Using the Literature to Construct Causal Models for Pharmacovigilance

Scott Malec [1], Assaf Gottlieb [1], Elmer V. Bernstam [1,2], Trevor Cohen [1]
[1] UTHealth School of Biomedical Informatics
[2] UTHealth Division of General Internal Medicine, McGovern Medical School

Abstract

Causal discovery methods provide a means to ascertain causal attribution from observational data. Causal modeling at scale requires a method to populate models with relevant domain knowledge. We propose to use the biomedical literature to perform feature selection for drug/adverse drug event (ADE) models with clinical observational data derived from electronic health records (EHR) as our primary input data source. We reason that spurious (non-causal) drug-ADE associations from co-occurrence-based analyses should diminish conditional on sets of validated confounders identified in the literature. To evaluate this hypothesis, we used a publicly available reference data set to test the proposed methodology with 4 ADEs and 399 drug-ADE pairs. We calculated baseline scores using the rank order regression coefficients each drug-ADE pair. We then identified confounding variable candidates for each drug-ADE pair using relationship constraints based on normalized predicates to search knowledge extracted from the literature in the publicly available SemMedDB repository. To determine eligibility for inclusion, we checked whether or not there were directed edges pointing to both the drug and the ADE. Finally, we tested whether associations from co-occurrence in the clinical data are diminished conditional on sets of permutations of confounders identified in the literature. Confounder yield rate was ~ 90%, indicating that our method successfully identified confounders in the observational data. Causal models attained aggregate performance improvements of ~ 0.07 area under the curve and reduced the False Discovery Rate from 0.50 to 0.38 over purely statistical models using unadjusted logistic regression.

Introduction

A body of evidence that demonstrates the utility of causal modeling is currently lacking in many core areas of biomedical informatics. In this paper, we demonstrate the utility of causal modeling methods at the task of performing EHR-based pharmacovigilance using domain knowledge derived from the literature to resolve the problem of identifying potential confounders. A confounder is present when an exogenous variable affects both a predictor and the outcome of interest. For a familiar example, one may wish to understand the association between smoking and lung cancer when there may be an underlying genetic predilection for both nicotine addiction and cancer **(1)**. As confounding variables may not be known at the outset (the "identification problem"), we propose a scalable method for confounding variable discovery (CVD). Our approach leverages domain knowledge extracted from the literature as a means to identify covariates with which to populate causal models.

Pharmacovigilance and FDA Adverse Event Reporting System (FAERS). The burden that adverse drug events place on our health system and the danger that such events pose to individuals motivate the current paper **(2)(3)**. After regulatory agencies release a pharmaceutical therapy to the market, these pharmaceuticals must be monitored, since not all conclusions can be transported from a test to a target population in randomized controlled trials **(4)**. The discipline concerned with the post-marketing surveillance of pharmaceuticals is known as pharmacovigilance (PV). Clinicians and pharmaceutical companies submit reports of adverse events to spontaneous reporting systems (SRSs) such as FAERS **(5)**. However, these data have limitations, such as incomplete clinical information, under-reporting of side-effects, and confounding bias **(6)**. An important issue is that there is no denominator with which to estimate the prevalence of the side-effects within the data. A current focus of attention is the use of Electronic Health Record (EHR) data, which provide a rich but imperfect record of routine clinical practice, as a complement to data from SRSs. However, these data present additional challenges, such as the inconsistent granularity of encoding, the presence of free text, and the prevalence of confounding variables **(6)(7)**.

Causal Discovery Methods. Most PV work utilizes disproportionality metrics such as the odds ratio, Gamma Poisson Shrinkage, and lasso shrinkage regression among others for the task of detecting drug-ADE "signal" from observational data **(7)(8)**. More recently, PV researchers have demonstrated the promise of meta-analytic techniques, wherein data from multiple sources are combined, so as to mitigate bias from any single source **(9)**. However, statistical analysis can only tell us that a correlation exists, not the direction of influence.

Causal discovery methods (in their current form as "causal Bayesian networks", where the "Bayesian" aspect encapsulates conditional dependencies indicative of causality) have been employed for almost three decades since the invention of the PC algorithm **(10)(11).** While such methods have been applied in the fields of oncology **(12)**,

neuroscience **(13)**, diagnostics (**14**), and epidemiology **(15)**, the adoption of such methods within PV has been sporadic, manifesting primarily in the form of renewed interest in instrumental variables* **(16)**.

