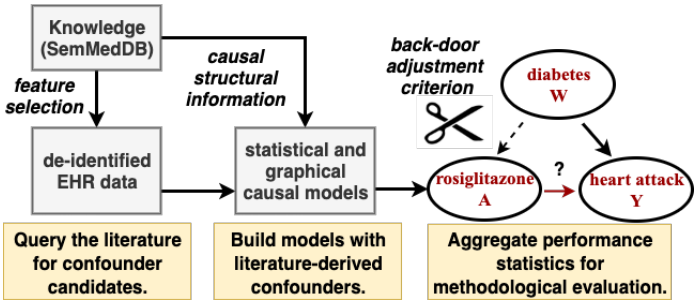
Graphical Abstract

Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance

Scott A. Malec,Elmer V. Bernstam,Peng Wei,Richard D. Boyce,Trevor Cohen



Highlights

Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance

Scott A. Malec,Elmer V. Bernstam,Peng Wei,Richard D. Boyce,Trevor Cohen

- Access to causal background knowledge is required for causal learning to scale to large datasets.
- We introduce a framework for identifying confounders to enhance causal inference from EHR.
- We search computable knowledge for indications TREATED BY the drug that CAUSE the outcome.
- Literature-derived confounders reduce confounding bias in EHR data.
- Structured knowledge helps interpret and explain data captured in clinical narratives.

Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance

Scott A. Malec[a,*], Elmer V. Bernstam[c], Peng Wei[b], Richard D. Boyce[a] and Trevor Cohen[d]

[a] University of Pittsburgh School of Medicine, Department of Biomedical Informatics, Pittsburgh, PA
[b] The University of Texas MD Anderson Cancer Center, Department of Biostatistics, Houston, TX
[c] University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, TX
[d] University of Washington, Department of Biomedical Informatics and Medical Education, Seattle, WA

ARTICLE INFO

Keywords:
Confounding bias
Confounder selection
Causal inference
Electronic health records
Pharmacovigilance

ABSTRACT

Introduction: Confounding bias threatens the reliability of observational studies and poses a significant scientific challenge. This paper introduces a framework for identifying confounding factors by exploiting literature-derived computable knowledge. In previous work, we have shown that semantic constraint search over computable knowledge extracted from the literature can be useful for reducing confounding bias in statistical models of EHR-derived observational clinical data. We hypothesize that adjustment sets of literature-derived confounders could also improve causal inference.

Methods: We introduce two methods (semantic vectors and string-based confounder search) that query the literature for potential confounders and use this information to build models from EHR-derived data to more accurately estimate causal effects. These methods search SemMedDB for indications TREATED BY the drug that is also known to CAUSE the adverse event. For evaluation, we attempt to rediscover associations in a publicly available reference dataset containing expected pairwise relationships between drugs and adverse events from empirical data derived from a corpus of 2.2M EHR-derived clinical notes. For our knowledge-base, we use SemMedDB, a database of computable knowledge mined from the biomedical literature. Using standard adjustment and causal inference procedures on dichotomous drug exposures, confounders, and adverse event outcomes, varying numbers of literature-derived confounders are combined with EHR data to predict and estimate causal effects in light of the literature-derived confounders. We then compare the performance of the new methods with naive (χ^2, reporting odds ratio) measures of association.

Results and Conclusions: Logistic regression with ten vector space-derived confounders achieved the most improvement with AUROC of 0.628 (95% CI: [0.556,0.720]), compared with baseline χ^2 0.507 (95% CI: [0.431,0.583]). Bias reduction was improved more often in modeling methods using more rather than less information, and using semantic vector rather than string-based search. We found computable knowledge useful for improving automated causal inference, and identified opportunities for further improvement, including a role for adjudicating literature-derived confounders by subject matter experts.

1. Introduction

This paper introduces a framework for automating causal inference from observational data by exploiting computable knowledge mined from the literature. Observational data, or data collected in non-randomized settings, contain a wealth of information for biomedical research. Such data are particularly important in settings where randomized controlled trials are infeasible, such as is the case with drug safety research, where a major goal is to prioritize associations empirically. Associations that are more likely to be genuinely causative should be prioritized for review. Unfortunately, confounding, endemic to such data, can induce misleading, non-causal associations. While the research question defines the exposure and outcome variables, the decision of which covariates to adjust for falls to the investigator. Classical criteria mandate, including all covariates correlating significantly with the exposure and the outcome [1]. Unfortunately, such an approach can introduce covariates that amplify bias rather than reduce it [2, 3, 4].

Recently, researchers have enumerated criteria for identifying adjustment sets emphasizing the role of causal knowledge for identifying covariates to reduce bias [5]. Unfortunately, it is infeasible to rely solely on human experts, since such expertise cannot scale to analyze large datasets or all available human knowledge. Consequently, the problem of how to access knowledge has been noted as an open research question [6]. Fortunately, knowledge resources and methods exist that could be useful for guiding the selection of confounders. The Semantic MEDLINE database, or SemMedDB, is one such resource [7]. The information in SemMedDB consists of pairs of biomedical entities, or concepts, connected by normalized predicates, e.g., "aspirin TREATS headache." We introduce and test methodological variants that combine computable knowledge with observational data. The idea is to use existing knowledge from previous discoveries to catalyze causal inference. More specifically, our methods query literature-derived computable knowledge to identify potential confounders for incorporation into statistical and graphical causal models. We then use these models for performing statistical and causal inference from data extracted from a corpus of EHR-derived free-text clinical narratives.

*Corresponding author
sam413@pitt.edu (S.A. Malec)
ORCID(s): 0000-0003-1696-1781 (S.A. Malec)

This paper introduces two methods for accessing background knowledge: string-based and semantic vector-based search. These methods are qualitatively distinct in how they store, represent, and retrieve information. We also explore whether or not knowledge representation affects performance, given how concepts are prioritized by the confounder search methods in search results. We also ask how varying the amount of literature-derived information affects bias reduction. Finally, we investigate whether or not computable knowledge is useful for informing causal inference. We compare using effect estimates from literature-informed graphical causal models with standard adjustment procedures with logistic regression.

2. Background

Adverse events arising from exposure to pharmaceutical drugs are both common and costly. With nearly half of the US population having been prescribed a prescription drug in any given month, the vast number of drug exposures drives the high prevalence of adverse events [8]. The annual financial cost of adverse event-related morbidity in the United States was estimated at 528.4 billion in 2016 alone [9], while another study noted that 16.88% of hospitalized patients experience an adverse drug reaction [10]. An adverse drug reaction is defined as an "appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medical product" [11], and is distinguished from other adverse events by the demonstration of a causal link to a drug. The discipline that helps adjudicate causal links between drugs and harmful side-effects is known as pharmacovigilance. Pharmacovigilance encompasses the development of procedures for collecting, summarizing, monitoring, detecting, and reviewing associations between drug exposures and health outcomes in general and adverse events in particular [12]. Causal links are established by the gradual accretion of evidence from observational data and by the relative strength of mechanistic explanations justifying biological plausibility [13].

Randomized controlled trials (RCTs) are the gold standard of demonstrating causality in medicine. However, the ability of RCTs to detect adverse drug reactions is limited by their size, cost, short duration, and ethical considerations [14, 15]. Accordingly, in the post-marketing phase after regulatory approval, regulatory bodies such as the Food and Drug Administration (FDA) in the United States and the European Medicines Agency in the European Union must rely largely on empirical data from non-randomized sources to prioritize drug safety signals for review. Traditionally, the primary data source for pharmacovigilance research has been spontaneous reporting systems such as the U.S. Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) [16, 17]. Spontaneous reporting systems aggregate data from patients, clinicians, and pharmaceutical companies. Unfortunately, these data present critical deficiencies, including lack of context, missing data, and no population-level denominator to estimate the prevalence of any association [18, 19].

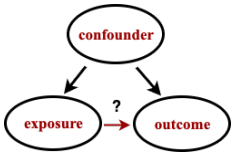


Figure 1: Illustration of confounding.

Consequently, there is a pressing need to advance methods that can more reliably detect adverse drug reactions from observational data. To addresses these and shortcomings, researchers have turned to other sources of empirical evidence such as social media [20, 21], claims [22, 23], and EHR data [24, 25, 26, 27], the focus of the present study. The FDA's ongoing Sentinel Initiative facilitates the federated search of structured data from EHR across the United States [28]. However, structured data may only provide an incomplete picture. Data embedded in unstructured free-text clinical narratives in EHR systems are known to contain a wealth of contextual information concerning routine clinical practice often absent from the fixed content fields in structured data [29]. The embedded contextual information may include indications of "temporal relations, severity and degree modifiers, causal connections, clinical explanations, and rationale" [30].

However, since such data were not derived from a controlled setting, variables of interest are subject to sources of influence outside of investigator control. The reliability of analytic conclusions from such data is highly sensitive to the quality of the assumptions used to analyze them [31, 32], including assumptions concerning how to adjust for sources of systematic bias, such as confounding bias. Confounding bias is induced when both the exposure and outcome of interest share a common cause that is not controlled for analytically. Confounders influence both the likelihood of the exposure and the outcome (Figure 1) [33].

If one knows which variables to control for, i.e., which variables are common causes, it may be possible to distinguish genuine causal relationships from confounder-induced associations [34]. Clearly, substantive, extra-statistical *a priori* subject-specific knowledge is critical for reliable causal inference [5, 32]. Many approaches for inferring causality have been developed to evade the requirement of subject-specific knowledge since it has not been clear how to access contextually relevant causal knowledge automatically. For example, an approach described in [35] performs meta-analytic techniques to cancel out bias from individual sources of data by combining EHR data with FAERS. Other approaches impute a pseudo-variable to absorb residual confounding between measured variables [36, 37, 38].

Computable knowledge mined from the biomedical literature in the form of structured information could be useful for identifying relevant biomedical concepts such as confounders given an exposure and a health outcome of interest. The focus of our ongoing research assumes the auspicious existence of computable knowledge mined from the litera-

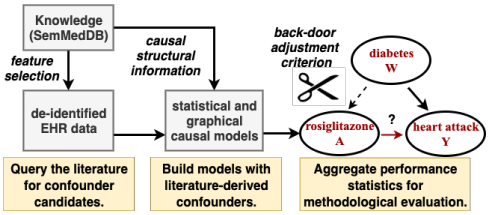


Figure 2: Schema for our literature-informed causal modeling toolkit.

ture to help select covariates to facilitate causal inference.

2.1. Components of a causal inference toolkit

In the next section, we identify the components necessary to automate feature selection for causal inference. We describe lessons learned from our prior work in this area and provide definitions required to understand the methods described, beginning with what we mean by causal inference. Figure 2 illustrates the components of our toolkit to access background knowledge from the literature to catalyze causal inference.

2.1.1. Causal inference

The goal of causal inference is to estimate an expectation (or population-level average) that quantifies the extent to which an intervention (such as a drug exposure) affects an outcome of interest (such as an adverse event). Because we can only observe one outcome at any time point, potential outcomes are important. We cannot directly observe counterfactual outcomes.

However, it may be possible to estimate the expectations of causal effects under certain conditions, even from observational data, a task referred to in the literature as causal inference. Causal inference is achieved by calculating the mean difference in potential outcomes across exposed and unexposed subgroups with similar pre-exposure characteristics, or, here, confounders. Such an operation can be expressed mathematically as the $do(.)$ operator. For a binary (non-dose-dependent) treatment $\mathbf{A}$ with effect $\mathbf{Y}$, the average treatment effect (ATE) can be estimated as a contrast between the exposed and unexposed groups given confounders $\mathbf{W}$. The lowercase a denotes the variable being fixed to set value. The ATE (henceforth denoted Δ) is expressed as the following equation:

$$E[Y|do(A)] = E[Y|a = 1, W] - E[Y|a = 0, W] \quad (1)$$

This equation defines the adjusted treatment effect, and adjusts for confounders $\mathbf{W}$. However, in order for Equation 1 to estimate causal effects reliably, these confounders must first be identified. Consequently, the next important piece is to figure out what features or variables $\mathbf{W}$ for which to adjust. The backdoor-criterion stipulates that causal inference is possible if a set of covariates $\mathbf{W}$ can be found that block all "backdoor" paths from $\mathbf{A}$ to $\mathbf{Y}$, then bias from confounding can be reduced or eliminated [34].

In the next section, we discuss knowledge resources and methods that can reason over large volumes of computable knowledge extracted from the published biomedical literature to provide convenient access to identify contextually relevant knowledge.

