statistical model is at least ten co-occurrences given the EHR data for both the drug and the adverse event.
2. **Construct confounding variable candidate-adjusted statistical models** using forward stepwise logistic regression and calculate the aggregated performance statistics with AUROC. The statistical models are constructed in descending order of co-occurrence count between drug, adverse event, and confounding variable candidate concepts (with an intersection threshold of ten) for inclusion in the statistical models. AUROC is calculated using the ground truth in the OMOP reference data set (Ryan, Schuemie, et al., 2013) and the ranked ordering of the regression coefficients from the generalized linear models of the drug/adverse event pairs.

In the course of building models iteratively in step 3 above, when the statistical model achieve a fit with a newly incorporated literature-derived covariate, that covariate is incorporated into subsequent models. Confounding variable candidates that are collinear with the exposure/drug and the adverse event may result in models that fail to converge, or achieve a fit. This process continues for each drug/adverse event pair until confounding variable candidates are exhausted for that discovery pattern. When such models fail to achieve convergence, the offending concept is added to a list of exclusions

for that pair so that it will not be included in subsequent builds, and it is retained for manual investigation for interesting patterns. If none of the models for a drug/adverse event pair converge using confounding variable candidates, then the score from the unadjusted, or baseline, logistic regression coefficient is used to calculate AUROC. The data analysis methods used in the current experiment are as follows both in the text and in **Figure 4** below:

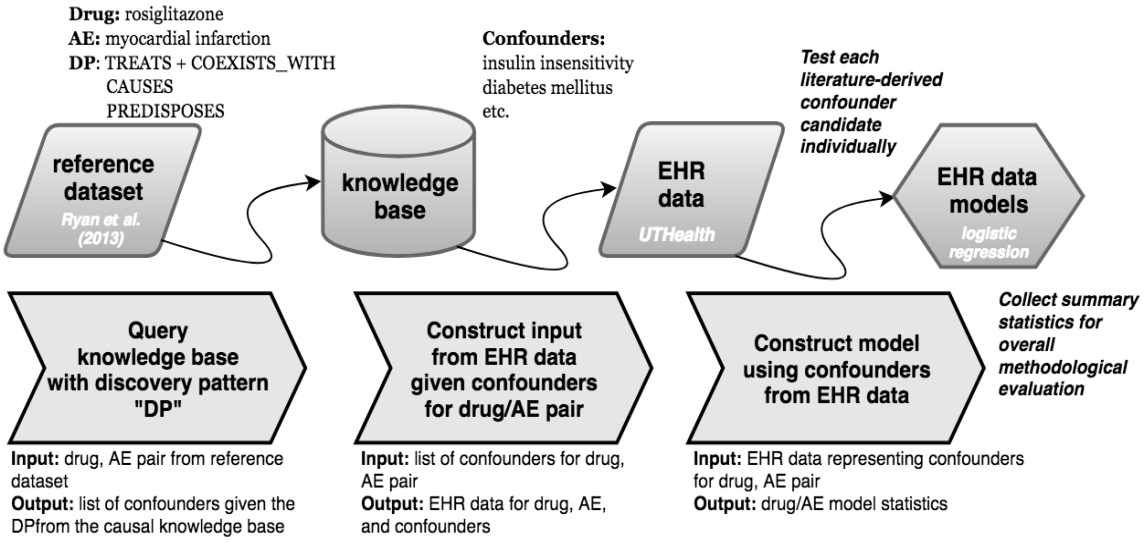


Figure 4. Illustration of LBD framework for statistical modeling. EHR data derived from the UTHealth clinical data warehouse was used as primary input and to test the literature-derived covariates. Literature-derived confounders from three discovery patterns will be included in statistical models of EHR data using forward stepwise logistic regression: **CAUSES**, **PREDISPOSES**, and **"drug TREATS x; x COEXISTS_WITH adverse_event"**.

4.3 Results

Table 4

Results for LBD-informed Statistical Modeling

		Complete Results			Constrained Results		
AEs	**DPs**	**Counts +/-**	**Baseline**	**Adjusted**	**Counts +/-**	**Baseline**	**Adjusted**
AKI	Caus			**0.5853**	11/13	0.6573	**0.6643**
	Pred	24 / 64	0.5547	0.584	NA	NA	NA
	Tcoe			**0.6126**	NA	0.6972	0.6125
ALI	Caus			0.515	44/14	0.5536	**0.5568**
	Pred	81/37	0.4957	**0.5297**	50/24	0.4992	0.4825
	Tcoe			0.492	58/25	0.509	**0.5303**
AMI	Caus			**0.5158**	24/41	0.6026	**0.6148**
	Pred	36/66	0.5112	**0.5196**	30/47	0.5319	**0.5574**
	Tcoe			0.5032	27/45	0.5687	**0.5835**
GIB	Caus			**0.6418**	20/48	0.6073	0.5792
	Pred	24/67	0.5643	**0.699**	20/49	0.5949	**0.6571**
	Tcoe			**0.7189**	20/50	0.5964	**0.69**

Note. Categories in far left denote adverse events adverse events (AE) in the OMOP reference data set. The next column denote discovery patterns. This table presents the baseline and adjusted AUROCs that were calculated from the ranked order of logistic regression models from each adverse event and discovery pattern combination. Caus = ”x **CAUSES** AE,” Pred = ”x **PREDISPOSES** AE,” Tcoe = ”drug **TREATS** x; x **COEXISTS_WITH** AE.” Counts = number of positive/negative examples. AUROCs in **bold** indicate that an adjusted model drug coefficient is higher than baseline.

There are two groupings of result data in **Table 4**, labeled *Complete Results* and *Constrained Results*. *Complete Results* indicates that the AUROCs have been calculated from the full data set without imposing any additional criteria.

In the *Constrained Results*, the following criteria were applied to calculate performance metrics values for each field per adverse event/discovery pattern row such that:

1.) All logistic regression models must have converged[35].

2.) The count for drug instances was ≥ 100 in the EHR data.

3.) The count for intersections was ≥ 10 between drug/adverse event pair.

4.) The calculations derive exclusively from cases where confounding variable candidates were included in the logistic regression models.

As a result, the count of positive and negative controls for the same adverse event will vary, since confounding variable candidates differ between discovery patterns.

Observations. There were ~55,000 instances in the EHR index for each of ALI, AMI, and GIB. AKI was the outlier with only ~5,000 instances. In the case of AKI for the "x **PREDISPOSES** adverse_event" discovery pattern, no co-occurring confounding variable candidates were identified so that no adjustment could be made. In the case of ALI, one might reason that the set is heavily weighted toward the positive examples, so little gain is to be had by adjusting with confounding variable candidates.

[35] Some covariates are colinear with predictor, outcome, or both causing perfect or quasi-perfect separation. This results in failure of convergence.

4.4 Discussion

Analysis of Constrained Results. In the Constrained Results, for models that include confounding variable candidate adjustments, performance improved in 8 of eleven cases. The "drug **TREATS** x; x **COEXISTS_WITH** adverse_event" discovery pattern improved performance in three or four cases, while the same can be said in two of four "x **PREDISPOSES** adverse_event." In only one case, AKI there was an improvement using the "x **CAUSES** adverse_event" discovery pattern[36].

Li et al. recently developed a data-driven confounding variable discovery method using a propensity score method (PSM) (Jackson et al., 2017; Rosenbaum & Rubin, Donald, 1983; Tatonetti et al., 2012) to identify factors associated with the treatment and another technique to identify factors associated with the outcomes (R. Harpaz et al., 2012) – in this case, rhabdomyolysis and pancreatitis. Overlaps between these two sets (of the "comorbidity" subtype in pharmacovigilance) were collected and processed using penalized regularization methods, specifically lasso regression (Y. Li et al., 2014). The results are not strictly comparable with this study as the input data and drug/adverse event pairs are different, but the best improvement was with the method above compared with only the PSM or risk factor scores, and when there were a sufficient number of exposures: at least 100. In another study in purely EHR-based pharmacovigilance, Li

[36] The "x **CAUSES** adverse_event" discovery pattern is an intriguing start, however. It will be re-deployed as the more reliable half of a dual predicate discovery pattern and figure prominently in discussions of later chapters.

combined FAERS with EHR data to obtain a 0.22 improvement in AUROC – from 0.51 to 0.73 (Y. Li, 2015).

Error Analysis. While it makes sense to use the single predicate discovery patterns that imply causation or risk factors, e.g., "x **CAUSES** adverse_event," "x **PREDISPOSES** adverse_event," respectively, concerning their ostensible causal association for adjusting negative controls, the results of my analysis did not support this intuition. Such discovery patterns uncover concepts that exist in the gray area between mediators, risk factors (Pearl's "indirect effects" predictors), and concepts which could manifest confounding effects, e.g., smoking with respect to a positive control drug and AMI (Judea Pearl, 2009). For example, the following cases of mediator-like confounding variable candidates from this discovery pattern were identified: stenosis, obstruction, thrombosis, and thrombus. Such concepts are suggestive of mediating concepts that relate to the causal mechanisms for AMI. Since mediators tend to be collinear with response variables, inclusion of such confounding variable candidates may be detrimental to performance (although means to correct for bias from such variables have been developed and have been explored at length elsewhere) (Aalen, Roysland, Gran, & Ledergerber, 2012; Richiardi et al., 2013; T. J. VanderWeele, 2012; Vanderweele, Vansteelandt, & Robins, 2014).

Wordclouds. I have generated word clouds below in **Figure 5**. The purpose of these wordclouds is to present examples of the types of concepts that were identified by LBD. The wordclouds were generated using R's wordcloud library (Fellows, 2014). The wordclouds are generated from the processing of EHR data. The size of a word in each

wordcloud indicates the prevalence of the concept. The first two wordclouds represent the confounding variable candidates of ALI, and GIB with the "x **CAUSES** adverse_event" discovery pattern. The third and fourth show the confounding variable candidates for the AMI groups using the "x **CAUSES** adverse_event" discovery pattern. The fifth word cloud represents the confounding variable candidates that were *excluded* when building the logistic models for GIB using "drug **TREATS** x; x **COEXISTS_WITH** adverse_event" discovery pattern. These literature-derived confounding variable candidates were identified in the literature, but failed to meet the inclusion criteria in the EHR data. The reader will notice that confounding variable candidates from single predicate discovery patterns (e.g., "x **CAUSES** adverse_event") that were associated only with the adverse event were of both the comorbidity and co-medication confounding subtypes, whereas confounding variable candidates from "drug **TREATS** x; x **COEXISTS_WITH** adverse_event" are constituted exclusively by comorbidities. The fourth word cloud (bottom row, left) is interesting in that the most prevalent confounding variable candidates are comorbidities and likely mediators of myocardial infarction.

Figure 5. Illustration of Wordclouds from statistical modeling discovery patterns. These word clouds represent confounding variable candidates for adverse events and discovery patterns with font size proportional to aggregated term frequency (across all drugs for a group [positive or negative control] of a given an adverse event and a discovery pattern) in the EHR index.

One perplexing confounding variable candidate from "drug **TREATS** x; x **COEXISTS_WITH** adverse_event" discovery pattern for acute myocardial infarction is fibroid tumor. Fibroid tumors of the gastrointestinal tract are relatively rare and usually appear on the uterus. However, review of the predication database suggested that at times anti-inflammatory agents may occur in **TREATS** relationships with fibroid tumors, as they are used to control the pain associated with this condition. Though the "drug **TREATS** x; x **COEXISTS_WITH** adverse_event" discovery pattern is intended to retrieve terms that are associated with both the drug and the adverse event, the underlying implementation involves vector superposition. Although I would anticipate terms that are bilaterally connected being retrieved first, terms that are unilaterally connected may still meet the threshold. Such spurious confounding variable candidates could be eliminated by using a higher threshold of associational strength than the 2.5 standard deviation level and by making a bilateral connection a prerequisite for retrieval[37].

Conclusion. With the aim of surmounting the obstacle of confounding, a phenomenon which diminishes the validity of information that can be extracted from observational data, I have proposed a scalable and computationally inexpensive LBD-based confounding variable discovery method. The evaluation shows that when there is sufficient support above random for an adverse event, i.e., AUROC 0.6, statistical models

[37] Recall that there is a multi-stage filtration of confounding variable candidates. First the confounding variable candidates must meet the distributional requirements in the PSI space. In vector space everything is connected, so concepts that are retrieved must score above 2.5 standard deviations as noted to be retrieved. Secondly, the confounding variable candidate that is identified with a score above the distributional threshold must meet additional requirements: they must be measured/recorded and those values must overlap at minimum ten times with both the predictor and outcome variables. Finally, confounding variable candidates cannot be colinear with the predictor and outcome variables as this will prevent the statistical models from converging ("perfect separation").

that incorporate adjustments for the influence of dual-predicate discovery pattern-derived confounding variable candidates exhibit modest (0.05 AUROC or higher) performance gains for the task of re-identifying drug/adverse event pairs from observational clinical data.

This study had several limitations. These include:

1.) The discovery patterns that were used may not have identified "true" confounders or mutual causes, but "alternate etiologies" and "mediators."
    a. An "*alternate etiology*" is another causal explanation for why something occurred, and it may be independent of the explanatory variable. It may not be useful for disentangling spurious associations due to statistical correlation.
    b. Like confounders, mediators are another type of intervening variable between a predictor variable **p** and an outcome of interest **o**. However, whereas a confounder **C** will betray this pattern: $\mathbf{p} \Leftarrow \mathbf{C} \Rightarrow \mathbf{o}$, a mediator **M** will have this pattern: $\mathbf{p} \Rightarrow \mathbf{M} \Rightarrow \mathbf{o}$.
2.) The best performing discovery pattern was a dual predicate discovery pattern. While the **TREATS** predicate offers strong evidence from the literature that the co-morbidity bridging term may be responsible for causing or increasing the likelihood of the exposure, the **COEXISTS_WITH** predicate does not provide a strong indication that its object is viable mechanistically, but it is associated by non-causal means. While such confounders may not be true

confounders, they do offer clues as to how strong confounding variable candidates might be identified.

The main lesson learned from this study is that dual predicate discovery patterns are powerful tools for identifying useful confounding variable candidates. In thinking about how the two types of intervening variables relate to each other (mediator and confounder), confounders of the mutual cause type were more easily understood and they performed better in the aggregate. This indicated the direction to proceed.

The next two studies expand upon the LBD component of this initial analysis and explore a more precise definition of confounder that was only hinted at in this chapter. However, mediators will make a return (albeit in refined form) in the discussion of future research in the final chapter, as mediators turn out to be vital components of future directions of research in LBD-informed modeling of EHR data.

The research that follows attempts to transcend associational methods. In the next chapter, an experiment is presented that explores ways in which asymmetric direction of influence inherent in causal predications can be leveraged in causal directed acyclic graphs instantiated with EHR data.

Chapter 5: Ars Combinatoria with Focal Sets of Potential Confounders

The material in this chapter was presented at the DMMI Workshop, part of the AMIA Symposium 2017 held in Washington D.C.:

Malec, S., Gottlieb, A., Bernstam, E., & Cohen, T. (2018). Using the Literature to Construct Causal Models for Pharmacovigilance. https://doi.org/10.29007/3rfr

In the previous chapter, the interpretation of an intervening variable was left quite open: it could either be a mediator or a confounder. Moving from "mere association" of logistic regression models, the experiment described below explores intervening variables more precisely. Specifically, the thought behind what the definition is and individual confounders behave or can be expected to behave collectively and individually become more precise and nuanced. This chapter introduces a new method to validate confounding variable candidates using the directionality inherent in the semantic predications, since graph structures with instantiated nodes, i.e., populated with values from the EHR data, must be compatible with those data.

5.1 Introduction

There can be said to be two distinct but related types of information encoded in causal graphs: *qualitative*[38] and *quantitative*. The *qualitative* aspect is more apparent: this may be readily ascertained by the presence or absence of directed edges, characterized as pointing in one or in the opposite direction. If a causal edge is present, influence will flow in the direction in which that edge is pointing and produce and increase (or prevent and decrease) the likelihood of the entity that is on the other side to occur. The other aspect is quantitative – this is the problem of estimating the magnitude of the strength of that edge, i.e., the "structural coefficient."

