EasyChair Preprint

№ 155

Using the Literature to Identify Confounders

Scott Malec

May 23, 2018

Using the Literature to Identify Confounders

Scott Malec, MLIS, MSIT*
School of Biomedical Informatics
University of Texas
Houston, TX 77030
scott.malec@uth.tmc.edu

Abstract

We introduce an approach to causal modeling that uses Literature-Based Discovery (LBD) to identify salient domain knowledge in observational data. Causal models represent a marriage between graph theory, probability, and domain knowledge. We hypothesize that the LBD paradigm can be applied to identify variables of interest for the automated construction of causal models of observational data, and that causal models thus informed will improve upon the performance of purely statistical techniques. We evaluated our hypothesis with a pharmacovigilance (PV) use case. In PV, the task is to discriminate between drug/side-effect signals and noise. We analyzed observational clinical data derived from electronic health records (EHR) and constructed causal models. We used logistic regression coefficients as our baseline and calculated estimated controlled direct effect from the LBD-informed causal models. Causal models improved upon unadjusted statistical models by 8.64% using Area under the Curve of the Receiver Operating Characteristic. Improving upon previous work in PV using EHR as the primary data source, our results establish the utility of the approach.

1 Introduction

In this study, we continue with our previous work refashioning the literature-based discovery (LBD) paradigm as a means to inform causal models for pharmacovigilance (PV), or the surveillance of post-marketing adverse drug events (ADEs) [1, 2]. We will address a limitation of our previous work by quantifying the effect of the endogenous variable (medication exposure) on the outcome of interest (ADE) [2].

Confounding is endemic to observational data. A confounder is present when an exogenous variable mutually influences both the predictor, or explanatory variable, and the outcome of interest [3]. For example, we may be interested in whether or not a drug causes gastrointestinal bleeding. To identify a confounder, we can search the literature for comorbidities that are treated by the drug which are also known to cause gastrointestinal bleeding, e.g., diabetes. As we have shown in our previous work, the identification of confounders can facilitate the process of **deconfounding**, (or "*screening off*" spurious associations from descriptive statistical correlation) in observational data, by imposing constraints from *a priori* domain knowledge on the topology of the causal graph [1, 2]. By incorporating confounders into causal models under a set of "vivid assumptions", one can perform experiments upon the resulting data generating model to test whether or not any influence from an explanatory variable becomes "blocked" [4]. **We hypothesize that causal models informed by LBD will improve upon the performance of purely statistical approaches.**

*Pre-Doctoral Fellow of the National Library of Medicine in Biomedical Informatics and Data Science.

2 Background

Adverse drug events impose a formidable onus upon health systems and individuals worldwide [5]. This danger serves as the primary impetus for the current study. After regulatory agencies such as the Food and Drug Administration (FDA) release a novel pharmaceutical therapy to market, these pharmaceuticals must be monitored. Clinicians and pharmaceutical companies submit reports of adverse events to spontaneous reporting systems (SRSs) such as FAERS in the United States and EudraVigilance in the E.U. However, these data have limitations, such as incomplete clinical information, under-reporting of side-effects, and selection bias. An important issue with SRSs is that they lack a denominator with which to calculate the prevalence of adverse events from the data alone. As a complement to data from SRSs, a current focus of attention among PV researchers is on the use of Electronic Health Record (EHR) data. Clinical notes can provide a plethora of detail of routine clinical practice. However, these data are not without additional challenges: inconsistent granularity of encoding, text processing overhead, and confounding.

Most PV work utilizes statistical methods (lasso shrinkage, meta-analysis) for the task of detecting drug-ADE signal from observational clinical data [6, 7]. Statistical analysis can only tell us that a correlation exists, not determine causality. As noted elsewhere, causal methods have been under-utilized in biomedicine [8]. A major hurdle to the adoption of causal modeling methods lies in the "identification" problem, or selection of relevant covariates, since this is labor intensive. Causal discovery at scale requires an automated method to populate these models. LBD provides the means to search for confounding variable candidates (CVCs) identified by the literature.