2.1.2. Literature-based discovery

Taking the biomedical literature for its input, the objective of literature-based discovery research (henceforth, LBD) is to reveal meaningful, but implicit connections between biomedical entities of interest [39]. The late Don Swanson pioneered LBD in his seminal work, discovering the potential of fish oil to treat Raynaud's syndrome [40], an example we will return to below. Much early work in LBD focused on investigating the strength of association between concepts and exploiting insights from information retrieval into concept co-occurrence patterns.

To further constrain the results returned by LBD systems, researchers have developed ways to exploit information concerning the nature of the relationships between biomedical concepts. Revisiting Swanson's original example with fish oil, researchers noticed that it was useful to pay attention not only to concepts themselves, but to information concerning *relationships between concepts*. For example, certain drug concepts are known to treat diseases and so the drug and the disease are related through a TREATS relationship. Other drug exposures may be associated with harmful outcomes, and so to CAUSE adverse events. In other cases, the connection between such concepts can be inferred indirectly from other mechanistic relationships.

To demonstrate how such knowledge can be useful, consider the example of Raynaud's disease, from Don Swanson's work [40]. Raynaud's is a circulatory disorder manifesting in skin discoloration affecting the extremities. Blood viscosity (loosely defined as "thickness" or "stickiness") is implicated as a mechanism in Raynaud's, [40, 41], with **increasing** viscosity in cold conditions thought to impede circulation to the peripheries causing them to appear white or blue, the primary symptom of Raynaud's. Swanson noticed that fish oil could have the effect of **decreasing** blood viscosity, thus leading to his therapeutic hypothesis that fish oil can treat Raynaud's, by countering the mechanisms of Raynaud's. Extrapolating from Swanson's example, researchers paid attention to how concepts were related to each other (e.g., A increases B, B decreases C in the example above), and were able to manually extrapolate useful patterns, called discovery patterns, that could generate biologically plausible hypotheses for novel therapies [42, 43]. Discovery patterns define semantic constraints for identifying concepts that relate to each other in particular ways [42]. We conjecture that discovery patterns may be useful for identifying which variables fulfill the backdoor criterion. Then, one could then enhance statistical and causal inference from observational data with literature-derived confounders.

2.1.3. SemMedDB - a causal knowledge resource

SemMedDB is a knowledge database deployed extensively in biomedical research and developed at the US National Library of Medicine. The knowledge contained in

SemMedDB consists of subject-predicate-object triples (or predications) extracted from titles and abstracts in MEDLINE [44] using the SemRep biomedical NLP system [44, 45, 46]. SemRep can be thought of as a machine reading utility for transforming biomedical literature into computable knowledge. Employing a rule-based syntactic parser enriched with domain knowledge, SemRep first uses the high precision MetaMap [47] (e.g. estimated at 83% in [45]) biomedical concept tagger to recognize biomedical entities (or concepts) in the Unified Medical Language System (UMLS). Next, SemRep categorizes how the recognized concepts are associated given a fixed set of normalized, pre-specified predicates (with thirty core predicate types) corresponding to relations of biomedical interest, e.g. TREATS, CAUSES, STIMULATES, INHIBITS [48, 45]. For example, the predication "ibuprofen-TREATS-inflammation_disorder" was extracted by SemRep from the source text: "Ibuprofen has gained widespread acceptance for the treatment of rheumatoid arthritis and other inflammatory disorders." To access knowledge, we utilize a semantic vector-based knowledge representation scheme which we use in the current paper and describe in the next subsection. To access knowledge, we utilize a semantic vector-based knowledge representation scheme which we use in the current paper and describe in the next subsection.

2.1.4. Predication-based Semantic Indexing (PSI)

Predication-based semantic indexing, or PSI, defines a scheme for encoding and performing approximate inference over large volumes of computable knowledge [49]. The basic premise of distributional semantics is that terms that appear in similar contexts tend to have similar meanings [50]. By encoding the contexts in which terms appear, methods of distributional semantics provide a natural way to extrapolate their semantics from a corpus.

PSI's approach to distributional semantics derives from the Random Indexing (RI) paradigm, wherein a semantic vector for each term is created as the (possibly weighted) sum of randomly instantiated vectors - which we will refer to as *elemental vectors* as they are not altered during training - representing the contexts in which it occurs. [51, 52]. To adapt the RI approach to the task of encoding concept-relation-concept triples extracted from the literature by SemRep, PSI adopts an approach that is characteristic of a class of representational frameworks collectively known as vector-symbolic architectures [53, 54, 55], or VSAs. VSAs were developed in response to a debate (with one side forcefully articulated in [56]) concerning the ability to represent hierarchical structures in connectionist models of cognition.

In RI, elemental vectors of high dimensionality (of dimensionality ≥ 1000), with a small number of non-zero values (≥ 10) which are set to either +1 or -1 at random, are generated for each context, where a context might represent a document or the presence of some other term in proximity to the term to be represented. The resulting vectors have a high probability of being approximately orthogonal, which ensures that each context has a distinct pattern that acts as a fingerprint for it. Alternatively, and as is the case with the current research, high-dimensional binary vectors (of dimensionality ≥ 10000 *bits*) can be employed as a unit of representation. In this case, vectors are initialized with an equal number of 0s and 1s, assigned at random. While these vectors are not sparse in the "mostly zero" sense, they retain the desirable property of approximate orthogonality [57], with orthogonality defined as a hamming distance of half the vector dimensionality.

VSAs provide an additional mechanism for encoding structured information by using what is known as a "binding operator", a nomenclature that suggests its application as a means to bind variables to values within the connectionist representational paradigm. The binding operator is an invertible operator that combines to vector representations a and b, to form a third c that is dissimilar to its component vectors. As this operation is invertible, a can be recovered from c using b, and vice-versa. Thus, a value can be retrieved from a variable bound to facilitate the encoding and retrieval of structured knowledge.

PSI is implemented in the open-source, and publicly available `Semantic Vectors` package written in Java [58], and in the context of SemMedDB is applied to concepts which are defined by normalized biomedical entities in the UMLS hierarchy, rather than the original terms as encountered in the biomedical source text.

The resulting models have been applied to a range of biomedical problems (as reviewed in [59]). As input, PSI accepts subject-predicate-object triples, or semantic predications, and transforms that input into a searchable vector space that may then be used to retrieve structured, computable knowledge, similarly to a search engine.

To build a knowledge base using PSI, the process begins by making an elemental vector, denoted $E(.)$, for each unique concept and each unique relation, in the corpus. PSI constructs semantic vectors, denoted $S(.)$, by superposing, denoted $+=$, the bound product of the elemental vectors of each relation and concept with which it co-occurs. The binding operation, denoted $\otimes$ is invertible, and its inverse is denoted $\oslash$. PSI encodes the semantic predication "aspirin TREATS headaches" like so:

$$S(\text{aspirin}) \mathrel{+}= E(\text{TREATS}) \otimes E(\text{headaches}) \tag{2}$$

Note that for directional predications, the inverse predicate is also encoded. For a predicate like "TREATS," the meaning would be "TREATED BY," here denoted $\text{TREATS}_{\text{INV}}$, thus:

$$S(\text{headaches}) \mathrel{+}= E(\text{TREATS}_{\text{INV}}) \otimes E(\text{aspirin}) \tag{3}$$

However, no such representation is made for predicates lacking "direction" such as COEXISTS_WITH and ASSOCIATED_WITH, as these predicates are their own inverse.

PSI can be used to perform approximate inference over the knowledge it has encoded after a semantic space has been constructed from the semantic predications. The `Semantic Vectors` package [60, 61] implements a VSA-based query

language that facilitates queries to the resulting vector space, which underlies the `EpiphaNet` system for literature-based discovery [62]. This query language provides a calculus for using vector algebra to perform logical inference over the computable knowledge contained within the vector space, enabling discovery pattern-based search.

In this calculus, $P(.)$ denotes an elemental predicate vector, $S(.)$ denotes a semantic vector for a concept, as noted earlier, and $E(.)$ denotes an elemental vectors for a concept. $*$ and $+$ denote binding (and its inverse, which are the same operation in binary vector implementations of PSI) and superposition, respectively.

PSI can also be used to generate discovery patterns automatically [63, 64]. Using `Semantic Vectors` syntax, we can both query the search space using discovery patterns, and generate discovery patterns from sets of paired cue terms. In [63], PSI was used to recapitulate the discovery pattern manually constructed in [43].

While generating patterns automatically for this work, we observed that certain discovery patterns produced results suggestive of common cause confounders of potential usefulness for statistical and causal inference. We noticed that a recurring discovery pattern is given a drug and an outcome was TREATS + COEXISTS_WITH, which suggests that rather than causing an outcome, a drug may treat a related comorbidity.

Confounders that lie along the path of an inferred discovery pattern can be retrieved by constructing queries using the same `Semantic Vectors` syntax, as illustrated by the example confounders identified in Table 1 from querying a PSI space for concepts that relate to the (drug) allopurinol and adverse event (AE) acute liver failure in particular ways. Note that the first two confounder discovery patterns (which have only one predicate) in Table 1 only retrieve concepts that are causally related to the outcome.

Table 1
This table illustrates concepts retrieved using discovery pattern search for the drug allopurinol and the adverse event acute liver failure. "AE" = adverse event. "-INV" = the inverse relation, i.e., "caused by."

Discovery pattern	Sample concepts retrieved
P(CAUSES−INV)$*S$(AE)	transplantation, embolism
P(PREDISPOSES−INV)$*S$(AE)	transplantation, embolism
S(drug)$*P$(TREATS)+ S(AE)$*P$(COEXISTS_WITH)	pericarditis, gout kidney failure

In previous work [26], we tested the extent to which literature-derived confounders could be used to accurately distinguish known causally related drug/adverse event pairs from other drug/event pairs with no known causal relationship by adjusting statistical models (multiple variable logistic regression) of data embedded in free-text clinical narrative [26]. Using the discovery patterns enumerated in Table 1, our goal was to see if including literature-derived covariates suggestive of confounding could reduce bias in data derived from free-text clinical narratives extracted from a large (de-identified) corpus of EHR data. For methodological evaluation, we used a publicly available reference dataset [65], containing labels drug/adverse event pairs about expected relationships, including negative control pairs for which no relationship is known to exist. Next, we integrated up to ten literature derived-covariates into statistical models of EHR data using multiple logistic regression [26].

We define performance as the ability of our methodological variants to mitigate confounding. We measure performance by building literature-informed models instantiated with EHR data and comparing these models with naive estimates of association such as reporting odds ratio and χ^2.

To summarize performance, we calculated Area under the ROC (AUROC) from the ranked order of coefficients from the literature-informed logistic regression models and compared these with their χ^2 baselines. Including literature-identified covariates resulted in a modest overall performance improvement of +.03 AUROC, depending on the statistical power of the available evidence used as input. A key finding was that the dual predicate discovery pattern TREATS+COEXISTS_WITH provided the most substantial performance improvement compared with single predicate discovery patterns. We reasoned that the dual predicate discovery pattern was better able to reduce confounding in the aggregate because it captures more information about both the exposure and outcome mechanisms, whereas single predicate discovery patterns only captured information about outcome mechanisms. This finding guides the choice of discovery patterns we have used in subsequent work, including the present paper.

In a follow-up study employing the same reference dataset, we again used PSI to identify causally relevant confounders to populate graphical causal models [66]. Graphical causal models represent variables as nodes and causal relationships as directed edges. We hypothesized that the structure learning algorithm would predict fewer causal edges in light of the literature-derived confounders for the negative controls than for the positive control relationships in the reference dataset. We used the causal semantics of the predicates to orient the directed edges in causal graphs. We used the TREATS+CAUSES discovery pattern to identify indications that are treated by the exposures treated, and that were also noted to cause the adverse drug reactions. We picked TREATS+CAUSES, since drugs that are prescribed for an indication would likely not be taken were it not for the indication for which they were prescribed. The top-ranked literature-derived confounder candidates were then incorporated into graphical causal models. To learn graph structure, we employed the Fast Greedy Equivalence Search algorithm (FGeS) [67] implemented in the TETRAD causal discovery system [68, 69] with default algorithm hyperparameters. FGeS is a causal structure learning algorithm that works by stochastically adding and subtracting edges until the graph's fit for the observed data is optimized. Each drug/adverse event pair was given a score determined by the ratio of causal edges between the exposure and the outcome in the presence

of all possible unique perturbations of five literature-derived confounders. Improvements in the order of +0.08 AUROC over baselines were noted from this experiment.

