Causal feature selection using a knowledge graph combining structured knowledge from the biomedical literature and ontologies: a use case studying depression as a risk factor for Alzheimer's disease

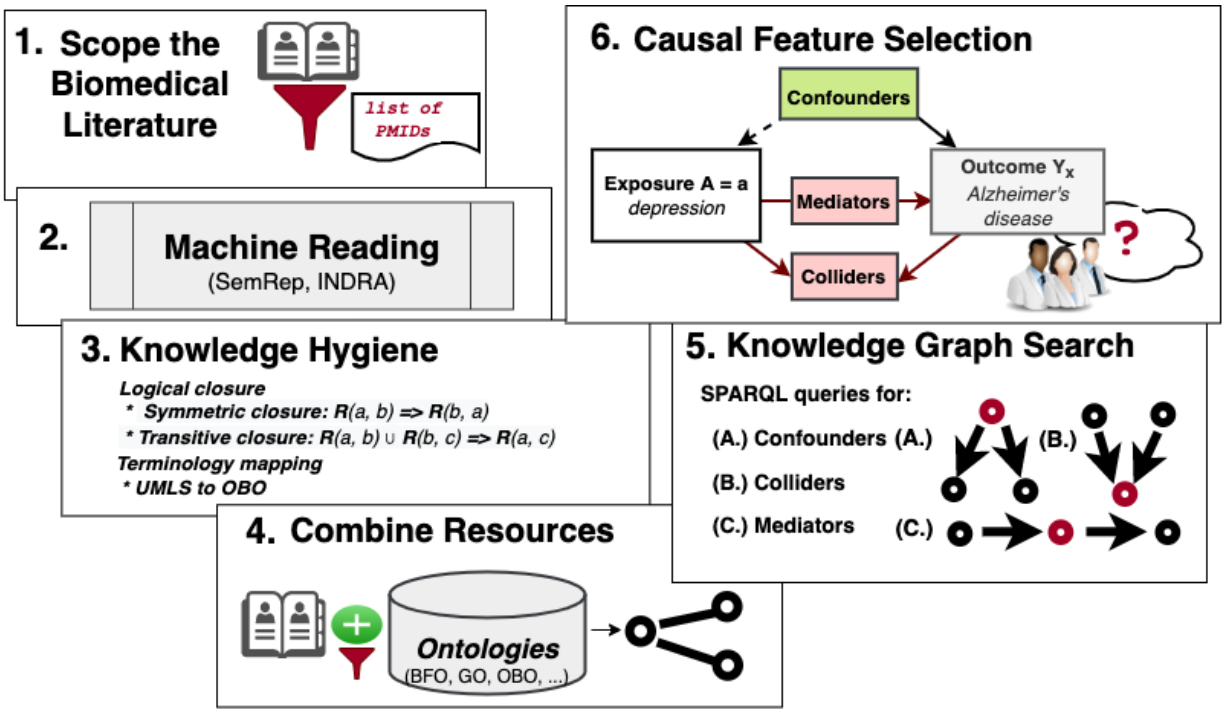
Scott A. Malec, PhD,[1] Sanya B. Taneja, MS,[2] Steven M. Albert, PhD, MS,[3]
C. Elizabeth Shaaban, PhD, MS, MPH,[4] Helmet T. Karim, PhD,[5,6] Arthur S. Levine, MD,[7,8]
Paul Munro, PhD,[9] Tiffany J. Callahan, PhD, MPH[10] Richard D. Boyce, PhD [1,2]

[1]Department of Biomedical Informatics, School of Medicine; [2]Intelligent Systems Program; [3]Department of Behavioral and Community Health Sciences, School of Public Health, [4]Department of Epidemiology, School of Public Health; [5]Department of Bioengineering; [6]Department of Psychiatry; [7]Department of Neurobiology, School of Medicine; [8]The Brain Institute, School of Medicine; [9]School of Computing and Information; University of Pittsburgh, Pittsburgh, PA USA
[10]Department of Biomedical informatics, Columbia University, New York, NY USA

Highlights

- *Knowledge of causal variables and their roles is essential for causal inference.*
- *We show how to search a knowledge graph (KG) for causal variables and their roles.*
- *The KG combines literature-derived knowledge with ontology-grounded knowledge.*
- *We design queries to search the KG for confounder, collider, and mediator roles.*
- *KG search reveals variables in these roles for depression and Alzheimer's disease.*

Graphical abstract:



Abstract

Introduction. *Causal feature selection entails identifying confounders that eliminate confounding bias when estimating effects from observational data. Traditionally, researchers employ expertise and literature review to identify confounders. Uncontrolled confounding from unidentified confounders threatens validity while conditioning on intermediate variables (mediators) weakens estimates, and conditioning on common effects (colliders) induces bias. Additionally, erroneously conditioning on variables playing multiple roles introduces bias. In a use case studying depression as a potential independent risk factor for Alzheimer's disease (AD), we introduce a novel knowledge graph application enabling causal feature selection from computable literature-derived knowledge and biomedical ontologies to address these challenges.*

Methods. *Using the output from three machine reading systems, we harmonized the computable knowledge extracted from a scoped literature corpus. Next, we applied logical closure operations to infer missing knowledge and mapped the outputs to target terminologies. We then combined the outputs with ontology-grounded resources using a robust KG framework developed by computational biologists. Next, we translated epidemiological definitions of confounder, collider, and mediator into queries for searching the KG and summarized the roles played by the variables identified. Finally, we analyze a selection of variables and reasoning paths in the search results.*

Results. *Confounder search yielded 128 confounders, including 58 phenotypes, 47 drugs, and 35 genes. Search also identified 23 collider and 16 mediator phenotypes. Only 31 of the 58 confounder phenotypes were found to behave exclusively as confounders. The remaining 27 phenotypes also play other roles, and 7 of the 21 confounders identified by both the KG and the literature were identified as being exclusively confounders. Stroke was an example of a variable playing all three roles.*

Discussion. *Our findings suggest that our KG application could augment human expertise, while confirming the complexity of selecting potential confounders for depression with AD. Imperfect concept mapping introduced errors, and the small literature corpus limited the scope of search results.*

Conclusion. *Our results suggest that our method may widely apply to causal feature selection. However, the search results need to be reviewed by human experts and tested empirically, and further work is required to optimize KG output for human consumption.*

Keywords: knowledge representation, management, or engineering; knowledge graphs; causal modeling; feature selection; Alzheimer's Disease; depression

1. Introduction

Randomized controlled trials, or RCTs, are still widely considered in medicine to be the gold standard for measuring the causal effect of an exposure on an outcome[1]. Randomization is key to removing bias because it severs the dependency of the exposure assignment on differences in the health dynamics of the study participants that also influence the outcome[2]. However, for pragmatic and ethical reasons, RCTs are not always feasible. For example, RCTs are impractical for studying longitudinal risk factors for chronic disease among the elderly since RCTs usually require a short duration and healthy subjects. Furthermore, RCTs often have issues with external validity in that it may not be possible to transport conclusions from the study to the target population[1]. Hence, observational study

designs have been developed to approximate randomization analytically when estimating the effects of non-randomized data. However, this capability comes at the expense of assumptions, including positivity (each participant has a positive probability of being exposed or unexposed), consistency (the exposure and levels thereof are clearly defined), and unconfoundedness (all variables affecting both the exposure and outcome have been observed and can be adjusted for to eliminate confounding bias). Under the presumption of complete knowledge of confounders, satisfying the assumption of unconfoundedness is possible. However, often knowledge is incomplete, and uncertainty remains. However, a sufficient subset of confounders can often be identified in realistic scenarios. In that case, it is possible to interpret these data causally by adjusting with those confounders. Thus, an essential step when the analytic goal is to estimate the effects from observational data is to identify variables sufficient to control for confounding[3]. This task, called causal feature selection, is the present study's focus.

The specific goal of causal feature selection is the identification of confounders, or common causes of the exposure and outcome. Conditioning on confounders can eliminate or reduce confounding bias by separating the effect of the exposure on the outcome from the effect of the confounders. Traditionally, the task of causal feature selection in epidemiology has fallen to human subject-matter experts to identify and carefully select variables for which to control[3–5] using a judicious mix of content-matter expertise, literature review, and guesswork[6–9]. Fortunately, the causal inference literature has produced exacting causal criteria to guide feature selection to populate a causal model (often in a directed acyclic graph or DAG) [3,10,11]. These causal criteria recommend which variables to include in an adjustment set. First and foremost of these criteria is the Principle of the Common Cause or PCC, first expressed by philosopher of science Hans Reichenbach[11]. Essentially, the PCC requires the inclusion of all common causes (confounders) of the exposure and the outcome. More recently, more advanced alternative criteria have been proposed for mitigating scenarios with unmeasured confounders[12]. Nevertheless, the successful application of these causal feature selection criteria to real-world problems depends on accurate, granular knowledge about etiological relationships relating an exposure to an outcome. However, the literature is vast (PubMed indexes over 1 million articles per year[13]), and time is short.

Moreover, variables do not always behave in expected ways. For example, conventional regression-based estimates fail in the phenomenon of treatment confounder feedback[14], confounders act simultaneously as confounders[15–17] and as intermediate variables called mediators. Conditioning on mediators can weaken measures of

effect. Additionally, variables called colliders that are the common effects of the exposure and outcome should not be used for conditioning because they induce collider bias that distorts measures of effect. By introducing non-causal dependency, collider bias can inflate a null effect to a positive or negative effect, or reverse the direction of effect[18]. In a well-known example of collider bias, a strong association was detected between locomotor disease and respiratory disease among hospitalized patients (odds ratio (OR) 4.09) that was absent when looking at data that included a mix of hospitalized patients and the general populations (OR 1.06)[19]. The roles (confounder, collider, and mediator) played by variables etiologically related to an exposure and an outcome are depicted in Figure 1. Because variables may not act in well-defined categories, it is important to be aware of potentially ill-behaved covariates when selecting features.

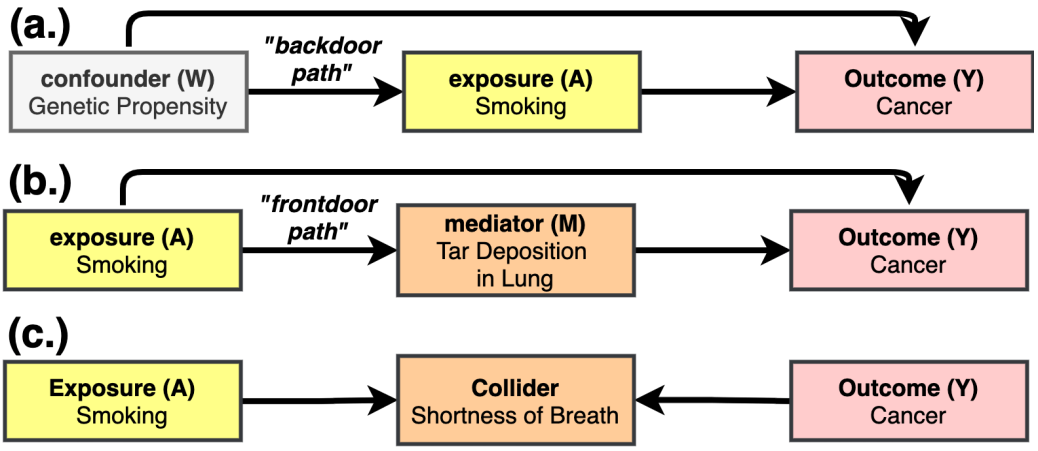


Figure 1. This figure depicts various exposures (A) and outcomes (Y) pairs, in yellow and pink, respectively, along with the following types of variables: **(a.)** a confounder, in gray (with node "*W*" denoting the confounder node); **(b.)** a mediator variable, in orange (with "*M*" indicating the mediator node); and **(c.)** a collider, or common effect of the exposure and outcome. (More background information is available in the Supplementary Materials in file **I.**, **Background S1**.)

The inability to ensure comprehensive knowledge can mean that important confounders are not controlled for[20,21] while inappropriate variables are[19,22–24]. Fortunately, it may be possible to better reduce confounding bias and prevent inducing bias from the inappropriate selection of variables by considering a more comprehensive set of potential confounders using computational tools. Harnessing computable causal knowledge made accessible by machine reading could improve the comprehensiveness of causal feature selection. Machine reading captures how different concepts are related by automatically extracting knowledge from the biomedical literature in subject-predicate-object triples, also called *semantic predications*, e.g., "`ibuprofen TREATS migraine`." There

is substantial literature on workflows deploying triples from machine reading systems in a broad range of biomedical applications from pharmacovigilance[25–28], cancer research[29,30], and drug repositioning[31,32]. However, machine reading output has potential weaknesses. For instance, the recall of machine reading is low relative to other natural language processing tasks since machine reading systems often rely on overly precise manually-crafted patterns[33]. Furthermore, machine reading output may be incomplete. For instance, while a machine reading system may accurately detect the above predication, critical implied information may be missed. For example, as human semantic reasoners proficient in English, it may be possible to infer that therapeutic potential extends to all NSAIDs, the broader category of drugs to which ibuprofen belongs, or that ibuprofen and the other NSAIDs could also treat migraine. Automatic semantic reasoners allow the machine to make such a leap. Other weaknesses include that machine reading systems can produce erroneous output that is unsuitable *as is* for downstream applications, e.g., causal feature selection[34].

Fortunately, biomedical ontologies contain authoritative, expert-curated representations of entities and relationships between those entities that are exploitable by computer programs. Causal knowledge in these ontologies has been used to inform causal models in AD research[35] and elsewhere[36–40]. This knowledge could be used to filter out erroneous assertions from machine reading. However, the knowledge of biomedical ontologies is incomplete due to the extensive manual effort to build and maintain them.

The network-like structure of computable literature-derived knowledge from the literature and ontology-grounded resources alike are often consolidated into a representation called a knowledge graph (KG). A KG is defined as a "graph-theoretic representation of human knowledge such that it can be ingested with semantics by a machine."[41] In biomedical KGs, entities such as disease phenotypes, genes, enzymes, ligands, and diseases are represented as nodes, while edges (or arrows) denote relationships (predicates) between the nodes. Certain KG implementations are capable of employing ontology-grounded information to prune erroneous inputs, and thus may overcome the limitations of relying exclusively on structured information extracted from the literature.

This paper introduces a novel causal feature selection method using a KG application combining literature-derived structured knowledge with the precise semantics of ontology-grounded resources. Our approach entails constructing and searching a large biomedical KG and provides a proof of concept in a use case identifying and refining a set of confounders for estimating the causal effects of depression on Alzheimer's Disease (AD).

AD is the fourth most common cause of death for those 75 years of age or older[42] and is a progressive, multifactorial neurodegenerative disease and the most common cause of dementia. Although most commonly associated with progressive memory loss, AD impacts cognitive ability across multiple domains[43], including executive and language function, sleep, mood, and personality. Over time, it reduces a person's autonomy and leads to death. Despite AD's significant and growing burden, major knowledge gaps remain in AD diagnosis, treatment, and prevention. From 2000 to 2018, the mortality rate for AD increased by 146%, with only one approved disease-modifying therapy. Depression is a common mental illness where the person focuses on negative emotions with symptoms of restlessness, inability to focus, irritability, fatigue, and suicidal thoughts, and may be accompanied by coping behavior, e.g., overeating[44,45].

We chose to study the relationship between depression and AD as a use case for the following reasons. First, the two conditions are multifactorial and heterogeneous in nature[46,46–48] and have overlapping clinical definitions and common risk factors and etiological mechanisms[46–50]. Hence, we expect the structure of the causal neighborhood of variables related to both depression and AD to be inherently challenging and complex. We would expect to find variables that play multiple roles. Second, the Lancet Commission on dementia identified depression among twelve other modifiable risk factors, reporting that 40% of AD diagnoses are attributable to these modifiable risk factors[51]. Thus, if depression were truly an independent risk factor, reducing the prevalence of depression could also reduce the disease burden of AD to a fraction of the population attributable risk. Third, However, controversy surrounds depression as either an independent and modifiable risk factor or a prodrome of AD in late life. Recent insights into the modifiable risk factors present to clinicians, public health officials, and individuals alike unprecedented opportunities for preventive interventions to dramatically reduce the disease burden of AD. Fourth, depression is a highly prevalent disorder. In 2018, it was the third most common mental illness; by 2030, it is expected to become the number one mental illness[45]. Thus, if the relationship between depression and AD is genuinely causal, intervening on depression could significantly impact public health. Finally, the depression and AD relationship is a good fit as a use case since the objective of our causal feature selection method could address a critical need to clarify the etiological status of depression for its effect on AD risk in future studies by furnishing a more comprehensive set of confounders.

This work aims to advance the state-of-the-art causal feature selection method that combines literature-derived structured knowledge with the precise semantics of ontology-grounded resources. The essential idea of this study is to use information extracted from the literature to define the domain of discourse (here, the set of variables potentially related to depression and AD) and then filter out the variables based on biological plausibility using an ontology-grounded KG. We hypothesize that:

- it is possible to identify plausible confounders, colliders, and mediators from structured knowledge distilled from the literature through an ontology-grounded KG (**H1**).
- there is a high degree of complexity regarding the variables linking depression and AD, due in part to the bi-directional nature of the relationship between depression and AD (**H2**).
- the reasoning pathways traversed searching the KG can produce reasonable hypotheses for the mechanisms underlying the relationships between the variables and clarify their meaning (**H3**).

In the next section, we describe the workflow for extracting and post-processing the structured knowledge from the literature, constructing the KG, and reporting and evaluating the KG search results. Then, we report the KG search results and evaluation. Finally, we interpret these results to clarify the strengths and weaknesses of the approach. The code and accompanying documentation we produced for this work are available on the Zenodo project repository, https://10.5281/zenodo.6785307.