In the experiment described in this chapter, the focus is on the qualitative detection of edges in the graph topology learned by causal discovery algorithms and imputed by causal propositions (predications) mapped from the literature. The operative hypothesis is that the presence of sets of literature-derived confounding variables (hereon to be referred to as "*focal sets*" of confounding variable candidates after Cheng (Cheng, n.d.; Cheng & Novick, 1990)) will diminish the detection of spurious drug/adverse event dependencies as encoded in graph structure in the aggregate.

However, since an AUROC cannot be calculated with binary values (the qualitative aspect of graphs mean that an edge is either absent or present), a means to

[38] Qualities are conceived as discrete, binary-valued: one either possesses an attribute or one does not, whereas one thinks of quantitative values as being continuous. At the stage of causal modeling that this experiment was drawn from, estimating the strength of influence between variables in causal models was not yet technically feasible. The experiment presented here is the result of a compromise between what I knew how to do and what I desired to accomplish.

derive continuous values was devised by using the literature and EHR to identify a focal set and using all unique combinations of that set as input to causal discovery algorithms. By dividing the number of resulting causal edges detected by the number of permutations, a metric with a continuous value was attained with which to calculate the aggregate performance of the causal discovery algorithm using AUROC. AUROC provides a means to weigh sensitivity (or recall: the ability to identify True Positives) against specificity, the capacity of a classifier to guard against the identification of spurious signals as positive. The evaluation metrics and procedure will be discussed in greater detail below.

5.2 Materials and Methods

Data Sources. To evaluate this hypothesis, the same pre-processed EHR data and reference dataset were used as in Chapter 4 (Ryan, Schuemie, et al., 2013; *UTHealth BIG*., 2017).

Discovery Pattern for Confounding Variable Discovery. Semantic Vectors provides an interface that permits searching PSI spaces for concepts that populate particular predicate discovery patterns, which I used to identify the most strongly associated confounding variable candidates for each drug/adverse event pair. I used the following discovery pattern to identify possible confounding variable candidates: **"**drug **TREATS** confounder; confounder **CAUSES adverse_event**.**"** For example, given rosiglitazone and gastrointestinal bleeding, "diabetes mellitus, non-insulin-dependent" was a result. This discovery pattern was chosen for its ability to identify co-morbidity

type "true"/"proper" confounders[39]; that is, mutual causes that increase the co-occurrence of both the explanatory and the outcome variables (Judea Pearl, 2009).

TETRAD and FGeS. TETRAD is an open source causal modeling and discovery toolkit written in Java that has been in continuous development at Carnegie Mellon University since the early 90s **(Scheines et al., 1998).** Depending on one's choice of algorithm, the input may be discrete, continuous, or mixed. The discrete version of the Fast Greedy Equivalence Search (FGeS) included with TETRAD with default parameters **(Ogarrio et al., 2016)** was used to infer graph structures from the EHR data.

FGeS recursively adds and then subtracts directed edges between nodes until the Bayesian Information Criterion[40] (*BIC*) is minimized. The output consists of a family of score-equivalent graphs which encode plausible dependency relationships given these data such that the orientation of the directed edges could not be determined from the data alone, i.e., graphs encoded by this structure have the same BIC score (Heckerman et al., 1995, 1995). However, background knowledge can be used to orient these edges, as causal predicates have inherent directionality. The resulting graph structure of a literature-derived confounding variable candidate should have a graph within this equivalence class with directed edges to both the drug and the adverse event ("confounder inclusion criterion").

---

[39] Proper confounder as per the Greenland et al. definition (S. Greenland et al., 1999; Weinberg, 1992).

[40] For continuous Gaussian data: $\mathbf{BIC} = \mathbf{n}\ \mathbf{ln}(\hat{\boldsymbol{\sigma}}_{\mathbf{e}}^{\mathbf{2}}) + \mathbf{k}\ \mathbf{ln}\ (\hat{\mathbf{L}})$ where $\hat{\boldsymbol{L}}$ = the argmax of the likelihood function, $\boldsymbol{x}$ = data, $\boldsymbol{n}$ = number of observations in $\boldsymbol{n}$, $\boldsymbol{k}$ = number of parameters {Citation}.

Combinatory Expansion of Confounding Variable Candidates. If I have a set of three confounding variable candidates, denoted **{A, B, C}**, this will result in seven unique combinations of focal sets: **A**, **B**, **C**, **AB**, **BC**, **AC**, and **ABC** (**AB** and **BA** are equivalents). At five confounding variable candidates per focal set, there are thirty-one unique combinations. I evaluated these because it is not known which combination of confounding variable candidates, if any, would cause spurious drug/adverse event graph dependencies to vanish.

Analysis of Observational Clinical Data. The steps were as follows:

1. **Confounding variable candidate identification.** Query PSI vector space for confounding variable candidates given each drug/adverse event pair for confounding variable candidates and extracted EHR data.
2. **Confounding variable candidate validation.** Use TETRAD/FGeS to construct and validate confounding variable candidates individually in ranked order of their relatedness to the PSI query, by testing each individually for directed edges to the drug/adverse event pair, and halting when five of the top-ranked confounding variable candidates for each drug/adverse event pair had been validated.
3. **Confounding variable candidate permutation.** Use combinatorial expansion to permute all combinations of confounders.
4. **Causal discovery.** For each of the confounding variable candidate permutations, use TETRAD/FGeS to determine whether or not a causal edge from predictor (drug) to outcome (adverse event) variable is present in the graph with minimal BIC.

5. **Aggregation.** Calculate the proportion of confounding variable candidate permutations for which a causal edge is present, and evaluate aggregate performance using AUROC from the ranked order drug/adverse event pairs per this proportion (such that those pairs with the highest proportion of causal edges across permutations will be most highly ranked).

Evaluation Procedure. To calculate and compare performance, I calculated the Area Under the Curve of the Receiver Operating Characteristic curve (AUROC), which is widely used to evaluate the performance of classifiers against a ground truth of positive and negative controls, based on the ranked order of a continuous estimate of the strength of predicated relationships. AUROC treats both Types I ("false negative") and Type II ("false positive") errors symmetrically. For baseline scores, I used correlation coefficients from logistic regression models without the incorporation of literature-derived confounding variables.

Evaluation Metrics. Causal model scores are calculated from the number of directed edges from the drug to the adverse event divided by the total number of permutations. The hypothesis is that the proportion of directed edges of positive cases in the reference dataset from the drug to their respective adverse events will be higher for the group of positive control drug/adverse event pairs than for the negative pairs in the reference dataset. In other words, the current study tests whether the associations from co-occurrence in the clinical data of a drug/adverse event pair are diminished conditional on sets of other concepts (i.e., validated confounding variable candidate permutations). **Figure 6** demonstrates how the scoring metric is applied:

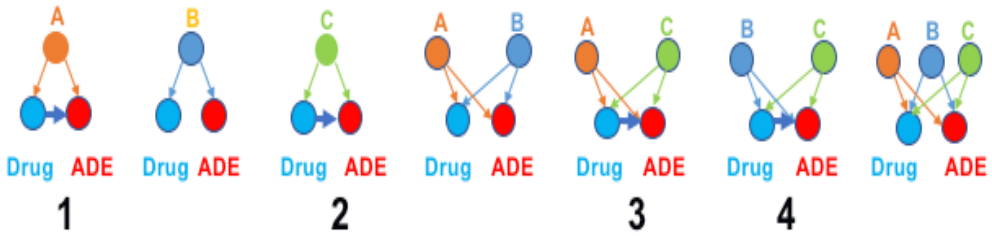


Figure 6. Illustration of Scoring Metric for Focal Set Permutations.

5.3 Results

Table 5

Results from Focal Set Permutation Experiment

Adverse events	Total	Baseline	Causal Models
AKI (Acute Kidney Injury)	24/64	0.5547	**0.6598**
ALI (Acute Liver Injury)	81/37	0.4957	**0.5449**
AMI (Acute Myocardial Infarction)	35/64	0.4946	**0.56**
GIB (Gastrointestinal Bleeding)	24/67	0.5643	**0.6912**
Total	164/232	0.504	**0.5813**

Note. This table presents the AUROCs as calculated from the aggregated logistic regression coefficients (Baseline) and the fraction of directed edges from the causal graphs. Pairs = number of test/control drug/adverse events from the Ryan reference dataset. AUROCs in bold indicate that the causal models outperformed the baseline models for that phenotype.

5.4 Discussion

Analysis of Results. 1915 out of 2124 total tested confounding variable candidates (for 399 drug/adverse event pairs each with five confounders) were both present in the clinical notes and passed validation, so the confounding variable candidate yield rate was 90%, indicating that LBD can identify confounders in clinical notes. As shown in **Table 5**, the overall aggregate performance boost that approached ~ 0.08 over statistical models confirms our hypothesis that the identification problem of confounding can be partially resolved by using the literature to inform feature selection[41] (an area that I have addressed earlier with using LBD for statistical models) **(Malec et al., 2016)**. The method performed the best when the baseline AUROC for drug safety signal was sufficiently above the level of noise (~ 0.5 AUROC). GIB, followed by AKI, had the best baseline AUROC. By contrast, MI and ALI hardly budged from noise to signal, indicating that the method requires a reliable initial baseline to be effective.

Practical Significance. Better detection methods in pharmacovigilance, if implemented, hold promise for improving public health and safety. For example, enhanced methods of drug/adverse event detection in observational clinical data could facilitate the prioritization of drug/adverse event relationships for critical review. Given the extent of the exposed population and the prevalence of adverse drug events, an improvement of even a few percentage points could have a substantial impact.

[41] Overall results were slightly higher (~ 0.08) when calculated using only pairs where drug occurrence ≥ 100 or 500. 0.75 is a desired level of significance desired typical for clinical applications.

In Chapter 4, LBD methods were used for feature selection of confounders to adjust for plausible confounding with the same set of clinical data. Both single predicate (**CAUSES** and **PREDISPOSES**) and dual predicate (**"drug TREATS x; x COEXISTS_WITH adverse_event"**) discovery patterns were deployed to identify confounding variable candidates. With the single predicate discovery patterns, the influence from the confounder was only exerted explicitly in the literature on the outcome and can be thought of as an alternative etiology. Dual predicate discovery patterns performed the best overall with a modest ~ 0.02 AUROC improvement over unadjusted models. My analysis was that the dual predicate discovery patterns identify confounding variable candidates that influence both predictor and outcome, fulfilling the graphical criteria (Judea Pearl, 2009). In the Chapter 4, I used a different dual predicate discovery pattern, and these results affirm previous observations about the utility of dual predicate discovery patterns for identifying confounders with the bonus that causal models with validated "true" confounders improved upon the performance of adjusted dual predicate statistical models. I reduced the False Discovery Rate (*FD*R), where $\mathbf{FDR} = \frac{\mathbf{True\ Positive}}{\mathbf{True\ Positive} + \mathbf{False\ Positive}}$, from 0.5 to 0.38 with causal models. Although the performance increase is substantive and better in general adjusted standard statistical models (for purely EHR-based pharmacovigilance), it does not approach the performance obtained in the work of Li et al. with extra-EHR data sources, where adjusted EHR statistics with adjusted FAERS performance improved upon adjusted FAERS and EHR performance alone **(Y. Li, Ryan, Wei, & Friedman, 2015a)**.

AUROCs improved across all adverse events with this chapter's method in contrast to the results from Chapter 4. In addition, the improvements were more considerable overall – an 0.08 improvement in the aggregate. Starting with similar baseline scores, Li was able to increase the AUROC considerably by integrating FAERS data (Y. Li, 2015). This indicates that a future direction of research should be in synthesizing data-driven confounding variable discovery methods with literature-derived covariates to see how this affects performance. Note that the discovery pattern of **"**drug **TREATS** confounder; confounder **CAUSES adverse_event"** only identifies comorbidities, while the discovery pattern in Chapter 4 identifies both comorbidities and co-medications.

Limitations of (and Lessons from) the Current Approach. One limitation is that my search for confounders was relatively shallow, having commenced confounder permutations after reaching five validated confounding variable candidates. Computational demand scales with the number of confounders. Five validated confounding variable candidates result in thirty-one permutations per drug/adverse event pair. Increasing to ten confounding variable candidates would leave 1032 permutations - which can take twelve hours to run on a Linux workstation with 64GB RAM and eight Xeon CPUs. I chose five confounding variable candidates because I could collect results for all drug/adverse event pairs for a single adverse event within a reasonable amount of time (7-8 hours overnight). There may be some theoretical justification for using a limited number of confounders as it is possible to "*overcontrol*" causal associations (Elwert & Winship, 2014; S. Greenland et al., 1999). For instance, increasing the number

of confounders to ten had the effect of reducing recall (recovering causal edges in the cases/True Positives), while increasing specificity (decreasing the number of False Positives).

An additional limitation arises from the available EHR data, which may not have a sufficient number of drug/adverse event co-occurrences, as the performance from analyses of FAERS data is usually better than results from any EHR data source **(Y. Li et al., 2015a)**.

One perplexing problem arose from three drug/adverse event pairs for myocardial infarction for which the proposed method could not identify any confounders (these are not included in **Table 5**). I suspect that confounders for myocardial infarction identified by my method, e.g., hypertension, coronary arteriosclerosis, metabolic syndrome, could have helped if incorporated into these models. Note that the discovery pattern limits result sets of potential confounders to comorbidities, although co-medications (for example, aspirin and acetaminophen for gastrointestinal bleeding and liver failure, respectively) often make exemplary confounder candidates, so there remains the question of the optimal mixture of confounder types. These factors (along with SemRep's low recall of ~ 0.64) may have impacted my system's performance by missing potential confounders **(Kilicoglu, Fiszman, Rosemblat, Marimpietri, & Rindflesch, 2010)**. Another consideration is that reference data sets, however essential to the scientific enterprise, may not be entirely accurate, as knowledge about drugs and their side-effects accumulates (Hauben, Aronson, & Ferner, 2016).

The implicit assumption in the calculating of the score of each drug/adverse event pair is that each permutation of the set of confounders is equally valid. This simplifying assumption, while likely invalid, was nonetheless useful as overall performance eclipsed that of the statistical models in Chapter 4. This could be addressed by using an information theoretic metric to encapsulate data/model goodness of fit, e.g., Bayesian Information Criterion, for each permutation of a confounder set, where each permutation could be assigned a weight.

The next chapter counters with a more principled approach to causal modeling, while building on the insights and essential lessons of what worked for LBD predictive modeling: respectively, dual predicate discovery patterns to identify focal sets of confounding variable candidates (Chapter 4) and the exploitation of the directionality inherent in causal predicates (Chapter 5). The experiment described in Chapter 6 infers the data generating process gleaned from the EHR data, whereby such generative causal models can be used to simulate the effects of interventions and to quantify average treatment effects.

Chapter 6: Quantifying the Average Treatment Effect from a Mutilated DAG

... [T]o explain an historical event is to find the reasons for its occurrence, and to do that is to re-enact the circumstances and states of mind of the actual agents whose actions brought about the event in question. The historian gives 'the' reasons for that event when he gives 'their' reasons for it.

Dov Gabbay, 2006, pg. 94

In the experiments described by the previous two chapters, the struggle was to first identify what species of intervening variable would be most helpful for the task of amplifying pharmacovigilance signals in EHR data. Dual predicate discovery patterns were found to be the most effective. These reasoning pathways proved effective for identifying not only concepts that were associated with both the predictor and outcome variables, but mutual causes of them both, i.e., the operative definition of confounder. This chapter presents an experiment to explore the frontiers of what such precisely defined confounders can do[42].

A causal model by definition describes a data generating process[43]. In this study, data generating processes are functional causal DAGs. As in the previous chapter, expert knowledge derived from the literature is directly mapped as a graph prior and instantiated with observational data derived from EHR. The novel feature of the current experiment is that the causal DAG generate simulated data from a mutilated posterior model with which to calculate average treatment effects, or ATE.

---

[42] Owing to their ability to answer counterfactual queries, Koller distinguishes the class of causal models discussed in this chapter as being "functional causal models" in contrast to plain causal models (Koller et al., 2009). In this chapter, techniques are described which can not merely describe conditional probability distributions with causal assumptions, but answer counterfactual queries.