2.2. The aim of the present study

The present study documents the current evolution of our framework for using computable knowledge extracted from the literature to facilitate more reliable causal inference from observational clinical data by reducing confounding bias. In our previous work, what was not clear is the extent to which the representation scheme affected the quality of the confounders and subsequently the performance of models. To probe this and other questions, we tested the following hypotheses:

- **[H1]**: that (overall) literature-informed models will reduce confounding bias in models of EHR-derived observational data and thereby improve causal inference from these data; *[premise: integrating confounders should reduce confounding bias]*
- **[H2]**: that incorporating more literature-derived confounders will improve performance over models with fewer such confounders; *[premise: models with fewer confounders may result in omitted variable bias]* and
- **[H3]**: that semantic vector-based discovery pattern confounder search (which compactly encode a vast array of information) will improve upon string-based search; *[premise: a compact representation incorporating global knowledge of causal mechanisms should better prioritize information]*, and finally
- **[H4]**: that computable knowledge is useful for informing causal inference. *[premise: estimates from causal graphs will reduce bias, not just variance.]*

This paper builds upon an active research program for performing inference across large volumes of knowledge. The primary contributions of this paper are to highlight how background knowledge can be used to 1.) elucidate specific confounding factors and 2.) facilitate causal inference from observational clinical data in a practical setting - that of drug safety.

3. Materials and Methods

In this section, we introduce the knowledge resources, the primary EHR-derived empirical data, and the tools and methods used to exploit computable knowledge to identify potential confounders. Our evaluation method was to compare the relative performance of baseline measures of association (ROR, and χ^2) with various adjusted statistical and literature-informed causal models. Encoding concept mentions discrete dichotomous (binary) variables and the labels from the reference dataset as ground truth. Our evaluation rests on the following assumptions:

- we expected effect estimates to be greater in magnitude for true positive drug/adverse event pairs than for the negative controls in the reference dataset; and
- we expected the causal effect estimates of the negative controls should approach zero.

Different drug/adverse event pairs can be expected to have a range of effect sizes: the true effect sizes of interest (such as one could collect under randomization) would range across varying intervals.

Working under the above assumptions, we measured performance by calculating Area under the Curve of the receiver-operating characteristic (AUROCs), Area under the Precision and Recall Curve (AUPRC), and Mean Average Precision at K (MAP-K) from the ranked regression coefficients and effect estimates given the ground truth expected labels from the reference dataset. The software and more extensive information and models developed for this paper are publicly available on the causalSemantics GitHub repository.

3.1. Extracting and representing clinical narrative

Following IRB approval and a data usage agreement, we obtained permission to use the same data as in our prior studies [26, 66] from the University of Texas Health Science Center clinical data warehouse [70, 71]. These data included a large random sample of 2.2 million free-text clinical narratives recorded during outpatient encounters involving approximately 364,000 individual patients during the years 2004 and 2012 in the Houston metropolitan area.

To reveal data in the clinical narratives for downstream analysis, the corpus of EHR data was processed using the MedLEE clinical natural language processing (NLP) system [72]. MedLEE encodes each concept it recognizes with a concept unique identifier (CUI) in the UMLS in machine-readable format [73, 48]. MedLEE can identify clinical concepts accurately from clinical notes, with a recall of 0.77 and a precision of 0.89 [72].

To extract document-level concept co-occurrence statistics from the MedLEE output, we next created an index using Apache Lucene [74] of normalized concepts mined from the MedLEE-processed corpus of clinical narratives. Next, we extracted document-by-concept binary arrays for each concept identified in the MedLEE output. We then store the resulting binary arrays for each concept recognized by the MedLEE NLP parser in compressed files stored locally on disk. These binary arrays represent whether a concept was mentioned in a particular document specific to the Lucene index. We utilize these binary arrays as our source of empirical data for our inference procedures.

3.2. Reference dataset

We used the popular reference dataset compiled by the Observational Medical Outcomes Partnership (OMOP) for performing methodological evaluation of novel drug safety methods [65]. The OMOP reference dataset includes 399 drug/adverse event pairs for four clinically important adverse events. To consolidate evidence which otherwise may have been diluted across synonyms, we used the UMLS metathesaurus to map between synonyms of the adverse events, and applied RxNorm mapping for synonym expansion at the

clinical drug ingredient level. For example, the generic concept of ibuprofen is encoded with a UMLS concept unique identifier (CUI) string of C0020740, while the specific concept that refers to a brand-name instance of Advil Ibuprofen Caplets is C0305170. We then applied these mappings to the EHR data. We used RxNorm to map from the more specific concept to the generic identifier of the pharmaceutical ingredient.

Since the OMOP reference dataset was published initially, varying degrees of accumulating evidence have cast doubt on the negative control status of certain drug/adverse event pairs. Hauben et al. published a list of mislabeled false negatives in the reference dataset [75] of negative control drug/adverse event pairs that have been implicated in adverse events from case reports, the literature, and pre-clinical studies. Correcting for the mislabeled false negatives noted by Hauben reduced the number of pairs for comparative evaluation. The number of pairs was reduced still further by factoring in limitations of the available empirical evidence.

We also appraised the statistical power of the available data to establish inclusion criteria about which drug/adverse event pairs to analyze. Peduzzi et al. studied the relationship between "events per variable," the accuracy of variance, and type I and II error [76]. Peduzzi found that variables with fewer than ten events per variable are unlikely to contribute to the methodological evaluation. As per the study of Peduzzi, we constrained which biomedical entities or concepts (drug exposures, adverse events, or confounders) to ten or more co-mentions. We have reported the number of drug/adverse event pairs that were compared in Table 2 (in parentheses), along with the number of drug/adverse event pairs in the original reference dataset (not in parentheses).

Table 2
This table presents the drug counts for each adverse event in the OMOP reference dataset. The number of drug/adverse event pairs was reduced by excluding misclassified pairs published by Hauben [75] and by the limited power of the available data in the EHR itself. The number in parentheses reflects the actual count of drug/adverse event pairs that we analyzed.

Adverse Event Type	+ Case	− Ctrl	Total
Acute kidney failure	(10) 24	(9) 64	(19) 88
Acute liver failure	(43) 81	(11) 37	(54) 118
Acute myocardial infarction	(15) 36	(23) 66	(38) 102
Gastrointestinal hemorrhage	(20) 24	(32) 67	(52) 91
Total	(88) 164	(75) 235	(163) 399

Preparing SemMedDB

We downloaded and imported the latest release (version 40) of SemMedDB into a local instance of the MySQL relational database system. This version contains 97 972 561 semantic predications extraced from 29 137 782 MEDLINE titles and abstracts.

We performed several operations upon it to tailor the information it contains for the requirements of this study. For example, if a treatment is known to cause an adverse event, physicians may avoid that treatment to eliminate the potential for undesirable outcomes. To emulate what knowledge was publicly available when the reference dataset was published (2013), we excluded predicates deriving from publications after December 31st, 2012. We also removed terms (stopwords) that occur ≥ 500,000 times, or were considered to be uninformative, e.g., patients, rattus norvegicus.

3.3. Searching SemMedDB for confounders

We developed and compared two variant methods for identifying a set of potential confounders by searching computable knowledge mined from the literature (SemMedDB). Both methods apply semantic constraint search using the `TREATS+CAUSES` discovery pattern, but rely on distinct knowledge representation frameworks, which we refer to henceforth as "string-based" and "semantic vector-based". To evaluate the effect of adding different amounts of literature-derived confounders, we used arbitrary a lower threshold of five and an upper threshold of ten for the number of confounders candidates.

3.3.1. String-based confounder search

Our first method is implemented in structured query language (SQL) and directly queries the predications table of the SemMedDB relational database. Each query takes a drug and an adverse event (called "focal concepts") and applies the `TREATS+CAUSES` discovery pattern search. The SQL query consists of two sub-queries - the first to obtain indications the drug treats and the second to obtain the indications that also cause the adverse event. The result set should contain a list of potential confounders that fulfill both of these semantic constraints. To find the best subset of confounders (that are associated with both the exposure and outcome), we developed a score to rank the confounders by the strength of their support as confounders in the literature. To score confounders, we calculated the product of the counts for each confounder given the number of citations from each arm of the discovery pattern query (the `TREATS` arm and the `CAUSES` arm). Results were next ranked in descending order of this product score. To screen out potential errors from machine reading, concepts with less than two mentions were excluded from the result set.

3.3.2. Semantic vector-based confounder search

Our second method transforms SemMedDB using a distributional representation scheme called PSI. Using the truncated version of SemMedDB described above as input, we derived a binary PSI space with 32,000 dimensions (in bits). We used inverse document frequency weighting to adjust for frequently occurring but uninformative predications. After the PSI model was trained, we then queried the resulting PSI space to identify confounder candidates for each drug/adverse event pair with discovery pattern search, retrieving the nearest neighboring elemental vector to the following composite query:

$$S(\text{rosiglitazone}) * P(\text{TREATS}) +$$

$$S(\text{myocardial_infarction}) * P(\text{CAUSES} - \text{INV})$$

After querying for confounders (up to set threshold of five or ten) that occur in the EHR data at least ten times with both the exposure and the outcome variables (following the heuristic of Peduzzi et al [76]), we next construct an input matrix with which to construct statistical and graphical models.

3.4. Assembling models from knowledge and data

As input, concept-by-observation matrices were constructed for each drug/adverse event in the reference dataset. In these matrices, each column represents a concept and each row represents the co-mentions of concepts extracted from the clinical narrative describing a patient's visit.

3.4.1. Literature-informed regression modeling

We applied off-the-shelf multiple variable logistic regression to the EHR data, adjusting for the literature-derived confounders, where Y = outcome (the ADR), A = exposure (drug), W = the set of confounders, with the Greek letters α, $\{\beta, \gamma\}$ and ϵ representing the intercept, regression coefficients, and an error term, respectively:

$$logit\{prob(Y = 1)\} = \alpha + \beta A + \sum_{i=1}^{k} (\gamma W_i) \tag{4}$$

In this paper, we use logistic regression as a comparator method to the exact causal inference method introduced in the next subsection. Regression estimates provide "guardrails" on the more advanced methods by providing a predictive check on our adjustments, since regression usually does a decent job at adjustment. For further discussion on the relationship between causal effect estimation and regression coefficients, see Chapter 6 in [77].

3.4.2. Literature-informed graphical causal modeling

To construct graphical causal models, we used the **bnlearn** R package [78]. The **bnlearn** package allows the user to incorporate variables and to define the relationships between variables. We exploited structural information from the literature to create "white lists" (lists of required edges) and "black lists" (lists of prohibited edges) to orient those edges. The white lists contain mandatory labeled edges between each of the confounders and the drug and adverse event, while black lists forbid effects from causing drug exposures.

We applied the Max-Min Hill-Climbing (MMHC) algorithm, first described by Tsamardinos et al. [79] and implemented in **bnlearn** [78] with default hyperparameter settings. MMHC is a hybrid structure learning algorithm that first uses constraint-based search to learn the dependency structure of a graph using the Max-Min Parents and Children algorithm. Next, it orients edges of the graph using the hill-climbing score-based search algorithm that optimizes the Bayesian information criterion score locally to find a structure that best fits the data and background knowledge. Next, we applied the maximum likelihood estimation (MLE) procedure within **bnlearn** to find the configuration of weights associated with each edge in the graph that is most likely given the observational data and the graph structure.

Once the structure and the parameters quantifying the edge strengths have been learned, the model is ready to answer questions of interest. We now describe how we used a classical exact causal inference procedure to estimate effects using back-door adjustment.

To estimate causal effects, we applied the junction tree algorithm, as implemented with default settings in the R package **gRain** [80]. The junction tree algorithm is a non-parametric estimation method that efficiently computes posterior probabilities by transforming the DAG into a tree structure which propagates updated values using the sum-product method across the graph.

The junction tree method is reasonably efficient because the graph is sparse (unsaturated) and all calculations are local. Next, we query the resulting object by telling it to "listen" to the adverse event node in the graph, and by fixing the value of the exposure/treatment to "1" and then to "0", and then subtracting the difference to obtain the ATE (Δ) (see Equation 1). More details about the derivation of the junction tree algorithm may be found in [81, 82].

3.5. Overview of the evaluation framework

We evaluated our literature-informed confounding variable identification framework by aggregating performance statistics across the following methodological variants:

- **more (literature-derived) information versus less**
- **string-based vs. semantic vector-based search**
- **regression (β) versus exact inference (Δ)**

The steps for evaluating our framework are outlined below and illustrated in Figure 3:

1. Query the literature for confounders using either string-based or semantic vector-based literature search.
2. Determine the eligibility for the inclusion of each confounder candidate in the order of its retrieval.
3. Build models (multiple logistic regression and graphical causal models) incorporating varying numbers of literature-derived confounders and using them for prediction and deriving effect estimates from the EHR-derived empirical data.