2. Methods and Materials

2.1 Overview

Figure 2 shows the workflow for this study. We first delineated the scope of the corpus of articles in PubMed by their PubMed identifiers. Next, we extracted triples from the titles and abstracts from the articles in the corpus using machine reading systems (SemRep[52–54] and INDRA[55]), and combined and harmonized the outputs. Next, since there may be information logically entailed by the extracted information, we applied forward-chaining over the extracted triples to infer implied triples based on transitive and symmetric properties of predicates extracted by machine reading. We refer to the structured knowledge in the post-processing stage as an *extraction graph*. The logically closed *extraction graph* was then combined with an ontology-grounded KG constructed with the PheKnowLator workflow developed by Callahan et al.[56–58]. To answer the "competency question" in Table 1, we

translate standard epidemiological criteria for confounders, colliders, and mediators into discovery patterns, or semantic constraints applied to identify new relations between concepts[59], and implement those discovery patterns as SPARQL queries for searching the KG. Note that we borrowed these patterns for confounders, colliders, and mediators from our previous research[60].

Lastly, we compile the search result, interpret reasoning paths produced by the search, and manually verify a selection of variables in PubMed. We provide details for each component in the paragraphs below.

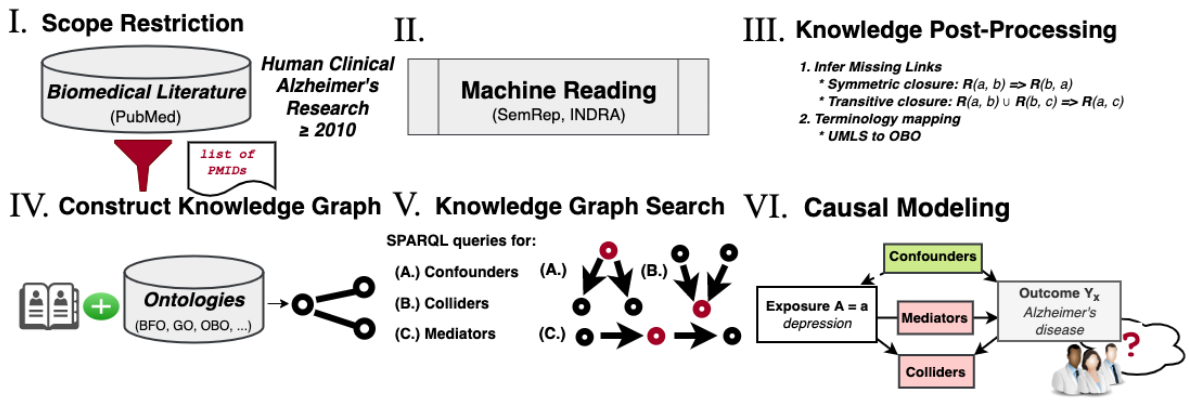


Figure 2. This figure depicts the principal stages of the workflow constructing the causal feature selection system, progressing from left to right and top to bottom by Roman numerals. **I.** Scope the literature to clinical studies investigating AD, published in 2010 or after. **II.** Obtain triples from the machine reading systems. **III.** Knowledge post-processing by performing logical closure operations with the CLIPS production rule system[61,62] to infer missing edges and terminology mapping. **IV.** Construction of the KG using the PheKnowLator platform merging the output of the machine reading systems with ontology-grounded information. **V.** Search the KG to identify relevant variables. **VI.** Compile the search results into a causal model analyzing KG search result output and comparing that output with the structured literature in SemMedDB and PubMed.

Table 1. Summary of the competency questions and approach to finding answers via KG search. The left column enumerates the competency questions, and the right column describes the graph search for answering the competency questions identifying confounders, colliders, and mediators. Discovery patterns for these variables were developed in the SPARQL graph querying language. Graph search was implemented as Dijkstra's shortest path search over the KG based on a query in the SPARQL graph querying language over the subgraph of the KG containing the literature-derived triples and edges inferred from the logical closure procedure. Next, the query uses the output from the first stage as the start and end nodes (universal resource identifiers or URIs) as the input of the queries over the ontology-grounded KG.

Competency questions	**Discovery patterns**
1. What are potential confounders of depression and AD?	**confounder** `CAUSES` depression; **confounder** `CAUSES` AD.
2. What are potential colliders of Depression and AD?	depression `CAUSES` **collider**; AD `CAUSES` **collider**.
3. What are potential mediators between Depression and AD?	depression `CAUSES` **mediator**; **mediator** `CAUSES` AD.

2.2 Scoping strategy defining the literature corpus

We consulted with a health sciences librarian to design the PubMed query listed below to limit the indexed literature's scope to human research related to AD-related factors. The broadness of the meaning of the phrase "AD-related factors" is apt here. The idea is to ingest a significant fraction of the biomedical entities (concepts) in the AD domain. We limited the search to 10 years before the search date (July 11, 2020) because 2010 was the first year referencing Pittsburgh compound B or PiB (first introduced in 2004[63]) with over 100 publications in PubMed, where PiB is a PET amyloid imaging agent that significantly improved AD diagnosis and research[64]. Arguably, PiB initiated a new era of AD research. Finally, we limited the scope of the literature solely to AD rather than include all causes of dementia because different types of dementia have different underlying etiological mechanisms and risk factors.

2.3 Machine reading systems

SemRep. SemRep is a natural language processing tool developed by the National Library of Medicine (NLM) for extracting semantic triples from biomedical literature[52]. SemRep uses a named entity recognition system called MetaMap[65–67]. MetaMap transforms mentions of particular biomedical entities identified as components of semantic triples to "concepts" in the Unified Medical Language System (UMLS) metathesaurus[52]. Each concept in the UMLS metathesaurus is assigned a unique handle called a concept unique identifier or CUI. SemRep infers relations between entities encountered in the biomedical text using a part-of-speech tagger and a semantic network informed by semantic constraints and outputs semantic triples. SemRep has been evaluated extensively over its existence (now almost two decades). Evaluations focusing on SemRep's intrinsic ability to extract semantic triples that accord with human judgment have yielded 0.55-0.9 precision, 0.24-0.42 recall, and 0.38-0.5 F_1[52]. For more extensive information on SemRep, see an overview of Semrep by Kilicoglu[52].

We used the triples present in the Semantic MEDLINE database, SemMedDB. We downloaded SemMedDB version 40 (on June 1st, 2020). We searched SemMedDB for the triples associated with the PubMed identifiers resulting from the literature scoping query.

INDRA. The Integrated Network and Dynamical Reasoning Assembler, or INDRA, is an automated machine reading and model assembly software. We use two machine reading systems developed within the

ecosystem, EIDOS and REACH. EIDOS is a rule-based open-domain machine reader that extracts causal and correlational relations between events mentioned in free-text[68,69]. Events are concepts identified by the machine reader. EIDOS extracts three predicate types - **`Event`**, **`Influence,`** and **`Association`**. Influence is a causal relationship between two events, while association represents a set of events associated with each other without being necessarily causal.

REading and Assembling Contextual and Holistic mechanisms from text, or REACH, detects mentions of protein modifying biochemical processes called post-translational modifications in text. Mass fluorescence spectroscopy has allowed for the precise characterization of different post-translational modifications and opened the possibility of discovering novel biomarkers, therapeutic strategies, or drug targets involving protein interactions. In addition, a burgeoning literature in AD research[79–81] has revealed that many processes involving proteins related to AD's pathogenesis are related to perturbations to processes involving post-translational modifications. REACH provides a window into these processes.

Table 2. This table displays the sampling of predicate types extracted by each machine reading system organized by semantic domain and predicate type. Note that the machine readers complement and overlap in their respective scientific focus domains. We utilized triples containing the predicates in bold.

READER	**Semantic Domains**	**Predicate Types Extracted**
SemRep	Clinical Medicine	**`TREATS, DIAGNOSES, PREVENTS, PROCESS_OF, PRODUCES`**
	Molecular Interactions	**`INTERACTS_WITH, INHIBITS, STIMULATES`**
	Disease Etiology	**`ASSOCIATED_WITH, CAUSES, PRECEDES, PREDISPOSES, MANIFESTATION_OF`**
	Pharmacogenomics	**`AFFECTS, AUGMENTS, COMPLICATES, DISRUPTS, INHIBITS, STIMULATES`**
	Static Relations	`ISA, PART_OF, LOCATION_OF, COEXISTS_WITH`
INDRA/EIDOS	General Domain	`EVENT`, **`ASSOCIATION, INFLUENCE`**
INDRA/REACH	Biochemical Reactions (involving covalent modification of a protein by attaching a functional group)	**`PHOSPHORYLATION, UBIQUITINATION, HYDROXYLATION, SUMOYLATION, GLYCOSYLATION, ACETYLATION, FARNESYLATION, RIBOSYLATION, METHYLATION`**
	Nested Events / "Regulations"	**`REGULATES, PROMOTES, INHIBITS, ACTIVATES`**

2.4 Knowledge post-processing

Knowledge hygiene is required to make the machine reading output more amenable to inference. Knowledge hygiene addresses issues with the validity of the extracted data, including redundancy. We performed the following steps on the triples extracted by the machine reading systems. First, triples were removed with phrases in the subject or object position with more than three space-delimited tokens, e.g., "alpha-blockers aggravation cognitive dementia patients," from the corpus since these phrases could not be condensed to a single concept. Second, we filtered triples containing all but the list of SemRep predicates listed in bold font in Table 2. We also removed duplicate triples. Third, we assigned a default probabilistic belief score of 0.8 to all SemRep triples. This value was chosen because it is the average published performance characteristic across SemRep various predicates[73,74]. Fourth, we mapped the REACH predicates to `RO:0002436 MOLECULARLY_INTERACTS_WITH`, except for phosphorylation, which was mapped to `RO:0002447 PHOSPHORYLATES` in the Relation Ontology (RO)[75]. We used the built-in INDRA preassembly module to remove duplicate triples and assign belief scores for the extracted triples from REACH and EIDOS with the INDRA belief module. From EIDOS, we included all unique `INFLUENCE` and `ASSOCIATION` triples. We also mapped all subjects and objects in the extracted INDRA triples to UMLS concepts using MetaMap[65,66] (version 1.8 with UMLS version 2018AA) and obtained the CUI, UMLS preferred name, and semantic type for the top-scoring match from MetaMap for each concept. All triples where the subject or object did not have the corresponding mapping in UMLS were removed from the corpus used to create the machine reading output extraction graph. Lastly, the triples from all three readers were converted to tab-separated files in a combined corpus that included one row for each subject, predicate, object (with UMLS mapping information), belief score, source sentence, and PubMed ID of the original text and combined the machine readers' output into an extraction graph (see Figure 2).

2.4.1 Logical closure over the extraction graph

Machine reading extracts predications from text without consideration of logical entailment. As a result, the extraction graph was initially incomplete regarding knowledge that could be inferred by applying simple logic. To address incompleteness, we mapped a subset of the machine readers' predicates to object properties in the Relation Ontology (RO)[76] which provided a logical definition for each predicate. The UMLS to RO mappings are available in

file **II.**, **Table S1.** and the analyses of logical properties of the predicates are in file **IV**. for each machine reading system. The latter contains annotations of predicates indicating transitivity, e.g., `CAUSES`, or symmetry, e.g., `COEXISTS_WITH`.) We then used the transitive, symmetric, and inverse properties present in RO in a procedure to infer new triples entailed by the extraction graph[77,78]. This procedure entails loading the initial extraction graph into the CLIPS forward-chaining production rule system [79] via the CLIPSPy Python library[62]. Simple rules using symmetric and transitive relationships belonging to RO predicates were applied to infer new triples and add them to the extraction graph. The CLIPS inference engine also tracked belief scores to manage uncertainty about the validity of inferred triples. New triples inferred from symmetric inference were assigned the same belief score as the asserted triple. The inference engine multiplied the belief score for each of the two asserted triples in transitive inference and assigned it to the inferred triple. All triples inferred from transitivity with a belief score of ≤ 0.6 were filtered out. The value 0.6 was chosen as a heuristic threshold to mimic a weakened transitivity property limiting transitive inference beyond one hop. For example, recalling that SemRep triples were assigned a belief score of 0.8, the first inferred triple from a transitive SemRep predicate would be retained because $0.8 * 0.8 = 0.64$. Subsequently inferred triples would be filtered out, since $0.8 * 0.8 * 0.8 < 0.6$.

2.4.2 PheKnowLator

Phenotype Knowledge Translator (PheKnowLator) is an Ecosystem and Python 3 library designed to construct semantically rich, large-scale biomedical KGs from biomedical ontologies and complex heterogeneous data[80]. We cloned the PheKnowLator 2.0 code and ran the ontology hygiene and data preparation documented in the project's Jupyter workbooks. PheKnowLator merged several ontologies represented using OWL[81] related to genes, diseases, proteins, chemicals, vaccines, and other ontologies of biological interest. Not that complete information about the resources used in PheKnowLator is available on GitHub[82]. Each class in the ontologies merged by the PheKnowLator process formed nodes in the KG. Thus, the object properties assigned to each class formed an edge in the KG. The edges were extended by integrating data sources that connected classes using object properties provided by the RO[76]. Finally, the resulting KG was saved as an OWL ontology and a Python NetworkX graph object[83].

2.4.3 Combining the Extraction Graph with PheKnowLator

Several steps were required to integrate the extraction graph with the PheKnowLator KG so that we could study its application in causal feature selection. The first step was to map the UMLS concept unique identifiers in the SemRep output to the OBO Foundry identifiers used by the PheKnowLator KG. We used existing mappings in the UMLS for the Gene Ontology and Human Phenotype Ontology classes supplemented by a query of the Bioportal Annotator service for other ontologies[78]. The REACH machine reader automatically generated mappings to OBO Foundry identifiers which we added to a mapping table. We also mapped SemRep predicates to the RO version used by the PheKnowLator workflow as SemRep predicates are not present in the RO. The final step was to create two copies of the PheKnowLator KG - one that was loaded into a graph data store (OpenLink Virtuoso version 7.20[84]) so that it could be queried using the SPARQL graph querying language, and the other that was saved as a Python NetworkX MultiDiGraph object for the application of path search algorithms. The extraction graph was loaded into the same graph data store as the PheKnowLator KG for querying.

2.4.4 Reweighting the edges to support path search

The PheKnowLator KG had features deriving from the knowledge representation that made querying and path search challenging without further processing. The ontologies were written in the OWL description logic language, which uses logical statements to define the semantics of ontology entities. Our approach used a reweighting procedure to distill the KG into biologically meaningful entities useful for our study. Therefore, we used the following steps and heuristics to remove OWL semantic statements and reweight the edges of the KG:

1.) **Apply OWL-NETS to the KG to simplify Web Ontology Language (OWL)-encoded biomedical knowledge into an easier-to-use network representation.** OWL-NETS replaced the more complex logical statements with simpler triples without losing biologically relevant information (for examples, see https://github.com/callahantiff/PheKnowLator/wiki/OWL-NETS-2.0)[85].

2.) **Remove all KG edges that specify a disjoint relationship between a subject and object in a triple.** A disjoint relationship in OWL specifies that the class extension of a class description has no members in common with the class extension of another class description[78]. While disjoint relationships are helpful for deductive closure of the KG using a description logic reasoner, they lead to uninformative results from a path search that seeks a mechanistic explanation.

3.) **Remove all edges in the KG that specify the domain of a relationship from the RO.** Many RO relations specify the ontology classes that comprise the relations' domain and range. Like the disjoint relationships mentioned earlier, including these edges helps maintain logical consistency during deductive closure but leads to uninformative results from a path search that seeks a mechanistic explanation.

4.) **Reweight each edge in the graph to influence the path search on predicates useful for mechanistic explanations.** The weighting strategy is based on observing the method used in PheKnowLator to represent triples of value for constructing a mechanistic explanation explained below:

   a.) **Assign default edge weights.** Since the PheKnowLator KG is not weighted by default, start by assigning a default weight of 2 to all edges. The value of 2 has no special significance except a value greater than 0.

   b.) **Favor hierarchical relationships.** Hierarchical relationships can be very informative for mechanistic explanations. Most of the classes represented in the ontologies that have been merged are hierarchical and are represented using the `rdfs:subClassOf` predicate. If a predicate relates a subject to an object by `rdfs:subClassOf`: weight = 0. Assignment of subClasOf weights to zero lowers the cost of including a hierarchical edge in a path search through the KG, which increases the prevalence of hierarchical edges in path search results.

   c.) **Favor edges providing a mechanistic triples RO.** We developed a simple heuristic to produce mechanistic triples and avoid triples that provide meta-knowledge about entities in an ontology. If the predicate relates a subject to an object by owl:someValuesFrom: weight = 0. Similarly, if the *object* is a RO entity and the predicate is owl:onProperty: weight = 0.

   d.) **Disfavor subjects and edges that primarily describe metadata**. If the *subject* is a RO entity or the edge is `owl:`equivalentClass: weight = 5.

5.) **Find the shortest path between biomedical entities of interest.** Use the Dijkstra shortest path algorithm[86] to identify the shortest paths between nodes *x* and *y* in the KG (many different paths share the same length). The Dijkstra algorithm is a variant of the best-first search that prioritizes finding the shortest path between potential nodes. There are many graph search methods, so we chose Dijkstra's shortest path algorithm as a

simple way to test proof-of-concept for identifying confounders, colliders, and mediators.

2.5 Evaluation

2.5.1 Answering competency questions

We report the confounders, colliders, and mediators for depression and AD identified by searching the KG. As our KG search implementation retains explanations in the form of a trace of nodes traversed by Dijkstra's shortest path search, we created visualizations of the reasoning pathways and analyzed and interpreted the biological plausibility of the reasoning pathways.

2.5.2 Analysis of search results

In addition to the pre-defined, "simple" roles (confounder, collider, and mediator, mentioned earlier), our compilation procedures allow us to identify variables with complex behavior, i.e., "combination" roles, which we define as additional role categories for variables with more than one causal role. The combination roles included confounder/mediators, confounder/colliders, and confounder/mediator/colliders. Our rationale here was to sort out the roles played by the variables and identify variables with complex behavior, of which other researchers ought to be aware. The combination roles were identified using set operations, e.g., set difference and intersection, over the list of variables for each role type. For example, to identify "Confounders only" variables, we perform a set difference operation over the list of confounders minus the variables that also appear in the collider and mediator lists.