[43] For more on data generating processes, see (Hendry & Doornik, 2014).

The methods evaluated in this chapter are the state-of-the-art refinement of my methodology and constitute a principled method for the incorporation of literature-derived confounding variable candidates into causal models.

6.1 Introduction

This study addresses the weaknesses of the *ad-hoc* permutation-based procedure described and evaluated in the previous chapter and documents the current stage of evolution of the methods developed for this dissertation. Specifically, this study assesses the ability of LBD-informed causal inference models to quantify the effect of the explanatory variable (drug exposure) on an outcome of interest (an adverse event) using observational clinical data derived from EHR. The identification of literature-derived confounders can facilitate the process of *de-confounding*, (or "screening off" spurious correlations from descriptive statistics) in observational data, by imposing constraints from *a priori* domain knowledge on the topology of the causal graph. By incorporating confounders into causal models, one can perform experiments/interventions with the resulting data generating model (J. Pearl et al., 2016; Judea Pearl, 2009; T. J. VanderWeele & Shpitser, 2013).

6.2 Technical Background

The following subsections provide an overview of the essential mathematical machinery and concepts required estimating ATE. These core components can be summarized as follows:

1.) *Graph factorizations and conditional independence* (section 6.2.1).

2.) *Counterfactual queries via graph mutilation* (section 6.2.2).

3.) *Mutilated graph simulations for* **ATE(drug ⇒ adverse_event)** (section 6.2.3).

Since the graphical approach for representing causal relationships (wherein nodes represent variables and directed edges represent causal relationships [or "dependencies" between variables]) is the most intuitive to grasp, the next section starts with graphs [44].

6.2.1 Graph factorizations and conditional independence

A Bayesian network consists of a network structure as directed edges that encode intra-node (variable) dependencies and a the conditional distribution of each nodes' adjacencies. When observed values of variables represented are populated, the network is said to be instantiated. If the joint density distribution inferred from these values are reflected in the anticipated network structure, that structure is said to be *compatible* with its *instantiation (Darwiche, 2009; Koller et al., 2009)*.

Consider a data set **A** that consists of a set of random variables **X** and is described by a directed acyclic graph **G**. The (causal) Bayesian network **B** = **(G, X)** and **Θ** denote the parameters of the global distribution of **X**, such that **Θ** is *i.i.d*. with **X**, so that **B** = **(G, Θ)** (and **Θ** can denote the sufficient statistics of appropriate marginal and joint distributions given **A**, e.g., Bernoulli/binomial/multinomial. if discrete, Gaussian/Poisson. if continuous, and so on) (Darwiche, 2009). The structure and parameter learning process

[44] Beyond being merely easy for humans to interpret, graphs have convenient mathematical properties that are useful for summarizing dependency relationships a set of random variables. In short, dependency relationships may be easily translated into either graphical or factorization form. These properties are summarized in great detail elsewhere (Koller et al., 2009; J. Pearl, 2014).

(noted above) is decomposable into the following factors (Koller et al., 2009; J. Pearl, 2014; M. Scutari & Denis, 2014; Marco Scutari, 2009) [**Equation 1**]:

$$P(B|A) = P(G,\Theta|A) = P(G|A)\,P(\Theta|G,A).$$

P(**G**|**A**), the first factor on the right side of **Equation 1**, denotes the structure (topology) of the graph and can be further decomposed as follows [**Equation 2**]:

$$P(G|A) = P(G)\,\mathbf{argmax}\int P(A|G,\Theta)\,P(\Theta|G)\,d\Theta$$ [45]

P(**G**) represents the "subgraph"[46] of the Bayesian network derived from domain (expert) knowledge as a *structural (graph) prior*. **G** encodes conditional probabilities of the random variables in **X** from data **A** parameterized by **Θ**.

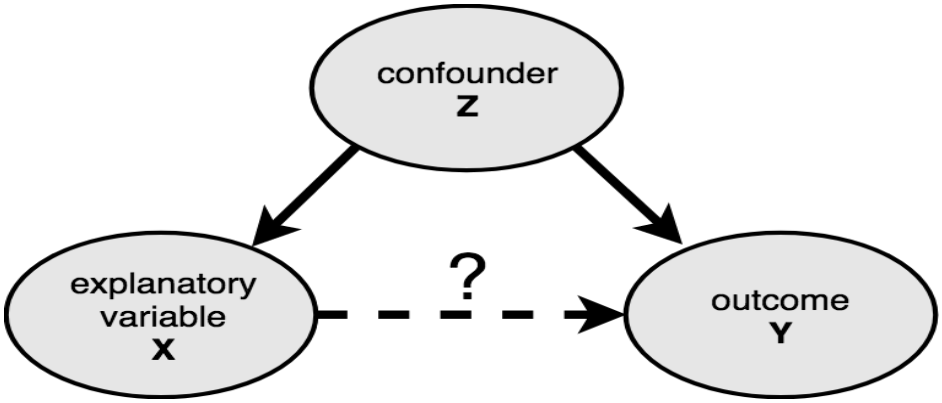


Figure 7. Illustration of a "true confounder" with a set of random variables {x, y, z}.

[45] Translated into plain English, this expression says: "find the values of parameter Θ that *maximize* the area under the curve (are the most likely) for graph G dataset A."

[46] The graph "subgraph" refers to how there may be unanticipated dependencies given the actual input data that were not recalled in the literature. In LBD research, this refers to the subset of dependencies identified in the literature (Cameron et al., 2013, 2015).

6.2.2 Counterfactual queries via graph mutilation.

Given a set of random variables **{x, y, z}**, where z is a confounder that influences both **x** and **y**, assume that we do not know the relationship between **x** and **y**. From *a priori* knowledge (or experience[47]) we know that the value of **z** determines **x** and **y**. Variable **z** is said to be their common "*parent,*" while **x** and **y** are **z**'s "*children*." Since **x** and **y** have a parent in common ("a mutual cause"), their values are statistically correlated. A directed ("causal") edge is drawn from **x** to **y** to represent the relationship about which we wish to know more. The task is to determine the likelihood of **x** to influence or determine the value of **y**, even in the presence of mutual cause **z** that influences them both as depicted in **Figure 8** below. Graph **G** is depicted below and the relationships between its variables encodes their dependencies (**G'** will be explained later).

[47] In the case of this dissertation: the source of *a priori* knowledge is the literature.

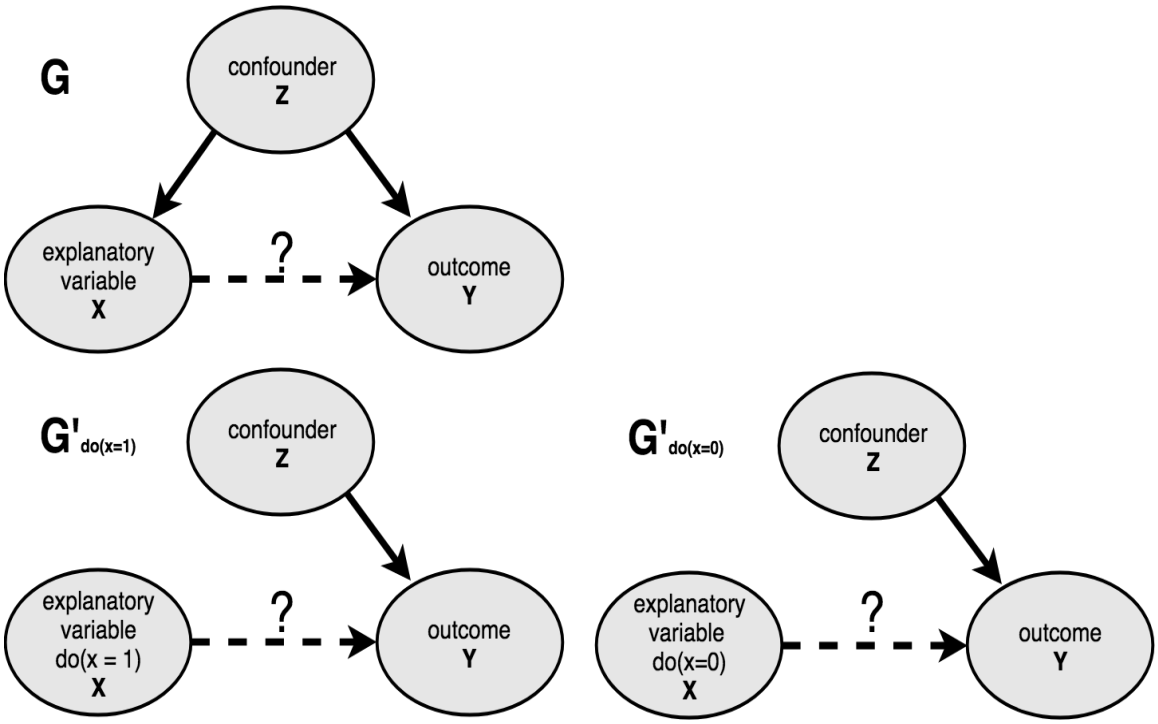


Figure 8. Illustration of graph mutilation/manipulation. To make the connection clear, in calculating the Average Treatment Effect, one is "severing" the relationship between **z** and **x**, cutting off the "*back-door path*".

G factorizes as the following joint probability distribution [**Equation 3**]:

$$\boldsymbol{P}\,\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\, = \boldsymbol{P}\,\boldsymbol{z}\, \boldsymbol{P}\,\boldsymbol{x}|\boldsymbol{z}\, \boldsymbol{P}\,\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z}\,$$

Since the task is to eliminate the noise emitted by the effect of z on x [**Equation 4**]:

$$\boldsymbol{P}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\,\boldsymbol{P}(\boldsymbol{x}|\boldsymbol{z})\, = \boldsymbol{P}\,\boldsymbol{z}\, \boldsymbol{P}\,\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z}\,$$

Note that the factorization of graph **G'** is represented in the factorization on the right-hand side of **Equation 4**. Given the dependencies depicted in the graph and its respective factorization properties, a question one might ask is: how might we estimate the magnitude of the relationship between **x** and **y**? To do this, the value of **x** can be fixed

such that it is no longer susceptible to influence from **z**. This "*fixing*" or "*setting constant*" the values of the explanatory variable **x** is the essence of the "***do(.)***" operator introduced below.

To determine the direct effect of **x** on **y**, one can *mutilate*[48] the graph by setting (randomizing) the values of **x** (to **1** and then to **0**), such that the post-intervention distribution to reflect **G'** above can be denoted by the following *truncated factorization*[49] (**P(x|z)** is dropped as **x** becomes parentless [independent of z] due to randomization) [**Equation 5**]:

$$P(z, y \mid do(x)) = P_{mutilated}(z)\, P_{mutilated}(y \mid x, z) = P(z)\, P(y \mid x, z)$$

By dividing Equation 3 by **P(x|z)** as per **Equation 4** and combining it with **Equation 5**, I obtain a telling pre- and post-intervention ratio [**Equation 6**]:

$$P(z, y \mid do(x)) = \frac{P(x, y, z)}{P(x \mid z)}$$

We can perform adjustment by marginalizing over **z** (J. Pearl et al., 2016) [**Equation 7**]:

$$P(z, y \mid do(x)) = \sum_{z} P(z)\, P(y \mid x, z)$$

Counterfactual and Potential Outcomes Framework. Confirming our intuition, the **P(x|z)** is critical for estimating the effect of **do(x)**. To nullify the dependency between the confounder and the explanatory variable, the explanatory variable **x** can be set to a value of **1** and then to **0**, where **1** indicates presence (or treatment/case) and **0** denotes absence (or placebo/control) (K. Lee, Small, & Rosenbaum, 2018; J. Pearl et al., 2016; Rosenbaum & Rubin, Donald, 1983; M. Scutari & Denis, 2014) [**Equation 8**]:

---

[48] Elsewhere, this is referred to as "graph surgery" (S. Greenland et al., 1999; Judea Pearl, 2009).

[49] The factorization is referred to as being "truncated" because the **P(x|z)** has been removed.

$$P\,y\,do\,x = E\,y\,x = 1 - E(y|x = 0)$$

Equation 8 is a means to express the severed dependency of **P(x|z)** in **G**'. When the graph is modified to remove the influence of **z** on **x**, a more accurate picture of the relationship between **x** and **y** emerges. This intuition that underlies the notion of the *"back-door" criterion* is that the graph is to permit the estimation of the causal influence of **x** on **y** irrespective of the influence of **z** on **x** (the "back-door path"). The apparent effect of **x** on **y** may or may not be spurious, and this procedure facilitates this estimation. The "*back-door criterion*" is discussed at length elsewhere (Morgan & Winship, 2015; Judea Pearl, 2009; T. J. VanderWeele & Shpitser, 2013).

Note the operations just described are "counterfactual" in that we are not using the joint distribution from the data, but estimating the relationships from a manipulated distribution. Recall that the relationship between **z** and **x** (nullifying the *"back-door"*) is severed and the values of **x** are fixed to **1** and **0**, so as to measure an average effect of **y** on **x**. **P(y | do(x))**, referred to as "average causal estimate" or "average treatment estimate", or *ATE*, is not the same as **P(y|x)** (association). However, they could be the same, in an hypothetical case, if the **x** and **y** are not confounded[50].

6.2.3 *Mutilated graph simulations for ATE($drug \Rightarrow adverse\_event$)*

[50] Note also that we cannot determine how an individual patient "**p**" would fare given a treatment since at one time slice patient p's outcome **Y** can only hold discrete values, **1** or **0**, not both. This is why the expectation or mean is calculated (the average treatment effect). In performing the ATE calculation, there is an explicit assumption that the patient has two values simultaneously. This is why the potential outcomes framework methods is referred to as being "*counterfactual*."

In probabilistic graphical models, the most common way to estimate the set of parameters Θ is to perform a maximum likelihood estimate. This finds the best set of edges given input data (and assumptions from domain knowledge about the orientation of the edges).

However, for causal models, one does not stop at an estimate from the joint density distribution but uses that distribution as the means to generate data from a perturbed or "*mutilated*" distribution. In this mutilated distribution, the predictor variable's values are fixed as per **Equation 8** above (to remove dependency to a common cause [the $\mathbf{P(x|z)}$ term in the graph factorization]).

A causal model specifies both *qualitative* (the edges/arcs between nodes/vertices in the graph) and *quantitative* relationships. The latter is embodied in another stream of influence on causal inference deriving from econometrics, more specifically of structural equations modeling (SEM) (Bollen, 2014; Bollen & Long, 1993; Goldberger & Duncan, 1973; J, 2008; Judea Pearl, 2009).

An example will illustrate a simple (causal) *SEM* as a *data generating process (Hendry & Doornik, 2014)*. Take variables $\mathbf{\{x, y, z\}}$ and presume that their respective relationships and parameters Θ have been learned from a set of observational data. Let us presume that these variables are related to each other as per **Figure 7** (from earlier in this chapter) and we define $\mathbf{z}$ as a discrete binary, such that:

$\mathbf{z} \sim \mathbf{Binomial}(\mathbf{1000}, \mathbf{0.2})$

$\mathbf{x} \sim \mathbf{f}(\mathbf{z})$

$\mathbf{y} \sim \mathbf{f}(\mathbf{z}, \mathbf{x})$

From these equations, we can see that **z** is exogenous, while **x** is dependent on the value of **z** and **y** is dependent on the values of both **x** and **z**. With this data generating model, the average effect of **x** on **y** may be estimated by shutting off the relationship between **x** and **z** by setting the value of **x** value to **1** and subsequently to **0** (severing the "back-door" path, i.e., the **P(x|z)** term in the graph factorization).