Given suitable input, causal (and most graphical) modeling takes place over two steps: first, represent anticipated inter-variable dependencies in terms of directed acyclic graph topology (with variables as nodes and dependencies as edges); second, learn the parameters of the structural equations that quantify these dependencies. However, causal discovery at scale requires an automated method to identify relevant background knowledge with which to populate this graph. Since we wish to "explain away" non-causal associations without introducing any additional noise, we seek to identify (and validate) confounders. More specifically, we need to search for confounding variable candidates (CVCs) indicated by the literature, which are known to influence both the predictor and the outcome, as illustrated in Figure 1. Once we have identified a set of CVCs, we need to validate these against clinical data with graphical criteria. By including such validated CVCs in causal models, we should improve our ability to discriminate between drug safety signal and noise.

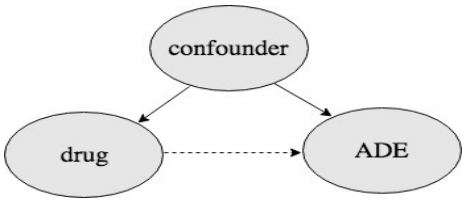

Figure 1. *Relationship between a confounder, drug/predictor, and ADE/outcome of interest. Note the directed edge from the confounder, as a parent node, to its two children. This graph topology signifies that the confounder influences the likelihood of both the drug and the ADE of occurring.*

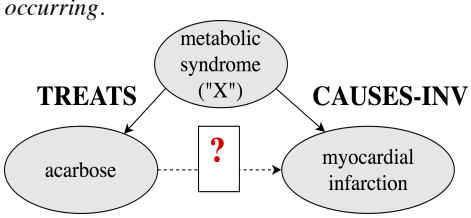

Figure 2. "**CAUSES-INV**" *may be read as "**is caused by**." Literature-Based Discovery identifies a co-morbidity that is known to cause the ADE and also increases the likelihood of a drug exposure (as a consequence of undergoing treatment). Here, the* "**TREATS**" *predicate can be causally interpreted as "to cause exposure."*

Literature-Based Discovery. **Literature-Based Discovery** (LBD) was first developed by Don Swanson as a means to discover therapeutically useful relationships from public knowledge **(17)**. Most automated implementations of Swanson's approach involve identifying implicit relationships using concept co-occurrence (e.g. drug-to-gene-to- disease) as a means of revealing novel therapeutic applications **(18).** More recently, LBD researchers have explored the idea of using semantic relations (extracted from the literature using natural language processing, or NLP) to constrain the search space of relevant associations **(19)**. For example, in the following pattern of semantic relationships "**drug TREATS 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, as in Figure 2. These patterns of relationships are known as "discovery patterns", or DPs **(18)**. As we have shown previously for statistical models, the Literature-Based Discovery (LBD) paradigm is a promising candidate for this task of mapping aspects of extra-statistical domain knowledge to observational data **(21)**. Incorporating LBD-derived cofounders into statistical models improved drug-ADE detection accuracy for side effects where the unadjusted signal had some predictive utility. LBD methods have also been applied to estimate the plausibility of drug-ADE relationships **(21)(22)**. In this paper, we refashion LBD methods to identify CVCs with which to populate causal models.

Materials and Methods

Processing EHR data. With IRB approval, we extracted a corpus of ~2.2 million electronic health records (EHR) concerning outpatient encounters for ~ 364,000 patients in the Houston metropolitan area between 2004-2012 from the UTHealth's clinical data warehouse **(23)**. MedLEE, a widely-used clinical Natural Language Processing (NLP) system, was used to identify and normalize concepts of interest in our EHR collection. MedLEE has been shown to perform accurately on clinical notes with recall of 0.77, and precision: ~ 0.89 for the task of extracting clinical concepts **(24)**. In addition to identifying types of entities ("problems" and "drugs"), MedLEE encodes each extracted concept with a concept unique identifier (CUIs) from the Unified Medical Language System (UMLS) **(25)**. We then extracted the concepts with Apache Lucene for document-level co-occurrence statistics. From the Lucene index, we obtained concept-by-concept arrays for each concept (drug, ADE, or CVC). Each concept (drug, ADE, or CVC) is then persisted and processed as a large sparse binary array (these arrays are subsequently used as input for the causal modeling algorithm, discussed below). A value of 1 or 0 represents the presence or absence, respectively, of that concept within a document in the index of clinical data. The arrays for a drug, ADE, and their confounders provide input matrices for the causal methods to be described.