2.5.3 Comparing search results

We sought to understand how our novel feature selection differs from complementary methods for causal feature selection using SQL queries to identify relevant variables from the literature explored in previous research. To this end, we compared the KG search results with results from searching the original scoped predictions in $\text{MachineReadingDB}_{\text{Scoped}}$ and $\text{SemMedDB}_{\text{Complete}}$. $\text{MachineReadingDB}_{\text{Scoped}}$ contained the original subset of structured knowledge restricted to human clinical studies on Alzheimer's disease published in 2010 or after that we used to construct the extraction graph. $\text{SemMedDB}_{\text{Complete}}$ contained the entire database of SemRep predicates extracted from titles and abstracts of articles published 2010 or after. We summarized the search results for simple and complex roles as we did in the previous subsubsection.

2.5.4 Manual verification

Finally, we performed several additional posthoc analyses. (**I.**) One of the authors (CES, PhD in Epidemiology and MS in Neuroscience) who is an expert in AD research examined the source sentences for three variables identified as being "confounders only" in both knowledge bases. We furnished the human expert with a set of source sentences from SemMedDB supporting the variables as confounders. Our overarching goal was to see if the reasoning pathways were reasonable and if the triples faithfully represented the semantics of the original assertion in the literature, and if the semantics, to the extent they were faithful, supported the case of the variable in each causal relationship. (**II.**) Next, we looked at the evidence supporting two variables that were found to play all three roles by again looking at the sentences. (**III.**) We then attempted to resolve contradictory information between the KG and SemMedDB about the roles of another variable. Our basic approach was to compare all of the evidence supporting a selection of variables for causal roles. To see if the search results made sense, we compiled the KG search results, including the variables and reasoning pathways, along with the triples and source sentences from the literature, which supported each variable for that role. We also searched PubMed to fill in gaps in our understanding. (**IV.**) Lastly, to determine the extent to which our methods recapitulate or even contradict existing studies, we used PubMed to review reported covariates in observational studies of depression as a risk factor for our use case. A subaim was to see to what extent KG search discovered confounders absent in reported covariates in published studies.

3. Results

Figure 3 shows the result of applying post-processing procedures to the extracted triples. 13,365 PubMed-indexed articles were returned from PubMed, resulting in 226,997 triples from machine reading. The number of triples remaining after executing the knowledge hygiene procedure was 94,934, or 41.8% of the original triples. The subject and object entities of the entire set of triples included 10,020 unique UMLS CUIs, of which 2,504 were mapped to ontological identifiers and merged using the PheKnowLator workflow (2,110 from Bioportal, 844 from the Gene Ontology[87,88] and Human Phenotype Ontology[89] mappings present in UMLS, and 394 from REACH). After filtering out the machine reading and inferred triples with a belief score of < 0.8 and mapping predicates to the RO, the total number of machine reading triples was 20,010, or 8.8% of the original total.

The KG from the PheKnowLator workflow contained 18,869,980 triples. Table 3 shows the counts of

inferred edges by predicate type. The most significant gains by predicate type `CAUSES_OR_CONTRIBUTES_TO_CONDITION` yielded the most novel inferred causal edges, followed by `CAUSAL_RELATION_BETWEEN_PROCESSES`. There were zero inferred edges for certain other predicate types because we did not define the logical properties of those predicates to allow for inference. Table 4 shows the count of edges for four triple types relating genes or variants of genes to health conditions.

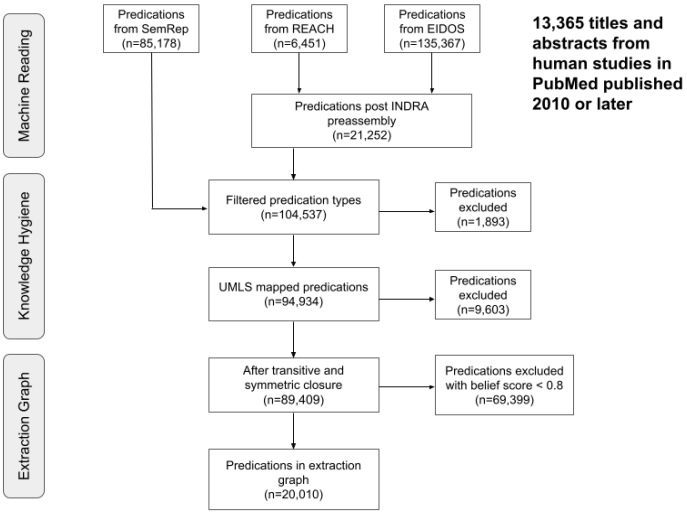


Figure 3. This figure illustrates merging the outputs from the machine reading systems, performing knowledge hygiene, and applying logical closure operations on the extraction graph.

Table 3. Counts of triples inferred using the logical closure operations given the symmetric and transitive properties of the literature-derived predicates from machine reading. Transitive and symmetric closure produced 73,046 triples after filtering out inferred triples with a belief score of < 0.6 and those without a mapping to the RO.

Predicate Type	**Machine Reading**	**Inferred**	**Symmetric (yes/no)**	**Transitive (yes/no)**
`CAPABLE_OF_INHIBITING_OR_PREVENTING PATHOLOGICAL_PROCESS`	683	0	No	Yes
`CAUSAL_RELATION_BETWEEN_PROCESSES`	3,103	20,945	No	Yes
`CAUSES_OR_CONTRIBUTES_TO_CONDITION`	2,003	45,697	No	Yes
`CORRELATED_WITH`	3,000	3,000	Yes	No
`EXISTENCE_OVERLAPS_WITH`	2,528	2,528	Yes	No
`INTERACTS_WITH`	1,236	1,236	Yes	No
`INVOLVED_IN_POSITIVE_REGULATION_OF`	457	0	No	No
`IS_SUBSTANCE_THAT_TREATS`	2,647	0	No	No
`MOLECULARLY_INTERACTS_WITH`	7	0	No	No
`NEGATIVELY_REGULATES`	816	0	No	No

PART_OF	1,355	6,929	No	Yes
PHOSPHORYLATES	18	0	No	No
PRECEDES	136	313	No	Yes
REGULATES	3,512	0	No	No

Table 4. The count of four causal predicates in the KG produced by PheKnowLator after running OWL-NETS. These predicates were chosen because they are essential to the competency questions shown in Table 1.

Predicate Type	**# of triples**
CAUSES_OR_CONTRIBUTES_TO_CONDITION	113,837
MOLECULARLY_INTERACTS_WITH	5,678,829
GENETICALLY_INTERACTS_WITH	3,246,417
CAUSES OR CONTRIBUTES TO CONDITION	113837

3.1 Answering competency questions

3.1.1 Question #1 - What are the potential confounders of Depression and AD?

The KG search found 126 unique variables potentially confounding the association between depression and AD, including 47 conditions or phenotypes, 43 drugs, and 35 genes or enzymes. Examples included apolipoproteins, arachidonic acid, chronic infectious disease, encephalopathies, estradiol, hypercholesterolemia, hypoglycemia, inflammatory response, interleukin-6, myocardial infarction, and non-alcoholic fatty liver disease. (The KG search found 141. Fifteen of the 141 confounders appeared to be redundant, e.g., *vascular endothelial growth factor A* (PR_000017284) and *vascular endothelial growth factor A (human)* (PR_P15692)). However, according to PRO, these concepts are not the same. The human version (PR_P15692) is a child of (PR_000017284), so we keep both in the KG. For the purposes of reporting in the tables, we have merged these concepts.

The shortest path lengths returned by Dijkstra's path search algorithm varied from 2 (15 paths from a confounder to depression / 16 paths from a confounder to AD) to 9 (2 paths, both from confounder to depression or confounder to AD). The paths for the confounder with the longest path lengths (*statin* (CHEBI_87631)) included paths that passed through the hub *molecular entity* (CHEBI_23367). The majority of confounders had paths to depression or AD with 4 or less path length (77 paths with <= 4 steps to depression and 80 paths with <= 4 steps to AD). Figure 4 depicts the reasoning path for inflammatory response.

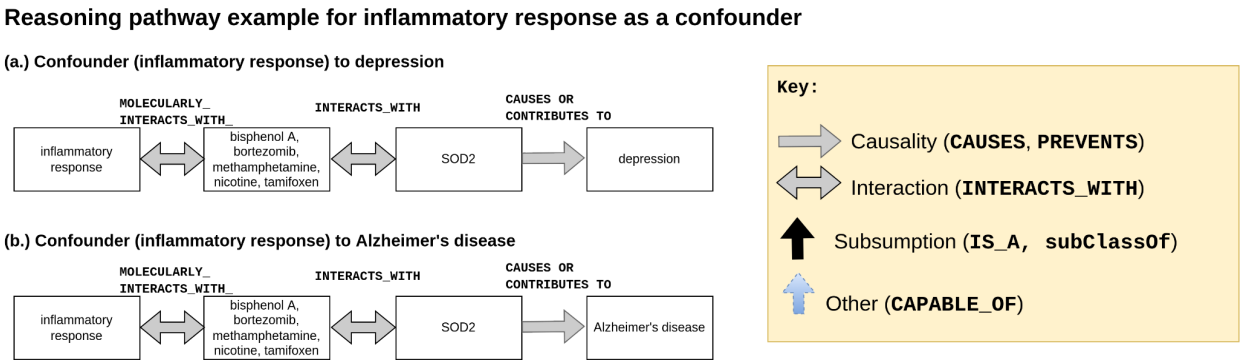


Figure 4. Sample reasoning pathway of inflammatory response, a confounder identified by KG search. SOD2 is superoxide dismutase 2, a gene that encodes a mitochondrial protein. SOD2 is associated with diabetes and gastric cancer[90]. SOD2 polymorphisms are associated with neurodegenerative diseases, including cognitive decline, stroke, and AD[91,92]. SOD2 is a gene that scavenges for reactive oxygen species (ROS) resulting from environmental exposures, e.g., bisphenol A, that result in inflammation[93]. Bisphenol A is found ubiquitously in plastics and is ingested from water bottles, packaged food, and many other sources[94–96].

3.1.2 Question #2 - What are potential colliders of Depression and AD?

The KG search found 28 unique potential colliders for depression and AD. These included amyotrophic lateral sclerosis, anemia, apraxias, atherosclerosis, brain hemorrhage, cerebral amyloid angiopathy, cerebrovascular accident, congestive heart failure, deglutition disorders, *dementia*, diabetes mellitus, diabetes mellitus, non-insulin-dependent, encephalitis, falls, *frontotemporal dementia*, ***homeostasis***, immune response, insulin resistance, ischemic stroke, malnutrition, neurofibrillary degeneration (morphologic abnormality), osteoporosis, parkinsonian disorders, pathogenesis, pneumonia, psychotic disorders, senile plaques, stroke, and tauopathies.

Note that in the list of colliders above, the KG search identified as potential colliders several variables we have emphasized by italicizing and underlining. The KG may have inferred dementia because AD is a specific category of dementia. Homeostasis may have been included in the results as a generic biological process that was not filtered out. Finally, it would not make sense that AD would cause frontotemporal dementia, the most common dementia in patients younger than 60. We drop supercategories like dementia and generic processes like homeostasis for our analysis since these can be easily identified and explained away from context as spurious.

There were two colliders identified using SPARQL for which no path could be found from depression to the collider in the KG created using the PheKnowLator workflow (*falls* (HP_0002527) and *neurofibrillary tangles* (HP_0002185)). Two of the 26 colliders with paths from depression in KG had multiple identifiers (*cerebral amyloid angiopathy* (HP_0011970 and DOID_9246) and *frontotemporal dementia* (DOID_10216, DOID_9255, and

HP_0002145)). Path search identified a total of 29 collider entities in KG.

For depression to the colliders, the path lengths of the shortest paths returned were 7 steps for 2 colliders, 8 for 26, and 9 for 1. The majority of colliders (24) resulted in 5 or more shortest paths returned. For AD to the colliders, the path step counts were slightly lower at 6 steps for 2 colliders, 7 for 26 colliders, and 8 for 1 collider. Similar to depression, 23 colliders have 5 or more shortest paths from AD.

As shown in Figure 5 below, the typical path pattern for both cases involved passing through the hub node *nervous system* (UBERON_0001016) and *nervous system process* (GO_0050877) that molecularly interacts with a drug or chemical that also interacts with a gene that causes or contributes to AD.

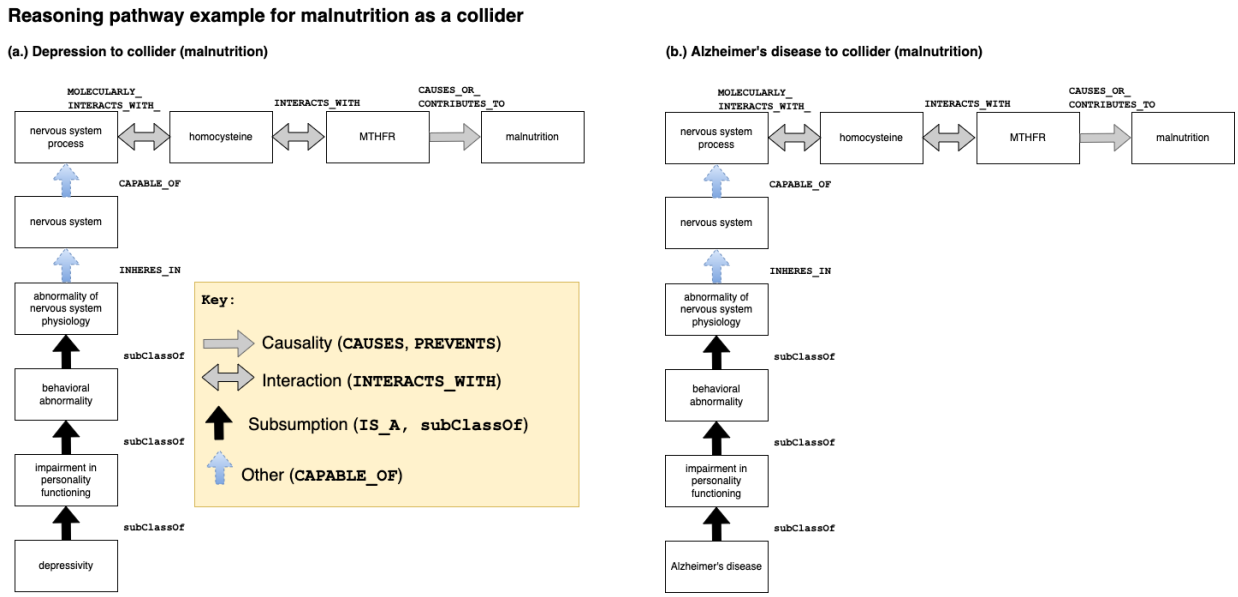


Figure 5. Sample reasoning pathway of malnutrition, a collider identified by KG search. The MTHFR gene is a gene that encodes the Methylenetetrahydrofolate reductase (MTHFR) enzyme that is critical for the metabolism of amino acids, including homocysteine[97]. MTHFR perturbations and mutations decrease the metabolism and elevate homocysteine. Elevated homocysteine is linked with blood clots and thrombosis events[98–100]. Homocysteine has been reported in the published literature as a confounder of depression and AD in at least one observational study[101]. Elevated homocysteine is also associated with low levels of vitamins B6 and B12 and folate.

3.1.3 Question #3 - What are potential mediators of Depression and AD?

The KG search found 18 unique potential mediators that were conditions or phenotypes for depression and AD. These included amyotrophic lateral sclerosis, anemia, atherosclerosis, cerebral amyloid angiopathy, cerebral hemorrhage, congestive heart failure, *dementia*, diabetes mellitus, dysphagia, insulin resistance, ischemic stroke, malnutrition, osteoporosis, parkinsonism, pathogenesis, stroke, tauopathy, type II diabetes mellitus. The shortest

path lengths from Dijkstra's path search algorithm were 8 for depression to mediators. There is more variety in path length from mediators to AD, with most path lengths between 4 and 6. As is shown in Figure 6 below, the typical path pattern for both cases involves associating the disorder (depression or a mediator) with a nervous system process that molecularly interacts with a chemical that causes or contributes to AD.

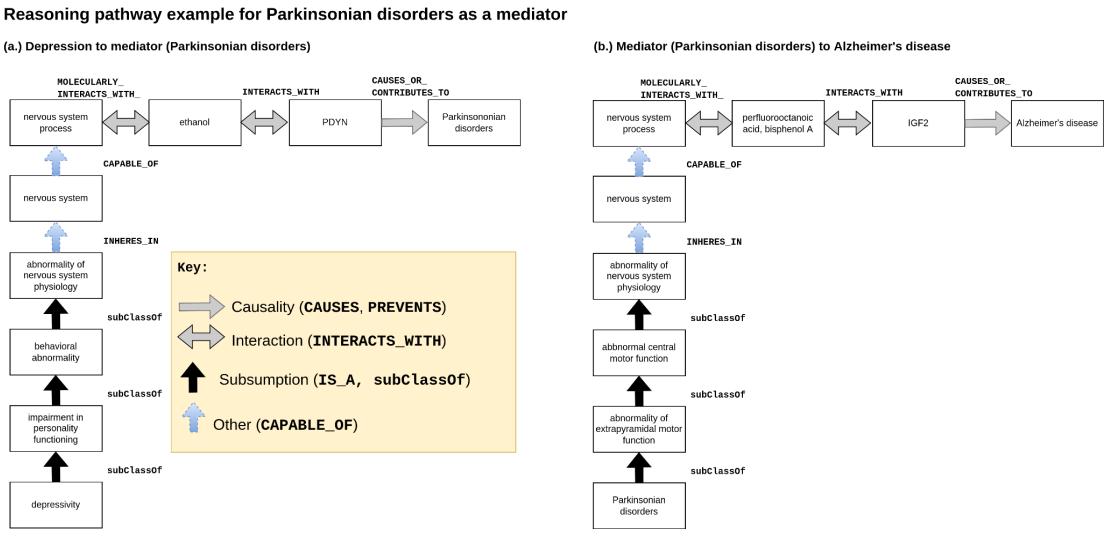


Figure 6. Sample reasoning pathway of Parkinsonian disorders, a potential mediator for Alzheimer's disease identified by KG search. Prodynorphin (PDYN) is a protein-producing gene. PDYN produces endogenous opioid peptides involved with motor control and movement[102] implicated in neurodegenerative diseases that modulate response to psychoactive substances, including cocaine and ethanol[103]. The gene insulin growth factor 2 (IGF2) is associated with development and growth. Perfluorooctanoic acid and bisphenol A exposure is associated with neurodegeneration resulting in AD[104]. Hippocampal IGF2 expression is decreased in AD-diagnosed patients[105]. Increased IGF2 improves cognition in AD mice models[106].

