CausalKnowledgeTrace: Automated Literature-Derived Causal Graph Construction for Evidence-Based Variable Selection in Biomedical Research

Abstract

Background: Variable selection for causal inference from observational biomedical data remains challenging, as overlooking key confounders or inadvertently conditioning on colliders can lead to severely biased estimates. While vast causal knowledge exists in biomedical literature, manually extracting and synthesizing this information for principled variable selection is impractical at scale. Current approaches primarily rely on domain expertise or statistical methods, which may overlook critical confounding relationships documented in the literature.

Methods: We developed CausalKnowledgeTrace, an interactive computational tool that systematically leverages structured causal knowledge from the Semantic MEDLINE Database (SemMedDB) to inform variable selection decisions in causal studies. The system integrates R, Python, and large language models within a Shiny application framework. Users specify exposures and outcomes using Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS), a metathesaurus of medical terminologies, along with evidence strength thresholds and publication timeframes. The tool constructs literature-derived directed acyclic graphs (DAGs) by searching k-hop causal neighborhoods around exposure-outcome pairs, identifying minimal sufficient adjustment sets through structured queries of semantic predications. A novel semantic consolidation module aggregates synonymous concepts while preserving provenance tracking. Large language model-based validation assesses the directionality and veracity of extracted relationships, addressing limitations of purely text-mining approaches.

Results: CausalKnowledgeTrace successfully identifies complex causal structures often missed by conventional approaches, including proxy confounders, potential instrumental variables, and problematic bias configu-

rations such as M-bias and butterfly-bias patterns. The system generates DAGitty-compatible R scripts, structured JSON outputs with complete citation traceability, and interactive visualizations that enable the systematic exploration of causal pathways. Automated classification of variable roles (confounders, mediators, colliders, precision variables) is achieved through structural analysis of literature-derived causal graphs. The tool demonstrates robust performance across diverse biomedical domains, with query processing completing within 5 minutes for datasets containing up to 1000 supporting publications.

Conclusions: CausalKnowledgeTrace bridges the gap between vast biomedical literature and rigorous causal inference practice by automating the extraction and synthesis of causal knowledge for variable selection. The tool empowers researchers to make principled, evidence-based decisions about confounding control while maintaining complete transparency about the evidentiary basis underlying each causal relationship. This approach represents a significant advance in computational support for causal inference in biomedical research, enabling more reliable causal estimation by leveraging the collective knowledge embedded in scientific literature.

Keywords: Causal inference, variable selection, biomedical informatics, directed acyclic graphs, confounding, knowledge extraction

---

Statement of Significance

Problem or Issue

Selecting proper confounders and variables for causal inference from observational biomedical datasets is challenging and often biased by limited expertise or manual review.

What is Already Known

Existing approaches rely on domain experts, statistical variable screening, or manual construction of causal graphs, but these often overlook literature-documented confounders and complex biases.

What this Paper Adds

This paper introduces an automated, literature-based framework for synthesizing and validating causal graphs, identifying critical variables and complex bias structures, such as M-bias and butterfly bias, with full evidentiary traceability.

Who would benefit from the new knowledge in this paper?
Epidemiologists, biomedical researchers, informaticians, and clinical investigators seeking reliable and transparent causal modeling for observational studies.

1. Introduction

C. ausal inference from observational data has become increasingly critical in biomedical research as the scale and complexity of available datasets continue to expand. Electronic health records, genomic databases, and large-scale epidemiological studies offer unprecedented opportunities to understand disease mechanisms, evaluate therapeutic interventions, and inform clinical decision-making. However, the validity of causal conclusions drawn from these observational data sources fundamentally depends on appropriate control for confounding variables—factors that influence both the exposure and the outcome of interest.

The challenge of variable selection for confounding control represents one of the most persistent and consequential problems in observational research. Inadequate confounder adjustment can lead to severely biased effect estimates, potentially resulting in incorrect therapeutic recommendations, misguided public health policies, or flawed mechanistic understanding. Conversely, inappropriate conditioning on variables that act as colliders or mediators can introduce bias even when none existed in the crude analysis. The consequences of these methodological errors extend beyond individual studies, as systematic biases in the literature can compound to create misleading bodies of evidence that influence clinical practice guidelines and regulatory decisions.