Causal discovery methods have been in existence since the late 1980s and represent a marriage between probability, graph theory, causal assumptions (faithfulness, causal Markov condition, absence of latent confounders), and domain knowledge [9, 3]. The modeling process takes place in two steps: first, represent anticipated inter-variable dependencies in terms of directed acyclic graph topology (nodes encode variables, edges dependencies); second, learn the parameters of the structural equations that quantify these relationships. Consider a data set **A** that consists of a set of random variables **X** and that is described by a directed acyclic graph **G**, where the Bayesian Network **B** = (**G**, **X**) and θ denotes the parameters of the global distribution of **X**, such that θ is *iid* with **X**, so that **B** = (**G**, θ) (and θ can denote the sufficient statistics of appropriate marginal and joint distributions given **A**, e.g. binomial if discrete, Gaussian if continuous). The structure and parameter learning process then can be decomposed into the following components [10]:

$$P(\mathbf{B}|\mathbf{A}) = P(\mathbf{G}, \theta|\mathbf{A}) = P(\mathbf{G}|\mathbf{A})P(\theta|\mathbf{G}, \mathbf{A}). \tag{1}$$

$P(\mathbf{G}|\mathbf{A})$ denotes the structure (topology) learning and can be further decomposed as follows:

$$P(\mathbf{G}|\mathbf{A}) = P(\mathbf{G})\,\mathrm{argmax}(\int P(\mathbf{A}|\mathbf{G}, \theta)P(\theta|\mathbf{G})d\theta) \tag{2}$$

where $P(\mathbf{G})$ represents the skeleton of the graph from domain knowledge as a prior.

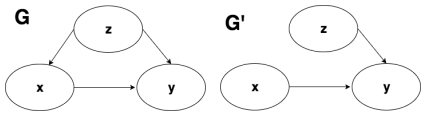

Figure 1: **G** and mutilated graph **G'**.

Given a set of random variables $\{x, y, z\} \in X$ such that $y \perp\!\!\!\perp x|z$ as in per **Figure 1**, where z is a confounder that influences both x and y. **G** factorizes as a the following joint (pre-intervention) probability distribution:

$$P(x, y, z) = P(z)P(x|z)P(y|x, z) \tag{3}$$

$$\frac{P(x, y, z)}{P(x|z)} = P(z)P(y|x, z) \tag{4}$$

To determine the direct effect of x on y, we will mutilate the graph by setting (randomizing) the values of x, such that the post-intervention distribution to reflect **G'** above can be denoted by the following truncated factorization ($P(x|z)$ is dropped as x becomes parentless):

$$P(z, y|do(x)) = P_{mutilated}(z)P_{mutilated}(y|x, z) = P(z)P(y|x, z) \tag{5}$$

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

$$P(z, y|do(x)) = \frac{P(x, y, z)}{P(x|z)} \tag{6}$$

Confirming our intuition, $P(x|z)$ will help us to estimate the effect of **do(x)**, i.e. fixing x's value to 1 and 0 as an "idealized experiment" (if given binary data) [10, 11]. We can perform adjustment by marginalizing over "z" [11]:

$$P(y|do(x)) = \sum_{z} P(z)P(y|x, z) \tag{7}$$

LBD was first developed by Don Swanson in the 1980s [12]. Historically, the target application of LBD has been to identify therapeutically useful relationships from publicly available knowledge. As we have shown previously for statistical models, the LBD paradigm is a promising candidate for this task of mapping aspects of extra-statistical domain knowledge to observational data [1]. Incorporating LBD-derived confounders into statistical models improved drug-ADE detection accuracy where the unadjusted signal had some predictive utility. Elsewhere, LBD methods have been utilized to assess the plausibility of drug-ADE associations [13]. In this study, we use LBD to identify covariates that we suspect will have graphs that are homomorphic with **G** in **Figures 1 and 2**.