3.2 Analysis of search results by causal role

3.2.1 Simple roles

Figure 7 illustrates the phenotype variables produced from our KG search. Aside from the conditions we analyzed, the KG search also identified pharmaceutical substances, genes, and enzymes as confounders. However, the shortest path parameters produced no search results from our queries for colliders and mediators belonging to those semantic categories. We report the complete set of variables the KG identified in the spreadsheet (see the worksheets entitled "Substances" and "GenesEnzymes") in the Supplementary Materials in file **III.**, **Worksheet S2,** and **Worksheet S3**.)

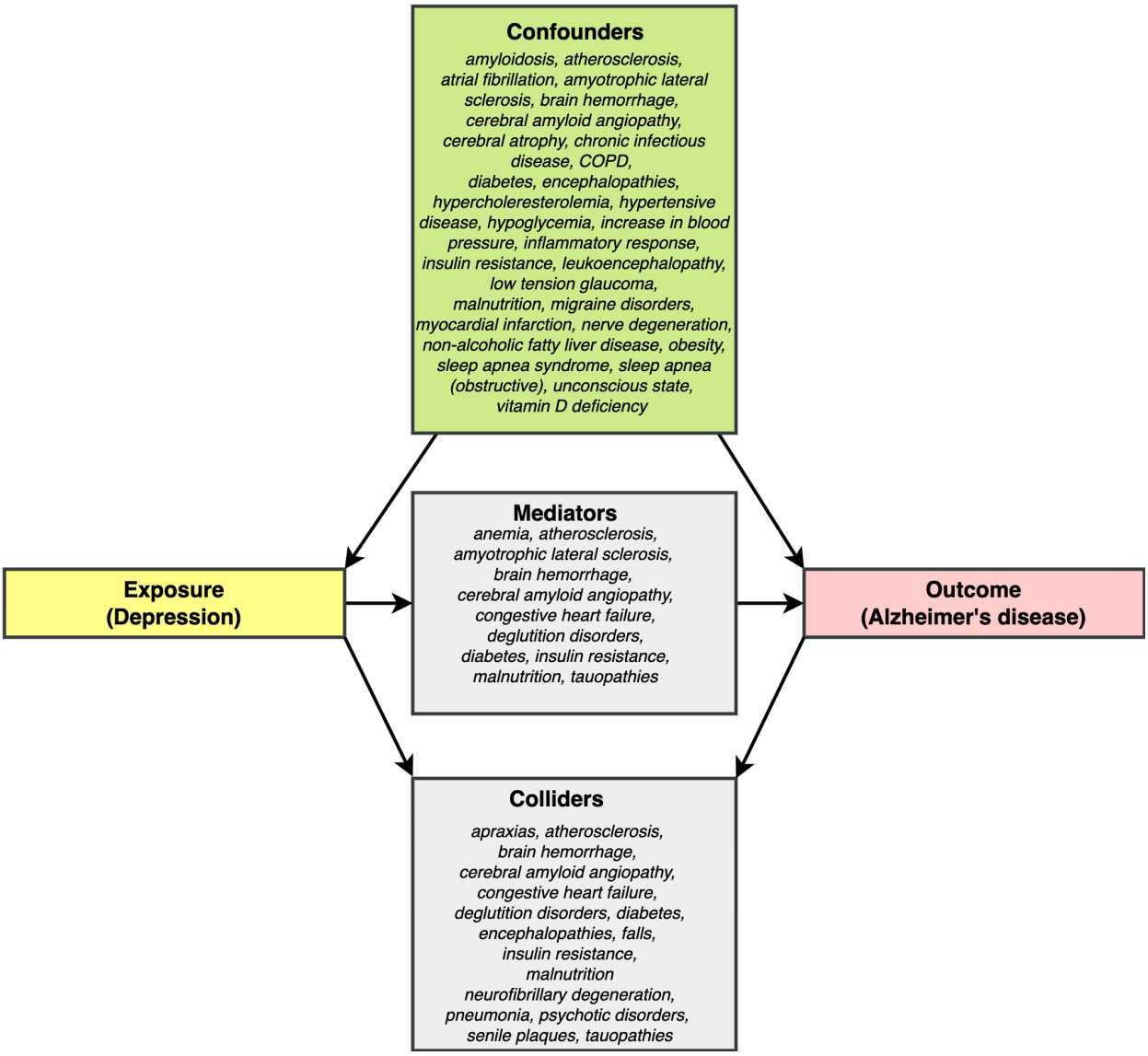


Figure 7. This diagram shows the variables identified using their causal roles relating to depression and AD by translating standard epidemiological definitions into patterns for querying the KG.

3.2.1 Complex roles

Figure 8 illustrates the same conditions from Figure 7, albeit in a more distilled form since we distinguish between variables playing single versus the different categories of combination roles.

Twenty-two distinct confounders and nine colliders were found to behave solely in one role. We could not find any variables that behave exclusively in the role of mediator. The rest of the variables fulfilled two or more causal roles. Of these, 11 variables fulfilled the causal criteria of all three causal roles.

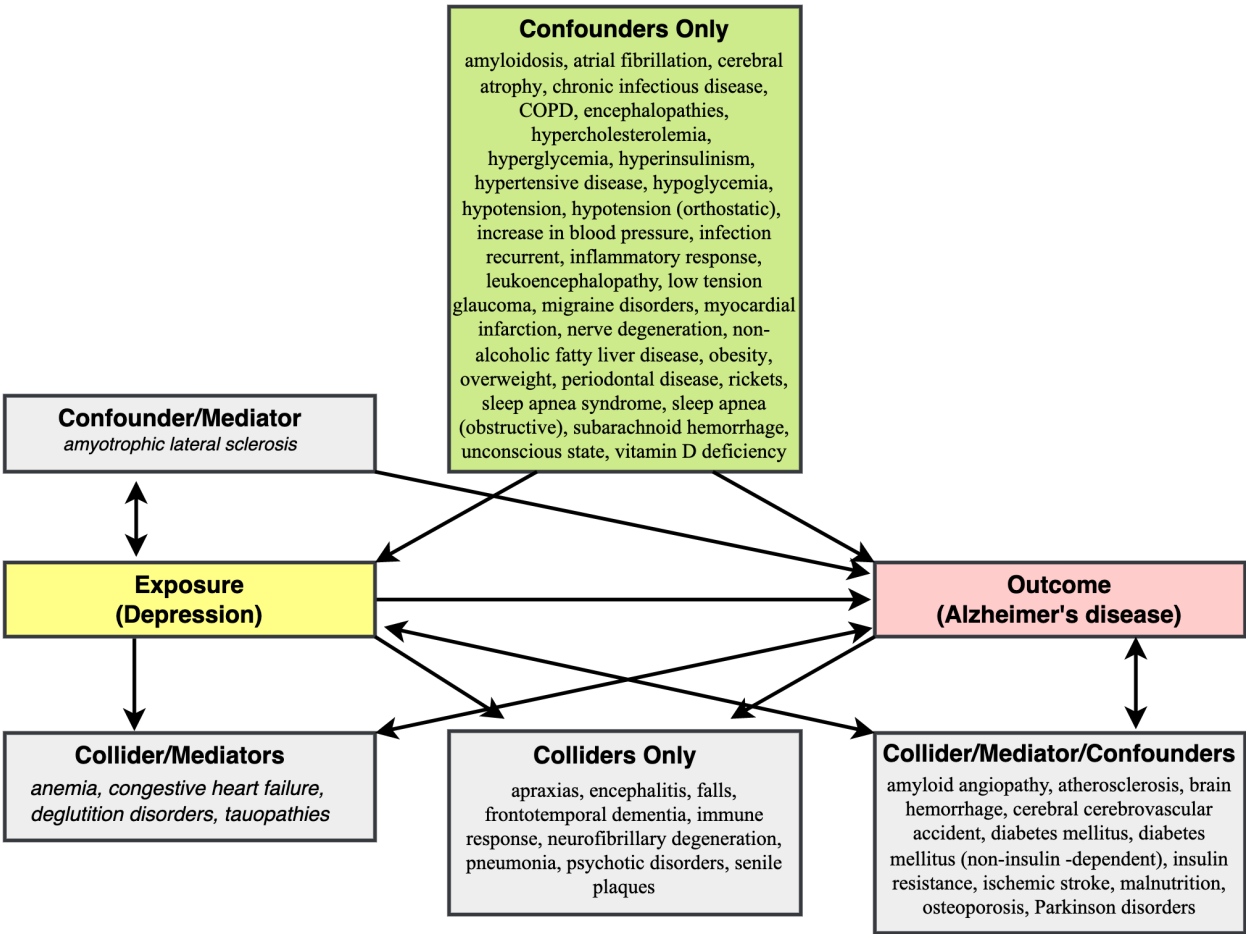


Figure 8. This diagram shows variables according to their roles, including single and hybrid role types for the relationship between depression (exposure) and AD (outcome). Each rectangular box in the figure lists all the conditions by single role (confounder, collider, or mediator) or combination roles (e.g., mediator/collider). For example, the box labeled "Confounders only" contains covariates that are exclusively confounders and were not found to act as a collider or mediator.

Note that while there is a literature describing the core roles (confounder, collider, and mediator), the literature for confounder/mediator (also called time-varying confounding and treatment-confounder feedback) is the only type of combination or hybrid variable for which a literature exists[15]. Finally, also note that the graphs in Figures 7 and 8 are not, strictly speaking, causal DAGs since they depict cycles. Since causal DAGs cannot represent cyclic relationships, new graph formalisms can represent latent variables and cycles (feedback loops) by generalizing the causal interpretation of DAGs, e.g., directed mixed graphs (DMGs)[107], chain graphs[108].

3.3 Comparison of KG versus SemMedDB search results

To save space in Table 5, we omitted columns for the INDRA machine reading systems (EIDOS and REACH) because there were no variables identified using information from those sources.

Table 5. Comparison with variables organized in rows by role (confounder, collider, or mediator) and columns organized by (knowledge graph (KG) and the two subsets of SemMedDB. The subsets (restricted to articles published in 2010 or after) include MachineReadingDB$_{Scoped}$ and SemMedDB$_{Complete}$. Concepts in bold appear in both the KG and SemMedDB$_{Complete}$. Underlined concepts appear in both SemMedDBs. To save space, we restrict our analysis of concepts retrieved to the semantic categories relating to findings and phenotypes. Strikethrough indicates that the variable appears in either the KG or SemMedDB$_{Complete}$ but not in the identical role(s).

	KG	**machineReadingDB$_{Scoped}$**	**SemMedDB$_{Complete}$**
Confounders	amyloidosis, amyotrophic lateral sclerosis, atherosclerosis, atrial fibrillation, brain hemorrhage, cerebral amyloid angiopathy, **cerebral atrophy**, **cerebrovascular accident**, ***chronic infectious disease***, COPD, **diabetes mellitus, diabetes mellitus (non-insulin -dependent)**, encephalopathies, **hypercholesterolemia**, **hyperglycemia**, **hyperinsulinism**, **hypertensive disease**, **hypoglycemia**, **hypotension**, ***hypotension (orthostatic)***, ***increase in blood pressure***, infection recurrent, inflammatory response, **insulin resistance**, **ischemic stroke**, leuko-encephalopathy, ***low tension glaucoma***, **malnutrition**, **migraine disorders**, **myocardial infarction**, nerve degeneration, non-alcoholic fatty liver disease, **obesity**, **osteoporosis**, **overweight**, ***Parkinson's disorders***, ***periodontal diseases***, rickets, **sleep apnea syndrome**, **sleep apnea (obstructive)**, unconscious state, **vitamin D deficiency**	**cerebrovascular accident**, deep brain stimulation, **diabetes mellitus, non-insulin-dependent***, diet (Mediterranean), hearing impairment, impaired cognition, metabolic syndrome, ***Parkinson's disease***	agitation, amputation, ~~anemia~~, blind vision, body weight decreased, caloric restriction, cancer treatment, cardiovascular diseases, cardiovascular risk factor, **cerebral atrophy**, **cerebrovascular accident**, cerebrovascular disorders, ***chronic disease***, **chronic infectious disease**, chronic stress, convulsions, coronary arteriosclerosis, coronary artery bypass surgery, death (finding), debridement, deep brain stimulation, degenerative disorder, degenerative polyarthritis, depressed mood, depressive symptoms, ***diabetes***, **diabetes mellitus, diabetes mellitus (non-insulin- dependent)**, diet, mediterranean, difficulty sleeping, disability, disrupted sleep, drowsiness, dyslipidemias, dyssomnias, endocrine system diseases, endothelial dysfunction, exercise, family history, feeling hopeless, fitness, folic acid deficiency, folic acid supplementation, frailty, functional disability, **glaucoma**, group exercise program, hearing impairment, heart diseases, heart failure, hyperactive behavior, **hypercholeresterolemia**, **hyperglycemia**, hypercysteinemia, **hyperinsulinism**, hyperlipidemia, hypersensitivity, **hypertensive disease**, hyperthyroidism, hypertrophy, **hypoglycemia**, hypotension, hypothyroidism, hysterectomy, illiteracy, impaired cognition, impulsivity, indifferent mood, infarction, lacunar, **insulin resistance**, intellectual functioning disability, **ischemic stroke**, kidney failure (chronic), lack of sensation, leisure activities, malaise, **malnutrition**, maternal history, memory training, memory skills training, metabolic syndrome, **migraine disorders**, moderate drinking, motor symptoms, multiple sclerosis, music therapy, **myocardial infarction**, narcotic use, need for isolation, neurodegenerative disorders, neurotic disorders, **obesity**, oestrogen therapy, **osteoporosis**, ***overweight***, oxidative stress, pain, ***Parkinson's disease***, ***periodontitis***, personal satisfaction, phototherapy, physical activity/inactivity, physical therapy, planning suicide, pollution, porphyromonas gingivalis, pre-eclampsia, psoriasis, psychotherapy, regular exercise, reproductive history, rest, rheumatoid arthritis, sedentary lifestyle, seizures, sensory stimulation, **sleep apnea syndromes**, **sleep apnea (obstructive)**, sleep deprivation, sleep disorders, sleep disturbances, sleep disturbances, sleeplessness, smoking, spinal puncture, stress, stress disorders, post-traumatic, stress, acute, transcranial alternating/direct current stimulation, transplantation, underweight, vascular diseases, virus diseases, visual impairment, **vitamin D deficiency**, ***waist circumference***, widowhood, work stress, worried
Colliders	~~anemia~~, apraxias, atherosclerosis, brain hemorrhage, cerebral amyloid angiopathy, **cerebrovascular accident**, congestive heart failure, deglutition disorders, **diabetes mellitus**, ***diabetes mellitus (non-insulin-dependent)***, encephalopathies, falls, fronto- temporal dementia, immune response, ~~insulin resistance~~, **ischemic stroke**, **malnutrition**, neurofibrillary degeneration, **osteoporosis**, ***Parkinson disorders***, **pneumonia**, **psychotic disorders**, senile plaques, tauopathies	NA	at risk for suicide, cardiovascular diseases, cerebral atrophy, **cerebrovascular accident**, **diabetes, diabetes mellitus**, epilepsy, feeling suicidal, ~~hypertensive disease~~, **ischemic stroke**, malignant neoplasms, **malnutrition**, ~~obesity~~, osteoporosis, ***Parkinson's disease***, **psychotic disorders**, rheumatoid arthritis, semantic impairment, sleep disturbances, suicide attempt
Mediators	amyotrophic lateral sclerosis, ~~anemia~~, atherosclerosis, brain hemorrhage, ***cerebral amyloid angiopathy***, **cerebrovascular accident**, ***congestive heart failure***, deglutition disorders, **diabetes mellitus**, ***diabetes mellitus (non-insulin-dependent)***, ~~insulin resistance~~, **ischemic stroke**, **malnutrition**, **osteoporosis**, ***Parkinson's disorders***, tauopathies	NA	blind vision, ***brain diseases***, cardiovascular diseases, cerebral atrophy, **cerebrovascular accident**, convulsions, ***diabetes***, **diabetes mellitus**, ***heart failure***, ~~hypercholesterolemia~~, ~~hypertensive disease~~, immune system diseases, **ischemic stroke**, **malnutrition**, metabolic syndrome, ~~migraine disorders~~, ~~obesity~~, **osteoporosis**, ***Parkinson's disease***, psoriasis, rheumatoid arthritis, semantic impairment, septicemia, sleep disturbances, stress disorders (post-traumatic), virus diseases, Wernicke encephalopathy

We see a moderately high concurrence between knowledge bases, but there are curious and perhaps telling patterns. For instance, among the discordant results, there is a pattern that concerns the semantic categories of output. Specifically, SemRep had better recall performance for variables that are potentially more difficult to establish mechanistically, e.g., psychiatric symptoms ("agitation," "dyssomnias," "feeling hopeless") or lifestyle factors (e.g., exercise, family history). Moreover, the concepts identified by KG search results are generally less redundant than those from searching the literature (e.g., diabetes, diabetes mellitus). Searching SemMedDB$_{Complete}$

yielded the most search results. Searching the KG yielded fewer search results. Searching MachineReadingDB$_{Scoped}$ yielded still fewer.