We defined performance in terms of the ability to correctly classify drug/adverse event pairs. Since the effect estimates of the negative controls should tend towards zero, we reason that the labeled pairs in the reference set will help us use the negative controls to diagnose and inform how well our methods are performing. We summarized performance by computing the area under the receiver-operating characteristic curve (AUROC), and the area under the precision-Recallrecall Curvecurve (AUPRC) - a method for summarizing the results best suited for when there is class label imbalance, and the mean average precision at K (MAP-K) - a method to describe the performance of a subset of predictions with the highest estimated probability -comes from

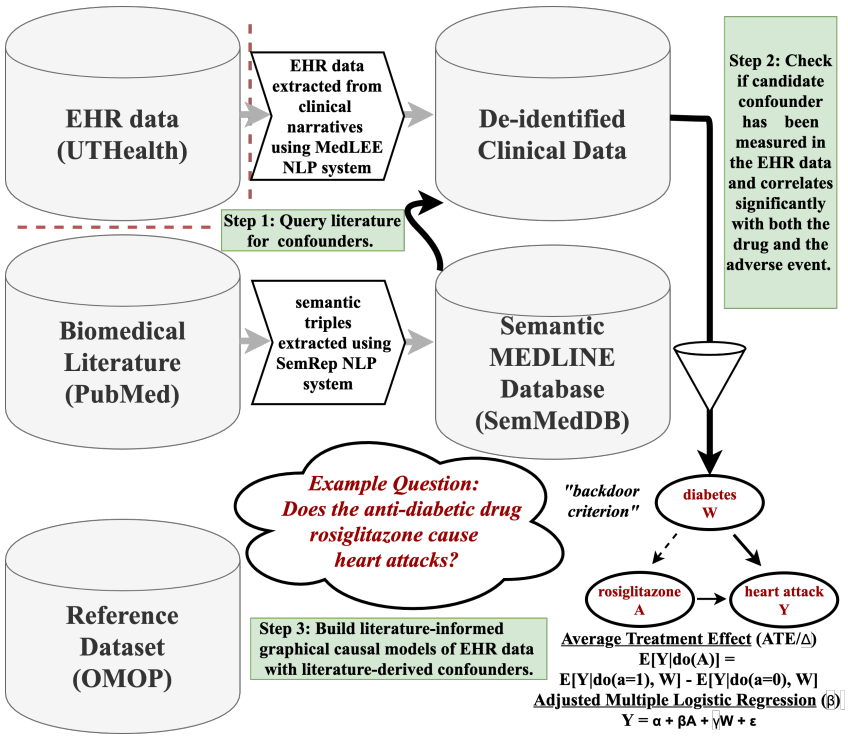


Figure 3: Illustration of our literature-informed modeling framework.

the following statistics: **baselines**: [χ^2, reporting odds ratio (ROR)] and **modeling scores**: the coefficients from performing multiple logistic regression (β) and average treatment effects (ATE) from the graphical causal models (Δ).

Finally, to obtain insight into overall performance, we needed to aggregate scores across adverse event types, albeit at the expense of fine detail. We applied the following weighting procedure since each adverse event has a different underlying prevalence in the sampled population. First, we calculated the mean effect estimates (whether β or Δ) and then to divided the estimates for each drug by that mean. The idea was to estimate a global AUROC based on the normalized per adverse event scores, which when divided by the mean were rendered comparable. The weighted scores were then combined into overall weighted metrics.

The PostgreSQL relational database system and various R statistical packages were used to analyze the data in this study. The list below enumerates that software packages used for this study: **R base** version 3.6, **gRain** version 1.3-3: exact causal inference [80], **RPostgreSQL** version 0.6-2: library for R connectivity with Postgres relational database, **pROC** version 1.16.1: ROC curves, **PRROC**: Precision-Recall curves, **tidyverse** version 1.3.0: data manipulation [83], and **bnlearn** version 4.5: graphical modeling and parameter estimation [84, 78]. We also used `Semantic Vectors` package version 5.9 [60, 61] (running Oracle Java 1.8.0_231).

4. Results

We begin by breaking down the results in terms of each performance metric and try to tease out what these are telling us about the strengths and weaknesses of each of the methodological variants. Then we will consider the implications for our hypotheses and future work, and conclude with the lessons learned. Note, however, that we present more complete results in the Supplementary Materials and on the causalSemantics GitHub repository.

Tables 3, 4, and 5 show the results for various performance metrics providing different perspectives on performance. The AUROC in Table 3 provides a global assessment of classifier performance irrespective of the classification threshold, while AUPRC in Table 4 is preferred with imbalanced reference datasets [85]. Next, in Table 5, we have MAP-K, which considers the top-ranked results, arguably the most important metric for practical purposes.

More extensive data derived from our analysis are available on the causalSemantics GitHub repository. The Material on GitHub includes Receiver-Operator Characteristic (ROC) and Precision-Recall curves along with the data tables, sample visualizations of the causal graphical models, and confounder sets for the drug/adverse event pairs from the reference dataset.

Performance is not consistent across all metrics. Below, we have provided summaries from our overall performance metrics, indicating clues about each methodological variant's strengths and weaknesses.

AUROC

As shown in Table 3, there was considerable variance across methodological variants with the AUROC metric. The methodological variant that performed well the most consistently was "β^5_{psi}" (or the variant using up to ten semantic vector-based confounders with multiple logistic regression), as shown in Figure 4 illustrating unweighted overall AUROCs.

AUPRC

AUPRC is preferred in analyzing classification problems where the classes are not well balanced between positive and negative labels. AUPRC scores range from 0.0 being the worst to 1.0 being the best, unlike AUROC scores, which usually range from 0.5 to 1.0. Although the subset of the reference dataset we analyzed was only moderately imbalanced overall, the class imbalance for each particular adverse event ranged from moderately imbalanced (acute kidney failure - with ten positive cases and eleven negative control drug/adverse event pairs) to strongly imbalanced (acute liver failure - with forty three positive cases and eleven negative control drug/adverse event pairs). Acute kidney failure had the most improvement using the AUPRC metric, while acute liver failure using the Δ^{10}_{psi} (with β^{10}_{psi} close behind) saw the best improvement across metrics for this adverse event.

MAP-K

As one might anticipate with methods intended to correct for false positives induced by an otherwise unmeasured confounding effect, consistent improvements in performance with adjustment are found with the MAP-K, which measures the accuracy of the top-ranked (most strongly predicted) results. Arguably, MAP-K is the most important metric for purposes of prioritizing signals, the primary application focus of this paper. The importance of MAP-K is apparent when comparing overall performance between the best baseline and best-adjusted models, with improvements of 0.05 and 0.1 with k=10 and 25, respectively. The best performance was observed with the β^{10}_{psi} models, which scored the highest MAP-K with k=25.

Performance Summary

Adjustment using the literature-derived confounders improved predictive and causal inference performance over naive baselines of association across all four adverse events. There were substantive improvements for particular adverse events, with increases in AUROC of 0.1 and 0.2 over the best baseline model with the best-adjusted models for gastrointestinal hemorrhage and acute kidney failure, respectively, and smaller but consistent improvements in AUPRC where best performance was always attained by one of the adjusted models.

5. Discussion

[H1]: Does literature-informed modeling reduce bias?

In most cases, the adjusted models show performance improvements over the unadjusted baseline measures of association. While there was a substantial reduction of bias, there is room for improvement. The overall improvement was consistent with, but not significantly better than that from previous work [26, 66]. We analyzed the distribution of the Δs and βs across the methodological variants.

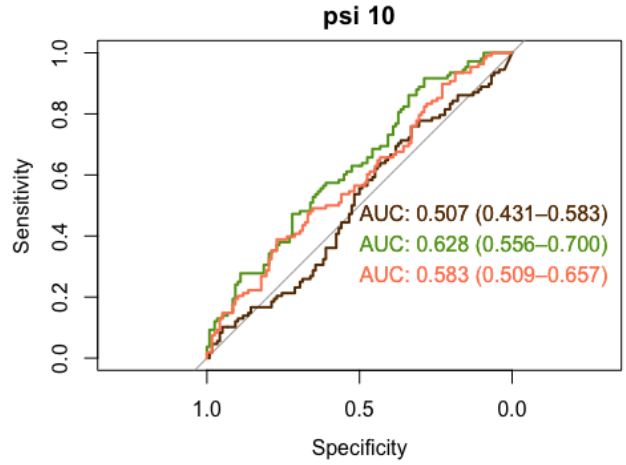


Figure 4: Unweighted AUROC for best results for literature-derived confounders (querying method = PSI, # confounders = 10) vs (baseline = χ^2). Brown = χ^2. Green = [logistic regression coefficient]. Orange = [average treatment effect/Δ]. The numbers to the right of the ROCs represent the 95% confidence intervals.

The mean Δs for the positive controls were higher than for the negative controls in the reference dataset. For example, for Δ^{10}_{psi}, the mean Δ for the negative controls was 0.03 (for reference, the mean Δ of the positive controls was 0.05). An ideal deconfounding method would reduce the Δs for the negative controls to zero.

[H2]: Does adding more literature-derived confounders versus fewer improve performance?

For two out of three metrics, as per Tables 4 and 5, the preponderance of high-performing methodological variants were those with ten literature-derived confounders. Considering there is a significant class imbalance between positive and negative controls for three out of four adverse events, we are inclined to favor the AUPRC over the AUROC score. The strong signal from the MAP-K metric at the higher threshold of confounders also supports this finding. From the standpoint of causal theory, what may be happening is that missing confounders at the lower threshold are being identified at the higher threshold. At the higher confounder threshold, confounders are discovered to partially resolve residual omitted variable bias [86].

[H3]: Which confounder search method (string-based versus semantic vector-based) results in better performing models?

We can observe interesting patterns comparing the individual adverse drug reaction summary results of string-based or semantic vector-based confounder search. With some exceptions, models informed by semantic vector-based

Table 3
Area under the ROC curve (AUROC). ROR = reporting odds ratio. χ^2 = chi squared. β = coefficient from multiple variable logistic regression. Δ = average treatment effect using graphical causal models. Results exceeding best baseline performance are in *boldface*. † indicates best performance for a side effect.

	Baselines		Statistical Models				Causal Models			
Adverse event type [+ ctrls, - ctrls]	ROR	χ^2	β^5_{sql}	β^5_{psi}	β^{10}_{sql}	β^{10}_{psi}	Δ^5_{sql}	Δ^5_{psi}	Δ^{10}_{sql}	Δ^{10}_{psi}
acute kidney failure [10+, 9-]	0.6222	0.4667	0.6	**0.7444**	0.5	**0.8111**†	0.5667	0.6222	0.5111	0.5
acute liver failure [43+, 11-]	0.5835	0.5814	**0.6004**	**0.6723**	0.4926	0.5243	0.4905	**0.6765**†	0.5433	0.5624
acute myocardial infarction [15+, 23-]	0.6261	0.6754†	0.658	0.5159	0.6	0.6725	0.4783	0.5217	0.6667	0.6464
gastrointestinal hemorrhage [20+, 32-]	0.5688	0.6	0.5672	**0.7078**	0.5484	0.55	0.5688	**0.7156**†	0.5203	0.4891
Weighted overall AUROC [88+, 75-]	0.5933	0.5959	**0.6032**	**0.6556**†	0.5363	**0.6005**	0.5215	**0.6466**	0.561	0.5513

Table 4
Area under the Precision Recall Curve (AUPRC).

	Baselines		Statistical Models				Causal Models			
Adverse Event Type [+ ctrls, - ctrls]	ROR	χ^2	β^5_{sql}	β^5_{psi}	β^{10}_{sql}	β^{10}_{psi}	Δ^5_{sql}	Δ^5_{psi}	Δ^{10}_{sql}	Δ^{10}_{psi}
acute kidney failure [10+, 9-]	0.4415	0.47	0.4316	0.3832	**0.4889**	**0.6785**†	0.4535	0.4276	**0.4781**	**0.5032**
acute liver failure [43+, 11-]	0.7554	0.7562	0.7464	0.7057	**0.7726**	**0.8055**	**0.7875**	0.7028	0.7303	**0.8281**†
acute myocardial infarction [15+, 23-]	0.3387	0.5277	0.4816	0.3743	0.4499	**0.5429**	0.4018	0.3735	**0.5631**†	0.4943
gastrointestinal hemorrhage [20+, 32-]	0.3243	0.308	0.3192	0.2744	**0.3694**	**0.3773**†	0.3189	0.2735	**0.3344**	**0.3397**
Weighted overall AUPRC [88+, 75-]	0.4841	0.5266	0.5117	0.4533	**0.5357**	**0.5929**†	0.5092	0.457	**0.5356**	**0.5566**

confounder search performed better string-based confounder search (as we expected). Another fall-out that we expected was the string-based search's conservative nature to result in missing coverage of many drug/adverse event pairs. This suspicion was confirmed after screening for synonyms and stopwords. In many cases, no confounders were available after screening out synonyms (of the exposure and outcome) and stopword-like concepts, e.g., patients, therapeutic procedure, Rattus norvegicus, (see Appendix A).