The results of this operation are a new "mutilated" set of equations that now look like this:

$\mathbf{z} \sim \mathbf{Binomial}(\mathbf{1000}, \mathbf{0.2})$

$\mathbf{x} \sim \{\mathbf{1}, \mathbf{0}\}$

$\mathbf{y} \sim \mathbf{f}(\mathbf{z}, \mathbf{x})$

To calculate the average treatment effect, or *ATE*, one is calculating an expectation or mean. This mean is not the mean of the observed, original joint density, but those of the mutilated models. Using these mutilated models (data generating processes) to simulate data, a conditional probability query can be performed with **x = 1** and **x = 0** on data generated from the new mutilated distributions, then one simply subtracts the results [**Equation 9**], where ***E*** is the *expectation* or *mean*[51]:

[51] $E = \frac{1}{n}\sum_{i=1}^{n}(y \mid x_i)$

$$\boldsymbol{ATE}(\mathbf{x} \Rightarrow \mathbf{y}) = \boldsymbol{E}\,\boldsymbol{y}\,|\,\boldsymbol{do}(\boldsymbol{x} = \mathbf{1})) - \boldsymbol{E}(\boldsymbol{y} \mid \boldsymbol{do}\,\boldsymbol{x} = \mathbf{0}|)$$

$$= \mathbf{1}|\boldsymbol{n}|\,\boldsymbol{i} = \mathbf{1}|\boldsymbol{n}|\,\boldsymbol{y}\,|\,\boldsymbol{x} = \mathbf{1}) - |\,\mathbf{1}|\boldsymbol{n}|\,\boldsymbol{i} =$$

$$\mathbf{1}|\boldsymbol{n}|\,\boldsymbol{y}\,|\,\boldsymbol{x} = \mathbf{0})\,|$$

To summarize: to compute the *ATE*, a mean is calculated from simulated data drawn from the joint density distribution of the two mutilated causal Bayesian network (where $\boldsymbol{x}$ is set to $\mathbf{1}$ and subsequently to $\mathbf{0}$)[52]. In graphical terms, it severs the backdoor path (flow of information from the confounder to the predictor. To define ATE in coarse terms, the ATE is the mean "delta" (change) that one could expect of an adverse event occurring from each increase drug exposure in that population; it is sometimes referred to as the stable unit treatment value assumption[53], or *SUTVA* (Cox, 1958).

In machine learning, sequential (data generating) processes such as those described above may be simulated using what is variously referred to as "*particle filtering*" or "*logic sampling*" methods (Darroch, Lauritzen, & Speed, 1980; Fung & Chang, 1990; M. Henrion, 1987; Max Henrion, 1988; Koller et al., 2009; M. Scutari & Denis, 2014). In logic sampling, one has a set of variables that are defined as parameterized distribution functions. To simulate data, the process runs forward. Depending on the implementation, there is a "*burn-in*" period of iterations to allow for the sampling to attain the intra-variable parametric targets (Darwiche, 2009; Koller et al., 2009; M. Scutari & Denis,

---

[52] The broad strokes of this method were first devised by Polish agronomist Jerzy Neyman who desired to answer "what if?" or "counterfactual" questions about which seeds to plant given what is known about their crop yields (Neyman, 1937; Rubin, 1990).

[53] Since the data are so coarse-grained, this is not to be taken at face value, but only in relative terms.

2014). For mutilated graphs an updated rendition of logic sampling is used called "likelihood weighting" (M. Scutari & Denis, 2014; Shwe & Cooper, 2013).

Recap: synthesizing the pieces. The steps for simulating data to perform such an estimate from EHR data are as follows:

1. Learn the graph structure with oriented edges between variables for each drug/adverse event pair in the OMOP reference dataset (Ryan, Schuemie, et al., 2013).
2. Given the data, learn the parameters Θ that define the intra-variable relationships.
3. Generate data using the mutilated versions of the causal models, fixing explanatory/drug exposure variable ("evidence") to **1** and **0**, and performing conditional probability queries on the simulated data.
4. Calculate the *ATE* from the mutilated model.
5. Aggregate the results to calculate AUROC from rank ordered ATEs for the OMOP reference data set.

Steps 1 and 2 of the procedure are described in section 6.2.1 and Steps 3 and 4 are described in sections 6.2.2 and 6.2.3.

6.2 Materials and Methods

6.2.1 Data and Knowledge Resources.

EHR data. To evaluate this hypothesis, the same pre-processed EHR data and reference dataset were used as in Chapter 4 and Chapter 5 (Ryan, Schuemie, et al., 2013; *UTHealth BIG.*, 2017).

Discovery Pattern for Confounding Variable Discovery. Semantic Vectors provides an interface that permits searching PSI spaces for concepts that populate particular predicate discovery patterns, which I used to identify the most strongly associated confounding variable candidates for each drug/adverse event pair. I used the following discovery pattern to identify possible confounding variable candidates: **"**drug **TREATS** confounder; confounder **CAUSES adverse_event"** (the same discovery pattern as in Chapter 5). This discovery pattern was chosen for its ability to identify mutual causes of both the explanatory and outcome variables in Chapter 5.

6.2.2 Methods

What follows in this section is a brief breakdown/overview. This is followed up with more details about each step of the procedure.

Core steps to evaluate ATE. The core steps of the method to quantify ATE are as follows:

1. **Query PSI vector space** for confounders in ranked order of relevance.
2. **Test each confounding variable candidate** for directed edges to both the drug and the adverse event using the clinical data, stopping after obtaining ten[54] valid confounding variable candidates for each drug/adverse event pair.

[54] I tested many variations of covariate threshold. In most cases, the more covariates identified, the better the models performed.

3. **Build predictive (statistical [multivariate logistic regression] and causal) models** for each drug/adverse event pair incorporating the validated LBD-derived confounders.
4. **Calculate the ATE using mutilated simulation data generated from the causal models for each drug/adverse event pair**.
5. **Calculate AUROCs** from rank ordered scores (baseline regression coefficient, adjusted regression coefficient, ATE) for the drug/adverse event pairs in the reference dataset.

The next subsections will break the novel procedures above into manageable parts.

6.2.2.1 Testing Confounding Variable Candidates.

There are two critical tasks after identifying confounding variable candidates concepts:

1.) Determine if confounding variable candidate appears in the data (if not, move to the next confounding variable candidate).

2.) Determine if the confounding variable candidate fulfills the graphical criteria as a confounder.

If the confounding variable candidate is not in the clinical data, it cannot be instantiated. However, if there are data for the identified concept, then it is tested. Simply stated, in order for a confounding variable candidate to be retained for the purposes of building the causal model, i.e., becomes part of the "causal story"/"focal set", it must first be instantiated with EHR data along with the drug and adverse event. If there is there is a subgraph in the resulting partial ancestral graph (PAG) that fulfills the pattern of: **drug**

⇐ confounder ⇒ adverse event, then the confounding variable candidate is retained for the next step (building the model).

To construct the causal models for testing confounding variable candidates, the hill climbing algorithm in the **bnlearn**[55] **R** package (M. Scutari & Denis, 2014; Marco Scutari, 2009). The hill climbing algorithm recursively adds and subtracts directed edges until the Bayesian Information Criterion is minimized. Note that the purpose in the current experiment is not so much to learn structure, but to assume that there is a relationship between the putative predictor/cause (drug exposure) and outcome (the adverse event). The mutual causes or literature-derived confounders are used to screen off the relationship between the drug and the adverse event.

The output consists of a family of score-equivalent graphs which encode plausible dependency relationships given these data such that the orientation of the directed edges could not be determined from the data alone, i.e., graphs encoded by this structure have the same BIC score (Heckerman et al., 1995, 1995). However, background knowledge can be used to orient these edges, as causal predicates have inherent directionality. The resulting graph structure that results from incorporating literature-derived confounding variable candidates have a graph within this equivalence class with directed edges to both the drug and adverse event ("confounder inclusion criterion").

---

[55] The primary attractiveness of the bnlearn package stems not from its Bayesian structure learning prowess, but its capacity to do most of the "functional" causal modeling toolbox sufficiently well and to run in batch processes from the command line. For a formal comparison of causal Bayesian structure learning packages available to the public see Ramsey et al. (J. D. Ramsey & Andrews, 2017).

6.2.2.2 Building causal models

Once the threshold for the number of confounding variable candidates has been reached (the number of covariates) for inclusion in the focal set, the next step is to use these concepts to populate a causal DAG/model. The causal DAG is instantiated with the variable values in the EHR data. In this step, EHR data and an initial skeleton of a graph (using the edge/arc orientations from the literature) is constructed.

To construct the LBD-informed causal models, the hill climbing algorithm in the **bnlearn R** package (M. Scutari & Denis, 2014; Marco Scutari, 2009) is used, as per the previous step, is used to learn the rest of the (inter-confounder) relationships and to fit a model using *maximum likelihood estimate* (MLE). MLE attempts to identify an optimal set of parameters given the graph structure and the observed data. These parameters may then be used to generate simulated data (as per section **6.2.3**).

The nodes of confounding variable candidates tend to be (but are not always) highly correlated, so there is lots of "cross-talk" between the confounders. An example graph generated from EHR data is provided below in **Figure 9**:

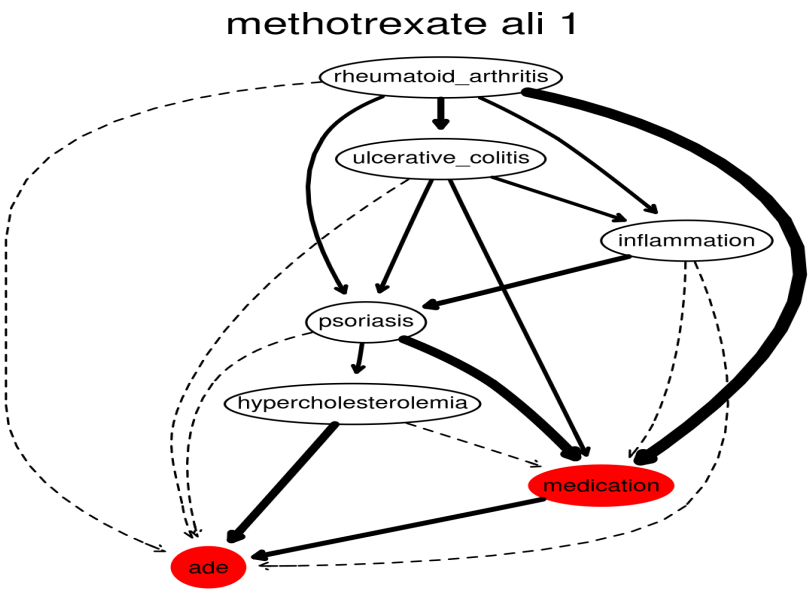


Figure 9. Illustration of a causal graph instantiated with EHR data. Drug = methotrexate, adverse event = acute liver failure ("ALI"). From the OMOP reference dataset (Ryan, Schuemie, et al., 2013), positive control. For clarity, the number of confounders has been reduced.

The graph provides an intuitive picture of some of the causal interactions between these variables. For example, it is clear from the graph that methotrexate is prescribed as a treatment for (therefore, observing it is *CAUSED BY*) psoriasis, ulcerative colitis, and rheumatoid arthritis: but not for hypercholesterolemia. Others are more difficult to interpret. For example, while one might anticipate raised cholesterol in advanced liver disease, hypercholesterolemia is not commonly considered as a possible cause for liver failure. Nonetheless, the ability to interrogate a visual representation of the relationships inferred by the model can be informative in interpreting the assertions of causality that emerge from it.

6.2.2.3 Calculate the ATE using mutilated simulation data

As mentioned earlier, a causal model not only describes a data generating process (that which has been learned from observational or experimental data or both), but a model that may be perturbed/mutilated/manipulated to see how the variables behave under duress. The next sub-subsections describe how to implement these simulations (Judea Pearl, 2009).

Simulating the data. Once a graph and its parameters have been learned from the data, the next step is to manipulate this model.

In a standard conditional probability query (CPQ), the CPQ generates data using the parameters of model, i.e., the joint density distribution learned from the observations. The values of the variables of interest and an "***n***" for the number of simulations may be applied as per the investigator's interest.

The R package ***bnlearn*** provides the means to perform mutilated conditional probability queries. In a mutilated CPQ, the investigator sets the "evidence", e.g., **drug = {1, 0}**, and specify an "event", e.g., adverse event **= 1**. By setting the evidence for drug to **1** and then to **0**, specifying the event to adverse event **= 1**, and running the query, and subtracting the two results (as per **Equation 8** and **Equation 9**), ***bnlearn***'s mutilated CPQ function can calculate an ATE for each drug/adverse event pair from the EHR data using weighted logic sampling simulation (Fung & Chang, 1990; Koller et al., 2009; M. Scutari & Denis, 2014). For the results below, 10⍰9⍰ simulated EHR record instances were generated from causal models derived from EHR data. The simulations were

implemented with the weighted logic sampling option in the **bnlearn R** package (M. Scutari & Denis, 2014; Marco Scutari, 2009).

Evaluation Procedure. Coefficients from logistic regression were used to calculate baseline scores. ATE was estimated by performing a conditional probability query on mutilated graphs for each drug/adverse event causal model, as per **Equation 8**. To evaluate performance, Area Under the Receiver Operating Characteristic curve (AUROC) was calculated based on the ranked order of the *ATE*s.

6.3 Results

Parameter estimates from causal models improved performance over logistic regression for three of four adverse events. Causal models improved upon unadjusted statistical models by 0.0795 AUROC.

Table 6

Results from Average Treatment Effect (ATE)

Adverse events	Total	Unadjusted	Adjusted	Causal Models
AKI (Acute Kidney Injury)	22/62	0.5672	0.5242	**0.6957**
ALI (Acute Liver Injury)	77/34	0.4798	**0.5092**	**0.594**
AMI (Acute Myocardial Infarction)	35/63	0.556	0.5424	0.5456
GIB (Gastrointestinal Bleeding)	24/64	0.5417	**0.5651**	**0.5866**
Overall	158/223	0.4782	**0.5054**	**0.5849**

Note. This table presents the AUROCs as calculated from the aggregated correlation coefficients from logistic regression (unadjusted and adjusted) and *ATE*s from the causal models. Pairs = number of test/control drug/adverse events from the OMOP reference dataset (Ryan, Schuemie, et al., 2013). AUROCs in **bold** indicate that the causal models outperformed the baseline models for that phenotype.

6.4 Discussion

Causal models. There were improvements with causal models in all four adverse events. The overall improvement was consistent and not significantly better than the results from the chapter previous. The improvements for GIB and AKI over their respective baselines were more modest for the causal models than for some of the regression models in **Chapter 4**. An unexpected and notable improvement came from ALI. Such an improvement for ALI has been absent in the results from the experiments in previous chapters.

The experiment in this chapter focuses on creating what Koller calls *functional* causal models (Koller et al., 2009). In a functional causal model, one does not stop with a picture or a joint density distribution, but takes these two (the graph and the distribution) as a data generating process for further exploration and as a means to explore hypothetical questions and scenarios that may not be present in the original/observed data. That is, the experiments in this chapter are not only data-driven, but model-driven. The nature of the improvement in this chapter is more conceptual than in the actualized improvement of performance. That is, the advance described allows for more complex models that will be addressed in the next chapter.

Logistic regression models. The adjusted logistic regression scores have been included to see how useful this discovery pattern is for logistic regression. The improvements overall were slight compared with causal models and did not fair well in comparison with the dual predicate discovery pattern **"drug TREATS x; x COEXISTS_WITH adverse_event"** applied to the EHR data in Chapter 4. This may be because the **"**drug **TREATS** confounder; confounder **CAUSES adverse_event"** discovery pattern identifies only comorbidities, while **"drug TREATS x; x COEXISTS_WITH adverse_event"** can also identify co-medication-type confounding variable candidates. This indicates that discovery patterns besides **"**drug **TREATS** confounder; confounder **CAUSES adverse_event"** should be explored to supplement co-morbidity-type confounders identified by that discovery pattern in future work in LBD-informed predictive modeling.

Li reported notable improvements using data-driven confounding discovery methods (Y. Li, Ryan, Wei, & Friedman, 2015b). Inasmuch as natural language processing (NLP) facilitates research in the text mining of EHR, NLP can also be a limiting factor. Confounders that are present and have been "measured" in the EHR may be absent on the literature side; conversely, not all useful confounders may have been identified in the literature. In light of this, logical follow-up would be to integrate data-driven confounding discovery methods with the current literature-based approach.

For the sake of comparison, the AUROCs of several EHR-based Pharmacovigilance methods has been included below in Table 7. Note that the performance patterns are not strictly comparable owing to different sample sizes and populations, but have been included here for convenience and to provide context. Note the similar starting points for both Li et al. (Y. Li, 2015; Y. Li et al., 2014) and UTH. The causal graphs improve upon the performance of the regression-based models, but fare poorly in comparison with meta-analysis. However, should the causal models be the starting point, meta-analysis should exceed the current state of the art of EHR-based pharmacovigilance barring non-linear effects.