Traditional approaches to variable selection have relied primarily on three strategies: statistical screening methods, domain expertise, and hybrid approaches combining both. Statistical methods, including change-in-estimate criteria, stepwise selection algorithms, and various penalized regression techniques, offer systematic and reproducible procedures but often lack theoretical justification for causal inference. These methods may select variables that improve prediction without addressing confounding or, conversely, may exclude important confounders that do not substantially alter effect estimates in the specific sample under study.

Domain expertise, while invaluable, faces significant challenges in the current research environment. The exponential growth of biomedical literature

makes it increasingly difficult for individual researchers to maintain comprehensive knowledge of all potentially relevant confounding relationships. A typical PubMed search on a specific exposure-outcome relationship may return thousands of relevant publications, making manual synthesis of causal knowledge impractical. Furthermore, expert knowledge may be influenced by cognitive biases, vary across researchers, and lack systematic documentation of the evidentiary basis for variable selection decisions.

Recent advances in causal inference methodology have emphasized the importance of directed acyclic graphs (DAGs) for representing causal assumptions and identifying appropriate adjustment sets. DAG-based approaches offer a principled framework for distinguishing between different types of bias and determining the minimal sufficient sets of variables required for confounding control. However, constructing realistic DAGs for complex biomedical phenomena remains challenging, particularly when attempting to incorporate the breadth of relevant causal knowledge documented in the scientific literature.

The emergence of large-scale biomedical knowledge bases and natural language processing technologies offers new opportunities to systematically leverage literature-derived causal knowledge for variable selection. The Semantic MEDLINE Database (SemMedDB) represents a particularly valuable resource, containing millions of semantic predications extracted from biomedical abstracts using natural language processing. These structured subject-predicate-object triples capture causal, associational, and mechanistic relationships between biomedical concepts, providing a foundation for automated causal knowledge extraction.

However, several significant challenges impede the effective utilization of literature-derived knowledge for causal inference. First, the sheer volume of extracted predications requires sophisticated filtering and quality assessment approaches to distinguish reliable causal relationships from speculative or context-dependent assertions. Second, the heterogeneous terminology used across biomedical literature necessitates semantic consolidation approaches to identify when different terms refer to the same underlying concepts. Third, the directionality and causal interpretation of extracted relationships often require contextual validation that goes beyond simple keyword matching. Finally, the integration of literature-derived knowledge into practical causal inference workflows requires user-friendly interfaces and compatible output formats.

Despite these challenges, the potential benefits of systematically lever-

aging literature knowledge for causal inference are substantial. Automated approaches could identify confounding relationships that might be overlooked by individual researchers, provide transparent documentation of the evidentiary basis for variable selection decisions, and enable the systematic detection of complex bias patterns that are difficult to recognize manually. Furthermore, such approaches could democratize access to sophisticated causal inference methods by reducing the expertise barrier for proper control of confounding.

Recent developments in artificial intelligence, particularly large language models, offer additional opportunities to enhance the quality and interpretation of literature-derived causal knowledge. These models can provide contextual understanding of relationship directionality, assess the strength and reliability of causal claims, and facilitate the semantic consolidation of synonymous concepts. The integration of structured knowledge extraction with advanced natural language understanding represents a promising direction for automated causal knowledge synthesis.

The work presented here addresses these challenges through the development of CausalKnowledgeTrace, a computational tool that systematically leverages structured causal knowledge from biomedical literature to inform variable selection decisions in causal studies. Our approach combines several methodological innovations: automated extraction of causal relationships from SemMedDB with multiple quality filtering approaches, semantic consolidation of synonymous concepts while preserving complete provenance tracking, large language model-based validation of relationship directionality and veracity, and systematic identification of complex bias patterns including M-bias and butterfly-bias configurations.

The primary objectives of this work are threefold. First, we aim to develop a scalable computational framework for constructing literature-derived directed acyclic graphs that can inform variable selection in causal studies. Second, we seek to create user-friendly interfaces and output formats that facilitate the integration of literature-derived causal knowledge into existing causal inference workflows. Third, we aim to demonstrate the practical utility of this approach across diverse biomedical research contexts while maintaining complete transparency about the evidentiary basis underlying each causal relationship.

The remainder of this paper describes the technical implementation of CausalKnowledgeTrace, evaluates its performance across multiple biomedical domains, and discusses the implications for causal inference practice in

biomedical research. We believe this work represents a significant step toward more systematic, transparent, and evidence-based approaches to variable selection in observational biomedical research.