The present paper partially lends supportive evidence for what Vanderweele calls the disjunctive cause criterion, or DCC [87, 5]. The DCC is a criterion for selecting covariates for which to adjust, and recommends for selection known determinants of either the exposure or the outcome, or both. Arguably, a major factor bolstering the performance of semantic vector-based search is the proportion of cue concept (drug or exposure) contexts occupied by the target (confounder) concept in the underlying knowledge representation. In contrast to the relatively brittle Boolean confounder search used with string matching, the cosine metric used to measure the similarity between vector space representations is continuous in nature, permitting partial match when only one of the two constraints is met. To the extent that PSI can pick either determinants of the exposure or the outcome or both, our framework is a step toward automating the DCC for causal inference.

Arguably, a key factor bolstering the performance of semantic vector-based search lies in its knowledge representation.

The question of how the representation of knowledge affects model performance is an important one. To extrapolate from our results, it is clear that how knowledge is

Table 5
Mean Average Precision at K. ROR = reporting odds ratio. χ^2 = chi squared. β = generalized linear models (multiple variable logistic regression). Δ = average treatment effect using graphical causal models. Results exceeding best baseline performance are in *boldface*. † indicates best performance for a side effect.

	Baselines		Statistical Models				Causal Models			
Adverse Event Type [+ ctrls, - ctrls]	ROR	χ^2	β^5_{sql}	β^5_{psi}	β^{10}_{sql}	β^{10}_{psi}	Δ^5_{sql}	Δ^5_{psi}	Δ^{10}_{sql}	Δ^{10}_{psi}
K = 10										
kidney failure, acute [10+, 9-]	0.4756	0.4921	0.4106	0.3068	**0.5378**	**0.7032**†	0.4768	0.4135	0.4756	**0.5653**
liver failure, acute [43+, 11-]	0.8955	0.8917	0.8389	0.7579	0.7972	0.8951	0.8803	0.725	0.6815	**0.9627**†
acute myocardial infarction [15+, 23-]	0.4889	0.5981	0.4974	0.4021	0.5193	**0.8529**†	0.4911	0.4155	**0.7296**	0.5883
gastrointestinal hemorrhage [20+, 32-]	0.5556	NA	NA	NA	0.1	0.325	0.125	NA	0.125	0.2
Weighted overall MAP [88+, 75-]	0.584	0.515	0.4874	0.4185	0.5201	**0.6378**†	0.4945	0.4113	0.4981	0.57
K = 25										
kidney failure, acute [10+, 9-]	0.4866	0.515	0.4751	0.4166	**0.5422**	**0.7277**†	0.5057	0.4665	**0.5229**	**0.5477**
acute liver failure [43+, 11-]	0.7993	0.7863	0.7767	0.7132	**0.8163**	**0.8764**	**0.8449**	0.7067	0.7085	**0.9051**†
acute myocardial infarction [15+, 23-]	0.3795	0.6056	0.54	0.4227	0.5092	**0.6715**†	0.455	0.4261	**0.6595**†	0.5618
gastrointestinal hemorrhage [20+, 32-]	0.4258†	0.2093	0.224	0.1937	0.3025	0.3739	0.255	0.1857	0.2361	0.2626
Weighted overall MAP [88+, 75-]	0.5458	0.5285	0.51	0.4452	**0.5488**	**0.651**†	0.5263	0.4471	0.5247	**0.5784**

represented and organized can affect measures of topical relevancy, which in turn affect the specific set of potential confounders retrieved by a query. Accordingly, the quality of the confounders affects the ability of the method to reduce confounding. One explanation for why semantic vector-based search usually performs better than string-based search is because the semantic vectors are normalized. Normalization prevents frequently occurring concepts from dominating the result set. This is not surprising, considering that very considerable effort has gone into making VSA models such as PSI able to perform approximate reasoning over large bodies of knowledge.

The purpose of LBD was to be used as a tool for proposing plausible, coherent hypotheses without being too tidy or logically consistent [63]. All aside, a possible research direction would be to combine the results of both string-based and semantic vectors-based confounder candidates for adjustment, or simply to use explicit co-occurrence on a constraint on semantic vector search.

[H4]: Are effect estimates from causal models versus multiple variable logistic regression less biased?

Literature-derived computable knowledge was found to be useful for informing causal inference (as implicated by the performance metrics). Logistic regression models (βs) bested (Δs) from the graphical causal models. To connect prediction and causal inference (estimation), a less biased estimate of the drug effect is likely to lead to better prediction performance.

While prediction and causal inference tasks are closely related, they are not the same: prediction optimizes by minimizing variance, whereas the objective of causal inference problems is to reduce or eliminate bias. Nevertheless, using regression with known confounders as regressors is a traditional way of performing causal inference, where the β in Equation 4 has been interpreted as the causal effect [88, 89, 90]. The causal effect estimates using the junction tree method were within the expected range of performance. Our analysis underlies the importance addressing issues raised in the limitation section of this paper. For the practical application of ideas in this paper, we recommend applying doubly-robust methods referenced in the limitations subsection. However, the exhaustive description and application of causal inference methods as an endpoint is beyond the scope of this paper.

Visualizations

We have included a sample graph in Figure 5. This figure provides a sample of the structure and content expressively modeled by the literature-informed graphical causal graph formalism and instantiated with EHR-derived observational clinical data from free-text clinical narratives. Noting the centrality of asthma in the graph, we searched the literature to find that asthma as an indication is associated with a two-fold increased risk of MI. While inactive asthma did not increase the risk of MI, individuals with active asthma had a higher odds of MI than those without asthma (adjusted OR: 3.18; 95% CI: 1.57 - 6.44) [91]. More such graphs are available in the GitHub repository.

5.1. Comparison with previous related work

For the sake of a coarse comparison, the AUROCs of several EHR-based Pharmacovigilance methods have been included in Table 6. Note that the performance patterns are not strictly comparable owing to different sample sizes and populations, but have been included here for convenience.

Table 6
We have included below AUROCs from aggregated summary statistics resulting from several exclusively EHR-based drug safety studies and component studies.

Li et al. (2014) ROR (naive)	Li et al. (2015) β (adjust.)	This study χ^2 (naive)	Malec (2016) β (adjust.)	Malec (2018) * (adjust.)	This study Δ (adjust.)	This study β (adjust.)
0.53	0.51	0.55	0.53	0.58	0.65	0.6466

In summary, our literature-informed graphical causal modeling framework resulted in superior performance compared with our previous purely EHR-based modeling efforts [26, 66], but fare poorly in comparison with Li et al. (2015) using meta-analysis [92] achieving 0.73 AUROC. However, modest the improvement, the success of such a principled approach using causal models applied to coarse cross-sectional data opens many doors for methodological refinement and future avenues of research.

5.2. Practical applications of literature-informed modeling

Domain knowledge improves the efficiency of causal learning tasks by:

- **Reducing the dimensionality of features**: the richness of EHR data introduces the "curse of dimensionality" problem, presenting a large number of potential covariates for which to adjust. By contrast, discovery pattern search can provide a parsimonious set of covariates vetted from background knowledge that is useful in many situations for explaining, controlling for, and reducing confounding bias.

- **Simplifying causal structure**: qualitative information about the orientation of variables in causal graphs simplifies the task of learning causal structure.

- **Providing *a priori* knowledge of causal order**: although time is not coded explicitly in cross-sectional data, *a priori* knowledge provides information about the likely ordering of events. In terms of graphs, when we assume that a biomedical entity is a confounder relative to exposure and an outcome, then it has a set topological structure (probabilistic and causal dependency). Without assumptions drawing from substantive knowledge, it is often impossible to determine the causal direction (such that it exists) from the data alone

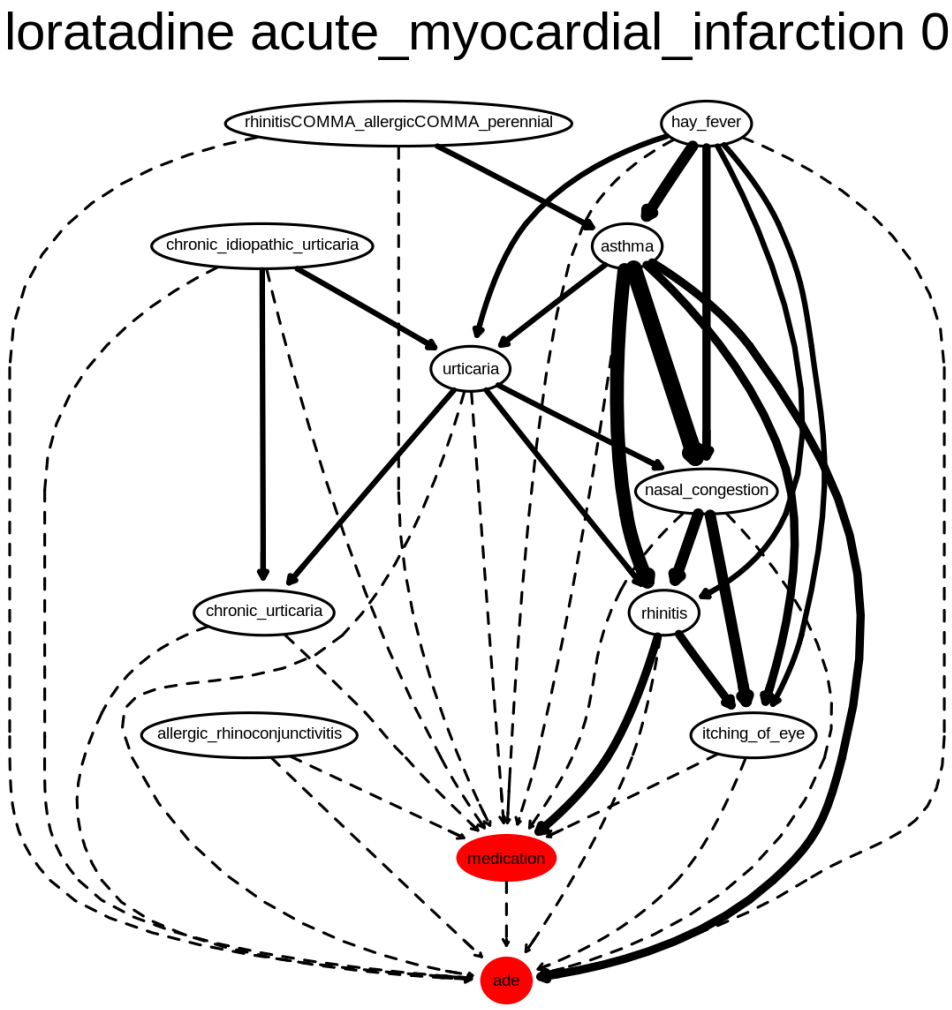


Figure 5: Graphical causal model using PSI at the 10 confounder threshold for loratadine, a negative case for acute myocardial infarction. The thickness of the edges indicates the strength of the observed relationship in the EHR-derived data.

Table 7
Discovery patterns for identifying different variable types.

DP query	Variable Type	Graph shape
X CAUSES ADR	**outcome mechanism**	ADR ← outcome_mechanisms
drug CAUSES X; **X CAUSES ADR**	**mediator**	drug → mediator → ADR *"Chain"*
drug CAUSES X; **AE CAUSES X**	**collider**	drug → collider ←ADR *"V-structure"*

[34], though much progress is being made in this area [93, 6].

Having subject knowledge to inform the selection of covariates with which to adjust is particularly critical with coarse formats such as cross-sectional data.

A Discovery pattern search could be useful for identifying variables with causal roles besides that of being a confounder. For example, the investigator may be interested in screening out certain types of variable such as colliders, or common effects of both the exposure and outcomes variables, that can amplify bias, or mediators, which lie along the causal chain from exposure to outcome. We have not screened for such variables in the current paper, we present the discovery patterns in Table 7 as a starting point for future research. Furthermore, although we have only used only the CAUSES and TREATS predicates, other potentially useful "causal" predicates, and related discovery patterns, exist. For example, other useful predicates in-

Table 8
Problematic source sentences. ARF = acute renal (kidney) failure.