Table 7

Performance Summary of EHR-based Pharmacovigilance Methods

AEs	UTH baseline	Li et al. baseline	Li et al. lasso	UTH Adjusted Stat Model	Causal Models Graphs	Causal Models ATE	Li et al. Meta-Analysis
Overall	0.50	0.53	0.51	0.51	0.58	0.58	0.73

Note. This table presents the AUROCs as calculated from the aggregated correlation coefficients from logistic regression (unadjusted and adjusted) and *scores* from the causal models from the OMOP reference dataset (Ryan, Schuemie, et al., 2013).

Models of 391 of the 399 drug/adverse event pairs in the OMOP reference dataset (Ryan, Schuemie, et al., 2013) were constructed. In some cases, confounding variable candidates were absent (as was the case for three drug/adverse event pairs for MI in the previous chapter). Having set the threshold to ten, the models of five drug/adverse event pairs were not able to constructed as the number of validated confounders did not reach the threshold.

This chapter has evaluated a method to calculate the Average Treatment Effect[56], or *ATE*, using EHR data. Calculating ATE is but one step toward harnessing the richness of EHR data for pharmacovigilance. The steps beyond ATE will be discussed in the next and final chapter.

Work on the methods described began with a vague idea that intervening variables having mechanistic causal relationships could be critical for

[56] The fundamental problem of causal inference is that causes cannot be directly observed (but they may be inferred from the average value of the response between exposed and unexposed population), as per Holland (Holland, 1986).

predictive modeling. The research has progressed from an amorphous, exploratory conception of "in-between" variables to an increasingly refined definition of confounding. The limitations, contributions, and directions of future work of the research program will be presented in the next chapter in detail.

Chapter 7: Summary, Contributions, and Limitations

In this, the concluding chapter of my dissertation, I will summarize my research, underscore its scientific contributions and limitations, and provide some ideas on how I plan to overcome those limitations in future work.

Innovations of Present Work. The first noteworthy accomplishment of this work is the development of a generalizable method through which knowledge extracted from the literature can be used to identify confounding variables for statistical and causal modeling. Confounding variable discovery was achieved by adapting methods from literature-based discovery (LBD) and specifically the use of "discovery patterns" (patterns of predicates that may indicate relationships of a particular type) to this novel task domain. A key finding from this novel component of the project is that when the vast majority of variables identified in this manner do, in fact, represent "true confounders," or mutual causes of both predictor and outcome variables (Judea Pearl, 2009).

In Chapter 4, I demonstrated that literature-derived confounders could be included in statistical (logistic regression) models to improve signal detection using a publicly available reference dataset.

In Chapter 5, I showed that literature-derived confounders could be incorporated into causal models and using the graph structure of the resulting models alone, there was a 7-8% improvement in signal detection over unadjusted baseline.

In Chapter 6, I presented a method to estimate average treatment effect, or *ATE*, of a drug to cause an adverse event given observational data derived from EHR and causal models informed by literature-derived confounders. The *ATE* was estimated by:

1.) learning the topology of the observational data given literature-derived confounders and accompanying subgraph structures inferred from semantic constraints;

2.) estimating model parameters (the joint distribution function) of the causal model that has been learned from the clinical data; and

3.) performing a conditional probability query on a "mutilated" version of the graph by fixing the confounder to drug/explanatory edges to **1** and **0** and subtracting these results.

Calculating the *ATE* had the effect of modestly improving performance over the purely "qualitative" approach to causal inference taken in Chapter 5. However modest the improvement, the success of such a principled approach using "functional causal models" applied to even cross-sectional data opens many doors for methodological refinement and future avenues of research.

7.1 Contributions

The contributions of this thesis to the discourse of biomedical informatics, pharmacovigilance, and causal inference are as follows:

1. **Novel domain of application of LBD methods.** Previous work in LBD has focused on generating novel therapeutic hypotheses, although recent

work has been done in the area of detecting drug/adverse event signals in the literature (Bruza & Weeber, 2008; Hristovski et al., 2006; Mower et al., 2016; Shang et al., 2014; Smalheiser, 2012). The application of these methods in general, and of distributed representations (PSI representations) of discovery patterns in particular, to the problem of identifying confounding variables for statistical and causal modeling is, to the best of my knowledge, entirely without precedent. The present work has demonstrated how causal knowledge embedded in the literature can map onto observational clinical data to inform causal models. This approach has the potential to eliminate a major bottleneck in the statistical and causal modeling of observational data – the need for domain experts to manually delineate a "causal story" describing variables of interest for each hypothesis to be evaluated (J. Pearl et al., 2016). This finding has broader implications for the science of informatics, as it shows that confounders identified through literature-based discovery are of practical utility in improving the accuracy of predictions made from such data.

2. **I show that LBD methods can indeed be used to identify confounding variables and that incorporating these automatically identified variables improved the accuracy of predictive models (both classical statistical and contemporary causal discovery methods).** The studies that I have presented have shown how if there is sufficient interaction between the cue terms of interest, the variables derived from the literature

suggestive of having the desirable property of being "confounders," or mutual causes of both the predictor and outcome variables, are useful for reducing confounding observational clinical data.

3. **The methods proposed may be generalized to other areas of biomedicine.** For example, these methods may be adapted to identify control groups for RCTs (Fokkema et al., 2017). Specifically, a promising area in which my approach to identifying confounders may be of value is in the area of identifying subgroups for stratification in Phase III clinical trials. That is to say that LBD-based identification of confounding subgroups/demographic cohorts could be used to augment the scientific scope or imagination of human experts and help to address unanticipated deficiencies in clinical trials, given for example their being carriers of specific alleles or having particular family medical histories. Some cohorts of the populate could be more informative or representative for the task of assessing safety and efficacy of pharmaceutical profiles before approval and release to market.
4. **Regarding a theoretical contribution, this dissertation demonstrates how domain knowledge may be used as a means to constrain hypothesis space so as to devise an explanation of empirical observational data.** The automatic generation of hypotheses using domain knowledge is tantamount to implementing what is referred to as tractable defeasible logic *in silico* (Gabbay & Woods, 2005). When one

considers the cognitive constraints of the human mind, limits of time and memory, it is remarkable what we as a species have been able to accomplish. There are some tasks where humans excel, yet it has been challenging to develop machines that can perform tractable abduction by arriving at a plausible if not an ideal explanation of what has been observed (what Herbert Simon refers to as "satisficing") (SIMON, 1956; Herbert A. Simon, 1955). It may be the cognitive limitations of human memory forces us to concentrate on a causally relevant subset of features (Cheng & Novick, 1990; Danks, 2014; C. N. Glymour, 2001; Holyoak & Cheng, 2011). To accomplish such a task requires knowledge of context and a database of causal knowledge that describes how the universe is organized. In the approach to causal modeling that I have described, the models apply semantic constraints given cue terms and process input EHR data to infer the causal structure, magnitude of relationship between those entities (the interpretation of the "parameter estimate"), and suggestions of alternate etiologies or explanations of that input data (H.A. Simon, 2012). Specifically, the contribution of this dissertation is to introduce practical constraints on the scope of variables upon which causal modeling is performed. The modeling of human cognition demands that we place constraints similar to those that human reasoners do. While we have a strong will to know and to learn, we also operate within constraints that are imposed by limited resources of time and energy. As is the case with

human reasoners, constraining causal reasoning in this manner offers significant computational advantages and paves a path to large-scale automated causal modeling.

5. **The present work demonstrates the utility of causal methods within a core area of biomedicine.** To the best of my knowledge, except for some research in the application of instrumental variables, the present work is the first attempt to introduce fully automated causal modeling methods within EHR-based pharmacovigilance.
6. **This work demonstrates the utility of *functional* causal models that can answer "what if" queries in biomedicine.** Causal models make assumptions explicit about a domain of knowledge if the form of directed acyclic graphs and the functional forms that define the nodes subsumed in them. Functional causal models instantiated with mixtures of observational and experimental data allow for scientists to be able to test the effects of hypothetical interventions and the falsifiability of their causal assumptions *in silico*.

7.2 Limitations

One limitation of the present study was that the representation of the input data was cross-sectional. In cross-sectional data, the primary unit of analysis is the individual EHR record. This coarse-grained, but simple data representation lacks the means to represent temporal constraints and patient-specific exposures. A study design that

incorporates patient-level longitudinal EHR data may address this limitation in future work (D, S, & K, 2008; Hoover & Demiralp, 2003; Moneta & Spirtes, 2006). An additional limitation arises from the available EHR data which may not have a sufficient number of cases (drug/adverse event co-occurrences), resulting in the poor performance of the baseline models, apparently a common experience (Y. Li, 2015; Y. Li et al., 2015b). As adequate performance without confounding adjustment appears to be a prerequisite to substantive improvement once confounders were accounted for, this limits the scope of application of my methods. This relatively poor performance with EHR data alone is not unique to the current work – performance from analyses of FAERS data is usually better than results from any EHR data source (Y. Li et al., 2015a). Another consideration is that reference data sets, however essential to the scientific enterprise, may not be entirely accurate, as knowledge about drugs and their side-effects accumulates – recent work has identified specific deficiencies in the reference set that I have utilized (Hauben et al., 2016). Current performance is not yet adequate to support a practically useful pharmacovigilance system and falls short of the performance obtained using other data sources, i.e., meta-analysis. It is not clear that this is only due to inadequate numbers of cases and resultant lack of statistical power: there may be other limitations of EHR data that come into play here. I believe that this limitation is due to the generally poor baseline performance, which:

1.) constrains the potential of confounding adjustment,

2.) is typical with EHR-only analyses, and

3.) has been addressed elsewhere successfully by combining EHR and FAERS data, which is likely to be a fruitful area for future research.

7.3 Future Work

In my future work, I will aim to utilize patient-level longitudinal representations of EHR data to address the limitations of cross-sectional causal.

Another idea to explore would the use of FAERS and EHR data together and the exploration of causal modeling using longitudinal patient-level data with structural vector autoregression (where patients become their own controls). I will also develop heuristics to allow for more in-depth confounder search, and explore the potential application of LBD methods to identify instrumental variables.

In any source of observational data, there is always likely to be confounding. Greedy Fast Causal Inference (GFCI), which first runs FGES and then prunes edges with a constraint-based structure learning algorithm, is an algorithm which can identify hidden latent confounding (Ogarrio et al., 2016; J. D. Ramsey & Malinsky, 2016; *TETRAD*, 2017). GFCI, in other words, could direct the LBD mechanism in automated causal inference to conduct a recursive search for confounding variable candidates until no further latent confounding is discovered (or up to a pre-determined threshold or search depth).

One challenge of integrating heterogeneous data is that it is likely that there is missing data. However, these missing data could be imputed and become useful if the pattern of missing data is not random, but systematic.

Finally, I have implemented a method to calculate the Average Treatment Effect, or *ATE*, using EHR data. Calculating ATE is but one step toward harnessing the richness of EHR data for pharmacovigilance. Two promising directions for building on the methods described (culminating) in Chapter 6 include: calculating controlled direct effects (M.A. Hernan & Robins, 2017; Petersen, Sinisi, & Laan, 2006; Robins & Greenland, 1992; Tchetgen Tchetgen & Vanderweele, 2014; T. VanderWeele, 2015; T. J. VanderWeele, 2012; Vanderweele et al., 2014) and/or implementing Cheng Models (wherein the average causal effect of known and established causes are included to normalize estimated treatment effects) (Cheng, 1997; Cheng & Novick, 1990; C. N. Glymour, 2001; Holyoak & Cheng, 2011).

7.4 Conclusion

Better detection methods in pharmacovigilance, if implemented appropriately, should result in better public health and less of an onerous burden stemming from adverse drug events. Enhanced methods for drug/adverse event detection using more granular data would also permit regulatory agencies to prioritize potentially causal drug/adverse event relationships for critical review more precisely owing to their severity and prevalence. Given the extent of the exposed population, an improvement of even a few percentage points could potentially save the lives of many thousands of our loved ones per year.

References

A, S. (1990). The simultaneous-equations model revisited. Statistical adequacy and identification. *Journal of Econometrics*, *44*, 87.

Aalen, O. O., Roysland, K., Gran, J. M., & Ledergerber, B. (2012). Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *175*(4), 831–861. https://doi.org/10.1111/j.1467-985X.2011.01030.x

Ali, M. S., Groenwold, R. H. H., Belitser, S. V., Pestman, W. R., Hoes, A. W., Roes, K. C. B., … Klungel, O. H. (2015). Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of Clinical Epidemiology*, *68*(2), 112–121. https://doi.org/10.1016/j.jclinepi.2014.08.011

Alvarez-Requejo, A., Carvajal, A., Begaud, B., Moride, Y., Vega, T., & Arias, L. H. (1998). Under-reporting of adverse drug reactions. Estimate based on a spontaneous reporting scheme and a sentinel system. *European Journal of Clinical Pharmacology*, *54*(6), 483–488.

Amirkhani, H., Rahmati, M., Lucas, P., & Hommersom, A. (2016). Exploiting Experts' Knowledge for Structure Learning of Bayesian Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2016.2636828

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings. AMIA Symposium*, 17–21.

Aronson, J. K. (2017). Post-marketing drug withdrawals: Pharmacovigilance success, regulatory problems. *Therapie*, *72*(5), 555–561. https://doi.org/10.1016/j.therap.2017.02.005

Backenroth, D., Chase, H., Friedman, C., & Wei, Y. (2016). Using Rich Data on Comorbidities in Case-Control Study Design with Electronic Health Record Data Improves Control of Confounding in the Detection of Adverse Drug Reactions. *PloS One*, *11*(10), e0164304. https://doi.org/10.1371/journal.pone.0164304

Bauer-Mehren, A., van Mullingen, E. M., Avillach, P., Carrascosa, M. D. C., Garcia-Serna, R., Pinero, J., … Furlong, L. I. (2012). Automatic filtering and substantiation of drug safety signals. *PLoS Computational Biology*, *8*(4), e1002457. https://doi.org/10.1371/journal.pcbi.1002457

Beebee, H., Hitchcock, C., & Menzies, P. (2009). *The Oxford Handbook of Causation*. OUP Oxford. Retrieved from https://books.google.com/books?id=xGnZtUtG-nIC

Bollen, K. A. (2014). *Structural Equations with Latent Variables*. Wiley. Retrieved from https://books.google.com/books?id=DPBjBAAAQBAJ

Bollen, K. A., & Long, J. S. (1993). *Testing Structural Equation Models*. SAGE Publications. Retrieved from https://books.google.com/books?id=FvIxxeYDLx4C

Bowden, R. J., & Turkington, D. A. (1985). *Instrumental Variables*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CCOL0521262410

Boyce, R. D., Ryan, P. B., Norén, G. N., Schuemie, M. J., Reich, C., Duke, J., … Dumontier, M. (2014). Bridging Islands of Information to Establish an Integrated Knowledge Base of Drugs and Health Outcomes of Interest. *Drug Safety*, *37*(8), 557–567. https://doi.org/10.1007/s40264-014-0189-0

Brock, T. D. (1961). Milestones in Microbiology. *Academic Medicine*, *36*(7), 847.