Table 6. Comparison of search results belonging to the semantic categories of findings and phenotypes organized in rows by variable role, including role combinations ("Confounders only," "Colliders only," or "Mediators only") and columns organized by (knowledge graph (KG) and SemMedDB$_{Complete}$. We ignore MachineReadingDB$_{Scoped}$ for this analysis since it did not yield colliders or mediators. Concepts in bold appear in both the KG and SemMedDB$_{Complete}$. Closely related hypernymic/hyponymic concepts (near matched) are italicized in bold. Concepts are underlined if they appear in both SemMedDBs. A strikethrough (e.g., ~~anemia~~) indicates disagreement between knowledge bases.

	KG	**SemMedDB$_{Complete}$**
Confounders only	amyloidosis, atrial fibrillation, ~~cerebral atrophy~~, **chronic infectious disease**, COPD, encephalopathies, ~~hypercholesterolemia~~, **hyperglycemia**, **hyperinsulinism**, ~~hypertensive disease~~, **hypoglycemia**, **hypotension**, ***hypotension (orthostatic)***, increase in blood pressure, ***infection recurrent***, inflammatory response, leukoencephalopathy, ***low tension glaucoma***, ~~migraine disorders~~, **myocardial infarction,** nerve degeneration, non-alcoholic fatty liver disease, ~~obesity~~, **overweight**, ***periodontal disease***, rickets, **sleep apnea syndrome**, **sleep apnea (obstructive)**, subarachnoid hemorrhage, unconscious state, **vitamin D deficiency**	agitation, amputation, ~~anemia~~, body weight decreased, caloric restriction, cancer treatment, cardiovascular risk factor, cerebrovascular disorders, ***chronic disease***, **chronic infectious disease**, chronic stress, coronary arteriosclerosis, coronary artery bypass surgery, death (finding), debridement, deep brain stimulation, degenerative disorder, degenerative polyarthritis, depressed mood, depressive symptoms, ~~diabetes mellitus, non-insulin-dependent, diet (mediterranean)~~, difficulty sleeping, disability, disrupted sleep, drowsiness, dyslipidemias, dyssomnias, endocrine system diseases, endothelial dysfunction, exercise, family history, feeling hopeless, fitness, folic acid deficiency, folic acid supplementation, frailty, functional disability, ***glaucoma***, group exercise program, hearing impairment, heart diseases, hyperactive behavior, **hyperglycemia**, hypercysteinemia, **hyperinsulinism**, hyperlipidemia, hypersensitivity, hyperthyroidism, hypertrophy, **hypoglycemia**, **hypotension**, hypothyroidism, hysterectomy, illiteracy, impaired cognition, impulsivity, indifferent mood, infarction, lacunar, ~~insulin resistance~~, intellectual functioning disability, kidney failure, chronic, lack of sensation, leisure activities, malaise, maternal history, memory training, memory skills training, moderate drinking, motor symptoms, multiple sclerosis, music therapy, **myocardial infarction**, narcotic use, need for isolation, neurodegenerative disorders, neurotic disorders, oestrogen therapy, **overweight,** oxidative stress, pain, ***periodontitis***, personal satisfaction, phototherapy, physical activity/inactivity, physical therapy, planning suicide, pollution, porphyromonas gingivalis, pre-eclampsia, psychotherapy, regular exercise, reproductive history, rest, sedentary lifestyle, seizures, sensory stimulation, **sleep apnea syndromes**, **sleep apnea, obstructive**, sleep deprivation, sleep disorders, sleep disturbances, sleeplessness, smoking, spinal puncture, stress, stress disorders, transcranial alternating/direct current stimulation, transplantation, underweight, vascular diseases, visual impairment, **vitamin D deficiency**, ***waist circumference***, widowhood, work stress, worried
Colliders only	apraxias, encephalitis, falls, frontotemporal dementia, immune response, neurofibrillary degeneration, pneumonia, **psychotic disorders**, senile plaques	at risk for suicide, epilepsy, feeling suicidal, malignant neoplasms, **psychotic disorders**, suicide attempt
Mediators only	NA	immune system diseases, septicemia, Wernicke encephalopathy
Confounder/ mediators	amyotrophic lateral sclerosis	blind vision, convulsions, ~~heart failure~~, ~~hypercholesterolemia~~, metabolic syndrome, ~~migraine disorders~~, psoriasis, stress disorders (post-traumatic), virus diseases
Collider/ confounders	encephalopathies	NA
Collider/ mediators	~~anemia~~, ~~congestive heart failure~~, deglutition disorders, tauopathies	brain diseases, semantic impairment
Collider/ confounder/ mediators	amyloid angiopathy, atherosclerosis, brain hemorrhage, **cerebral cerebrovascular accident**, **diabetes mellitus,** ***diabetes mellitus (non-insulin -dependent)***, ~~insulin resistance~~, **ischemic stroke**, **malnutrition**, **osteoporosis**, ***Parkinson disorders***, *stroke*	cardiovascular diseases, ~~cerebral atrophy~~, **cerebrovascular accident**, ***diabetes***, **diabetes mellitus**, ~~hypertensive disease~~, **ischemic stroke**, **malnutrition**, ~~obesity~~, **osteoporosis**, ***Parkinson's disease***, rheumatoid arthritis, sleep disturbances

The two primary patterns that stand out in Table 6 are the relatively high degree of concordance between the knowledge bases regarding confounders and combination variables that play all three roles (confounder, collider, and mediator) on the one hand. From this point on, we call these variables playing a combination of roles, such as stroke, chimeras. In Greek mythology, a chimera is a creature having the head of a lion, the body of a goat, and the tail of a snake. From the perspective of valid causal inference, the chimera is a dangerous variable category because conditioning on chimeras can potentially bias analysis. Thus, chimeras are important to identify to avoid introducing bias.

On the other hand, the recall performance for the other types of exclusive and the combination variable is poor. For example, anemia appears in "Mediator only" in the KG search results but in "Confounders only" in the $SemMedDB_{Complete}$ search results. Encephalopathies was another concept where the knowledge bases were in disagreement.

3.4 Manual qualitative evaluation of search results

3.4.1 Confounders

One of the authors (CES, PhD in Epidemiology and MS in Neuroscience) who is an expert in AD research examined the support for three variables identified as being "confounders only" in both knowledge bases: chronic infectious disease, obstructive sleep apnea (OSA), and vitamin D deficiency. The source sentences are available in the Supplementary Materials in **File V. *ConfounderSourceSentences.xlsx***. In the view of the content expert, all three variables seemed biologically plausible common causes of both depression and AD. OSA seemed to have the most and strongest evidence. However, evidence for it was based on meta-analyses and reviews but not the source studies. While there is a larger body of evidence on infections and AD, the amount and quality of the evidence seemed less for vitamin D deficiency and chronic infectious disease. The expert noted the difficulty in assessing the evidence quality from one sentence.

3.4.1 Stroke as a chimeric variable

Stroke was identified as chimeras by both knowledge bases. The support for stroke as a chimera was substantive, though with a caveat where the interpretation of the meaning of stroke depends on the stroke subtype. To review the meaning of stroke, stroke results from a reduced blood supply to the brain from either constriction, e.g., ischemic stroke, or hemorrhage, e.g., hemorrhagic stroke, intracerebral hemorrhage. There are also less immediately devastating stroke-like events called transient ischemic attacks, that presage future more serious stroke events. Ischemic stroke is a known risk factor for both depression[109–111] and AD[112–114]. Although the KG search classified both stroke and the ischemic stroke subtype as chimeras, presumably including both subtypes under the umbrella term or hypernym of stroke, the evidence was stronger for ischemic stroke as a chimera. However, evidence for AD as a risk factor was weaker for ischemic than for hemorrhagic stroke[113]. In the meta-analysis of Waziri et al. based on four national studies and assessing the relative risk of hemorrhagic vs. ischemic stroke in AD

vs. non-AD patients and applying matching criteria, the relative risk of AD patients with similar non-AD controls was 1.41 (95% CI 1.23-1.64) for hemorrhagic stroke but 1.15 (95% CI 0.89-1.48) for ischemic stroke. PubMed search did not yield any specific information about the hemorrhagic stroke subtype as a confounder. The source sentences from SemMedDB for stroke are included in the Supplementary Materials in **File VI.** ***ChimeraSourceSentences.xlsx***.

3.4.2 Resolving contradictions between SemMedDB and KG search

As noted previously in Results, anemia, a disease characterized by low hemoglobin levels, appears as a "mediator/collider" in the KG results but exclusively as a confounder in the $\text{SemMedDB}_{\text{Complete}}$ results. Table 7 shows the support from SemMedDB for anemia causing depression was provided by two **PREDISPOSES** triples, and the support for anemia causing AD was supplied by a single **CAUSES** triple. In terms of conveying causative force, **PREDISPOSES** is weaker than **CAUSES** and implies that anemia as a risk factor may be confounded or otherwise attenuated as a risk factor.

Table 7. Support from SemMedDB for anemia as a confounder of depression as a risk factor for AD. The support is divided into two components: **(a.)** the depression -> anemia component and **(b.)** the anemia -> AD component.

Support in the structured literature for anemia as a confounder for depression as a risk factor for Alzheimer's disease			
(a.) Anemia -> Depression			
PMID	**Triples**	**Source text from titles and abstracts in the literature**	**Comment**
22307099[115]	Anemia PREDISPOSES depressive disorder	Anaemia predicted raised depression scores 3 weeks later independently of age, gender, marital status, educational attainment, smoking, Global Registry of Acute Cardiac Events (GRACE) risk scores, the negative mood in hospital and history of depression (p=0.003).	This study focused on a cohort of patients with the acute coronary syndrome. Moreover, this statement insinuates a statistical prediction rather than a strictly causal role for anemia. However, the paper did find anemia to be predictive of depression.
32559605[116]	Anemia PREDISPOSES depressive disorder	However, no safety signal regarding anemia for DNG users could be detected, whereas a slight increase in depression risk cannot be excluded but might be explained by baseline severity of endometriosis or unknown country-specific confounding variables.	This paper was focused on hormonal endometriosis treatments rather than looking at anemia and depression directly. However, patients with anemia taking Dienogest (DNG) to treat endometriosis were at slightly increased risk of developing depression, though this may be partially explainable by confounding.
(b.) Anemia -> AD			
PMID	**Triple**	**Sentence**	
29954452[117]	Anemia CAUSES Alzheimer's disease	Red blood cell indices and anemia as causative factors for cognitive function deficits and Alzheimer's disease	Source text has a precise causal interpretation. This paper used Mendelian randomization using UK BioBank data to investigate the relationship between hemoglobin level and anemia and cognitive performance, concluding that the relationship was likely causal.

Figures 9 and 10 present the support from the KG for anemia as a collider and mediator, respectively, of depression and AD.

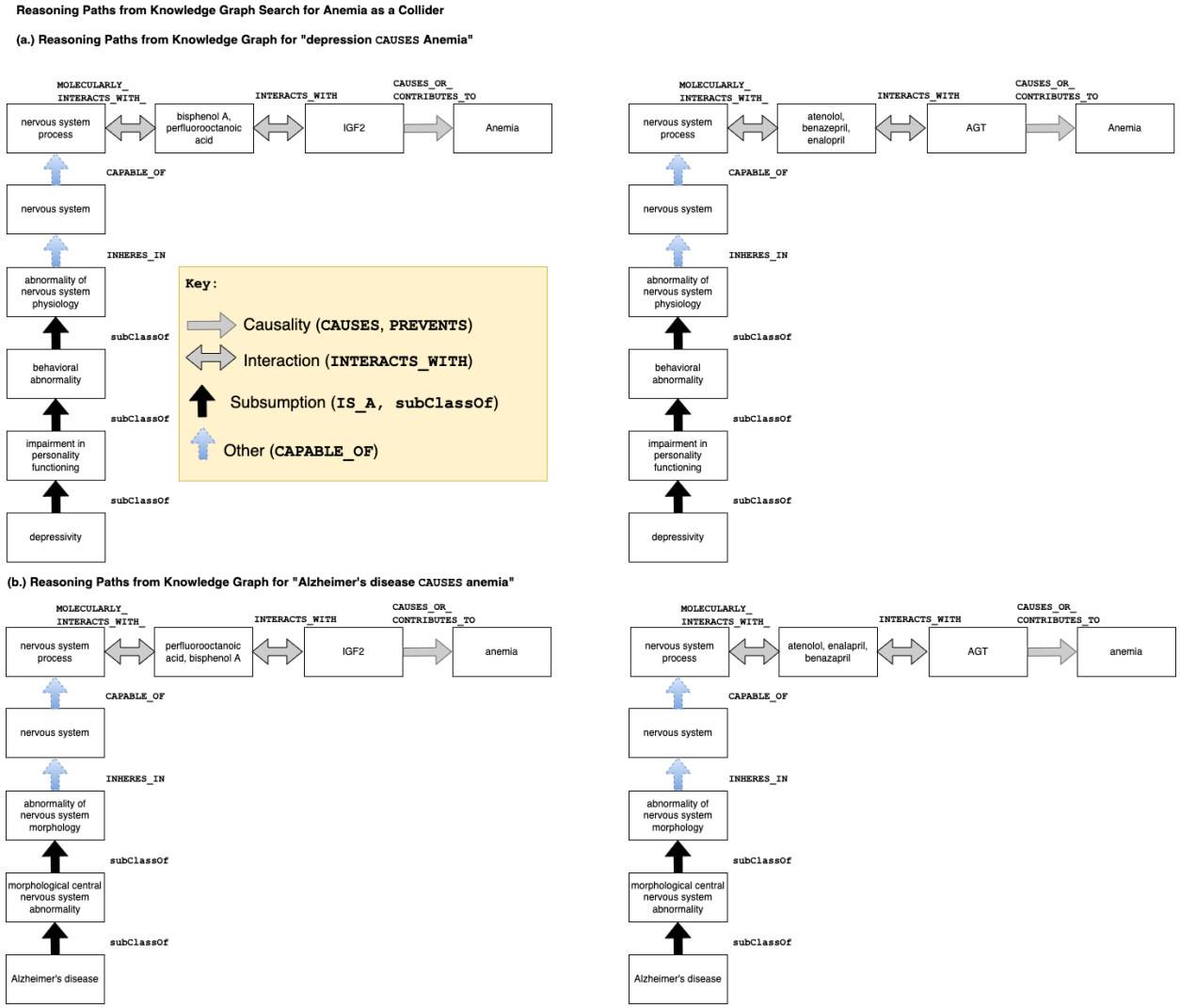


Figure 9. Support in the KG for anemia as a collider for depression and AD. Both reasoning paths on the left **(a.)** and **(b.)** show how IGF2 mediates between depression and environmental exposures (e.g., perfluorooctanoic acid and bisphenol A) on the one hand and anemia on the other. Angiotensinogen (AGT) is the name of a gene and enzyme coded by that gene involved in blood pressure and maintaining fluid and electrolyte hemeostasis[118]. Both reasoning paths on the right **(a.)** and **(b.)** show how AGT mediates between AD and atenolol, benazepril, and enalapril, which are drugs known collectively as angiotensin-converting enzyme inhibitors, or (ACEIs), on the one hand, and anemia on the other. These drugs are known to cause anemia[119]. Numerous studies indicate AD as a potential cause of anemia[120,121].

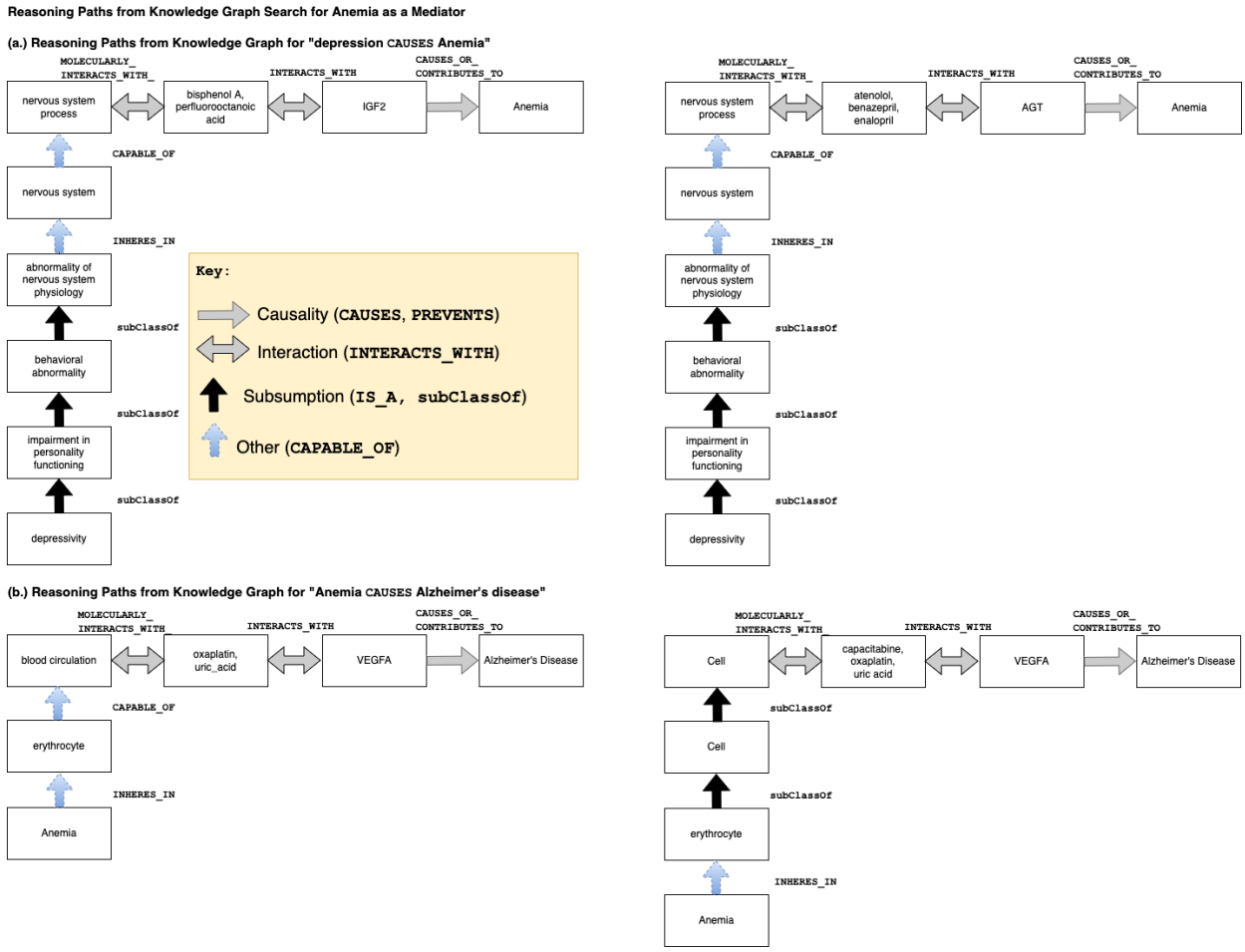


Figure 10. Support in the KG for anemia as a mediator for depression and AD. Note that **(a.)** contains the same information as Figure 9 **(a.)**. Vascular endothelial growth factor A (VEGFA) promotes the proliferation and migration of vascular endothelial cells and is upregulated in tumorogenesis[122]. VEGFA is also upregulated in AD[123].