Problem	Source → Target	Confounder	Source sentence
Speculation	ibuprofen $\rightarrow_{\textbf{TREATS}}$ **X**	Benign prostatic hypertrophy	**(PMID: 15947693)** 'Further research must be done to investigate the potential use of ibuprofen in patients with BPH and examine if JM-27 expression in patients with BPH may stratify individuals who may be most responsive to pharmacological treatment.'
Negative evidence	ketoconazole $\rightarrow_{\textbf{TREATS}}$ **X**	Hypercalcemia	**(PMID: 11033844)** 'However, deterioration of renal function during ketoconazole administration as well as failure of hypercalcemia to be affected during short-term ketoconazole treatment suggest that this drug might not be appropriate for acute treatment of hypercalcemic sarcoidosis.'
Hedging	**X** $\rightarrow_{\textbf{CAUSES}}$ ARF	Toxic Epidermal Necrolysis / Steven's-Johnson Syndrome	**(PMID: 19155617)** 'CONCLUSION: ARF, the need for dialysis, and late hypokalemia could be the consequences of SJS/TEN.'
Hedging	ketoconazole $\rightarrow_{\textbf{TREATS}}$ X	Tuberculosis	**(PMID: 17644711)** 'Further investigation is necessary to determine the role of KTC in the treatment of TB.', '17644711'
Case report	X $\rightarrow_{\textbf{CAUSES}}$ ARF	Tuberculosis	**(PMID: 2386602)** 'A 78 year old male patient who had been treated by haemodialysis for 17 years for renal failure, secondary to tuberculosis, is reported.'

clude: PREDISPOSES, AFFECTS, STIMULATES, PREVENTS, INHIBITS, and PRODUCES.

The discovery patterns listed above are "manually-designed" discovery patterns. Methods exist that can automatically generate discovery patterns directly from the distributional semantics of PSI spaces [63, 64], but exploring other discovery patterns is beyond the scope of the present paper.

5.3. Error analysis

To gain an understanding of where the modeling procedures went awry and gather ideas on how to address these issues in subsequent work, we interrogated our models to see where adjustment procedures failed.

We bring the case of ketorolac and acute myocardial infarction to attention. Initially, we thought that the listing of the right platysma (a facial muscle) demonstrated the power of PSI to infer relationships based on the global similarity of structured knowledge. Although there were no results in PubMed or SemMedDB linking right platysma to ketorolac and acute myocardial infarctions, Keterolac was found to be a useful adjunct to Botox treatment to reduce discomfort (after facial injection). Ketorolac was also studied in an RCT for biliary colic pain.

However, we note that PSI can only draw inferences based on global similarity of structured knowledge when deliberately asked to (using e.g. two-predicate path queries from semantic vector cue to semantic vector target). The results of the current search method do not benefit from vector similarity, as we are retrieving elemental vectors, which are, by definition, dissimilar.

Another perspective could be that this suggestion may have resulted from the random overlap between elemental vectors. However, it is helpful that the empirical data can often be used to correct for machine-reading or information retrieval errors in causal models. By contrast, string-based search often yielded confounder candidates that were overly general. This was less often the case with PSI, presumably because statistical weighting was used to deliberately limit the influence of frequently occurring terms during construction of the vector space.

As per table 8, the explanation we considered for the poor performance we observed in some cases was that some of the TREATS relationships mined by SemRep in SemMedDB were occasionally more suggestive of potential non-standard uses that are not as yet FDA approved. It would be unlikely that a patient would be prescribed the drug for that particular (confounder candidate) indication (though it could still affect the outcome). Also, we noted other cases of hedging, the use of case reports, and anecdotal evidence. Recent work on SemRep has focused on assigning confidence scores capturing the factuality of extracted SemRep triples [94]. These results suggest a promising path for constraining SemMedDB predications further, especially now as the values are disseminated with the latest releases of SemMedDB.

Treatment-confounder feedback

A significant limitation of the present work is that it does not consider the time-varying behavior of covariates. For the practical purpose of reducing confounding bias, we made simplifying assumptions. Most notable among these assumptions was that the behavior of the confounder candidates was stable. However, the relationship between variables can change over time. For example, a confounder can behave like a mediator. Mediators are links on the causal chain between exposure and outcome: $A_{exposure} \rightarrow$ **mediator** $\rightarrow Y_{outcome}$. The modeling problem associated with the issue of time-varying covariates is called treatment-confounder feedback. Treatment-confounder feedback can

result in "overcontrol bias," wherein estimates are biased toward the null.

One discovery pattern search-based solution might be to filter for potential mediators, and problematic confounders with potential confounder-treatment feedback (mediation) behavior would be by querying the literature using the discovery pattern in Table 7 for mediators.

Note that a similar logic could guide using discovery patterns to exclude colliders, the discovery pattern for which has also been included in the table. An approach to leave for future work to control for treatment-confounder feedback bias would be to use the above procedure to exclude mediators.

Further, potentially problematic confounders could still be included in the analysis by (longitudinally) truncating all but the first instance of the confounder per patient, and by performing sensitivity analysis using bootstrapping procedures.

Advanced estimation frameworks such as that of targeted learning have been shown to be robust to model misspecification by employing data-adaptive procedures [95, 96]. The targeted learning inference framework combines a propensity score model with an outcome regression to optimize causal effect estimates and can be further enhanced with ensemble machine learning. We have begun to use these methods in our current work, but have not reported the results here.

5.4. Limitations and themes for future work

Ideally, we would wish that our methods for studying observational data would possess sufficient rigor to approach the level of scientific confidence of an RCT. We have enumerated additional limitations of the present study with a view toward future work in this area to bridge gaps in the following areas:

- **Temporality:** one fundamental limitation of our approach stems from coarseness of cross-sectional data. Cross-sectional data represent a "snapshot in time" rather than temporality. A study design that incorporates patient-level longitudinal EHR data may address this limitation in future work. Also, since different observations made of the same patient will capture information about biological processes that unfold through time, some observations are from the same patient, and thus not all samples will be independently distributed. We want to explore adapting our methods to longitudinal data in future work along the lines explored in [97, 98].

- **Data hygiene:** in this study did not carry out advanced phenotyping procedures or correct for missing-not-at-random data or selection bias. Although we have no evidence to prove that such factors negatively impacted our methods' performance, we will be going forward to adopt more rigorous approaches to validating exposures and disease phenotypes [99]. Furthermore, we have not reconciled the data underlying each synonym into a single representation for each overarching biomedical entity of interest representing potential confounder concepts. The study in this paper used empirical data extracted solely from the unstructured free-text narrative. Ideally, phenotyping algorithms for defining exposure, outcome, and confounders would consider data in both structured fields and unstructured free-text in the EHR systems.

- **More robust effect estimation:** more advanced techniques exist for estimating causal effects with more desirable statistical properties than what has been presented here (e.g., G-methods [100, 101, 102], TMLE estimators [95], effect estimators such as Effect of the Treatment on the Treated (ATT) for situations where treated subgroup may have distinct background characteristics compared with untreated.

- **Knowledge representation:** updated literature-based discovery and distributional representation methods (e.g., Embedding of Semantic Predications [ESP], a neural-probabilistic extension of the PSI model [103], [104]). The targeted learning framework would benefit the methods outlined in this paper, as these tools natively embed sensitivity analysis and bootstrapping procedures typical of causal machine learning. Moreover, as long as either the exposure or outcome mechanisms are adequately modeled with substantive knowledge, estimates are proven to be robust.

- **Other discovery patterns:** it is probably not the case that a single discovery pattern is sufficient to capture all confounders. More research is needed in this area to expand search through other pathways.

- **Confounder hygiene:** it is crucial to know your confounder. That is, researchers need to familiarize themselves with the local causal structure of biomedical entities involved in a working hypothesis. In this way, unruly covariates such as colliders can be filtered out from adjustment sets. While automated tools such as those described in this paper may be useful, caution is required and expert adjudication is often ultimately necessary to filter out noise. The literature recommends sensitivity analysis using bootstrapping and other procedures [105, 106].

- **The deconfounder:** there have been notable efforts afoot to create a variable called a deconfounder [37, 107] that can substitute for substantive knowledge of the causal structure relative to the exposure and the outcome. It would be interesting to combine the best from both simulated and empirical components.

- **Developments in machine reading:** The extent of knowledge itself further limits the ability to control for confounding using knowledge, the capability to process such knowledge into a usable form, and the limitations on the representation of the variable's state and whether the variable was measured at all. To this

end, we are exploring combining knowledge from the SemRep reading system with knowledge extracted using the INDRA system. The INDRA system can translate scientific prose directly into executable graphical models [108, 109]. The SemRep system (soon to be released in Java) is being upgraded with exciting features, including factuality levels (potentially useful for improving "knowledge hygiene" and identifying contradictory claims [110]) and end-user extensibility [46].

- **New information sources:** We have also processed Special (drug) Product Labels using SemRep, which may be another valuable source of information on drug safety [111].
- **Principled adjustment set selection:** Another limitation of our work is that the threshold on the number of confounders in the adjustment set was arbitrary. The selection of optimal subsets of confounders is a distinct and active research area referred to as causal feature selection [112]. However, optimizing feature selection for causal inference not the focus of this paper.

Notwithstanding confounding adjustment, residual bias may remain from unmeasured, mismeasured, or omitted variables [86], as well as from other forms of systematic bias [113] along with random noise. For example, selection bias can be induced by unmeasured covariates such as socioeconomic class, which can determine who receives treatment and determine the relative health of the patient receiving that treatment [113]. However, we assume that confounding and other forms of bias can be reduced, but not eliminated.

5.5. Conclusion

This paper introduced a generalizable framework for helping solve a ubiquitous problem (confounding) and expands upon our previous work combining computable knowledge from the literature with observational data to reduce confounding [26, 66]. We used advanced methods to guide the analysis of observational data, but fell short given the limitations listed above.

We found that incorporating literature-derived confounders improved causal inference using a publicly available reference dataset with true labels for drug/adverse event pairs. We also found that generally including more rather than fewer literature-derived confounders improved performance when incorporated into either statistical and causal models.

In conclusion, we have demonstrated that our knowledge integration methods can improve the ability to detect genuine pharmacovigilance signals from observational clinical data by reducing confounding bias, though leaving ample room for methodological refinement. Our framework can be easily adapted to help in other areas with only a modicum of difficulty. The development of more powerful tools for reducing confounding bias could have a potentially significant public health impact by facilitating more efficient screening review of drug safety signals and allowing for more rigorous observational studies, tantamount to conducting "pragmatic trials" [114]. The ability to efficiently learn causal relationships by leveraging existing causal knowledge opens up the potential of realizing the value of large datasets to accelerate the discovery of new knowledge. Finally, tools that help provide insight into causal mechanisms will permit scientists to reverse-engineer nature with more trenchant clues that may lead ultimately to efficacious preventive interventions, to better treatments and, eventually, even to cures.

Acknowledgments: This research was supported by the US National Library of Medicine grants: R01 LM011563, 2 T15 LM007093−26, 5 T15 LM007059−32, NIH/BD2K supplement R01 LM011563−02S1, NCATS Grant U54 TR002804, Cancer Prevention Research Institute of Texas (CPRIT) Precision Oncology Decision Support Core RP150535, and CPRIT Data Science and Informatics Core for Cancer Research (RP170668). Many thanks to readers Sofia Triantafillo, Harry Hochheiser, Doug Landslittel, and the causal reading group at Carnegie Mellon University for helpful suggestions on earlier drafts of this manuscript.