Brookhart, M. A., Rassen, J. A., & Schneeweiss, S. (2010). Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and Drug Safety*, *19*(6), 537–554. https://doi.org/10.1002/pds.1908

Brookhart, M. A., Stürmer, T., Glynn, R. J., Rassen, J., & Schneeweiss, S. (2010). Confounding control in healthcare database research: challenges and potential approaches. *Medical Care*, *48*(6 Suppl), S114-120. https://doi.org/10.1097/MLR.0b013e3181dbebe3

Brookhart, M. A., Wang, P. S., Solomon, D. H., & Schneeweiss, S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology (Cambridge, Mass.)*, *17*(3), 268–275. https://doi.org/10.1097/01.ede.0000193606.58671.c5

Bruza, P., & Weeber, M. (2008). *Literature-based Discovery*. Springer Berlin Heidelberg. Retrieved from https://books.google.com/books?id=niMgUkzU42cC

Cameron, D., Bodenreider, O., Yalamanchili, H., Danh, T., Vallabhaneni, S., Thirunarayan, K., … Rindflesch, T. C. (2013). A graph-based recovery and

decomposition of Swanson's hypothesis using semantic predications. *Journal of Biomedical Informatics*, *46*(2), 238–251. https://doi.org/10.1016/j.jbi.2012.09.004

Cameron, D., Kavuluru, R., Rindflesch, T. C., Sheth, A. P., Thirunarayan, K., & Bodenreider, O. (2015). Context-driven automatic subgraph creation for literature-based discovery. *Journal of Biomedical Informatics*, *54*, 141–157. https://doi.org/10.1016/j.jbi.2015.01.014

Cartwright, N. (1994). No Causes in, No Causes out. Retrieved from http://oxfordindex.oup.com/view/10.1093/0198235070.003.0003, //oxfordindex.oup.com/view/10.1093/0198235070.003.0003

Cartwright, N. (2004). Nature's capacities and their measurement. Retrieved from http://www.oxfordscholarship.com/oso/public/content/philosophy/0198235070/toc.html

Cartwright, N. (2007). Are RCTs the Gold Standard? *BioSocieties*, *2*(1), 11–20. https://doi.org/10.1017/S1745855207005029

CFM 14 | Cowles Foundation for Research in Economics. (n.d.). Retrieved June 15, 2018, from https://cowles.yale.edu/cfm-14

Chen, Y., & Briesacher, B. A. (2011). Use of Instrumental Variable in Prescription Drug Research with Observational Data: A Systematic Review. *Journal of Clinical Epidemiology*, *64*(6), 687–700. https://doi.org/10.1016/j.jclinepi.2010.09.006

Cheng, P. W. (1997). From Covariation to Causation: A Causal Power Theory, 39.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*(4), 545–567. https://doi.org/10.1037/0022-3514.58.4.545

Chickering, D. M. (2015). Selective Greedy Equivalence Search: Finding Optimal Bayesian Networks Using a Polynomial Number of Score Evaluations. Retrieved June 27, 2018, from https://arxiv.org/abs/1506.02113

Chu, T., Scheines, R., & Spirtes, P. L. (2013). Semi-Instrumental Variables: A Test for Instrument Admissibility. *ArXiv:1301.2261 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1301.2261

Cohen, T., Schvaneveldt, R. W., & Rindflesch, T. C. (2009). Predication-based semantic indexing: permutations as a means to encode predications in semantic space. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2009*, 114–118.

Cohen, T., Schvaneveldt, R., & Widdows, D. (2010). Reflective Random Indexing and indirect inference: a scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, *43*(2), 240–256. https://doi.org/10.1016/j.jbi.2009.09.003

Cohen, T., Whitfield, G. K., Schvaneveldt, R. W., Mukund, K., & Rindflesch, T. (2010). EpiphaNet: An Interactive Tool to Support Biomedical Discoveries. *Journal of Biomedical Discovery and Collaboration*, *5*, 21–49.

Cohen, T., & Widdows, D. (2017). Embedding of semantic predications. *Journal of Biomedical Informatics*, *68*, 150–166. https://doi.org/10.1016/j.jbi.2017.03.003

Cohen, T., Widdows, D., Schvaneveldt, R. W., Davies, P., & Rindflesch, T. C. (2012). Discovering discovery patterns with Predication-based Semantic Indexing. *Journal of Biomedical Informatics*, *45*(6), 1049–1065. https://doi.org/10.1016/j.jbi.2012.07.003

Cohen, T., Widdows, D., Schvaneveldt, R. W., & Rindflesch, T. C. (n.d.). Logical Leaps and Quantum Connectives: Forging Paths through Predication Space, 8.

Collier, R. (2017). Electronic health records contributing to physician burnout. *CMAJ : Canadian Medical Association Journal*, *189*(45), E1405-1406. https://doi.org/10.1503/cmaj.109-5522

Commissioner, O. of the. (n.d.). Data Mining at FDA - Data Mining at FDA -- White Paper [WebContent]. Retrieved June 25, 2018, from https://www.fda.gov/ScienceResearch/DataMiningatFDA/ucm446239.htm

confound | Definition of confound in English by Oxford Dictionaries. (n.d.). Retrieved June 12, 2018, from https://en.oxforddictionaries.com/definition/confound

Cooper, G. F., & Yoo, C. (2013). Causal Discovery from a Mixture of Experimental and Observational Data. *ArXiv:1301.6686 [Cs]*. Retrieved from http://arxiv.org/abs/1301.6686

Cooper, Gregory F. (1984). *NESTOR: A Computer-Based Medical Diagnostic Aid that Integrates Causal and Probabilistic Knowledge.* Stanford University, Palo Alto, California. Retrieved from http://www.dtic.mil/get-tr-doc/pdf?AD=ADA152046

Corn and hog correlations / by Sewall Wright. v.1300(1925). - Full View | HathiTrust Digital Library | HathiTrust Digital Library. (n.d.). Retrieved June 14, 2018, from https://babel.hathitrust.org/cgi/pt?id=uiug.30112019239976;view=1up;seq=1

Cox, D. R. (1958). *Planning of experiments*. Wiley. Retrieved from https://books.google.com/books?id=sZFQAAAAMAAJ

D, H. K., S, J., & K, J. (2008). Allowing the data to speak freely: The macroeconometrics of the cointegrated vector autoregression. *The American Economic Review*, *98*, 251.

Danks, D. (2014). *Unifying the Mind: Cognitive Representations as Graphical Models*. MIT Press. Retrieved from https://books.google.com/books?id=cQFzBAAAQBAJ

Darroch, J. N., Lauritzen, S. L., & Speed, T. P. (1980). *Annals of Statistics*, *8*(null), 522.

Darwiche, A. (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511811357

Davies, N. M., Smith, G. D., Windmeijer, F., & Martin, R. M. (2013). COX-2 selective nonsteroidal anti-inflammatory drugs and risk of gastrointestinal tract complications and myocardial infarction: an instrumental variable analysis. *Epidemiology (Cambridge, Mass.)*, *24*(3), 352–362. https://doi.org/10.1097/EDE.0b013e318289e024

de Bie, S., Coloma, P. M., Ferrajolo, C., Verhamme, K. M. C., Trifiro, G., Schuemie, M. J., … Sturkenboom, M. C. J. M. (2015). The role of electronic healthcare record databases in paediatric drug safety surveillance: a retrospective cohort study.

British Journal of Clinical Pharmacology, *80*(2), 304–314. https://doi.org/10.1111/bcp.12610

Diaz-Garelli, J.-F., Bernstam, E. V., Mse, null, Rahbar, M. H., & Johnson, T. (2015). Rediscovering drug side effects: the impact of analytical assumptions on the detection of associations in EHR data. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, *2015*, 51–55.

Dore, D. D., Trivedi, A. N., Mor, V., & Lapane, K. L. (2009). Association Between Extent of Thiazolidinedione Exposure and Risk of Acute Myocardial Infarction. *Pharmacotherapy*, *29*(7), 775–783. https://doi.org/10.1592/phco.29.7.775

DuMouchel, W., Ryan, P. B., Schuemie, M. J., & Madigan, D. (2013). Evaluation of disproportionality safety signaling applied to healthcare databases. *Drug Safety*, *36 Suppl 1*, S123-132. https://doi.org/10.1007/s40264-013-0106-y

Duncan, O. D. (1975). *Introduction to Structural Equation Models*. (null, Ed.) (Vol. null).

Dziurosz-Serafinowicz, P. (2012). Common Cause Abduction: Its Scope and Limits. *Filozofia Nauki*, *20*(4).

Edlinger, D., Sauter, S. K., Rinner, C., Neuhofer, L. M., Wolzt, M., Grossmann, W., … Gall, W. (2014). JADE: a tool for medical researchers to explore adverse drug events using health claims data. *Applied Clinical Informatics*, *5*(3), 621–629. https://doi.org/10.4338/ACI-2014-04-RA-0036

Edwards, I. R., & Aronson, J. K. (2000). Adverse drug reactions: definitions, diagnosis, and management. *Lancet (London, England)*, *356*(9237), 1255–1259. https://doi.org/10.1016/S0140-6736(00)02799-9

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499. https://doi.org/10.1214/009053604000000067

EHRIntelligence. (2018, January 16). EHR Use, Administrative Burden Accelerating Physician Burnout. Retrieved June 1, 2018, from https://ehrintelligence.com/news/ehr-use-administrative-burden-accelerating-physician-burnout

Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, *40*(1), 31–53. https://doi.org/10.1146/annurev-soc-071913-043455

Eshleman, R., & Singh, R. (2016). Leveraging graph topology and semantic context for pharmacovigilance through twitter-streams. *BMC Bioinformatics*, *17*(Suppl 13), 335. https://doi.org/10.1186/s12859-016-1220-5

Evans, D., Chaix, B., Lobbedez, T., Verger, C., & Flahault, A. (2012). Combining directed acyclic graphs and the change-in-estimate procedure as a novel approach to adjustment-variable selection in epidemiology. *BMC Medical Research Methodology*, *12*, 156. https://doi.org/10.1186/1471-2288-12-156

Federal Drug Administration Adverse Event Reporting System. (2017). Retrieved from http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/

Fellows, I. (2014). wordcloud: Word Clouds (Version 2.5). Retrieved from https://CRAN.R-project.org/package=wordcloud

Florez, H., Reaven, P. D., Bahn, G., Moritz, T., Warren, S., Marks, J., … Emanuele, N. (2015). Rosiglitazone treatment and cardiovascular disease in the Veterans Affairs Diabetes Trial. *Diabetes, Obesity & Metabolism*, *17*(10), 949–955. https://doi.org/10.1111/dom.12487

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2017). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0971-x

Freedman, D. A. (2004). Graphical models for causation, and the identification problem. *Evaluation Review*, *28*(4), 267–293. https://doi.org/10.1177/0193841X04266432

Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association: JAMIA*, *11*(5), 392–402. https://doi.org/10.1197/jamia.M1552

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Friedman, N. (2013). Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks, 31.

Fung, R., & Chang, K.-C. (1990). Weighing and Integrating Evidence for Stochastic Simulation in Bayesian Networks. In *Machine Intelligence and Pattern Recognition* (Vol. 10, pp. 209–219). Elsevier. https://doi.org/10.1016/B978-0-444-88738-2.50023-3

Gabbay, D. M., & Woods, J. (2005). *A Practical Logic of Cognitive Systems: The Reach of Abduction: Insight and Trial*. Elsevier Science. Retrieved from https://books.google.com/books?id=CFuM9cNIFMEC

Gahr, M., Eller, J., Connemann, B. J., & Schonfeldt-Lecuona, C. (2016). Subjective Reasons for Non-Reporting of Adverse Drug Reactions in a Sample of Physicians in Outpatient Care. *Pharmacopsychiatry*, *49*(2), 57–61. https://doi.org/10.1055/s-0035-1569291

Gavril, F. (1977). Some NP-complete problems on graphs. *Proceedings of the 11th Conference on Information Sciences and Systems*.

Gayler, R. W. (2004). Vector Symbolic Architectures answer Jackendoff's challenges for cognitive neuroscience. *CoRR*, *abs/cs/0412059*. Retrieved from http://arxiv.org/abs/cs/0412059

Geiger, D. (1990). *Graphoids: A Qualitative Framework for Probabilistic Inference* (PhD Thesis). University of California at Berkeley, Berkeley, CA, USA.

Glymour, C. N. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. A Bradford Book. Retrieved from https://books.google.com/books?id=x8GDLz9dhhIC

Glymour, C., Scheines, R., & Spirtes, P. (2014). *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Elsevier Science. Retrieved from https://books.google.com/books?id=iA_jBQAAQBAJ

Goldberger, A. S., & Duncan, O. D. (1973). *Structural Equation Models in the Social Sciences*. (null, Ed.) (Vol. null).

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, *111*(1), 3–32. https://doi.org/10.1037/0033-295X.111.1.3

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, *29*(4), 722–729.

Greenland, S., & Morgenstern, H. (2001). Confounding in health research. *Annual Review of Public Health*, *22*, 189–212. https://doi.org/10.1146/annurev.publhealth.22.1.189

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology (Cambridge, Mass.)*, *10*(1), 37–48.

Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, *15*(3), 413–419.

Greenland, Sander, & Robins, J. M. (2009). Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations : EP+I*, *6*, 4. https://doi.org/10.1186/1742-5573-6-4

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science*, *35*(8), 1407–1455. https://doi.org/10.1111/j.1551-6709.2011.01203.x

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384. https://doi.org/10.1016/j.cogpsych.2005.05.004

Haavelmo, T. (1944). *Econometrica*, *12*(Supplement), 1.

Haerian, K., Varn, D., Vaidya, S., Ena, L., Chase, H. S., & Friedman, C. (2012). Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clinical Pharmacology and Therapeutics*, *92*(2), 228–234. https://doi.org/10.1038/clpt.2012.54

Han, B., Xie, R., Wu, S., Li, L., & Zhu, L. (2015). A case study on the identification of confounding factors for gene disease association analysis. *Cancer Biomarkers : Section A of Disease Markers*, *15*(3), 267–280. https://doi.org/10.3233/CBM-150462

Haneuse, S. (2016). Distinguishing Selection Bias and Confounding Bias in Comparative Effectiveness Research. *Medical Care*, *54*(4), e23-29. https://doi.org/10.1097/MLR.0000000000000011

Hanssen, N. M. J., Westerink, J., Scheijen, J. L. J. M., van der Graaf, Y., Stehouwer, C. D. A., & Schalkwijk, C. G. (2018). Higher Plasma Methylglyoxal Levels Are

Associated With Incident Cardiovascular Disease and Mortality in Individuals With Type 2 Diabetes. *Diabetes Care*. https://doi.org/10.2337/dc18-0159

Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology and Therapeutics*, *91*(6), 1010–1021. https://doi.org/10.1038/clpt.2012.50

Harpaz, Rave. (2014). A time-indexed reference standard of adverse drug reactions, *1*. https://doi.org/10.1038/sdata.2014.43

Harpaz, Rave, DuMouchel, W., LePendu, P., Bauer-Mehren, A., Ryan, P., & Shah, N. H. (2013). Performance of Pharmacovigilance Signal Detection Algorithms for the FDA Adverse Event Reporting System. *Clinical Pharmacology and Therapeutics*, *93*(6), 10.1038/clpt.2013.24. https://doi.org/10.1038/clpt.2013.24

Harpaz, Rave, DuMouchel, W., Schuemie, M., Bodenreider, O., Friedman, C., Horvitz, E., … Shah, N. H. (2017). Toward Multimodal Signal Detection of Adverse Drug Reactions. *J. of Biomedical Informatics*, *76*(C), 41–49. https://doi.org/10.1016/j.jbi.2017.10.013

Harpaz, Rave, DuMouchel, W., & Shah, N. H. (2015). Comment on: "Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance". *Drug Safety*, *38*(1), 113–114. https://doi.org/10.1007/s40264-014-0245-9

Hasford, J., Goettler, M., Munter, K.-H., & Muller-Oerlinghausen, B. (2002). Physicians' knowledge and attitudes regarding the spontaneous reporting system for adverse drug reactions. *Journal of Clinical Epidemiology*, *55*(9), 945–950.

Hauben, M., Aronson, J. K., & Ferner, R. E. (2016). Evidence of Misclassification of Drug-Event Associations Classified as Gold Standard "Negative Controls" by the Observational Medical Outcomes Partnership (OMOP). *Drug Safety*, *39*(5), 421–432. https://doi.org/10.1007/s40264-016-0392-2

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, *20*(3), 197–243. https://doi.org/10.1023/A:1022623210503

Hendry, D. F., & Doornik, J. A. (2014). *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*. MIT Press. Retrieved from https://books.google.com/books?id=EunMAwAAQBAJ

Henricks, W. H. (2011). "Meaningful use" of electronic health records and its relevance to laboratories and pathologists. *Journal of Pathology Informatics*, *2*. https://doi.org/10.4103/2153-3539.76733

Henrion, M. (1987). *Should we use probability in uncertain inference systems?* (null, Ed.) (Vol. null).