As shown in Figure 10, anemia is indeed in a complex relationship. Bisphenol A, an endocrine disruptor and an everyday exposure, disrupts the iron cycle, and low iron decreases dopamine which causes low mood and depression causes brain atrophy associated with cognitive decline. Tentatively, our verdict is that anemia is a potential confounder and possibly a mediator, though the evidence provided is not very strong.

Searching PubMed, we found strong support for AD causing anemia with an adjusted OR = 3.41 (95% CI 1.68, 6.92) reported in Faux et al.[120], and was explained mechanistically with decreased plasma iron levels in AD[121]. We also found support for anemia causing AD. A recent paper using a robust study design explored the relationship between anemia and AD[124]. The likely causal relationship between anemia and depression is well known[125,126]. Most surprisingly, we found support for depression causing anemia. Vulser et al. found a dose-response relationship between depression severity and anemia[127].

3.4.2 Covariates reported in observational studies of depression and AD

Table 8 shows that there is a high degree of inconsistency regarding which confounders are adjusted for in the published literature investigating depression and AD. While demographic variables such as age and sex are consistently reported, there is considerable inconsistency in the other variables.

Table 8. Examples of reported confounders from published studies investigating depression and AD (or dementia).

PMID	Exposure	Outcome	Confounders
27138970[128]	depression and dementia	Dementia	"model 1 adjusted for age and sex only, and model 2 adjusted additionally for APOEε4 carrier status, educational level, body-mass index, smoking, alcohol consumption, cognitive score, use of antidepressants, and prevalent disease status at baseline"
33726788[129]	depression and cerebrovascular disease	Dementia	age, sex, residential area, income level, comorbidities (myocardial infarction, congestive heart failure, peripheral vascular disease, chronic pulmonary disease, connective tissue disorder, connective tissue disorder, peptic ulcer, mild liver disease, uncomplicated and complicated diabetes, hemiplegia, non-metastatic solid cancer, moderate or severe liver diseases, metastatic solid cancer)
23154050[130]	Depression, mediated by AD pathology	AD	baseline cognition, age, education, the total number of medications, and brain weight at autopsy
26746105[131]	depression	AD	age, sex, education, history of head trauma, and apolipoprotein E status
19485655[132]	depression	dementia	age, gender, zygosity, and the number of years of education completed

5. Discussion

5.1 Interpretation of results

Our findings suggest that our novel causal feature selection method can identify plausible confounder, collider, and mediator variables for depression and AD (**H1**). A substantial number of variables were found to play multiple roles. The extent to which specific candidate confounders could behave in unexpected ways that violate the strict definition of confounding further confirms the complexity of variables required for studying the relationship between depression and AD (**H2**). KG search produced interpretable mechanistic explanations in the form of reasoning paths (**H3**). Finally, the KG and $\text{SemMedDB}_{\text{Complete}}$ search results did not always agree about either the variables or the roles, and there were potentially essential confounders identified by human subject-matter experts that were not identified by our method. Note however that SemMedDB is not a reference standard and that the status of the variables identified can only be determined by expert review.

In contrast to systems based on assertions from the literature, where source sentences provide human-auditable provenance, the explanatory capability of KG search is provided by the reasoning pathways. The

reasoning pathways provide supporting evidence by containing a trace of the nodes and predicates traced executing the SPARQL query. The reasoning pathway furnishes a mechanistic explanation that is in essence a hypothesis or theory justifying the causal role(s) identified for that variable, filtered through ontology-grounded resources.

Upon inspection, the reasoning pathways, which informally represent mechanistic hypotheses, seemed to be biologically plausible. Although there were problems from the reasoning pathways from the including of generic nodes, we felt that it had a strong explanation capability that could adequately clarify how variables were inferred from the source ontologies.

The review by a human expert in AD confirmed the likelihood that the variables identified, including chronic infectious disease, obstructive sleep apnea, and vitamin D deficiency, identified by both the KG and $\text{SemMedDB}_{\text{Complete}}$, were likely confounders. However, in our inexhaustive, exemplary review of reported covariates from observational studies investigating depression and AD, these variables were not included as covariates. We interpret this to mean that causal feature selection method achieved a major objective in discovering new confounders, beyond those that have been reported in existing observational studies. We also note that there was inconsistency in the reported covariates, indicating a opportunities to standardize confounder adjustment. For discussion in greater depth on the inconsistency of reported covariates, we refer interested readers here[133–135].

In our comparative evaluation, comparing outputs between the KG and $\text{SemMedDB}_{\text{Complete}}$, we found a moderate but substantial concurrence between knowledge bases regarding the variables identified and the causal roles played by those variables. We found contradictions, and a significant portion of the KG and $\text{SemMedDB}_{\text{Complete}}$ results did not co-occur. A partial explanation is that the KG and the structured literature-derived knowledge derive from completely different processes, origins, and purposes. The KG uses information derived from granular, ontology-grounded knowledge and $\text{SemMedDB}_{\text{Complete}}$ relies on the output of advanced NLP systems. Some concepts appeared in the KG that did not appear in $\text{SemMedDB}_{\text{Complete}}$, e.g., inflammatory response (although a literature search produces promising evidence supporting inflammatory response as a confounder[136–140]), and vice versa, e.g., leukoencephalopathy. $\text{SemMedDB}_{\text{Complete}}$ had a better recall of lifestyle-related confounders. The KG excelled at biological plausibility. Arguably, the KG and the literature knowledge bases complement one another in their coverage. $\text{SemMedDB}_{\text{Complete}}$ will have a different set of assertions than what $\text{MachineReadingDB}_{\text{Scoped}}$ has or that make sense through the distillation process searching the combined KG. These different sets of assertions can

contradict one another, resulting in topologies for the variable/nodes that reflect those contradictory assertions. Of this information, it is not known what is erroneous or not until these components undergo verification by humans.

Human experts suggested reasonable confounders not included in the KG search results. For example, two external correspondents with backgrounds in neurology noted two important confounders missing from the list of confounders identified by the KG and the literature: pesticides and adverse childhood experiences. The scoping of the literature also likely impacted the results. For example, the pesticide was present when searching $SemMedDB_{Complete}$. In addition, childhood trauma has been noted as a potential risk factor for depression and AD[141]. However, "adverse childhood event" was missing from $SemMedDB_{Complete}$, further suggesting the importance of a human-in-the-loop process for identifying potential confounders.

To the best of our knowledge, the literature has identified covariates that act as either confounder/mediators or confounder/colliders, but not all three roles, i.e., "the chimera." We verified the plausibility of stroke as an example of a chimera for depression and AD. Since chimeras are a new category of variable, there has been no work to our present knowledge has been performed to identify chimeras explicitly.

5.2 Related Work

There is a vast, and still growing, literature on integrating and using forms of computable biomedical knowledge. For example, related work has been done in defining ontologies for representing knowledge for neurodegenerative diseases[142], identifying erroneous information by applying transitive closure over causal predicates[142], modeling argument structure[143], detecting contradictions in statements extracted from the literature[144,145], drug repurposing pipelines[31,32,34,146,147], and elucidating mechanisms in systems and molecular biology[39]. Simply put, the vast body tangentially related work would be impossible to condense here. However, in the interest of space, since that literature is so vast, we will confine discussion to a review related work regardless of application area on the problem of pruning spurious nodes resulting from erroneous information.

Uncritical use of computable knowledge can result in a causal graph with low-precision[34]. Thus, the development of methods for filtering out erroneous information is arguably a major hindrance to automating causal feature selection, and has spurred much creativity developing novel approaches to address this problem. For example, Hu et al. used information from authoritative ontologies to orient edges of a graphical causal model

studying risk factors for AD[35]. In the approach of Nordon et al., edges were removed if the variables were not correlated in their corpus of electronic medical record data[34]. Incompleteness of knowledge was another major problem, and approaches exist that attempt to solve both problems by using concept similarity measures to filter out erroneous edges, while inferring new edges. Zhang et al. report using neural graph embedding method PubMedBioBert[32]. Malec et al. report used cutoff scores for string-based and vector-based confounder search[60,148].

Our novel causal feature selection method builds on and differs from previous research mentioned in the following regards. In contrast to the previous work, our method combines the breadth of literature-derived information with the precision of ontology-grounded resources. In addition, we introduce refinements to confounder search that allows us to identify a comprehensive set of variables with which to adjust without inducing bias (by identifying and filtering out potential colliders and mediators).

The present work builds on the idea of a discovery pattern, an essential notion to our approach. Originating in the field of literature-based discovery, an area pioneered by Don Swanson, discovery patterns were originally developed to identify potential therapeutic hypotheses[149,150] and were manually constructed. More recently, fresh applications for discovery patterns have been identified and methods have been developed to infer discovery patterns automatically[151,152]. Discovery patterns can be and have been implemented in any querying or knowledge representation language that takes into account how concepts relate to one another, including SPARQL[153], SQL[154], and advanced methods for querying distributed semantic representations[60,148,155,156] of structured knowledge.

Related to the present work, in its explanatory aspect, would be a paper by Elsworth et al. using literature-derived triples to identify intermediate concepts[157]. However, the objective of their paper was focused on identifying mediating mechanisms from a theoretical perspective rather than the context of identifying contextual knowledge relevant for data-driven studies. Other notable work related to explanation would be the EpiphaNet project, an interactive knowledge discovery system developed by Cohen et al.[158] EpiphaNet allows scientists to navigate relationships that interest them in a visual and interactive way.

Other prior work has applied structure learning methods to observational data to recapitulate known and unknown edges between the biomedical entities under the models[159], and providing automated cogent explanatory pathways relating drugs and adverse events[28].

5.3 Strengths and limitations

A significant strength of the present work were utilizing ontology-grounded knowledge resources in the KG. The ontology-grounded KG helped prune erroneous information and providing mechanistic explanations in the form of reasoning paths that describe how the variables identified were related to depression and AD.

A methodological limitation may have arisen in our search procedure. In the interest of search efficiency, we constrained the max length of our search. The KG search also identified genes, enzymes, and drug exposures as potential confounders. However, the KG search only produced output for phenotypes playing collider and mediator roles. Hence, for illustration purposes, our analysis focused primarily on confounders that were conditions or phenotypes. We suspect that the threshold applied to limit search depth may have constrained the results, resulting in lower recall than a deeper search in these categories. Several variables in these categories were recapitulated in our literature review in Table 8, e.g., homocysteine, APOEε4 carrier status.

The present study successfully incorporated output from different machine reading systems with distinct but complementary purposes, beginning with their respective focus domains. For example, SemRep excels at extracting pharmacogenomic and clinical knowledge, whereas INDRA/REACH excels with detailed mechanistic knowledge of proteins, thus facilitating the linkage between detailed mechanistic and clinical knowledge. However, we could not capitalize on this knowledge because of imperfect terminology mapping. This form of information loss arises when a concept appears in the source terminology that is missing in the target. As seen in the PRISMA diagram in Figure 3, many concepts identified in the literature were lost in translation undergoing the terminology mapping procedure. There was also information loss because of a lack of granularity in the mappings we were able to find. For example, all of the post-translational modifications were mapped to the predicate `BIOMOLECULARLY_INTERACTS_WITH`, which is not unidirectional in its meaning- molecularly interacts with is associational. As a result, we may have lost information about directionality in some of the predicates. In other cases, inferences may have been erroneous.

Another limitation was the small size of the literature corpus and the small size likely had negative side effects on the recall of variables and roles. Also, since we scoped the inputs from the literature-derived computable knowledge to focus exclusively on AD and the identification of confounders requires comprehensive knowledge of

both the exposure and outcome mechanisms, we likely missed triples having to do with our exposure, depression, limiting the ability to identify confounders. A larger and more representative sample of the knowledge in the literature could partially address the missing bits of knowledge.

There is always the risk of latent and unmeasured confounding regardless of the causal feature selection method. Knowledge, like all things human, is never complete or perfec. We can only do our best.

5.4 Future directions

The manual verification exercises demonstrated how KG and structured literature search can be fruitfully combined with manual review searching PubMed to adjudicate KG search output. However, human experts would likely require more than just source sentences to adjudicate the causal roles of identified variables. More work needs to be done to provide more comprehensive contextual information about the source material and to operationalize creating diagrams from the reasoning paths. In the present work, we built the diagrams manually. In future work, we will develop tools that automatically compile and generate causal graphs and reasoning path diagrams after collecting requirements from human content-experts. Explanatory power and ease of use are critical for engaging with human content-matter experts who cannot be expected to be the most technologically savvy.

We are currently preparing a manuscript on our work with epidemiologists of aging, testing the effect of adjusting for confounding under different feature selection criteria, including variables from the “Confounders only” category. The variables in this category could be said to be a potentially strong starting point for an adjustment sets. These concepts fulfill the definition of a confounder as a mutual cause of the exposure (depression) and the outcome (AD) of interest without being effects of either depression or AD and have been vetted by two distinct causal feature selection methods but require review by human content-matter experts.

In another project, we have obtained Institutional Review Board approval to initiate a pilot project gathering input from human subject-matter experts concerning variables identified via our workflow to derive a reference standard for a subset of modifiable risk factors. These efforts will help us refine our method and our means of conveying information to subject-matter experts for review. An automated explanation capability would significantly facilitate the investigation of causal feature selection methods to facilitate review.

Additionally, in parallel work, we are working on a method extending our work here by addressing the problem of "M"-bias. "M"-bias is a type of collider bias, wherein the bias is amplified by inducing a spurious dependency between the exposure and the outcome conditioning on a collider[160,161]. In "M"-bias, there are confounders that are themselves effects of two or more other confounders. To reduce bias in this scenario, causal relationships between the confounders must be considered.

5.5 Practical Implications

Recent work has advocated for the development of processes for the systematic review of potential confounders by subject-matter experts with the objective of the creation of field-specific databases of confounders[162]. reporting causal DAGs in observational research[162]. Furthermore, Tennant et al. suggest reporting variables measured in observational studies and variables that researchers would have liked to have measured but were lacking in their current study[163]. They also suggest using a standardized format to report causal DAGs such as DAGitty[164]. Such reporting standards, employing FAIR (fair, accessible, interoperable, and reusable) data principles[165] ensure shareability and, if adopted, could make observational studies more rigorous and reproducible and allow researchers to build more concretely on each other's work, making incremental, but strident advances upon each iteration to answer difficult causal questions from large sets of non-randomized patient health data "across sites and studies" as suggested by Blum et al.[162].

Observational studies that employ a more comprehensive, refined adjustment set and standardization of adjustment sets by expert review could improve the rigor of observational studies (by standardizing adjustment, a point raised above) and bolster the precision of causal effect estimates. This information could then be applied to facilitate causal inference at scale over large centralized and federated data stores, such as PCORnet[166,167] and *allofus*[168].

Lastly, the following bullet points represent the contributions of this work: (1.) contributes information on potential confounders to consider when designing rigorous observational studies of depression and AD; (2.) provides explanatory pathways providing mechanistic hypotheses for the causal edges identified; and (3.) shows how semantic and causal inference can be partially combined.

6. Conclusion

We found that the KG search can identify potential confounder, collider, and mediator variables related to depression as the exposure and AD as the outcome. Our method partially addressed the limitations of relying exclusively on human subject-matter experts and the incompleteness and inaccuracy of literature-derived structured knowledge. KGs such as we have described in this paper are a natural way to represent such knowledge and a convenient way to automate reasoning over this knowledge to generate hypotheses about confounders. This work also shows how semantic inference (over subsumptive and transitive causal relationships) can enhance causal inference and demonstrates the possibility of programmatically unifying these modes of inference. The complexity of variables in this domain, e.g., chimeras, powerfully highlights the value of a combined machine-human strategy, where machines can excel at searching for knowledge linking clinical phenotypes with mechanistic molecular processes at scale, and humans can focus on understanding the appropriateness of the variables identified for their context. In particular, a workflow that would augment the unique strengths of human judgment, intuition, and experience with input from computational tools for harnessing computable knowledge could help scientists construct more rigorous observational studies that could help prioritize public health interventions and discover strategies for preventing or delaying the progression of AD with greater odds of success. The methods described herein can be easily extended to answer new questions in other areas. However, more work must be done to improve the suitability of the KG search for causal feature selection. Future research can build on these methods, advance the rigor of observational studies and prioritize public health interventions with a greater likelihood of success.

Supplementary Information

Supplementary Information accompanies this paper as described below.