---

* An instrumental variable, by contrast with a confounder, influences the cause, but not the outcome, except through the error term of the cause. For more discussion see: Chu TJ, Scheines R, Spirtes PL. Semi-instrumental variables: a test for instrument admissibility. *Eprint arXiv*:1301.2261; 2013.

Representing Knowledge from the Literature. The purpose of SemRep, a publicly available NLP system developed at the US National Library of Medicine (NLM), is 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 by a predicate such as "**TREATS**" or "**CAUSES**" **(26).** SemRep is optimized for precision, and Kilicoglu *et al*. report this at 0.745, with recall of 0.640, based on a human annotated gold standard **(27)**. SemMedDB, used as our knowledge base for accessing the biomedical literature, is a publicly-available NLM product that contains the output of SemRep processing of the entirety of MEDLINE. All extracted predications can be retrieved from a MySQL database in the following form: **ARGUMENT_0 + PREDICATE + ARGUMENT_1 (28)**. Such representations make domain knowledge amenable to computation.

Constructing a knowledge Base. To construct our knowledge base, we applied Predication-based Semantic Indexing (PSI) **(18)(29)** to SemMedDB, our knowledge repository. 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. Our hypothesis is that by refashioning the LBD paradigm to identify CVCs in the literature, we can improve upon the performance for detecting drug-ADE relationships over standard statistical models. Details on the theoretical and methodological underpinnings of PSI have been described at length elsewhere **(30)**. For the current work, we used 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 **(28)**. A 48,000-dimensional binary vector PSI space was built using the Semantic Vectors package (version 5.9) **(31)**. We excluded a small number of predicates that indicate negation (e.g., **DOES_NOT_TREAT**), as well as terms ("stop words") with occurrence $\geq$ 500,000.

Discovery Pattern for CVD. The Semantic Vectors package **(31)** provides an interface that permits searching PSI spaces for concepts that populate particular predicate pathways, which we used to identify the most strongly associated CVCs for each drug/ADE pair. We used the following DP to identify CVCs: **"**drug **TREATS** confounder; confounder **CAUSES-INV** ADE.**"** Given acarbose (used to treat diabetes mellitus) and myocardial infarction, "metabolic disorder" was one of the results (Figure 2). The order in which these covariates are retrieved reflect 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, angiodysplasia. Further down the list, confounders become less specific: peptic ulcers, diabetes.

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 **(10).** Depending on one's choice of algorithm, input may be discrete, continuous, or mixed. We used the discrete version of the Fast Greedy Equivalence Search (FGeS) that is included with TETRAD with default parameters **(32)**. FGeS recursively adds and then subtracts directed edges between nodes until the Bayesian Information Criterion is minimized. Output consists of a Markov equivalence class or family of graphs which encode plausible dependency relationships given these data. An equivalence class may have undirected edges. However, background knowledge can be used to orient these edges, as causal predicates have inherent directionality. The resulting graph structure should be similar to Figures 1 and 2, with directed edges pointing from the confounder to both the drug and ADE ("confounder inclusion criterion").

Reference Set and Data Collection. 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 **(33)**. 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 mapped and expanded drug/ADE synonyms to make the EHR data amenable to additional processing. We used RxNorm for drug synonyms at the clinical drug level and we assigned the reference set's Observational Medical Outcomes Partnership (OMOP) ADE to UMLS CUIs **(34)(35)**.

Combinatory Expansion of CVCs. If we have a set of three CVCs, denoted {A, B, C}, this will result in seven unique combinations: A, B, C, AB, BC, AC, and ABC (AB and BA are equivalent). We evaluated all of these, because we did not know which combination of CVCs will cause spurious directed edges from the drug to the ADE to vanish.

Analysis of Observational Clinical Data. The core steps of our approach were as follows, as per Figure 3:

1. **We queried PSI vector space for the top 50 CVCs in the literature**, result set in ranked order of relevance.
2. **We used TETRAD/FGeS to validate CVCs**, testing each CVC for directed edges to both the drug and ADE (graphical criteria) using the clinical data. We stopped after obtaining five valid CVCs for each pair.
3. **We built causal models using all unique permutations of the tested LBD-identified confounders**.