Henrion, Max. (1988). Propagating Uncertainty in Bayesian Networks by Probabilistic Logic Sampling. In J. F. Lemmer & L. N. Kanal (Eds.), *Machine Intelligence and Pattern Recognition* (Vol. 5, pp. 149–163). North-Holland. https://doi.org/10.1016/B978-0-444-70396-5.50019-4

Hernan, M.A., & Robins, J. M. (2017). *Causal Inference*. Taylor & Francis. Retrieved from https://books.google.com/books?id=_KnHIAAACAAJ

Hernan, Miguel A., Hernandez-Diaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology (Cambridge, Mass.)*, *15*(5), 615–625.

Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R. O., Bernstam, E. V., … Saltz, J. H. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*, *51*(8 Suppl 3), S30-37. https://doi.org/10.1097/MLR.0b013e31829b1dbd

Hill, A. B. (1965). *Proceedings of the Royal Society of Medicine*, *58*(null), 295.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: the new synthesis. *Annual Review of Psychology*, *62*, 135–163. https://doi.org/10.1146/annurev.psych.121208.131634

Hoover, K. D., & Demiralp, S. (2003). Searching for the Causal Structure of a Vector Autoregression. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.388840

Hristovski, D., Friedman, C., Rindflesch, T. C., & Peterlin, B. (2006). Exploiting semantic relations for literature-based discovery. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, 349–353.

Hume, D. (1740a). *A Treatise on Human Nature*. (null, Ed.) (Vol. null).

Hume, D. (1740b). *A Treatise on Human Nature*. (null, Ed.) (Vol. null).

Hume, D. (1748). *An Inquiry Concerning Human Understanding*. (null, Ed.) (Vol. null).

Hurricane Maria Contributed to Nearly 5,000 Deaths, Researchers Say - Scientific American. (n.d.). Retrieved June 26, 2018, from https://www.scientificamerican.com/article/hurricane-maria-contributed-to-nearly-5-000-deaths-researchers-say/

ISO/TR 20514:2005(en), Health informatics — Electronic health record — Definition, scope and context. (n.d.). Retrieved June 16, 2018, from https://www.iso.org/obp/ui/#iso:std:iso:tr:20514:ed-1:v1:en

J, H. (2008). Econometric causality. *International Statistical Review*, *76*, 1.

Jackson, J. W., Schmid, I., & Stuart, E. A. (2017). Propensity Scores in Pharmacoepidemiology: Beyond the Horizon. *Current Epidemiology Reports*, *4*(4), 271–280. https://doi.org/10.1007/s40471-017-0131-y

Johnson, S. B., Bakken, S., Dine, D., Hyun, S., Mendonça, E., Morrison, F., … Stetson, P. (2008). An Electronic Health Record Based on Structured Narrative. *Journal of the American Medical Informatics Association : JAMIA*, *15*(1), 54–64. https://doi.org/10.1197/jamia.M2131

Kanerva, P. (1994). *The Spatter Code for Encoding Concepts at Many Levels*.

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random Indexing of Text Samples for Latent Semantic Analysis. In *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 103–6). Erlbaum.

Kilicoglu, H., Fiszman, M., Rosemblat, G., Marimpietri, S., & Rindflesch, T. (2010). Arguments of Nominals in Semantic Interpretation of Biomedical Text. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*

(pp. 46–54). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W10-1906

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., & Rindflesch, T. C. (2012). SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics (Oxford, England)*, *28*(23), 3158–3160. https://doi.org/10.1093/bioinformatics/bts591

Koller, D., Friedman, N., & Bach, F. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press. Retrieved from https://books.google.com/books?id=dOruCwAAQBAJ

Kovesdy, C. P., & Kalantar-Zadeh, K. (2012). OBSERVATIONAL STUDIES VS. RANDOMIZED CONTROLLED TRIALS: AVENUES TO CAUSAL INFERENCE IN NEPHROLOGY. *Advances in Chronic Kidney Disease*, *19*(1), 11–18. https://doi.org/10.1053/j.ackd.2011.09.004

Lanza, A., Ravaud, P., Riveros, C., & Dechartres, A. (2016). Comparison of Estimates between Cohort and Case–Control Studies in Meta-Analyses of Therapeutic Interventions: A Meta-Epidemiological Study. *PLoS ONE*, *11*(5). https://doi.org/10.1371/journal.pone.0154877

Lee, K., Small, D. S., & Rosenbaum, P. R. (2018). A powerful approach to the study of moderate effect modification in observational studies. *Biometrics*. https://doi.org/10.1111/biom.12884

Lee, P. H., & Burstyn, I. (2016). Identification of confounder in epidemiologic data contaminated by measurement error in covariates. *BMC Medical Research Methodology*, *16*, 54. https://doi.org/10.1186/s12874-016-0159-6

Levin, B., & Hovav, M. R. (2005). *Argument Realization*. Cambridge University Press. Retrieved from https://books.google.com/books?id=msi9a50gHVYC

Levy, S. D., & Gayler, R. (2008). Vector Symbolic Architectures: A New Building Material for Artificial General Intelligence. In *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (pp. 414–418). Amsterdam, The Netherlands, The Netherlands: IOS Press. Retrieved from http://dl.acm.org/citation.cfm?id=1566174.1566215

Lewallen, S., & Courtright, P. (1998). Epidemiology in Practice: Case-Control Studies. *Community Eye Health*, *11*(28), 57–58.

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow. *Noûs*, *13*(4), 455–476.

Li, W., Jiang, Z., Geng, Z., & Zhou, X.-H. (2018). Identification of causal effects with latent confounding and classical additive errors in treatment. *Biometrical Journal. Biometrische Zeitschrift*, *60*(3), 498–515. https://doi.org/10.1002/bimj.201700048

Li, Y. (2015). Combining Heterogeneous Databases to Detect Adverse Drug Reactions. Retrieved from https://academiccommons.columbia.edu/catalog/ac:189526

Li, Y., Ryan, P. B., Wei, Y., & Friedman, C. (2015a). A Method to Combine Signals from Spontaneous Reporting Systems and Observational Healthcare Data to Detect Adverse Drug Reactions. *Drug Safety*, *38*(10), 895–908. https://doi.org/10.1007/s40264-015-0314-8

Li, Y., Ryan, P. B., Wei, Y., & Friedman, C. (2015b). A Method to Combine Signals from Spontaneous Reporting Systems and Observational Healthcare Data to Detect Adverse Drug Reactions. *Drug Safety*, *38*(10), 895–908. https://doi.org/10.1007/s40264-015-0314-8

Li, Y., Salmasian, H., Vilar, S., Chase, H., Friedman, C., & Wei, Y. (2014a). A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records. *Journal of the American Medical Informatics Association: JAMIA*, *21*(2), 308–314. https://doi.org/10.1136/amiajnl-2013-001718

Li, Y., Salmasian, H., Vilar, S., Chase, H., Friedman, C., & Wei, Y. (2014b). A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records. *Journal of the American Medical Informatics Association: JAMIA*, *21*(2), 308–314. https://doi.org/10.1136/amiajnl-2013-001718

Lin, K. J., & Schneeweiss, S. (2016). Considerations for the analysis of longitudinal electronic health records linked to claims data to study the effectiveness and safety of drugs. *Clinical Pharmacology and Therapeutics*, *100*(2), 147–159. https://doi.org/10.1002/cpt.359

Liu, J., Zhao, S., & Zhang, X. (2016). An ensemble method for extracting adverse drug events from social media. *Artificial Intelligence in Medicine*, *70*, 62–76. https://doi.org/10.1016/j.artmed.2016.05.004

Mackie, J. L., & Press, O. U. (1980). *The Cement of the Universe: A Study of Causation*. Clarendon Press. Retrieved from https://books.google.com/books?id=KhJ99xSUkPAC

Madigan, D., Ryan, P. B., & Schuemie, M. (2013). Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. *Therapeutic Advances in Drug Safety*, *4*(2), 53–62. https://doi.org/10.1177/2042098613477445

Malec, S. A., Wei, P., Xu, H., Bernstam, E. V., Myneni, S., & Cohen, T. (2016). Literature-Based Discovery of Confounding in Observational Clinical Data. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2016*, 1920–1929.

Meek, C. (2013). Causal Inference and Causal Explanation with Background Knowledge. *ArXiv:1302.4972 [Cs]*. Retrieved from http://arxiv.org/abs/1302.4972

Mill, J. S. (1843). *A System of Logic*. (null, Ed.) (Vol. null).

Moneta, A., & Spirtes, P. (2006). *Graphical Models for the Identification of Causal Structures in Multivariate Time Series Models* (Vol. 2006). https://doi.org/10.2991/jcis.2006.171

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press. Retrieved from https://books.google.com/books?id=Q6YaBQAAQBAJ

Mower, J., Subramanian, D., Shang, N., & Cohen, T. (2016). Classification-by-Analogy: Using Vector Representations of Implicit Relationships to Identify Plausibly

Causal Drug/Side-effect Relationships. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2016*, 1940–1949.

Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T., & Moore, R. (2011). Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association: JAMIA*, *18*(4), 441–448. https://doi.org/10.1136/amiajnl-2011-000116

Nesi, T. (2008). *Poison Pills: The Untold Story of the Vioxx Drug Scandal*. St. Martin's Press. Retrieved from https://books.google.es/books?id=iyS8w5hBZ3MC

News & Events > 50 Years: The Kefauver-Harris Amendments. (2013, March 7). Retrieved June 12, 2018, from https://web.archive.org/web/20130307165433/https://www.fda.gov/Drugs/NewsEvents/ucm320924.htm

Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *236*(767), 333–380.

Neyman, J., Iwaszkiewicz, K., & Kolodziejczyk, S. (1935). *Supplement of Journal of the Royal Statistical Society*, *2*(null), 107.

Norén, G. N., Bergvall, T., Ryan, P. B., Juhlin, K., Schuemie, M. J., & Madigan, D. (2013). Empirical performance of the calibrated self-controlled cohort analysis within temporal pattern discovery: lessons for developing a risk identification and analysis system. *Drug Safety*, *36 Suppl 1*, S107-121. https://doi.org/10.1007/s40264-013-0095-x

Observational Medical Outcomes Partnership (OMOP). (2017). Retrieved from http://omop.org

Ogarrio, J. M., Spirtes, P., & Ramsey, J. (2016). A Hybrid Causal Search Algorithm for Latent Variable Models. *JMLR Workshop and Conference Proceedings*, *52*, 368–379.

Oliveira, J. L., Lopes, P., Nunes, T., Campos, D., Boyer, S., Ahlberg, E., … van der Lei, J. (2013). The EU-ADR Web Platform: delivering advanced pharmacovigilance tools. *Pharmacoepidemiology and Drug Safety*, *22*(5), 459–467. https://doi.org/10.1002/pds.3375

Pearl, J. (2014). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier Science. Retrieved from https://books.google.com/books?id=mn2jBQAAQBAJ

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Wiley. Retrieved from https://books.google.com/books?id=L3G-CgAAQBAJ

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books. Retrieved from https://books.google.com/books?id=9H0dDQAAQBAJ

Pearl, Judea. (2009a). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, Judea. (2009b). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, Judea. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, *6*(2), Article 7. https://doi.org/10.2202/1557-4679.1203

Perez Garcia, M., & Figueras, A. (2011). The lack of knowledge about the voluntary reporting system of adverse drug reactions as a major cause of underreporting: direct survey among health professionals. *Pharmacoepidemiology and Drug Safety*, *20*(12), 1295–1302. https://doi.org/10.1002/pds.2193

Petersen, M., Sinisi, S. E., & Laan, M. J. van der. (2006). Estimation of direct causal effects. *Epidemiology*, *17 3*, 276–284.

Pierce, C. E., Bouri, K., Pamer, C., Proestel, S., Rodriguez, H. W., Van Le, H., … Dasgupta, N. (2017). Evaluation of Facebook and Twitter Monitoring to Detect Safety Signals for Medical Products: An Analysis of Recent FDA Safety Alerts. *Drug Safety*, *40*(4), 317–331. https://doi.org/10.1007/s40264-016-0491-0

Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A. K., Walley, T. J., … Breckenridge, A. M. (2004). Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ : British Medical Journal*, *329*(7456), 15–19.

Pugh, C. A., Bronsvoort, B. M. de C., Handel, I. G., Summers, K. M., & Clements, D. N. (2014). What can cohort studies in the dog tell us? *Canine Genetics and Epidemiology*, *1*. https://doi.org/10.1186/2052-6687-1-5

Ramsey, J. D. (2015). Scaling up Greedy Equivalence Search for Continuous Variables. *CoRR*, *abs/1507.07749*. Retrieved from http://arxiv.org/abs/1507.07749

Ramsey, J. D., & Andrews, B. (2017). A Comparison of Public Causal Search Packages on Linear, Gaussian Data with No Latent Variables. *ArXiv:1709.04240 [Cs]*. Retrieved from http://arxiv.org/abs/1709.04240

Ramsey, J. D., & Malinsky, D. (2016). Comparing the Performance of Graphical Structure Learning Algorithms with TETRAD. *ArXiv:1607.08110 [Stat]*. Retrieved from http://arxiv.org/abs/1607.08110

Ramsey, J., Zhang, J., & Spirtes, P. L. (2012). Adjacency-Faithfulness and Conservative Causal Inference. *ArXiv:1206.6843 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1206.6843

Ranganath, R., & Perotte, A. (2018). Multiple Causal Inference with Latent Confounding. *ArXiv:1805.08273 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1805.08273

Reeves, D., Springate, D. A., Ashcroft, D. M., Ryan, R., Doran, T., Morris, R., … Kontopantelis, E. (2014). Can analyses of electronic patient records be independently and externally validated? The effect of statins on the mortality of patients with ischaemic heart disease: a cohort study with nested case–control analysis. *BMJ Open*, *4*(4). https://doi.org/10.1136/bmjopen-2014-004952

Reichenbach, H. (2012). *The Direction of Time*. Dover Publications. Retrieved from https://books.google.com/books?id=qjTrZ2vuQ9sC

Richiardi, L., Bellocco, R., & Zugna, D. (2013). Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology*, *42*(5), 1511–1519. https://doi.org/10.1093/ije/dyt127

Robins, J., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3 2*, 143–155.

Rosenbaum, P. R., & Rubin, Donald, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1990). [On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.] Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. *Statistical Science*, *5*(4), 472–480. https://doi.org/10.1214/ss/1177012032

RxNorm. (2017). Retrieved from https://www.nlm.nih.gov/research/umls/rxnorm/

Ryan, P. B., Schuemie, M. J., Welebob, E., Duke, J., Valentine, S., & Hartzema, A. G. (2013a). Defining a reference set to support methodological research in drug safety. *Drug Safety*, *36 Suppl 1*, S33-47. https://doi.org/10.1007/s40264-013-0097-8

Ryan, P. B., Schuemie, M. J., Welebob, E., Duke, J., Valentine, S., & Hartzema, A. G. (2013b). Defining a reference set to support methodological research in drug safety. *Drug Safety*, *36 Suppl 1*, S33-47. https://doi.org/10.1007/s40264-013-0097-8

Ryan, P. B., Stang, P. E., Overhage, J. M., Suchard, M. A., Hartzema, A. G., DuMouchel, W., … Madigan, D. (2013). A comparison of the empirical performance of methods for a risk identification system. *Drug Safety*, *36 Suppl 1*, S143-158. https://doi.org/10.1007/s40264-013-0108-9

S, W. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press. Retrieved from https://books.google.com/books?id=2ug9DwAAQBAJ

Samuels, J. G., McGrath, R. J., Fetzer, S. J., Mittal, P., & Bourgoine, D. (2015). Using the Electronic Health Record in Nursing Research: Challenges and Opportunities. *Western Journal of Nursing Research*, *37*(10), 1284–1294. https://doi.org/10.1177/0193945915576778

Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1998). The TETRAD Project: Constraint Based Aids to Causal Model Specification. *Multivariate Behavioral Research*, *33*(1), 65–117. https://doi.org/10.1207/s15327906mbr3301_3

Schuemie, M. J. (2011). Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. *Pharmacoepidemiology and Drug Safety*, *20*(3), 292–299. https://doi.org/10.1002/pds.2051

Scutari, M., & Denis, J. B. (2014). *Bayesian Networks: With Examples in R*. CRC Press. Retrieved from https://books.google.com/books?id=jS3cBQAAQBAJ

Scutari, Marco. (2009). Learning Bayesian networks with the bnlearn R package. *ArXiv Preprint ArXiv:0908.3817*.