A. Appendix A - Stopwords

A.1. Stopterms

The following terms were excluded from being confounder candidates on account of the terminological vagueness or other reasons. adhesions; adolescent; adult; adverse drug effect; adverse effects; adverse event; adverse reactions; agent; animals; antibodies; antigens; apoptosis; application procedure; assay; assessment procedure; assessment procedure; bacteria; biopsy; blood; body tissue; boys; canis familiaris; capsule; cattle; cell line; cell membrane; cells; cerebrovascular accident; child; chronic disease; clinical research; cohort; color; complication; congenital abnormality; contrast media; control groups; country; detection; diagnosis; disease; dna; elderly; embryo; entire hippocampus; enzymes; excision; extracellular; family; family suidae; felis catus; fibroblasts; follow-up; fracture; fume; functional disorder; genes; girls; growth; house mice; human; implantation procedure; implantation procedure; individual; induction; infant; infant; infiltration; injection procedure; injection procedure; injury; intervention regimes; jersey cattle; lesion; macaca mulatta; macrophage; magnetic resonance imaging; male population group; malignant neoplasm of breast; malignant neoplasms; management procedure; management procedure; medical imaging; membrane; microrna; mild adverse event; monkeys; monoclonal antibodies; mothers; mus; muscle; nonhuman primates; obesity; obstruction; operative surgical procedures; organ; participant; pathogenesis; patient; patient state; patients; persons; pharmaceutical preparations; pharmacophore; pharmacotherapy; placebos; plants; plasma; primates; procedures; prophylactic treatment; proteins; protoplasm; psychopharmacologic agent; rabbits; radiation therapy; rats; rats; rattus norvegicus; receptor; rna; rodent; screening procedure; screening procedure; screw; serum; sloths; solutions; stimulation procedure; stimulation procedure; study models; substance; sup-

plementation; symptoms; syndrome; techniques; test result; therapeutic procedure; toxic effect; transplantation; treatment aids; treatment protocols; voluntary workers; water; woman; young child

References

[1] G. Heinze, C. Wallisch, D. Dunkler, Variable selection âĂŞ A review and recommendations for the practicing statistician, Biometrical Journal. Biometrische Zeitschrift 60 (3) (2018) 431–449. doi:10.1002/bimj.201700067.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5969114/

[2] S. R. Cole, R. W. Platt, E. F. Schisterman, H. Chu, D. Westreich, D. Richardson, C. Poole, Illustrating bias due to conditioning on a collider, International journal of epidemiology 39 (2) (2010) 417–420. doi:10.1093/ije/dyp334.
URL https://www.ncbi.nlm.nih.gov/pubmed/19926667

[3] F. Elwert, C. Winship, Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable, Annual review of sociology 40 (2014) 31–53. doi:10.1146/annurev-soc-071913-043455.
URL https://www.ncbi.nlm.nih.gov/pubmed/30111904

[4] M. A. LuqueâĂŘFernandez, M. Schomaker, B. Rachet, M. E. Schnitzer, Targeted maximum likelihood estimation for a binary treatment: A tutorial, Statistics in Medicine 37 (16) (2018) 2530–2546. doi:10.1002/sim.7628.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7628

[5] T. J. VanderWeele, Principles of confounder selection, European Journal of Epidemiology 34 (3) (2019) 211–219. doi:10.1007/s10654-019-00494-6.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6447501/

[6] P. Spirtes, K. Zhang, Causal discovery and inference: concepts and recent methodological advances, Applied informatics 3 (2016) 3–3. doi:10.1186/s40535-016-0018-x.
URL https://www.ncbi.nlm.nih.gov/pubmed/27195202

[7] H. Kilicoglu, M. Fiszman, G. Rosemblat, S. Marimpietri, T. Rindflesch, Arguments of Nominals in Semantic Interpretation of Biomedical Text, in: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 46–54.
URL http://www.aclweb.org/anthology/W10-1906

[8] J. Che, K. C. Malecki, M. C. Walsh, A. J. Bersch, V. Chan, C. A. McWilliams, F. J. Nieto, Overall Prescription Medication Use Among Adults: Findings from the Survey of the Health of Wisconsin, WMJ : official publication of the State Medical Society of Wisconsin 113 (6) (2014) 232–238.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6095699/

[9] J. H. Watanabe, T. McInnis, J. D. Hirsch, Cost of Prescription Drug-Related Morbidity and Mortality., The Annals of pharmacotherapy 52 (9) (2018) 829–837. doi:10.1177/1060028018765159.

[10] A. Miguel, L. F. Azevedo, M. AraÃžjo, A. C. Pereira, Frequency of adverse drug reactions in hospitalized patients: a systematic review and meta-analysis, Pharmacoepidemiology and Drug Safety 21 (11) (2012) 1139–1154. doi:10.1002/pds.3309.
URL https://doi.org/10.1002/pds.3309

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

[12] J. K. Aronson, M. Hauben, A. Bate, Defining 'surveillance' in drug safety., Drug safety 35 (5) (2012) 347–357. doi:10.2165/11597590-000000000-00000.

[13] I. R. Edwards, Considerations on causality in pharmacovigilance., The International journal of risk & safety in medicine 24 (1) (2012) 41–54. doi:10.3233/JRS-2012-0552.

[14] N. Cartwright, Are RCTs the Gold Standard?, BioSocieties 2 (1) (2007) 11–20. doi:10.1017/S1745855207005029.
URL http://www.palgrave-journals.com/doifinder/10.1017/S1745855207005029

[15] J. Sultana, P. Cutroneo, G. TrifirÃš, Clinical and economic burden of adverse drug reactions, Journal of pharmacology & pharmacotherapeutics 4 (Suppl 1) (2013) S73–S77. doi:10.4103/0976-500X.120957.
URL https://www.ncbi.nlm.nih.gov/pubmed/24347988

[16] C. f. D. E. a. Research, FDA Adverse Event Reporting System (FAERS) Public Dashboard, FDA.
URL https://bit.ly/35dAAFy

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

[18] M. Perez Garcia, A. Figueras, 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) (2011) 1295–1302. doi:10.1002/pds.2193.

[19] L. Wang, M. Rastegar-Mojarad, Z. Ji, S. Liu, K. Liu, S. Moon, F. Shen, Y. Wang, L. Yao, J. M. Davis Iii, H. Liu, Detecting Pharmacovigilance Signals Combining Electronic Medical Records With Spontaneous Reports: A Case Study of Conventional Disease-Modifying Antirheumatic Drugs for Rheumatoid Arthritis., Frontiers in pharmacology 9 (2018) 875. doi:10.3389/fphar.2018.00875.

[20] C. E. Pierce, K. Bouri, C. Pamer, S. Proestel, H. W. Rodriguez, H. Van Le, C. C. Freifeld, J. S. Brownstein, M. Walderhaug, I. R. Edwards, N. Dasgupta, Evaluation of Facebook and Twitter Monitoring to Detect Safety Signals for Medical Products: An Analysis of Recent FDA Safety Alerts., Drug safety 40 (4) (2017) 317–331. doi:10.1007/s40264-016-0491-0.

[21] R. Eshleman, R. Singh, Leveraging graph topology and semantic context for pharmacovigilance through twitter-streams., BMC bioinformatics 17 (Suppl 13) (2016) 335. doi:10.1186/s12859-016-1220-5.

[22] G. Trifiro, J. Sultana, A. Bate, From Big Data to Smart Data for Pharmacovigilance: The Role of Healthcare Databases and Other Emerging Sources., Drug safety 41 (2) (2018) 143–149. doi:10.1007/s40264-017-0592-4.

[23] S. Nojiri, [Bias and confounding: pharmacoepidemiological study using administrative database]., Yakugaku zasshi : Journal of the Pharmaceutical Society of Japan 135 (6) (2015) 793–808. doi:10.1248/yakushi.15-00006.

[24] LePendu P, Iyer S V, BauerâĂŘMehren A, Harpaz R, Mortensen J M, Podchiyska T, Ferris T A, Shah N H, Pharmacovigilance Using Clinical Notes, Clinical Pharmacology & Therapeutics 93 (6) (2013) 547–555. doi:10.1038/clpt.2013.47.
URL https://doi.org/10.1038/clpt.2013.47

[25] J. M. Banda, A. Callahan, R. Winnenburg, H. R. Strasberg, A. Cami, B. Y. Reis, S. Vilar, G. Hripcsak, M. Dumontier, N. H. Shah, Feasibility of Prioritizing Drug-Drug-Event Associations Found in Electronic Health Records., Drug safety 39 (1) (2016) 45–57. doi:10.1007/s40264-015-0352-2.

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

[27] X. Wang, G. Hripcsak, M. Markatou, C. Friedman, 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) (2009) 328–337. doi:10.1197/jamia.M3028.

[28] R. E. Behrman, J. S. Benner, J. S. Brown, M. McClellan, J. Woodcock, R. Platt, Developing the Sentinel System–a national resource for evidence development, The New England Journal of Medicine 364 (6) (2011) 498–499. doi:10.1056/NEJMp1014427.

[29] T. K. Colicchio, J. J. Cimino, CliniciansâĂŹ reasoning as reflected in electronic clinical note-entry and reading/retrieval: a systematic review and qualitative synthesis, Journal of the American Medical Informatics Association 26 (2) (2018) 172–184.
URL https://doi.org/10.1093/jamia/ocy155

[30] S. B. Johnson, S. Bakken, D. Dine, S. Hyun, E. MendonÃğa, F. Morrison, T. Bright, T. Van Vleck, J. Wrenn, P. Stetson, An Electronic Health Record Based on Structured Narrative, Journal of the American Medical Informatics Association : JAMIA 15 (1) (2008) 54–64. doi:10.1197/jamia.M2131.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2274868/

[31] J.-F. Diaz-Garelli, E. V. Bernstam, M. H. Rahbar, 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 (2015) 51–55, publisher: American Medical Informatics Association.
URL https://pubmed.ncbi.nlm.nih.gov/26306235

[32] M. A. Hernan, S. Hernandez-Diaz, M. M. Werler, A. A. Mitchell, Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology., American journal of epidemiology 155 (2) (2002) 176–184. doi:10.1093/aje/155.2.176.

[33] T. J. VanderWeele, I. Shpitser, On the definition of a confounder, Annals of statistics 41 (1) (2013) 196–220.

[34] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd Edition, Cambridge University Press, Cambridge, 2009. doi:10.1017/CBO9780511803161.
URL http://ebooks.cambridge.org/ref/id/CBO9780511803161

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

[36] C.-S. Wang, P.-J. Lin, C.-L. Cheng, S.-H. Tai, Y.-H. Kao Yang, J.-H. Chiang, Detecting Potential Adverse Drug Reactions Using a Deep Neural Network Model, Journal of Medical Internet Research 21 (2). doi:10.2196/11016.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6381404/

[37] Y. Wang, D. M. Blei, The Blessings of Multiple Causes, arXiv:1805.06826 [cs, stat]ArXiv: 1805.06826.
URL http://arxiv.org/abs/1805.06826

[38] R. Ranganath, A. Perotte, Multiple Causal Inference with Latent Confounding, arXiv:1805.08273 [cs, stat]ArXiv: 1805.08273.
URL http://arxiv.org/abs/1805.08273

[39] P. Bruza, M. Weeber, Literature-based Discovery, Information Science and Knowledge Management, Springer Berlin Heidelberg, 2008.
URL https://books.google.com/books?id=niMgUkzU42cC

[40] D. R. Swanson, Fish oil, Raynaud's syndrome, and undiscovered public knowledge., Perspectives in biology and medicine 30 (1) (1986) 7–18.

[41] N. R. Smalheiser, Rediscovering Don Swanson: the Past, Present and Future of Literature-Based Discovery, Journal of data and information science (Warsaw, Poland) 2 (4) (2017) 43–64. doi:10.1515/jdis-2017-0019.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5771422/

[42] D. Hristovski, C. Friedman, T. C. Rindflesch, B. Peterlin, Exploiting semantic relations for literature-based discovery, AMIA ... Annual Symposium proceedings. AMIA Symposium (2006) 349–353.

[43] C. B. Ahlers, M. Fiszman, D. Demner-Fushman, F.-M. Lang, T. C. Rindflesch, Extracting semantic predications from Medline citations for pharmacogenomics, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (2007) 209–220.

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

[45] T. C. Rindflesch, 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) (2003) 462–477. doi:10.1016/j.jbi.2003.11.003.
URL https://linkinghub.elsevier.com/retrieve/pii/S1532046403001175

[46] H. Kilicoglu, G. Rosemblat, M. Fiszman, D. Shin, Broad-coverage biomedical relation extraction with SemRep, BMC Bioinformatics 21 (1) (2020) 188. doi:10.1186/s12859-020-3517-7.
URL https://doi.org/10.1186/s12859-020-3517-7

[47] D. Demner-Fushman, W. J. Rogers, A. R. Aronson, MetaMap Lite: an evaluation of a new Java implementation of MetaMap, Journal of the American Medical Informatics Association 24 (4) (2017) 841–844, publisher: Oxford Academic. doi:10.1093/jamia/ocw177.
URL https://academic.oup.com/jamia/article/24/4/841/2961848

[48] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic acids research 32 (Database issue) (2004) D267–D270. doi:10.1093/nar/gkh061.
URL https://www.ncbi.nlm.nih.gov/pubmed/14681409

[49] T. Cohen, R. W. Schvaneveldt, T. C. Rindflesch, Predication-based Semantic Indexing: Permutations as a Means to Encode Predications in Semantic Space, AMIA Annual Symposium Proceedings 2009 (2009) 114–118.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815384/

[50] Z. S. Harris, Distributional Structure, WORD 10 (2-3) (1954) 146–162, publisher: Routledge _eprint: https://doi.org/10.1080/00437956.1954.11659520. doi:10.1080/00437956.1954.11659520.
URL https://doi.org/10.1080/00437956.1954.11659520

[51] M. Sahlgren, An introduction to random indexing, in: In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005, 2005, pp. 1–9.

[52] P. Kanerva, J. Kristoferson, A. Holst, Random Indexing of Text Samples for Latent Semantic Analysis, in: In Proceedings of the 22nd Annual Conference of the Cognitive Science Society, Erlbaum, 2000, pp. 103–6.