Evaluation Procedure. To evaluate the performance of our method, we calculated the Area Under the Receiver Operating Characteristic curve (AUROC), which is widely used to compare 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 predicted relationships. For baseline scores, we used the coefficients from logistic regression without adjustment with literature-derived confounding variables. To score causal models, we divide the fraction of directed edges out of the total number of permutations from the drug to the ADE for each drug/ADE pair. For example, if 7 directed edges resulted from 31 unique permutations (of 5 validated confounders), then the score for that drug-ADE pair would be 7/31, or 0.2258. We reason that the proportion of correct directed edges will be higher for the group of positive drug-ADE pairs than for the negative drug-ADE pairs in the reference data set. In other words, we are testing whether associations from co-occurrence in the clinical data of a drug-ADE pair are diminished conditional on sets of CVC permutations.

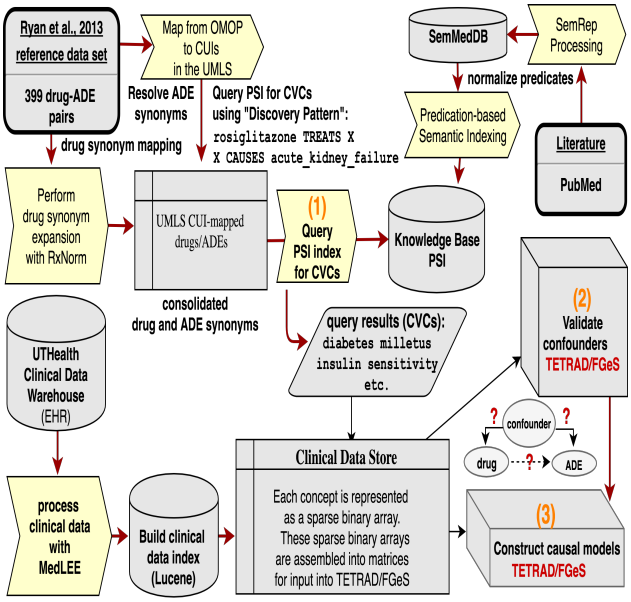

Figure 3. *Workflow. Numbers in* **bold orange** *=core method steps.*

Results and Discussion

Analysis of Results. 1915 out of 2124 total tested CVCs (for 399 drug-ADE pairs each with 5 confounders) were both present in the clinical notes and passed validation, so the CVC yield rate was 90%, indicating that LBD can identify confounders in clinical notes. The overall aggregate performance boost that approached ~ 0.07[†] over statistical models confirms our hypothesis that the identification problem of confounding can be partially resolved by using the literature to inform feature selection (an area that we have addressed earlier with using LBD for statistical models) **(21)**. Our 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 our method requires a strong baseline to be effective.

Table 1. Pairs=number of test/control drug-ADE from the reference data set. AUROCs are calculated from coefficients from logistic regression ("Baseline") and the proportion of directed edges ("Causal Models").

ADE	Pairs (+/-)	Baseline	Causal Models
AKI	24 / 64	0.5547	**0.6598** (95% CI: 0.5251-0.7946)
ALI	81 / 37	0.4957	**0.5449** (95% CI: 0.4385-0.6513)
GIB	24 / 67	0.5643	**0.6912** (95% CI: 0.5182-0.7828)
MI	35 / 64	0.4946	**0.56** (95% CI: 0.4292-0.6549)
ALL	164 / 232	0.504	**0.571** (95% CI: 0.5157-0.6262)

Practical Significance. Better detection methods in PV, if implemented, hold promise for improving public health and safety. For example, enhanced methods of drug-ADE detection in observational clinical data could facilitate the prioritization of drug-ADE 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 large impact.

In our previously published work, we use LBD methods for feature selection of confounders to adjust for plausible confounding with the same set of clinical data **(21)**. In that work, we used both single predicate (**CAUSES-INV** and **PREDISPOSES-INV**) and dual predicate (**TREATS+COEXISTS_WITH-INV**) DPs. With the single predicate DPs, the influence from the confounder was only exerted explicitly in the literature on the outcome and can be thought of as an alternative etiology. We found that the dual predicate DPs performed the best overall with a modest ~ 0.02 AUROC improvement over unadjusted models. Our analysis was that the dual predicate patterns identify CVCs that influence both predictor and outcome, fulfilling the graphical criteria **(11)**. In the present article, we used a different

[†] Overall results were slightly higher (~ 0.08) when calculated using only pairs where drug occurrence ≥ 100 or 500.

dual predicate pattern and our results affirm our previous observation with the bonus that causal models with validated "true" confounders improved upon the performance of adjusted dual predicate statistical models. We reduced the False Discovery Rate (FDR), where $\text{FDR} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$, from 0.5 to 0.38 with causal models. Although the performance increase is substantive **(7)** and better in general adjusted standard statistical models (for purely EHR-based PV), 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 performance alone **(9)**.