Scutari, Marco, Vitolo, C., & Tucker, A. (2018). Learning Bayesian Networks from Big Data with Greedy Search: Computational Complexity and Efficient Implementation.

Semantic Vectors. (n.d.). Retrieved from https://github.com/semanticvectors/semanticvectors

SemMedDB. (2017). Retrieved from http://skr3.nlm.nih.gov/SemMedDB/

SemMedDB. (n.d.). Retrieved from http://skr3.nlm.nih.gov/SemMedDB/

Shang, N., Xu, H., Rindflesch, T. C., & Cohen, T. (2014). Identifying plausible adverse drug reactions using knowledge extracted from the literature. *Journal of Biomedical Informatics*, *52*, 293–310. https://doi.org/10.1016/j.jbi.2014.07.011

Shrier, I., & Pang, M. (2015). Confounding, effect modification and the odds ratio: Common misinterpretations. *Journal of Clinical Epidemiology*, *68*(4), 470–474. https://doi.org/10.1016/j.jclinepi.2014.12.012

Shwe, M., & Cooper, G. F. (2013). An Empirical Analysis of Likelihood-Weighting Simulation on a Large, Multiply-Connected Belief Network. *ArXiv:1304.1141 [Cs]*. Retrieved from http://arxiv.org/abs/1304.1141

SIMON, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138.

Simon, H.A. (2012). *Models of Discovery: and Other Topics in the Methods of Science*. Springer Netherlands. Retrieved from https://books.google.com/books?id=iCvpCAAAQBAJ

Simon, Herbert A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99. https://doi.org/10.2307/1884852

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, *39*(5), 1–13.

Smalheiser, N. R. (2012). Literature-based discovery: Beyond the ABCs. *Journal of the American Society for Information Science and Technology*, *63*(2), 218–224. https://doi.org/10.1002/asi.21599

Smalheiser, N. R. (2017). Rediscovering Don Swanson: the Past, Present and Future of Literature-Based Discovery. *Journal of Data and Information Science (Warsaw, Poland)*, *2*(4), 43–64. https://doi.org/10.1515/jdis-2017-0019

Spirtes, P., Glymour, C., & Scheines, R. (2012). *Causation, Prediction, and Search*. Springer New York. Retrieved from https://books.google.com/books?id=oUjxBwAAQBAJ

Spirtes, P. L. (2013). Detecting Causal Relations in the Presence of Unmeasured Variables. *ArXiv:1303.5754 [Cs]*. Retrieved from http://arxiv.org/abs/1303.5754

Spirtes, Peter, & Glymour, C. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review*, *9*(1), 62–72. https://doi.org/10.1177/089443939100900106

Spirtes, Peter, & Zhang, K. (2016). Causal Discovery and Inference: Concepts and Recent Methodological Advances. *Applied Informatics*, *3*, 3. https://doi.org/10.1186/s40535-016-0018-x

Stephens, T., & Brynner, R. (2009). *Dark Remedy: The Impact Of Thalidomide And Its Revival As A Vital Medicine*. Basic Books. Retrieved from https://books.google.com/books?id=9IGyL1Cwy08C

Suzuki, E., Tsuda, T., Mitsuhashi, T., Mansournia, M. A., & Yamamoto, E. (2016). Errors in causal inference: an organizational schema for systematic error and random error. *Annals of Epidemiology*, *26*(11), 788-793.e1. https://doi.org/10.1016/j.annepidem.2016.09.008

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, *30*(1), 7–18.

Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, *31*(4), 526–557.

Swanson, D. R. (1989). A second example of mutually isolated medical literatures related by implicit, unnoticed connections. *Journal of the American Society for Information Science. American Society for Information Science*, *40*(6), 432–435. https://doi.org/10.1002/(SICI)1097-4571(198911)40:6<432::AID-ASI5>3.0.CO;2-#

Talbot, J. C. C., & Aronson, J. K. (Eds.). (2012). *Stephens' detection and evaluation of adverse drug reactions: principles and practice* (6th ed). Chichester, West Sussex, UK: John Wiley & Sons.

Tatonetti, N. P., Ye, P. P., Daneshjou, R., & Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, *4*(125), 125ra31. https://doi.org/10.1126/scitranslmed.3003377

Tchetgen Tchetgen, E. J., & Vanderweele, T. J. (2014). Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology (Cambridge, Mass.)*, *25*(2), 282–291. https://doi.org/10.1097/EDE.0000000000000054

TETRAD. (2017). Retrieved from http://www.phil.cmu.edu/tetrad/current.html

Thagard, P. (2018). *How Scientists Explain Disease*. Princeton University Press. Retrieved from https://books.google.com/books?id=0f9ZDwAAQBAJ

The Unified Medical Language System (UMLS). (2017). Retrieved from http://www.nlm.nih.gov/research/umls/

Thygesen, L. C., Pottegard, A., Ersboll, A. K., Friis, S., Sturmer, T., & Hallas, J. (2017). External adjustment of unmeasured confounders in a case-control study of benzodiazepine use and cancer risk. *British Journal of Clinical Pharmacology*, *83*(11), 2517–2527. https://doi.org/10.1111/bcp.13342

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Trifiro, G., Coloma, P. M., Rijnbeek, P. R., Romio, S., Mosseveld, B., Weibel, D., … Sturkenboom, M. (2014). Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *Journal of Internal Medicine*, *275*(6), 551–561. https://doi.org/10.1111/joim.12159

Trifiro, Gianluca, Sultana, J., & Bate, A. (2018). From Big Data to Smart Data for Pharmacovigilance: The Role of Healthcare Databases and Other Emerging Sources. *Drug Safety*, *41*(2), 143–149. https://doi.org/10.1007/s40264-017-0592-4

Uhler, C., Raskutti, G., Bühlmann, P., & Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, *41*(2), 436–463. https://doi.org/10.1214/12-AOS1080

UTHealth BIG. (2017). Retrieved from http://redcap.uth.tmc.edu/cdwstats/stats-mpi.htm

Valente, M. J., Pelham, W. E., Smyth, H., & MacKinnon, D. P. (2017). Confounding in statistical mediation analysis: What it is and how to address it. *Journal of Counseling Psychology*, *64*(6), 659–671. https://doi.org/10.1037/cou0000242

VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press. Retrieved from https://books.google.com/books?id=K6cgBgAAQBAJ

VanderWeele, T. J. (2012). Mediation analysis with multiple versions of the mediator. *Epidemiology (Cambridge, Mass.)*, *23*(3), 454–463. https://doi.org/10.1097/EDE.0b013e31824d5fe7

VanderWeele, T. J., & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, *67*(4), 1406–1413. https://doi.org/10.1111/j.1541-0420.2011.01619.x

VanderWeele, T. J., & Shpitser, I. (2013). On the definition of a confounder. *Annals of Statistics*, *41*(1), 196–220.

Vanderweele, T. J., Vansteelandt, S., & Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*

(Cambridge, Mass.), *25*(2), 300–306. https://doi.org/10.1097/EDE.0000000000000034

Wachter, R., & Goldsmith, J. (2018, March 30). To Combat Physician Burnout and Improve Care, Fix the Electronic Health Record. Retrieved June 1, 2018, from https://hbr.org/2018/03/to-combat-physician-burnout-and-improve-care-fix-the-electronic-health-record

Wang, C., Dominici, F., Parmigiani, G., & Zigler, C. M. (2015). Accounting for Uncertainty in Confounder and Effect Modifier Selection when Estimating Average Causal Effects in Generalized Linear Models. *Biometrics*, *71*(3), 654–665. https://doi.org/10.1111/biom.12315

Wang, G., Jung, K., Winnenburg, R., & Shah, N. H. (2015). A method for systematic discovery of adverse drug events from clinical notes. *Journal of the American Medical Informatics Association : JAMIA*, *22*(6), 1196–1204. https://doi.org/10.1093/jamia/ocv102

Wang, X., Hripcsak, G., Markatou, M., & Friedman, C. (2009). Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. *Journal of the American Medical Informatics Association : JAMIA*, *16*(3), 328–337. https://doi.org/10.1197/jamia.M3028

Wasserman, L. (2013). *All of Statistics: A Concise Course in Statistical Inference*. Springer New York. Retrieved from https://books.google.com/books?id=qrcuBAAAQBAJ

Weinberg, C. R. (1992). *Sponsored by the Society for Epidemiologic Research Toward a Clearer Definition of Confounding*.

Widdows, D., & Cohen, T. (2015). Reasoning with Vectors: A Continuous Model for Fast Robust Inference. *Logic Journal of the IGPL*, *23*(2), 141–173. https://doi.org/10.1093/jigpal/jzu028

Woodward, J. (2016). The problem of variable choice. *Synthese*, *193*(4), 1047–1072. https://doi.org/10.1007/s11229-015-0810-5

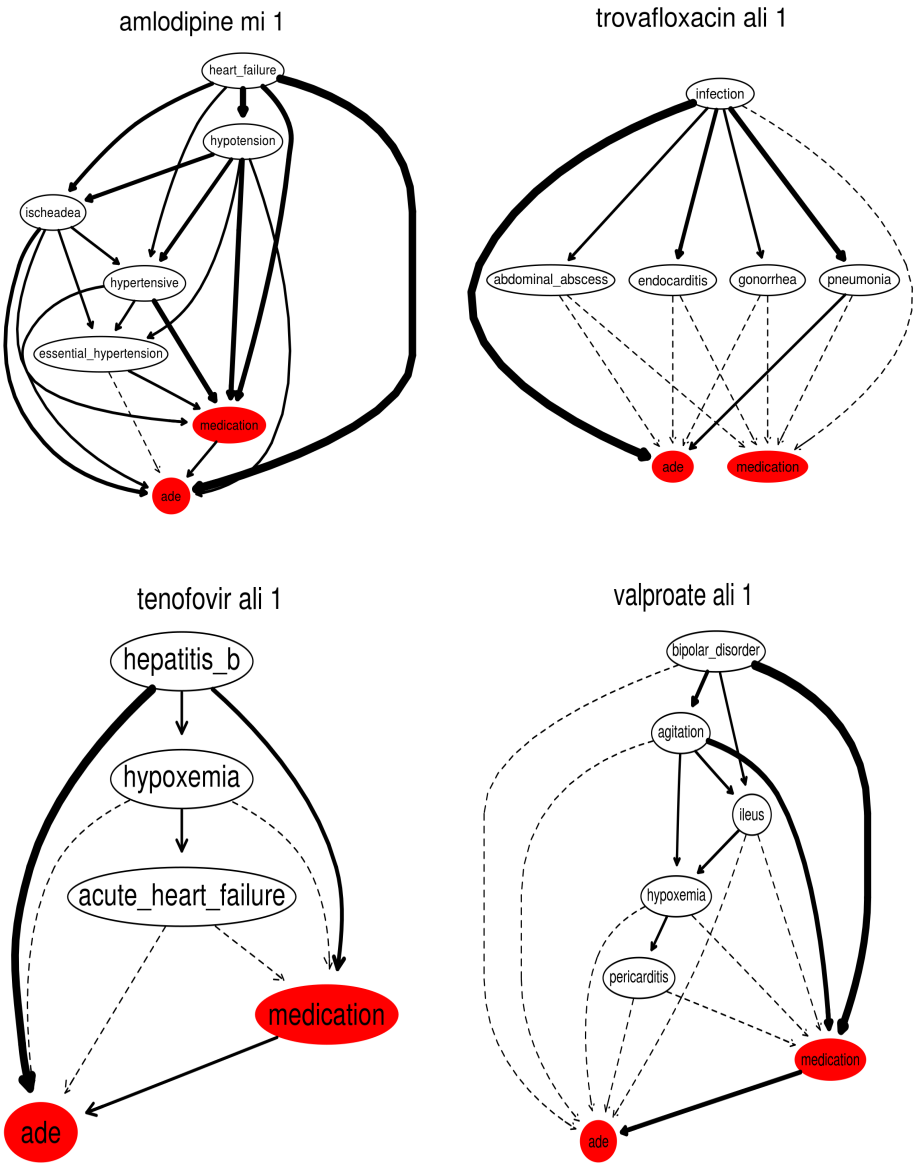
Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, *172*(16–17), 1873–1896. https://doi.org/10.1016/j.artint.2008.08.001

Zhang, J., & Spirtes, P. L. (2012). Strong Faithfulness and Uniform Consistency in Causal Inference. *ArXiv:1212.2506 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1212.2506

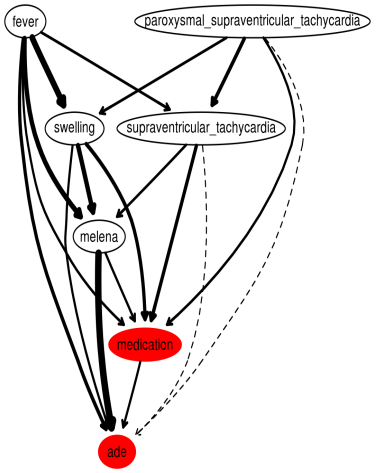
Zorych, I., Madigan, D., Ryan, P., & Bate, A. (2013). Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Statistical Methods in Medical Research*, *22*(1), 39–56. https://doi.org/10.1177/0962280211403602

Appendix: Causal Graph Examples

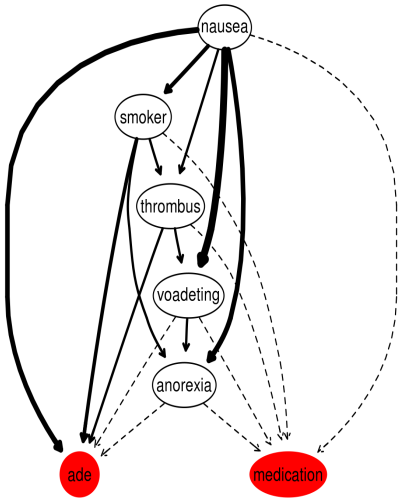
Stochastically selected graphs of causal models instantiated with EHR data.



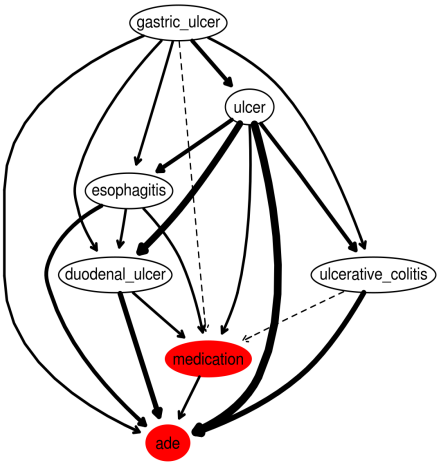
adenosine gib 0



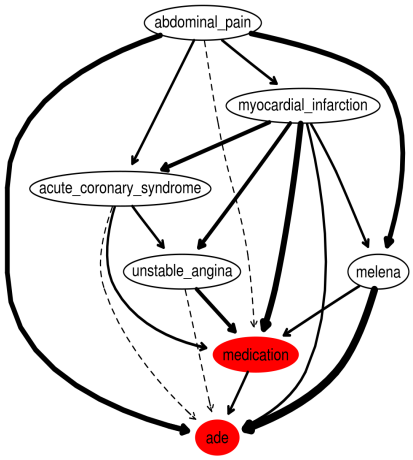
tetrahydrocannabinol mi 0



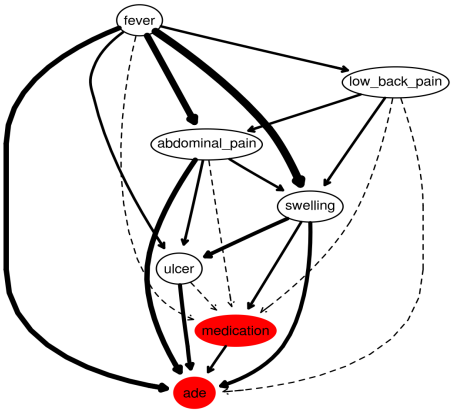
sucralfate gib 0



clopidogrel gib 1



indomethacin gib 1



clindamycin mi 0