Additional files: File I. ***Appendix_A.docx***. **Background S1.** Confounders, Colliders, Mediators. **Background S1.** Transitive and symmetric closure. **File II.** ***Appendix_B.docx***. **Listing S1.** PubMed scoping query. **Table S1.** Mappings for UMLS CUIs to RO. **Table S2.** UMLS CUI definitions for depression and AD. **Figure S1.** Example confounder reasoning path. **Figure S1.** Example collider reasoning path. **Figure S1.** Example mediator reasoning path. **Table S3.** Complete analysis of KG search with SemMedDBs. **Table S4.** Comparison of confounders - search results belonging to the gene or enzyme, protein, or semantic categories. **Table S5.** Comparison of confounders - search results belonging to the drug or hormone semantic categories. **File III.** ***AdjustmentSetDepression2AD.xlsx***. **Worksheet S1.** Conditions. **Worksheet S2.** Substances. **Worksheet S3.** GenesEnzymes. **File IV.** ***predicatePropertyAnalysis.xlsx***. This file contains our analysis of the logical properties of symmetry and transitivity for the predicates from the machine reading systems. **File V.** ***ConfounderSourceSentences.xlsx***. This file contains source sentences for three potential confounders that have not been reported as covariates in published observational studies looking at depression and AD. **File VI.** ***ChimeraSourceSentences.xlsx***. This file contains source sentences for stroke as a chimeric variable for depression and AD.

Acknowledgments:

We would like to thank health sciences librarian Helena M. Vonville for assisting in designing the PubMed query to scope the literature. Previous versions of this paper benefited from the suggestions and comments from Professors Gregory F. Cooper and Howard Aizenstein.

Funding

The US National Library of Medicine grants supported this research, including R01 LM011563, 5 T15 LM007059-32, NIH/BD2K supplement R01 LM011563-02S1, and K99 LM013367-01A1; US National Institutes of Aging grant T32 AG055381; and the University of Pittsburgh's Momentum Funds. The views expressed here do not necessarily reflect the views of the National Library of Medicine of the National Institutes of Health.

Availability of data and materials

The data supporting these findings are available from this project's Zenodo repository (https://10.5281/zenodo.6785307). No restrictions apply to the public availability of these data.

Authors' Contributions

SAM had full access to all of the data in this study and took full responsibility for the integrity and accuracy of the data analysis. RDB supervised the study. SBT and RDB conceived of the graph querying methods with SAM's input. RDB and SAM conceived and executed the initial study design and SAM analyzed the study data, interpreted study findings, and drafted the manuscript. TJC provided invaluable input in constructing the KG. SMA, CES, HTK, ASL, TJC, and PWM critically reviewed drafts of the manuscript. The corresponding author attests that all listed authors meet the authorship criteria and that no other meeting the criteria have been omitted. The authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Cartwright N. Are RCTs the Gold Standard? BioSocieties. 2007 Mar;2(1):11–20.
2. Cartwright N. A philosopher's view of the long road from RCTs to effectiveness. The Lancet. 2011 Apr 23;377(9775):1400–1.
3. VanderWeele TJ, Shpitser I. On the definition of a confounder. Ann Stat. 2013 Feb;41(1):196–220.
4. Groenwold RHH, Sterne JAC, Lawlor DA, Moons KGM, Hoes AW, Tilling K. Sensitivity analysis for the effects of multiple unmeasured confounders. Ann Epidemiol. 2016 Sep 1;26(9):605–11.
5. Pearl J. Causality: Models, Reasoning, and Inference [Internet]. 2nd ed. Cambridge: Cambridge University Press; 2009 [cited 2017 Jul 21]. Available from: http://ebooks.cambridge.org/ref/id/CBO9780511803161
6. Patorno E, Glynn RJ, Hernandez-Diaz S, Liu J, Schneeweiss S. Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. Epidemiol Camb Mass. 2014 Mar;25(2):268–78.
7. Talbot D, Diop A, Lavigne-Robichaud M, Brisson C. The change in estimate method for selecting confounders: A simulation study. Stat Methods Med Res. 2021 Sep;30(9):2032–44.
8. Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, Franklin JM. Variable Selection for Confounding Adjustment in High-dimensional Covariate Spaces When Analyzing Healthcare Databases. Vol. 28. United States; 2017.
9. Shpitser I, VanderWeele T, Robins JM. On the Validity of Covariate Adjustment for Estimating Causal Effects. :10.
10. VanderWeele TJ. Principles of confounder selection. Eur J Epidemiol. 2019;34(3):211–9.
11. Arntzenius F. Reichenbach's Common Cause Principle. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy [Internet]. Fall 2010. Metaphysics Research Lab, Stanford University; 2010 [cited 2019 Dec 10]. p. 1. Available from: https://plato.stanford.edu/archives/fall2010/entries/physics-Rpcc/
12. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. Biometrics. 2011 Dec;67(4):1406–13.
13. MEDLINE PubMed Production Statistics [Internet]. U.S. National Library of Medicine; [cited 2022 Jul 3].

Available from: https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html
14. Hernan MA, Robins JM. Causal Inference [Internet]. Taylor & Francis; 2017. (Chapman & Hall/CRC Monographs on Statistics & Applied Probab). Available from: https://books.google.com/books?id=_KnHIAAACAAJ
15. Karim ME, Tremlett H, Zhu F, Petkau J, Kingwell E. Dealing with Treatment-confounder Feedback and Sparse Follow-up in Longitudinal studies - Application of a Marginal Structural Model in a Multiple Sclerosis Cohort. Am J Epidemiol. 2020 Oct 30;
16. Jackson JW. Diagnostics for confounding of time-varying and other joint exposures. Epidemiol Camb Mass. 2016 Nov;27(6):859–69.
17. Robins J, Hernan M. Estimation of the causal effects of time-varying exposure. In: Longitudinal Data Analysis. 2008. p. 553–99.
18. Sackett DL. Bias in analytic research. J Chronic Dis. 1979;32(1–2):51–63.
19. Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. Annu Rev Sociol. 2014 Jul;40:31–53.
20. Steiner PM, Kim Y. The Mechanics of Omitted Variable Bias: Bias Amplification and Cancellation of Offsetting Biases. J Causal Inference. 2016 Sep;4(2):20160009.
21. Abrahamowicz M, Bjerre LM, Beauchamp ME, LeLorier J, Burne R. The missing cause approach to unmeasured confounding in pharmacoepidemiology. Stat Med. 2016 Mar 30;35(7):1001–16.
22. Grätz M. When Less Conditioning Provides Better Estimates: Overcontrol and Collider Bias in Research on Intergenerational Mobility [Internet]. Working Paper Series. Stockholm University, Swedish Institute for Social Research; 2019 Jun [cited 2020 Jun 16]. (Working Paper Series). Report No.: 2/2019. Available from: https://ideas.repec.org/p/hhs/sofiwp/2019_002.html
23. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010 Apr;39(2):417–20.
24. Banack HR, Kaufman JS. From bad to worse: collider stratification amplifies confounding bias in the "obesity paradox". Eur J Epidemiol. 2015 Oct;30(10):1111–4.
25. Mower J, Subramanian D, Shang N, Cohen T. Classification-by-Analogy: Using Vector Representations of Implicit Relationships to Identify Plausibly Causal Drug/Side-effect Relationships. AMIA Annu Symp Proc AMIA Symp. 2016;2016:1940–9.
26. Mower J, Cohen T, Subramanian D. Complementing Observational Signal with Literature-derived Distributed Representations for Post-marketing Drug Surveillance. Drug Saf. 2020 Jan;43(1):67–77.
27. Mower J, Subramanian D, Cohen T. Learning predictive models of drug side-effect relationships from distributed representations of literature-derived semantic predications. J Am Med Inform Assoc. 2018 Jul 11;ocy077–ocy077.
28. Shang N, Xu H, Rindflesch TC, Cohen T. Identifying plausible adverse drug reactions using knowledge extracted from the literature. J Biomed Inform. 2014 Dec;52:293–310.
29. Fathiamini S, Johnson AM, Zeng J, Araya A, Holla V, Bailey AM, et al. Automated identification of molecular effects of drugs (AIMED). J Am Med Inform Assoc JAMIA. 2016 Jul;23(4):758–65.
30. Cohen T, Widdows D, De Vine L, Schvaneveldt R, Rindflesch TC. Many Paths Lead to Discovery: Analogical Retrieval of Cancer Therapies. In: Busemeyer JR, Dubois F, Lambert-Mogiliansky A, Melucci M, editors. Quantum Interaction. Berlin, Heidelberg: Springer; 2012. p. 90–101. (Lecture Notes in Computer Science).
31. Nordon G, Koren G, Shalev V, Horvitz E, Radinsky K. Separating Wheat from Chaff: Joining Biomedical Knowledge and Patient Data for Repurposing Medications. Proc AAAI Conf Artif Intell. 2019 Jul 17;33(01):9565–72.
32. Zhang R, Hristovski D, Schutte D, Kastrin A, Fiszman M, Kilicoglu H. Drug repurposing for COVID-19 via knowledge graph completion. J Biomed Inform. 2021 Mar;115:103696.
33. Müller HM, Kenny EE, Sternberg PW. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLoS Biol. 2004 Nov;2(11):e309.
34. Nordon G, Koren G, Shalev V, Kimelfeld B, Shalit U, Radinsky K. Building Causal Graphs from Medical Literature and Electronic Medical Records. Proc AAAI Conf Artif Intell. 2019 Jul 17;33(01):1102–9.
35. Hu H, Kerschberg L. Improved Causal Models of Alzheimer's Disease. In: 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC). 2021. p. 274–83.
36. Besnard P, Cordier M o, Moinard Y. Inferring Causal Explanations. Hunt Parsons Eds. 1999;99:55–67.
37. Friedman S, Magnusson I, Sarathy V, Schmer-Galunder S. From Unstructured Text to Causal Knowledge

Graphs: A Transformer-Based Approach. ArXiv220211768 Cs [Internet]. 2022 Feb 23 [cited 2022 Feb 25]; Available from: http://arxiv.org/abs/2202.11768
38. Thomas PD, Hill DP, Mi H, Osumi-Sutherland D, Van Auken K, Carbon S, et al. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat Genet. 2019 Oct;51(10):1429–33.
39. Tripodi IJ, Callahan TJ, Westfall JT, Meitzer NS, Dowell RD, Hunter LE. Applying knowledge-driven mechanistic inference to toxicogenomics. Toxicol Vitro Int J Publ Assoc BIBRA. 2020 Aug;66:104877.
40. Sarvet AL, Stensrud MJ. Without Commitment to an Ontology, There Could Be No Causal Inference. Epidemiology. 2022 May;33(3):372–8.
41. Kejriwal M. Domain-Specific Knowledge Graph Construction [Internet]. Springer International Publishing; 2019. Available from: https://books.google.com/books?id=naHrvwEACAAJ
42. Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet. 2020 Oct 17;396(10258):1204–22.
43. Komarova NL, Thalhauser CJ. High Degree of Heterogeneity in Alzheimer's Disease Progression Patterns. PLOS Comput Biol. 2011 Nov 3;7(11):e1002251.
44. Association AP. DSM 5 Diagnostic and statistical manual of mental disorders. DSM 5 Diagn Stat Man Ment Disord. 2013;947 p.-947 p.
45. Bains N, Abdijadid S. Major Depressive Disorder. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 [cited 2022 Jun 5]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK559078/
46. Cai N, Choi KW, Fried EI. Reviewing the genetics of heterogeneity in depression: operationalizations, manifestations and etiologies. Hum Mol Genet. 2020 Sep 30;29(R1):R10–8.
47. Karim HT. The Elusive "White Whale" of Treatment Response Prediction: Leveraging the Curse of Heterogeneity in Late-Life Depression. Am J Geriatr Psychiatry Off J Am Assoc Geriatr Psychiatry. 2021 Apr 9;S1064-7481(21)00288-8.
48. Devi G, Scheltens P. Heterogeneity of Alzheimer's disease: consequence for drug trials? Alzheimers Res Ther [Internet]. 2018 Dec 19 [cited 2019 Oct 2];10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6300886/
49. Butters MA, Young JB, Lopez O, Aizenstein HJ, Mulsant BH, Reynolds CF, et al. Pathways linking late-life depression to persistent cognitive impairment and dementia. Dialogues Clin Neurosci. 2008;10(3):345–57.
50. Koenig AM, Bhalla RK, Butters MA. Cognitive functioning and late-life depression. J Int Neuropsychol Soc JINS. 2014 May;20(5):461–7.
51. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet Lond Engl. 2020 Aug 8;396(10248):413–46.
52. Kilicoglu H, Rosemblat G, Fiszman M, Shin D. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics. 2020 May 14;21(1):188.
53. Kilicoglu H, Rosemblat G, Fiszman M, Rindflesch TC. Constructing a semantic predication gold standard from the biomedical literature. BMC Bioinformatics. 2011 Dec 20;12:486.
54. Rindflesch TC, Blake CL, Fiszman M, Kilicoglu H, Rosemblat G, Schneider J, et al. Informatics Support for Basic Research in Biomedicine. ILAR J. 2017 Jul 1;58(1):80–9.
55. Gyori BM, Bachman JA, Subramanian K, Muhlich JL, Galescu L, Sorger PK. From word models to executable models of signaling networks using automated assembly. Mol Syst Biol [Internet]. 2017 Nov 24 [cited 2020 Aug 14];13(11). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5731347/
56. VanderWeele T. Explanation in Causal Inference: Methods for Mediation and Interaction [Internet]. Oxford University Press; 2015. Available from: https://books.google.com/books?id=K6cgBgAAQBAJ
57. Aalen OO, Roysland K, Gran JM, Ledergerber B. Causality, mediation and time: a dynamic viewpoint. J R Stat Soc Ser A Stat Soc. 2012 Oct;175(4):831–61.
58. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. Int J Epidemiol. 2013 Oct;42(5):1511–9.
59. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA Annu Symp Proc AMIA Symp. 2006;349–53.
60. Malec SA, Wei P, Bernstam EV, Boyce RD, Cohen T. Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance. J Biomed Inform. 2021 May 1;117:103719.

61. Cafasso M. noxdafox/clipspy [Internet]. 2020 [cited 2020 Aug 8]. Available from: https://github.com/noxdafox/clipspy
62. CLIPS: A Tool for Building Expert Systems [Internet]. [cited 2022 Mar 20]. Available from: http://www.clipsrules.net/
63. Klunk WE, Engler H, Nordberg A, Wang Y, Blomqvist G, Holt DP, et al. Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. Ann Neurol. 2004 Mar;55(3):306–19.
64. Weiner MW, Aisen PS, Jack CR, Jagust WJ, Trojanowski JQ, Shaw L, et al. The Alzheimer's Disease Neuroimaging Initiative: Progress report and future plans. Alzheimers Dement J Alzheimers Assoc. 2010 May;6(3):202-11.e7.
65. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17–21.
66. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May;17(3):229–36.
67. Semantic Types and Groups - MetaMap documentation [Internet]. [cited 2022 Feb 21]. Available from: https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/documentation/SemanticTypesAndGroups.html
68. Sharp R, Pyarelal A, Gyori B, Alcock K, Laparra E, Valenzuela-Escárcega MA, et al. Eidos, INDRA, & Delphi: From Free Text to Executable Causal Models. Proc 2019 Conf North Am Chapter Assoc Comput Linguist Demonstr. 2019;6.
69. clulab/eidos [Internet]. Computational Language Understanding Lab (CLU Lab) at University of Arizona; 2020 [cited 2020 Aug 14]. Available from: https://github.com/clulab/eidos
70. Bazrgar M, Khodabakhsh P, Mohagheghi F, Prudencio M, Ahmadiani A. Brain microRNAs dysregulation: Implication for missplicing and abnormal post-translational modifications of tau protein in Alzheimer's disease and related tauopathies. Pharmacol Res. 2020;155:104729.
71. Marcelli S, Corbo M, Iannuzzi F, Negri L, Blandini F, Nistico R, et al. The Involvement of Post-Translational Modifications in Alzheimer's Disease. Curr Alzheimer Res. 2018 22;15(4):313–35.
72. Schaffert LN, Carter WG. Do Post-Translational Modifications Influence Protein Aggregation in Neurodegenerative Diseases: A Systematic Review. Brain Sci [Internet]. 2020 Apr 11 [cited 2020 Oct 13];10(4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7226274/
73. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec;36(6):462–77.
74. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinforma Oxf Engl. 2012 Dec 1;28(23):3158–60.
75. Relation Ontology [Internet]. [cited 2022 Mar 20]. Available from: https://obofoundry.org/ontology/ro.html
76. oborel/obo-relations [Internet]. oborel; 2020 [cited 2020 Aug 8]. Available from: https://github.com/oborel/obo-relations
77. Horrocks I, Patel-schneider PF. Knowledge Representation and Reasoning on the Semantic Web: OWL.
78. OWL Web Ontology Language Reference [Internet]. [cited 2022 Mar 28]. Available from: https://www.w3.org/TR/owl-ref/
79. NASA Technical Reports Server (NTRS) [Internet]. [cited 2020 Aug 8]. Available from: https://ntrs.nasa.gov/citations/19910014730
80. Callahan T. PheKnowLator [Internet]. 2019. Available from: https://doi.org/10.5281/zenodo.3401437
81. OWL - Semantic Web Standards [Internet]. [cited 2020 Aug 20]. Available from: https://www.w3.org/OWL/
82. v2 Data Sources · callahantiff/PheKnowLator Wiki [Internet]. GitHub. [cited 2022 Jun 1]. Available from: https://github.com/callahantiff/PheKnowLator
83. NetworkX — NetworkX documentation [Internet]. [cited 2020 Aug 21]. Available from: https://networkx.github.io/
84. OpenLink Software: Virtuoso Homepage [Internet]. [cited 2022 Jun 7]. Available from: https://virtuoso.openlinksw.com/
85. Callahan TJ, Baumgartner WA, Bada M, Stefanski AL, Tripodi I, White EK, et al. OWL-NETS: Transforming OWL Representations for Improved Network Inference. Pac Symp Biocomput Pac Symp Biocomput. 2018;23:133–44.
86. Dijkstra EW. A note on two problems in connexion with graphs. Numer Math. 1959 Dec 1;1(1):269–71.
87. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–9.