Limitations of (and Lessons from) the Current Approach. One limitation is that our search for confounders was relatively shallow, having commenced confounder permutations after reaching five validated CVCs. Computational demand scales with the number of confounders. Five validated CVCs result in 31 permutations per drug-ADE pair. Increasing to 10 CVCs would leave 1032 permutations – which can take 12 hours to run on a Linux workstation with 64GB RAM and 8 Xeon CPUs. We chose 5 CVCs because we could collect results for all drug-ADE pairs for a single ADE within a reasonable amount of time (7-8 hours overnight). An additional limitation arises from our available EHR data, which may not have a sufficient number of drug-ADE co-occurrences, as the performance from analyses of FAERS data is usually better than results from any EHR data source **(9)**. One perplexing problem arose from three drug-ADE pairs for myocardial infarction for which the proposed method could not identify any confounders (these are not included in Table 1). We suspect that confounders for myocardial infarction identified by our method, e.g., hypertension, coronary arteriosclerosis, metabolic syndrome, could have helped if incorporated into these models. Note that the discovery pattern that we deployed 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 our system's performance by missing potential confounders **(27)**. Another consideration is that reference data sets, however essential to the scientific enterprise, may not be perfectly accurate, as knowledge about drugs and their side-effects accumulates. In future work, we plan on estimating parameters ("arc strengths") of causal models that include all confounders to avoid the inefficient confounder permutation procedure.

Conclusion

We have developed a scalable CVD method that identifies "true confounders" by leveraging existing NLP tools and knowledge resources that improves the signal to noise ratio in observational clinical data in causal models. Our method results in notable (~ 0.07 AUROC) overall performance gains for the task of re-identifying drug-ADE pairs from observational clinical data in comparison with unadjusted statistical models from a reference data set, thus demonstrating the utility of causal modeling methods for EHR-based PV. Finally, we suspect that our method may be applicable in other areas of biomedicine for which observational data is admissible as input.

Acknowledgments

We would like to thank (in no specific order) Frank Manion, Rory Lettvin, Lex Frieden, Sándor Darányi, Swaroop Gantela, Richard Boyce, Harry Hochheiser, Clark Glymour, and the anonymous reviewers for feedback. 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. Stolley PD. When genius errs: R.A. Fisher and the lung cancer controversy. Am J Epidemiol. 1991 Mar 1;133(5):416-425; discussion 426-428.
2. Talbot JCC, Aronson JK, editors. Stephens' detection and evaluation of adverse drug reactions: principles and practice. 6th ed. Chichester, West Sussex, UK: John Wiley & Sons; 2012. 732 p.
3. National Action Plan for Adverse Drug Event Prevention [Internet]. [cited 2017 Jul 22]. Available from: https://health.gov/hcq/pdfs/ade-action-plan-508c.pdf https://health.gov/hcq/pdfs/ade-action-plan-508c.pdf
4. Cartwright N. Are RCTs the Gold Standard? BioSocieties. 2007 Mar;2(1):11–20.
5. Federal Drug Administration Adverse Event Reporting System [Internet]. [cited 2017 Jul 21]. Available from: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/
6. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013 Aug;51(8 Suppl 3):S30-37.
7. Li Y, Salmasian H, Vilar S, Chase H, Friedman C, Wei Y. A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records. J Am Med Inform Assoc JAMIA. 2014 Apr;21(2):308–14.
8. DuMouchel W, Ryan PB, Schuemie MJ, Madigan D. Evaluation of disproportionality safety signaling applied to healthcare databases. Drug Saf. 2013 Oct;36 Suppl 1:S123-132.