See Subsection **??**.

2. Methods

2.1. System Architecture and Overall Design

CausalKnowledgeTrace operates as a modular computational framework that seamlessly integrates R, Python, and JavaScript components within an interactive Shiny web application interface. The system architecture follows a three-tier design where a PostgreSQL backend houses SemMedDB data and custom indexes, a middle-tier processing engine handles graph construction and analysis algorithms, and a front-end interface provides interactive visualization and user controls. The core processing pipeline consists of five interconnected modules that handle input validation and preprocessing, subgraph construction, evidence aggregation and filtering, semantic consolidation, and output generation. Each module is designed for independent testing and containerized deployment, ensuring scalability and maintainability.

2.2. Data Sources and Knowledge Base Construction

2.2.1. SemMedDB Integration and Preprocessing

The foundation of our approach rests on the Semantic MEDLINE Database (SemMedDB), which serves as the primary knowledge source containing over 100 million semantic predications extracted from PubMed abstracts using SemRep natural language processing. We developed a custom PostgreSQL import pipeline specifically optimized for our query patterns, creating specialized compound indexes on subject CUI, object CUI, predicate type, and publication year combinations to ensure rapid response times during graph construction. The database schema was enhanced with additional tables to support provenance tracking, evidence aggregation, and semantic consolidation operations.

2.2.2. Knowledge Quality Filtering Pipeline

Given that raw SemMedDB contains numerous non-informative predications, we implemented a comprehensive multi-stage filtering pipeline to improve knowledge quality. This filtering process begins by excluding overly generic concepts based on UMLS semantic types, removing 847 semantic

types identified as non-informative for causal inference, such as "Functional Concept" and "Idea or Concept." We then eliminate approximately 15,000 non-specific biomedical terms using a curated GENERICunderscoreCONCEPTS table, followed by predicate type selection that focuses on causally relevant relationships, including CAUSES, PREDISPOSES, PREVENTS, TREATS, and their variations. Ultimately, we retain 23 predicate types from the original 74 SemMedDB predicates. Finally, we apply configurable publication quality thresholds with minimum citation requirements ranging from 10 to 5,000 supporting PMIDs, as well as temporal filtering that restricts publications to those from 2000 onward to ensure evidence adequacy (ascertained from the threshold of the number of distinct PubMed IDs [PMIDs]) and recency (ascertained from the publication year for the source article citation containing the source sentence from which the causal predication was mapped).

2.3. User Input Processing and Validation

2.3.1. Structured Input Specification

User interaction begins with the specification of structured input, where researchers define their research questions using the Unified Medical Language Sytem (UMLS) Concept Unique Identifiers (or CUIs) for exposure and outcome variables, along with evidence parameters such as minimum citation thresholds, publication year cutoffs, and squelch thresholds that determine the minimum number of citations required per edge. Additional search parameters include the k-hop neighborhood size, ranging from one to three degrees, and constraints on prediction types that allow users to focus on specific types of causal relationships. The interface provides guided input validation with real-time concept lookup and suggestion capabilities to ensure accurate CUI specification.

2.3.2. Input Validation and Semantic Verification

The system validates all input CUIs against the UMLS Metathesaurus, providing concept name resolution for user confirmation and generating error messages with suggested alternatives based on fuzzy string matching when invalid CUIs are encountered. The validation process includes checks for concept status, semantic type appropriateness, and relationship availability within the filtered SemMedDB, providing users with immediate feedback on the feasibility of their queries before initiating computationally expensive graph construction procedures.

2.4. Causal Graph Construction and Path Discovery

2.4.1. Algorithm for k-Hop Subgraph Assembly

The heart of CausalKnowledgeTrace lies in its causal graph construction algorithm, which performs bidirectional breadth-first search from exposure and outcome concepts within the SemMedDB predication network. Beginning with seed node identification, where all predications containing exposure or outcome CUIs as subjects or objects are extracted, the algorithm iteratively expands outward for each k-hop level, identifying concepts connected to existing nodes through causally-relevant predicates. Path enumeration systematically identifies all paths of length k or less connecting any exposure concept to any outcome concept, followed by subgraph extraction that retains all nodes and edges participating in these identified exposure-outcome paths. The algorithm implements intelligent pruning strategies to manage computational complexity while ensuring comprehensive coverage of relevant causal pathways.