88. Gene Ontology Resource [Internet]. Gene Ontology Resource. [cited 2022 Mar 20]. Available from: http://geneontology.org/
89. Human Phenotype Ontology [Internet]. [cited 2022 Mar 20]. Available from: https://hpo.jax.org/app/
90. SOD2 superoxide dismutase 2 [Homo sapiens (human)] - Gene - NCBI [Internet]. [cited 2022 May 28]. Available from: https://www.ncbi.nlm.nih.gov/gene/6648
91. Flynn JM, Melov S. SOD2 in Mitochondrial Dysfunction and Neurodegeneration. Free Radic Biol Med. 2013 Sep;62:10.1016/j.freeradbiomed.2013.05.027.
92. Wiener HW, Perry RT, Chen Z, Harrell LE, Go RCP. A polymorphism in SOD2 is associated with development of Alzheimer's disease. Genes Brain Behav. 2007 Nov;6(8):770–5.
93. Nguyen M, Sabry R, St. John . Elizabeth J, Favetta LA. Bisphenol A and S, but Not F, Alter Oxidative Stress Levels in Spermatozoa. J Endocr Soc. 2021 May 1;5(Supplement_1):A483–4.
94. Braun: Early-life exposure to EDCs: role in childhood... - Google Scholar [Internet]. [cited 2022 Jun 6]. Available from: https://scholar.google.com/scholar_lookup?title=Early-life%20exposure%20to%20EDCs%3A%20role%20in%20childhood%20obesity%20and%20neurodevelopment&publication_year=2017&author=J.M.%20Braun
95. Choi YJ, Lee YA, Hong YC, Cho J, Lee KS, Shin CH, et al. Effect of prenatal bisphenol A exposure on early childhood body mass index through epigenetic influence on the insulin-like growth factor 2 receptor (IGF2R) gene. Environ Int. 2020 Oct 1;143:105929.
96. Bisphenol A: How the Most Relevant Exposure Sources Contribute to Total Consumer Exposure - Von Goetz - 2010 - Risk Analysis - Wiley Online Library [Internet]. [cited 2022 Jun 6]. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1539-6924.2009.01345.x?casa_token=g9CVKHGnY7EAAAAA%3ARU__TZPsV8Fmu_6Qz259EJKD5PsDX5fbWC7wUCDA3sCcCZ3WY9rPuD46POLpkGjM26Ed6ndOSs28RA
97. MTHFR methylenetetrahydrofolate reductase [Homo sapiens (human)] - Gene - NCBI [Internet]. [cited 2022 May 28]. Available from: https://www.ncbi.nlm.nih.gov/gene/4524
98. Dwyer R, Skrobot OA, Dwyer J, Munafo M, Kehoe PG. Using Alzgene-like approaches to investigate susceptibility genes for vascular cognitive impairment. J Alzheimers Dis JAD. 2013;34(1):145–54.
99. Lemche E. Early Life Stress and Epigenetics in Late-onset Alzheimer's Dementia: A Systematic Review. Curr Genomics. 2018 Nov;19(7):522–602.
100. Varga EA, Sturm AC, Misita CP, Moll S. Homocysteine and MTHFR Mutations. Circulation. 2005 May 17;111(19):e289–93.
101. Saczynski JS, Beiser A, Seshadri S, Auerbach S, Wolf PA, Au R. Depressive symptoms and risk of dementia: the Framingham Heart Study. Neurology. 2010 Jul 6;75(1):35–41.
102. Sgroi S, Capper-Loup C, Paganetti P, Kaelin-Lang A. Enkephalin and dynorphin neuropeptides are differently correlated with locomotor hypersensitivity and levodopa-induced dyskinesia in parkinsonian rats. Exp Neurol. 2016 Jun;280:80–8.
103. PubChem. PDYN - prodynorphin (human) [Internet]. [cited 2022 May 29]. Available from: https://pubchem.ncbi.nlm.nih.gov/gene/PDYN/human
104. Kwintkiewicz J, Nishi Y, Yanase T, Giudice LC. Peroxisome Proliferator–Activated Receptor-γ Mediates Bisphenol A Inhibition of FSH-Stimulated IGF-1, Aromatase, and Estradiol in Human Granulosa Cells. Environ Health Perspect. 2010 Mar;118(3):400–6.
105. Pascual-Lucas M, Viana da Silva S, Di Scala M, Garcia-Barroso C, González-Aseguinolaza G, Mulle C, et al. Insulin-like growth factor 2 reverses memory and synaptic deficits in APP transgenic mice. EMBO Mol Med. 2014 Oct;6(10):1246–62.
106. Mellott TJ, Pender SM, Burke RM, Langley EA, Blusztajn JK. IGF2 ameliorates amyloidosis, increases cholinergic marker expression and raises BMP9 and neurotrophin levels in the hippocampus of the APPswePS1dE9 Alzheimer's disease model mice. PloS One. 2014;9(4):e94287.
107. Forré P, Mooij JM. Markov Properties for Graphical Models with Cycles and Latent Variables [Internet]. arXiv; 2017 Oct [cited 2022 Jun 5]. Report No.: arXiv:1710.08775. Available from: http://arxiv.org/abs/1710.08775
108. Lauritzen SL, Richardson TS. Chain graph models and their causal interpretations. J R Stat Soc Ser B Stat Methodol. 2002;64(3):321–48.
109. Das J, G K R. Post stroke depression: The sequelae of cerebral stroke. Neurosci Biobehav Rev. 2018 Jul;90:104–14.

110. Lewin-Richter A, Volz M, Jöbges M, Werheid K. Predictivity of Early Depressive Symptoms for Post-Stroke Depression. J Nutr Health Aging. 2015 Aug;19(7):754–8.
111. Medeiros GC, Roy D, Kontos N, Beach SR. Post-stroke depression: A 2020 updated review. Gen Hosp Psychiatry. 2020 Oct;66:70–80.
112. Lo Coco D, Lopez G, Corrao S. Cognitive impairment and stroke in elderly patients. Vasc Health Risk Manag. 2016;12:105–16.
113. Zhou J, Yu JT, Wang HF, Meng XF, Tan CC, Wang J, et al. Association between stroke and Alzheimer's disease: systematic review and meta-analysis. J Alzheimers Dis JAD. 2015;43(2):479–89.
114. Vijayan M, Reddy PH. Stroke, Vascular Dementia, and Alzheimer's Disease: Molecular Links. J Alzheimers Dis JAD. 2016 Sep 6;54(2):427–43.
115. Steptoe A, Wikman A, Molloy GJ, Kaski JC. Anaemia and the development of depressive symptoms following acute coronary syndrome: longitudinal clinical observational study. BMJ Open. 2012;2(1):e000551.
116. Moehner S, Becker K, Lange JA, von Stockum S, Heinemann K. Risk of depression and anemia in users of hormonal endometriosis treatments: Results from the VIPOS study. Eur J Obstet Gynecol Reprod Biol. 2020 Aug;251:212–7.
117. Winchester LM, Powell J, Lovestone S, Nevado-Holgado AJ. Red blood cell indices and anaemia as causative factors for cognitive function deficits and for Alzheimer's disease. Genome Med. 2018 Jun 28;10(1):51.
118. AGT angiotensinogen [Homo sapiens (human)] - Gene - NCBI [Internet]. [cited 2022 Jun 6]. Available from: https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=183
119. Ajmal A, Gessert CE, Johnson BP, Renier CM, Palcher JA. Effect of angiotensin converting enzyme inhibitors and angiotensin receptor blockers on hemoglobin levels. BMC Res Notes. 2013 Nov 4;6:443.
120. Faux NG, Rembach A, Wiley J, Ellis KA, Ames D, Fowler CJ, et al. An anemia of Alzheimer's disease. Mol Psychiatry. 2014 Nov;19(11):1227–34.
121. Hare DJ, Doecke JD, Faux NG, Rembach A, Volitakis I, Fowler CJ, et al. Decreased plasma iron in Alzheimer's disease is due to transferrin desaturation. ACS Chem Neurosci. 2015 Mar 18;6(3):398–402.
122. Mao Y, Liu X, Song Y, Zhai C, Zhang L. Data from: VEGF-A/VEGFR-2 and FGF-2/FGFR-1 but not PDGF-BB/PDGFR-β play important roles in promoting immature and inflammatory intraplaque angiogenesis [Internet]. Dryad; 2019 [cited 2022 Jun 9]. p. 266587 bytes. Available from: http://datadryad.org/stash/dataset/doi:10.5061/dryad.8tt0411
123. Harris R, Miners JS, Allen S, Love S. VEGFR1 and VEGFR2 in Alzheimer's Disease. J Alzheimers Dis JAD. 2018;61(2):741–52.
124. Jeong SM, Shin DW, Lee JE, Hyeon JH, Lee J, Kim S. Anemia is associated with incidence of dementia: a national health screening study in Korea involving 37,900 persons. Alzheimers Res Ther. 2017 Dec 6;9(1):94.
125. Shafi M, Taufiq F, Mehmood H, Afsar S, Badar A. Relation between Depressive Disorder and Iron Deficiency Anemia among Adults Reporting to a Secondary Healthcare Facility: A Hospital-Based Case Control Study. J Coll Physicians Surg--Pak JCPSP. 2018 Jun;28(6):456–559.
126. Hosseini SR, Zabihi A, Ebrahimi SH, Jafarian Amiri SR, Kheirkhah F, Bijani A. The Prevalence of Anemia and its Association with Depressive Symptoms among Older Adults in North of Iran. J Res Health Sci. 2018 Dec 3;18(4):e00431.
127. Vulser H, Wiernik E, Hoertel N, Thomas F, Pannier B, Czernichow S, et al. Association between depression and anemia in otherwise healthy adults. Acta Psychiatr Scand. 2016 Aug;134(2):150–60.
128. Mirza SS, Wolters FJ, Swanson SA, Koudstaal PJ, Hofman A, Tiemeier H, et al. 10-year trajectories of depressive symptoms and risk of dementia: a population-based study. Lancet Psychiatry. 2016 Jul;3(7):628–35.
129. Jang YJ, Kang C, Myung W, Lim SW, Moon YK, Kim H, et al. Additive interaction of mid- to late-life depression and cerebrovascular disease on the risk of dementia: a nationwide population-based cohort study. Alzheimers Res Ther [Internet]. 2021 [cited 2022 Mar 13];13. Available from: https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC7968260/
130. Royall DR, Palmer RF. Alzheimer's disease pathology does not mediate the association between depressive symptoms and subsequent cognitive decline. Alzheimers Dement J Alzheimers Assoc. 2013 May;9(3):318–25.
131. Herbert J, Lucassen PJ. Depression as a risk factor for Alzheimer's disease: Genes, steroids, cytokines and neurogenesis - What do we need to know? Front Neuroendocrinol. 2016 Apr;41:153–71.
132. Brommelhoff JA, Gatz M, Johansson B, McArdle JJ, Fratiglioni L, Pedersen NL. Depression as a risk factor or prodromal feature for dementia? Findings in a population-based sample of Swedish twins. Psychol

Aging. 2009 Jun;24(2):373–84.
133. Stang PE, Ryan PB, Overhage JM, Schuemie MJ, Hartzema AG, Welebob E. Variation in choice of study design: findings from the Epidemiology Design Decision Inventory and Evaluation (EDDIE) survey. Drug Saf. 2013 Oct;36 Suppl 1:S15-25.
134. Patel CJ, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol. 2015 Sep;68(9):1046–58.
135. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results - R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, B. A. Nosek, 2018 [Internet]. [cited 2022 Jul 10]. Available from: https://journals.sagepub.com/doi/10.1177/2515245917747646
136. Abbott A. Depression: the radical theory linking it to inflammation. Nature. 2018 May 29;557(7707):633–4.
137. Lee CH, Giuliani F. The Role of Inflammation in Depression and Fatigue. Front Immunol. 2019 Jul 19;10:1696.
138. Kinney JW, Bemiller SM, Murtishaw AS, Leisgang AM, Salazar AM, Lamb BT. Inflammation as a central mechanism in Alzheimer's disease. Alzheimers Dement Transl Res Clin Interv. 2018 Sep 6;4:575–90.
139. Hampel H, Caraci F, Cuello AC, Caruso G, Nisticò R, Corbo M, et al. A Path Toward Precision Medicine for Neuroinflammatory Mechanisms in Alzheimer's Disease. Front Immunol [Internet]. 2020 Mar 31 [cited 2021 May 12];11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7137904/
140. Lyra e Silva NM, Gonçalves RA, Pascoal TA, Lima-Filho RAS, Resende E de PF, Vieira ELM, et al. Pro-inflammatory interleukin-6 signaling links cognitive impairments and peripheral metabolic alterations in Alzheimer's disease. Transl Psychiatry. 2021 Apr 28;11(1):1–15.
141. Tani Y, Fujiwara T, Kondo K. Association Between Adverse Childhood Experiences and Dementia in Older Japanese Adults. JAMA Netw Open. 2020 Feb 5;3(2):e1920740.
142. Kahn CE. Transitive closure of subsumption and causal relations in a large ontology of radiological diagnosis. J Biomed Inform. 2016 Jun 1;61:27–33.
143. Ciccarese P, Wu E, Wong G, Ocana M, Kinoshita J, Ruttenberg A, et al. The SWAN biomedical discourse ontology. J Biomed Inform. 2008 Oct;41(5):739–51.
144. Rosemblat G, Fiszman M, Shin D, Kilicoglu H. Towards a characterization of apparent contradictions in the biomedical literature using context analysis. J Biomed Inform. 2019;98:103275.
145. Alamri A. The Detection of Contradictory Claims in Biomedical Abstracts [Internet] [phd]. University of Sheffield; 2016 [cited 2019 Aug 15]. Available from: http://etheses.whiterose.ac.uk/15893/
146. Nordon G, Gottlieb L, Radinsky K. Chemical and Textual Embeddings for Drug Repurposing. :6.
147. Hsieh KL, Plascencia-Villa G, Lin KH, Perry G, Jiang X, Kim Y. Deep Learning for Alzheimer's Disease Drug Repurposing using Knowledge Graph and Multi-level Evidence [Internet]. medRxiv; 2021 [cited 2022 May 2]. p. 2021.12.03.21267235. Available from: https://www.medrxiv.org/content/10.1101/2021.12.03.21267235v1
148. Malec SA, Wei P, Xu H, Bernstam EV, Myneni S, Cohen T. Literature-Based Discovery of Confounding in Observational Clinical Data. AMIA Annu Symp Proc AMIA Symp. 2016;2016:1920–9.
149. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986 Autumn;30(1):7–18.
150. Smalheiser NR. Rediscovering Don Swanson: the Past, Present and Future of Literature-Based Discovery. J Data Inf Sci Wars Pol. 2017 Dec;2(4):43–64.
151. Cohen T, Widdows D, Schvaneveldt RW, Davies P, Rindflesch TC. Discovering discovery patterns with Predication-based Semantic Indexing. J Biomed Inform. 2012 Dec;45(6):1049–65.
152. Cohen T, Widdows D, Schvaneveldt R, Rindflesch TC. Finding Schizophrenia's Prozac Emergent Relational Similarity in Predication Space. In: Song D, Melucci M, Frommholz I, Zhang P, Wang L, Arafat S, editors. Quantum Interaction. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 48–59.

153. Hristovski D, Kastrin A, Dinevski D, Rindflesch TC. Constructing a Graph Database for Semantic Literature-Based Discovery. Stud Health Technol Inform. 2015;216:1094.
154. Malec SA, Bernstam EV, Wei P, Boyce RD, Cohen T. Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance. medRxiv [Internet]. 2020; Available from: https://www.medrxiv.org/content/early/2020/07/10/2020.07.08.20113035.1
155. Cohen T, Widdows D. Embedding of semantic predications. J Biomed Inform. 2017 Apr;68:150–66.
156. Cohen T, Schvaneveldt RW, Rindflesch TC. Predication-based Semantic Indexing: Permutations as a Means to Encode Predications in Semantic Space. AMIA Annu Symp Proc. 2009;2009:114–8.
157. Elsworth B, Dawe K, Vincent EE, Langdon R, Lynch BM, Martin RM, et al. MELODI: Mining Enriched Literature Objects to Derive Intermediates. Int J Epidemiol. 2018 Apr 1;47(2):369–79.
158. Cohen T, Whitfield GK, Schvaneveldt RW, Mukund K, Rindflesch T. EpiphaNet: An Interactive Tool to Support Biomedical Discoveries. J Biomed Discov Collab. 2010;5:21–49.
159. the Alzheimer's Disease Neuroimaging Initiative, Shen X, Ma S, Vemuri P, Simon G. Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology. Sci Rep. 2020 Dec;10(1):2975.
160. Ding P, Miratrix L. To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias. ArXiv14080324 Math Stat [Internet]. 2014 Aug 1 [cited 2019 Oct 19]; Available from: http://arxiv.org/abs/1408.0324
161. Flanders WD, Ye D. Limits for the Magnitude of M-bias and Certain Other Types of Structural Selection Bias. Epidemiol Camb Mass. 2019 Jul;30(4):501–8.
162. Ioannidis JPA, Tan YJ, Blum MR. Limitations and Misinterpretations of E-Values for Sensitivity Analyses of Observational Studies. Ann Intern Med. 2019 Jan 1;170(2):108–11.
163. Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. Int J Epidemiol. 2021 Apr 1;50(2):620–32.
164. Textor J, Zander B van der, Ankan A. dagitty: Graphical Analysis of Structural Causal Models [Internet]. 2021 [cited 2021 Aug 3]. Available from: https://CRAN.R-project.org/package=dagitty
165. Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, et al. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov Today. 2019 Apr 1;24(4):933–8.
166. Forrest CB, McTigue KM, Hernandez AF, Cohen LW, Cruz H, Haynes K, et al. PCORnet® 2020: current state, accomplishments, and future directions. J Clin Epidemiol. 2021 Jan;129:60–7.
167. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014 Jul 1;21(4):578–82.
168. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, et al. The "All of Us" Research Program. N Engl J Med. 2019 Aug 15;381(7):668–76.