9. Li Y, Ryan PB, Wei Y, Friedman C. A Method to Combine Signals from Spontaneous Reporting Systems and Observational Healthcare Data to Detect Adverse Drug Reactions. Drug Saf. 2015 Oct;38(10):895–908.
10. Scheines R, Spirtes P, Glymour C, Meek C, Richardson T. The TETRAD Project: Constraint Based Aids to Causal Model Specification. Multivar Behav Res. 1998 Jan 1;33(1):65–117.
11. Pearl J. Causality: Models, Reasoning, and Inference [Internet]. 2nd ed. Cambridge: Cambridge University Press; 2009 [cited 2017 Jul 21]. Available from: http://ebooks.cambridge.org/ref/id/CBO9780511803161
12. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, et al. An Integrated Approach to Uncover Drivers of Cancer. Cell. 2010 Dec;143(6):1005–17.
13. Ramsey JD, Hanson SJ, Hanson C, Halchenko YO, Poldrack RA, Glymour C. Six problems for causal inference from fMRI. NeuroImage. 2010 Jan 15;49(2):1545–58.
14. Cooper, Gregory F. NESTOR: A Computer-Based Medical Diagnostic Aid that Integrates Causal and Probabilistic Knowledge. [Internet]. [Palo Alto, California]: Stanford University; 1984 [cited 2017 Jul 22]. Available from: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA152046
15. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiol Camb Mass. 1999 Jan;10(1):37–48.
16. Ertefaie A, Small DS, Flory JH, Hennessy S. A tutorial on the use of instrumental variables in pharmacoepidemiology. Pharmacoepidemiol Drug Saf. 2017 Apr;26(4):357–67.
17. Smalheiser NR, Swanson DR. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Comput Methods Programs Biomed. 1998 Nov;57(3):149–53.
18. Smalheiser NR. Literature-based discovery: Beyond the ABCs. J Am Soc Inf Sci Technol. 2012 Feb;63(2):218–24.
19. Cohen T, Widdows D, Schvaneveldt RW, Davies P, Rindflesch TC. Discovering discovery patterns with Predication-based Semantic Indexing. J Biomed Inform. 2012 Dec;45(6):1049–65.
20. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA Annu Symp Proc AMIA Symp. 2006;349–53.
21. Malec SA, Wei P, Xu H, Bernstam EV, Myneni S, Cohen T. Literature-Based Discovery of Confounding in Observational Clinical Data. AMIA Annu Symp Proc AMIA Symp. 2016;2016:1920–9.
22. Shang N, Xu H, Rindflesch TC, Cohen T. Identifying plausible adverse drug reactions using knowledge extracted from the literature. J Biomed Inform. 2014 Dec;52:293–310.
23. UTHealth BIG. [Internet]. [cited 2017 Jul 21]. Available from: http://redcap.uth.tmc.edu/cdwstats/stats-mpi.htm
24. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc JAMIA. 2004 Oct;11(5):392–402.
25. The Unified Medical Language System (UMLS). [Internet]. [cited 2017 Jul 21]. Available from: http://www.nlm.nih.gov/research/umls/
26. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec;36(6):462–77.
27. Kilicoglu H, Fiszman M, Rosemblat G, Marimpietri S, Rindflesch T. Arguments of Nominals in Semantic Interpretation of Biomedical Text. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing [Internet]. Uppsala, Sweden: Association for Computational Linguistics; 2010. p. 46–54. Available from: http://www.aclweb.org/anthology/W10-1906
28. SemMedDB [Internet]. [cited 2017 Jul 21]. Available from: http://skr3.nlm.nih.gov/SemMedDB/
29. Cohen T, Schvaneveldt RW, Rindflesch TC. Predication-based semantic indexing: permutations as a means to encode predications in semantic space. AMIA Annu Symp Proc AMIA Symp. 2009 Nov 14;2009:114–8.
30. Widdows D, Cohen T. Reasoning with Vectors: A Continuous Model for Fast Robust Inference. Log J IGPL. 2015 Oct;23(2):141–73.
31. Semantic Vectors [Internet]. [cited 2017 Jul 21]. Available from: https://github.com/semanticvectors/semanticvectors
32. Ogarrio JM, Spirtes P, Ramsey J. A Hybrid Causal Search Algorithm for Latent Variable Models. JMLR Workshop Conf Proc. 2016 Aug;52:368–79.
33. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf. 2013 Oct;36 Suppl 1:S33-47.
34. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc JAMIA. 2011 Aug;18(4):441–8.
35. Observational Medical Outcomes Partnership (OMOP) [Internet]. [cited 2017 Jul 21]. Available from: http://omop.org