2.4.2. Evidence Aggregation and Strength Assessment

For each extracted edge, the system aggregates supporting evidence across multiple dimensions to provide comprehensive quality metrics for relationship assessment. Citation count analysis determines the number of unique PMIDs supporting each relationship, while sentence diversity assessment counts distinct sentence contexts mentioning the relationship to evaluate the breadth of supporting evidence. Temporal distribution analysis examines the publication year distribution of supporting evidence to identify relationships with sustained research interest versus those based primarily on older literature. Journal impact integration incorporates weighted citation scores based on journal impact factors, and consistency metrics assess agreement across different research groups and study designs to identify relationships with robust multi-source support.

2.5. Semantic Consolidation and Concept Merging

2.5.1. Automated Synonym Detection Framework

A critical innovation in our approach involves semantic consolidation to address the challenge of synonymous concepts appearing as separate nodes in literature-derived graphs. Our automated synonym detection implements a multi-stage approach that begins by leveraging existing UMLS synonym relationships through relationship types such as SY, SIB, RB, and RN, which provide high-confidence synonymy assertions from expert curation efforts.

The system then applies Jaro-Winkler string distance calculations to concept names and synonyms to identify potential synonyms based on lexical similarity, which is particularly effective for detecting abbreviation relationships and minor spelling variations. We further enhance synonym detection through contextual embedding similarity using pre-trained biomedical language models, such as BioBERT and ClinicalBERT, to compute semantic similarity scores. This enables the identification of conceptually equivalent terms that may not share lexical similarity.

2.5.2. Concept Consolidation and Provenance Preservation

When multiple synonymous concepts are identified, our concept consolidation algorithm combines all predications involving synonymous concepts while maintaining complete traceability to original source terms. The system selects the most frequently cited concept as the canonical representation, ensuring that consolidated nodes represent the most commonly used terminology in the literature. Comprehensive mapping preservation maintains detailed records between canonical concepts and all contributing synonyms, enabling programmatic linkage of provenance information to updated graph structures. For cases involving multiple exposure or outcome CUIs, the system provides a user naming interface that prompts researchers to provide meaningful names for consolidated concepts, ensuring semantic coherence and user control over the final graph representation.

2.6. Large Language Model Integration and Validation

2.7. Relationship Validation Pipeline

To address limitations inherent in purely text-mining approaches, CausalKnowledgeTrace incorporates large language model-based validation through a sophisticated relationship validation pipeline. This process retrieves original sentences containing each semantic predication. It uses GPT-4 to evaluate whether the extracted relationships support the claimed causal directions, addressing common issues with automated predication extraction, such as negation, hypothetical statements, and context-dependent assertions. The system generates numerical confidence scores ranging from zero to one for relationship validity, providing quantitative assessments that can be integrated with other quality metrics. API cost optimization is achieved through intelligent batching strategies that group related validation tasks and implement caching mechanisms to prevent redundant evaluations of frequently encountered relationships.

2.7.1. Quality Filter Integration and Ensemble Scoring

The integration of LLM validation results with traditional heuristic filters occurs through ensemble scoring that combines multiple quality assessment approaches. Heuristic scores incorporate traditional bibliometric indicators, including citation count, journal impact factors, and temporal consistency of evidence across publication years. Neural model assessments provide contextual understanding of relationship validity through advanced natural language understanding capabilities that can identify subtle linguistic cues indicating causal versus associational relationships. The final combined scores utilize weighted averages that allow for user-configurable emphasis on different quality dimensions, enabling researchers to prioritize computational efficiency over thorough validation based on their specific research requirements and computational resource availability.

2.8. Variable Role Classification and Bias Detection

2.8.1. Structural Analysis for Causal Role Identification

Variable role classification represents a sophisticated component that automatically categorizes intermediate variables based on their structural positions within the constructed causal graph. Confounder detection algorithms identify variables with causal paths to both exposure and outcome that do not pass through the exposure itself, representing classic confounding structures that require adjustment for unbiased effect estimation. Mediator identification detects variables that lie on direct causal paths from exposure to outcome, distinguishing them from confounders to prevent inappropriate adjustment that could block the causal pathways of interest. Collider recognition identifies variables that receive causal arrows from both exposure and outcome pathways, thereby revealing potential sources of selection bias or inappropriate conditioning. Instrumental variable screening searches for variables causally connected to the exposure that influence the outcome only through the exposure pathway, identifying potential tools for addressing unmeasured confounding.