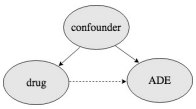

Figure 2: Classic confounder with "forking" directed edges.

3 Materials and Methods

To derive our data set, we used a reference set of curated drug-ADE associations that was developed by Ryan and his colleagues as a standard for evaluating PV methods [14]. This reference set includes 399 drug/ADE pairs and 4 ADEs with both positive (drug-ADE relationships supported by the literature and other sources, including package labeling events) and negative (drug-ADE relationships without support) control groups per ADE. The four ADEs are as follows: acute kidney injury (AKI), acute liver injury (ALI), gastrointestinal bleeding (GIB), and acute myocardial infarction (MI). These ADEs were chosen for their importance to PV and their impact on financial and personal cost. We refer interested readers to detailed descriptions of the pre-processing steps in our previous work [1, 2].

We extracted a corpus of approximately 2.2 million electronic health records (EHR) concerning outpatient encounters for 364,000 patients in the Houston metropolitan area between 2004-2012 from the UTHealths clinical data warehouse [15]. We used MedLEE, a clinical Natural Language Processing system, to normalize concepts in our EHR corpus [16]. Next, we then extracted the concepts with Apache Lucene for document-level co-occurrence statistics for convenience. From this index, we obtained document-by-concept arrays. Each concept (drug, ADE, or CVC) is persisted as a large sparse binary array. 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.

The publicly available SemRep NLP system was developed to identify and normalize relationships between concepts expressed in the biomedical literature, resulting in sets of semantic predications, each consisting of a pair of UMLS concepts connected through a predicate such as **TREATS**, **CAUSES** [17]. SemMedDB is a publicly-available database product that contains the SemRep output from processing of the entirety of MEDLINE. SemMedDB was used for accessing the biomedical literature. Domain knowledge is retrieved as triple stores: **ARGUMENT0 + PREDICATE + ARGUMENT1**. Such representations make domain knowledge amenable to computation.

We applied Predication-based Semantic Indexing (PSI) to SemMedDB to construct our knowledge base. Our LBD methods are discussed at length elsewhere [18, 1, 2]. In the present study, PSI is

used to facilitate rapid ranked order retrieval of concepts that fulfill semantic constraints through particular predicates [19]. We used the following DP to identify CVCs: drug **TREATS** confounder; confounder **CAUSES-INV** ADE. LBD yields not only covariates, but the skeleton of a graph, denoted by the factorization $P(\mathbf{G})$ as per **Equation 2**.

We used the hill climbing algorithm in the **bnlearn** R package [10] . Hill climbing recursively adds and subtracts directed edges until the Bayesian Information Criterion is minimized.

The core steps of our approach were as follows:

1. Query PSI vector space for confounders in ranked order of relevance.
2. Test each CVC for directed edges to both the drug and ADE using the clinical data.
3. Build causal models for each drug-ADE pair using the LBD-identified confounders.

For baseline scores, we used the coefficients from logistic regression. We performed parameter estimation using conditional probability query on the mutilated graph for each drug-ADE causal model, as per **Equation 7**. To evaluate performance, we calculated the Area Under the Receiver Operating Characteristic curve (AUROC) based on the ranked order of the scores.

4 Results

Parameter estimates from causal models improved performance over logistic regression for all four ADEs. Causal models improved upon unadjusted statistical models by 8.64% using Area under the Curve of the Receiver Operating Characteristic.

5 Discussion

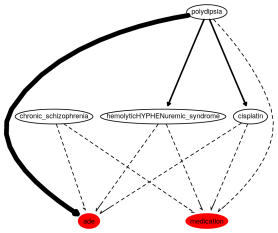

Figure 3: Causal graph for clozapine (- ctrl / AKI). Width indicates relationship strength.