[53] T. Plate, Holographic Reduced Representation: Distributed Representation for Cognitive Structures, CSLI lecture notes, CSLI Publications, 2003.
URL https://books.google.com/books?id=cKaFQgAACAAJ

[54] P. Kanerva, The Spatter Code for Encoding Concepts at Many Levels, in: M. Marinaro, P. G. Morasso (Eds.), ICANN âĂŹ94, Springer, London, 1994, pp. 226–229. doi:10.1007/978-1-4471-2097-1_52.

[55] R. W. Gayler, Vector Symbolic Architectures answer Jackendoff's challenges for cognitive neuroscience, CoRR abs/cs/0412059.
URL http://arxiv.org/abs/cs/0412059

[56] J. A. Fodor, Z. W. Pylyshyn, Connectionism and cognitive architecture: A critical analysis, Cognition 28 (1-2) (1988) 3–71. doi:10.1016/0010-0277(88)90031-5.
URL http://linkinghub.elsevier.com/retrieve/pii/0010027788900315

[57] P. Kanerva, Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors, Cognitive Computation 1 (2) (2009) 139–159. doi:10.1007/s12559-009-9009-8.
URL http://link.springer.com/10.1007/s12559-009-9009-8

[58] D. Widdows, T. Cohen, SemanticVectors creates semantic WordSpace models from free natural language text.: semanticvec-

tors/semanticvectors, original-date: 2015-03-14T17:39:37Z (May 2019).
URL https://github.com/semanticvectors/semanticvectors

[59] D. Widdows, T. Cohen, Reasoning with Vectors: A Continuous Model for Fast Robust Inference, Logic journal of the IGPL 23 (2) (2015) 141–173. doi:10.1093/jigpal/jzu028.

[60] D. Widdows, K. Ferraro, Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008.
URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/300_paper.pdf

[61] D. Widdows, T. Cohen, The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics, in: 2010 IEEE Fourth International Conference on Semantic Computing, IEEE, Pittsburgh, PA, USA, 2010, pp. 9–15. doi:10.1109/ICSC.2010.94.
URL http://ieeexplore.ieee.org/document/5628804/

[62] T. Cohen, G. K. Whitfield, R. W. Schvaneveldt, K. Mukund, T. Rindflesch, EpiphaNet: An Interactive Tool to Support Biomedical Discoveries, Journal of Biomedical Discovery and Collaboration 5 (2010) 21–49.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2990276/

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

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

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

[66] S. Malec, A. Gottlieb, E. Bernstam, T. Cohen, Using the Literature to Construct Causal Models for Pharmacovigilance, Easychair Preprints.Number: 158 Publisher: EasyChair. doi:10.29007/3rfr.
URL https://easychair.org/publications/preprint/X6kk

[67] J. D. Ramsey, Scaling up Greedy Equivalence Search for Continuous Variables, CoRR abs/1507.07749.
URL http://arxiv.org/abs/1507.07749

[68] J. D. Ramsey, B. Andrews, A Comparison of Public Causal Search Packages on Linear, Gaussian Data with No Latent Variables, arXiv:1709.04240 [cs]ArXiv: 1709.04240.
URL http://arxiv.org/abs/1709.04240

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

[70] E. V. Bernstam, Big-Arc Home, library Catalog: sbmi.uth.edu.
URL https://sbmi.uth.edu/uth-big/

[71] H. Saitwal, D. Qing, S. Jones, E. V. Bernstam, C. G. Chute, T. R. Johnson, Cross-terminology mapping challenges: a demonstration using medication terminological systems, Journal of Biomedical Informatics 45 (4) (2012) 613–625. doi:10.1016/j.jbi.2012.06.005.

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

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

[74] Apache, Apache Lucene - Welcome to Apache Lucene.
URL https://lucene.apache.org/

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

[76] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, A. R. Feinstein, A simulation study of the number of events per variable in logistic regression analysis, Journal of Clinical Epidemiology 49 (12) (1996) 1373–1379. doi:10.1016/s0895-4356(96)00236-3.

[77] J. Angrist, J. Pischke, Mastering 'Metrics: The Path from Cause to Effect, Princeton University Press, 2014.
URL https://books.google.com/books?id=dEh-BAAAQBAJ

[78] M. Scutari, J. Denis, Bayesian Networks: With Examples in R, Chapman & Hall/CRC Texts in Statistical Science, CRC Press, 2014.
URL https://books.google.com/books?id=jS3cBQAAQBAJ

[79] I. Tsamardinos, L. E. Brown, C. F. Aliferis, The max-min hill-climbing Bayesian network structure learning algorithm, Machine learning 65 (1) (2006) 31–78.

[80] S. HÃÿjsgaard, gRain: Graphical Independence Networks (Oct. 2016).
URL https://CRAN.R-project.org/package=gRain

[81] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, 1988, google-Books-ID: AvNID7LyMusC.

[82] D. J. Spiegelhalter, Probabilistic Reasoning in Expert Systems, American Journal of Mathematical and Management Sciences 9 (3-4) (1989) 191–210. doi:10.1080/01966324.1989.10737262.
URL https://doi.org/10.1080/01966324.1989.10737262

[83] H. Wickham, Tidyverse.
URL https://www.tidyverse.org/

[84] M. Scutari, Learning Bayesian networks with the bnlearn R package, arXiv preprint arXiv:0908.3817.

[85] J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, in: Proceedings of the 23rd international conference on Machine learning, ICML '06, Association for Computing Machinery, Pittsburgh, Pennsylvania, USA, 2006, pp. 233–240. doi:10.1145/1143844.1143874.
URL https://doi.org/10.1145/1143844.1143874

[86] P. M. Steiner, Y. Kim, The Mechanics of Omitted Variable Bias: Bias Amplification and Cancellation of Offsetting Biases, Journal of causal inference 4 (2) (2016) 20160009. doi:10.1515/jci-2016-0009.
URL https://www.ncbi.nlm.nih.gov/pubmed/30123732

[87] T. J. VanderWeele, I. Shpitser, A new criterion for confounder selection, Biometrics 67 (4) (2011) 1406–1413. doi:10.1111/j.1541-0420.2011.01619.x.
URL https://www.ncbi.nlm.nih.gov/pubmed/21627630

[88] A. S. Blinder, Wage Discrimination: Reduced Form and Structural Estimates, The Journal of Human Resources 8 (4) (1973) 436–455, publisher: [University of Wisconsin Press, Board of Regents of the University of Wisconsin System]. doi:10.2307/144855.
URL https://www.jstor.org/stable/144855

[89] R. Oaxaca, Male-Female Wage Differentials in Urban Labor Markets, International Economic Review 14 (3) (1973) 693–709, publisher: [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University]. doi:10.2307/2525981.
URL https://www.jstor.org/stable/2525981

[90] S. Wright, Corn and Hog Correlations, Department bulletin, U.S. Department of Agriculture, 1925.
URL https://books.google.com/books?id=vVFIMQAACAAJ

[91] D. W. Bang, C.-I. Wi, E. N. Kim, J. Hagan, V. Roger, S. Manemann, B. Lahr, E. Ryu, Y. J. Juhn, Asthma Status and Risk of Incident Myocardial Infarction: A population-based case-control study, The journal of allergy and clinical immunology. In practice 4 (5) (2016)

917–923. doi:10.1016/j.jaip.2016.02.018.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5010477/

[92] Y. Li, Combining Heterogeneous Databases to Detect Adverse Drug Reactions, Drug Safety.
URL https://academiccommons.columbia.edu/catalog/ac:189526

[93] G. F. Cooper, C. Yoo, Causal Discovery from a Mixture of Experimental and Observational Data, arXiv:1301.6686 [cs]ArXiv: 1301.6686.
URL http://arxiv.org/abs/1301.6686

[94] H. Kilicoglu, G. Rosemblat, T. C. Rindflesch, Assigning factuality values to semantic relations extracted from biomedical research literature, PloS One 12 (7) (2017) e0179926. doi:10.1371/journal.pone.0179926.

[95] M. J. van der Laan, S. Rose, Targeted Learning: Causal Inference for Observational and Experimental Data, Springer New York, 2011.
URL https://books.google.com/books?id=RGnSX5aCAgQC

[96] M. J. van der Laan, S. Rose, Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies, Springer International Publishing, 2018.
URL https://books.google.com/books?id=vKFTDwAAQBAJ

[97] K. D. Hoover, S. Demiralp, Searching for the Causal Structure of a Vector Autoregression, SSRN Electronic Journaldoi:10.2139/ssrn.388840.
URL http://www.ssrn.com/abstract=388840

[98] A. Moneta, Graphical causal models and VARs: an empirical assessment of the real business cycles hypothesis, Empirical Economics 35 (2) (2008) 275–300.
URL https://doi.org/10.1007/s00181-007-0159-9

[99] R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo, C. H. Wiggins, N. Elhadad, Learning probabilistic phenotypes from heterogeneous EHR data., Journal of biomedical informatics 58 (2015) 156–165. doi:10.1016/j.jbi.2015.10.001.

[100] J. Robins, A new approach to causal inference in mortality studies with a sustained exposure periodâĂŤapplication to control of the healthy worker survivor effect, Mathematical Modelling 7 (9) (1986) 1393–1512. doi:10.1016/0270-0255(86)90088-6.
URL http://www.sciencedirect.com/science/article/pii/0270025586900886

[101] J. M. Robins, S. Greenland, F.-C. Hu, Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome, Journal of the American Statistical Association 94 (447) (1999) 687–700. doi:10.2307/2669978.
URL https://www.jstor.org/stable/2669978

[102] J. M. Robins, S. Greenland, Estimability and estimation of excess and etiologic fractions., Statistics in medicine 8 (7) (1989) 845–859.

[103] T. Cohen, D. Widdows, Embedding Probabilities in Predication Space with Hermitian Holographic Reduced Representations, in: QI, 2015. doi:10.1007/978-3-319-28675-4_19.

[104] T. Cohen, D. Widdows, Embedding of semantic predications., Journal of biomedical informatics 68 (2017) 150–166. doi:10.1016/j.jbi.2017.03.003.

[105] P. Ding, L. Miratrix, To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias, arXiv:1408.0324 [math, stat]ArXiv: 1408.0324.
URL http://arxiv.org/abs/1408.0324

[106] X. Wang, Y. Jiang, N. R. Zhang, D. S. Small, Sensitivity analysis and power for instrumental variable studies., Biometricsdoi:10.1111/biom.12873.

[107] Y. Wang, D. Liang, L. Charlin, D. M. Blei, The Deconfounded Recommender: A Causal Inference Approach to Recommendation, arXiv:1808.06581 [cs, stat]ArXiv: 1808.06581.
URL http://arxiv.org/abs/1808.06581

[108] R. Sharp, A. Pyarelal, B. Gyori, K. Alcock, E. Laparra, M. A. Valenzuela-EscÃąrcega, A. Nagesh, V. Yadav, J. Bachman, Z. Tang, H. Lent, F. Luo, M. Paul, S. Bethard, K. Barnard, C. Morrison, M. Surdeanu, Eidos, INDRA, & Delphi: From Free Text to Executable Causal Models, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. (2019) 6.

[109] K. Sayed, C. A. Telmer, A. A. Butchy, N. Miskov-Zivanov, Recipes for Translating Big Data Machine Reading to Executable Cellular Signaling Models, arXiv:1706.04117 [q-bio]ArXiv: 1706.04117.
URL http://arxiv.org/abs/1706.04117

[110] A. Alamri, The Detection of Contradictory Claims in Biomedical Abstracts, phd, University of Sheffield (Dec. 2016).
URL http://etheses.whiterose.ac.uk/15893/

[111] S. A. Malec, R. D. Boyce, Exploring Novel Computable Knowledge in Structured Drug Product Labels, AMIA Summits on Translational Science Proceedings 2020 (2020) 403–412.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233092/

[112] V. Lagani, G. Athineou, A. Farcomeni, M. Tsagris, I. Tsamardinos, Feature Selection with the R Package MXM: Discovering Statistically-Equivalent Feature Subsets, arXiv:1611.03227 [q-bio, stat]ArXiv: 1611.03227.
URL http://arxiv.org/abs/1611.03227

[113] T. R. Vetter, E. J. Mascha, Bias, Confounding, and Interaction: Lions and Tigers, and Bears, Oh My!, Anesthesia & Analgesia 125 (3) (2017) 1042–1048. doi:10.1213/ANE.0000000000002332.
URL http://journals.lww.com/00000539-201709000-00046

[114] M. A. HernÃąn, B. C. Sauer, S. HernÃąndez-DÃŋaz, R. Platt, I. Shrier, Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses, Journal of clinical epidemiology 79 (2016) 70–75, edition: 2016/05/27. doi:10.1016/j.jclinepi.2016.04.014.
URL https://pubmed.ncbi.nlm.nih.gov/27237061