This excursion into parameter estimation from interventions on observational data improved upon previous work in both statistical and causal modeling for EHR-based PV. This implies that our method is useful for screening off spurious associations. Causal models have additional advantages in providing visual explanations of the data generating processes that can account for patterns in large observational datasets. One question we hope to explore in future work is the extent to which interactions between confounders are themselves confounded, as per **Figure 3**. One limitation of the present study is that is was cross-sectional in nature, so granularity may be lost in exchange for simplicity. We aim to address this and other limitations with longitudinal patient-level analysis.

6 Conclusion

We have demonstrated that the feasability of estimating parameters from cross-sectional observational clinical data using a minimal set of confounders and have improved upon previous results in EHR-based PV. We suspect that our method could be useful for any field where observational clinical data is admissible and there exists a structured repository of causal knowledge.

Acknowledgments

This work was supported by the Brown Foundation, NIH NCATS grants UL1 TR000371 and UL1 TR001105, NLM R01-LM011563, and by a training fellowship from the Gulf Coast Consortia, on the NLM Training Program in Biomedical Informatics and Data Science T15 LM007093.

References

[1] S.A Malec, P Wei, E Bernstam, S Myneni, and T Cohen. Using literature to identify confounding variables in clinical observational data. *Proceedings of the AMIA Symposium, Chicago, 2016*, 24:118–173, 2016.

[2] S.A Malec, A. Gottlieb, E.V. Bernstam, , and T. Cohen. Using the literature to construct causal models for pharmacovigilance. *Data Mining Health Information, 2017, Wash, D.C.*, 2017.

[3] J. Pearl. *Causality: Models, Reasoning, Inference, 2nd Edition*. New York: Cambridge University Press, 2009.

[4] I. Shpitser, T. VanderWeele, and J.M. Robins. On the validity of covariate adjustment for estimating causal effects. *Arxiv*, March 2012.

[5] J. Aronson and J. Talbot. Stephens detection and evaluation of adverse drug reactions: principles and practice. 2012.

[6] Y. Li, H. Salmasian, S. Vilar, H. Chase, C. Friedman, and Y. Wei. A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records. *J Am Med Inform Assoc*, 21(2):308–14, 2014.

[7] Y. Li, R.B. Ryan, Y Wei, and C Friedman. A method to combine signals from spontaneous reporting systems and observational healthcare data to detect adverse drug reactions. *Drug Saf.*, 38(10):895–908, October 2015.

[8] S. Kleinberg and G. Hripcsak. A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics*, 44(6):1102 – 1112, 2011.

[9] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):67–72, 1992.

[10] M. Scutari and J.-B. Denis. *Bayesian Networks with Examples in R*. Chapman and Hall, Boca Raton, 2014. ISBN 978-1-4822-2558-7, 978-1-4822-2560-0.

[11] J. Pearl, M. Glymour, and N.P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, New York, 2016.

[12] D. Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives Biological Medicine*, 30(1):7–18, 2010.

[13] N. Shang, H. Xu, T. Rindfleisch, and T. Cohen. Identifying plausible adverse drug reactions using knowledge extracted from the literature. *J. Biomed. Inform.*, 52:293–310, 2014.

[14] P.B. Ryan, M.J. Schuemie, E. Welebob, J. Duke, S. Valentine, and A. Hartzema. Defining a reference set to support methodological research in drug safety. *Drug Saf*, 36:S33–47, 2013.

[15] University of Texas Health Science Center in Houston, School of Biomedical Informatics. UTHealth BIG. https://sbmi.uth.edu/uth-big/, 2016. [Online; accessed 22-Nov-2016].

[16] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Association*, 11(5):392–402, 2010.

[17] T. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477, 2003.

[18] T. Cohen, R.W. Schvaneveldt, and T.C. Rindflesch. Predication-based semantic indexing: Permutations as a means to encode predications in semantic space. In *AMIA Annual Symposium Proceedings*, volume 2009, page 114. American Medical Informatics Association, 2009.

[19] D. Hristovski, C. Friedman, T. Rindflesch, and B. Peterlin. The semantic vectors package: New algorithms and public tools for distributional semantics. In *AMIA Annual Symposium Proceedings*, pages 349–353, 2006.