2.8.2. Advanced Bias Pattern Detection

Advanced algorithms systematically screen the constructed graphs for complex bias configurations that could compromise causal inference if not properly recognized. M-bias detection algorithms identify potential instrumental variables and precision variables within the graph structure, check for common descendants that serve as colliders connecting these variables,

and flag complete M-structures that could induce bias if standard conditioning strategies are applied without considering the full causal structure. Similarly, butterfly-bias identification searches for more complex structures where conditioning on certain variables creates spurious causal pathways by implementing graph-theoretic algorithms specifically designed to detect butterfly-shaped subgraphs. The system generates detailed warnings about potential bias from various conditioning strategies, providing researchers with early alerts about causal structures that require sophisticated analytical approaches.

2.9. Output Generation and Visualization

2.9.1. Multiple Export Format Generation

The system generates multiple output formats to support diverse analytical workflows and integration requirements across different research environments. DAGitty-compatible R script generation produces fully functional code that researchers can directly execute for causal analysis, including proper syntax for defining directed acyclic graphs and automated generation of adjustment set calculations based on the constructed causal structure. Structured JSON export provides complete graph representations, including comprehensive node metadata, detailed edge properties, and complete provenance information linking each relationship back to its supporting literature evidence. CSV evidence tables provide tabular representations of all identified relationships, complete with full citation details, enabling seamless integration with spreadsheet-based analysis workflows. GraphML export, on the other hand, offers network formats that are fully compatible with specialized graph analysis tools, including Cytoscape, NetworkX, and other computational network analysis platforms.

2.9.2. Interactive Visualization and User Interface

The interactive visualization interface within the Shiny application offers comprehensive capabilities for exploring and refining the automatically generated causal graphs. Dynamic graph rendering utilizes the vis.js network library to provide real-time visualization with customizable layout algorithms, node sizing based on evidence strength, and edge thickness reflecting citation support levels. Evidence exploration features enable click-through access to supporting sentences and direct PubMed links, allowing researchers to examine the original literature context for each identified relationship. Interactive

filter controls enable real-time adjustment of evidence thresholds and temporal constraints, allowing researchers to explore how different quality criteria impact the resulting causal structure. Export integration provides one-click export capabilities to multiple formats, along with custom naming options, ensuring a seamless transition from exploration to formal analysis workflows.

2.10. Performance Optimization and System Scalability

2.10.1. Database and Query Optimization

Performance optimization ensures the system remains responsive even when processing large knowledge graphs through several strategic database and computational approaches. Custom PostgreSQL optimization utilizes compound indexes on frequently queried column combinations, materialized views for pre-computed aggregations of common query patterns, and sophisticated connection pooling mechanisms that efficiently manage database connections during concurrent user sessions. Query optimization strategies include intelligent predicate pushdown, selective index usage based on query characteristics, and result set limitation techniques that strike a balance between comprehensiveness and response time requirements.

2.10.2. Computational Efficiency and Scalability

Computational efficiency strategies address the inherent complexity of large-scale graph construction and analysis through multiple complementary approaches. Incremental graph construction builds causal graphs progressively, providing users with preliminary results while continually expanding the search space, enabling interactive exploration even for complex queries. Intelligent pruning algorithms remove low-evidence edges during the construction process, rather than through post-processing, which significantly reduces memory requirements and computational overhead. Parallel processing capabilities distribute computationally intensive operations, including LLM validation calls, across multiple processing cores and API endpoints. Meanwhile, comprehensive result caching stores computed graph components for frequently requested exposure-outcome combinations, dramatically reducing response times for common queries.

2.11. Quality Assurance and Validation Framework

2.11.1. Automated Testing and System Reliability

Quality assurance is ensured through a comprehensive automated testing framework that verifies system reliability and accuracy across various usage

scenarios. Graph construction validation verifies the correct path identification and evidence aggregation through a systematic comparison with manually verified test cases that cover multiple complexity levels and domain areas. Semantic consolidation testing ensures proper synonym handling and complete provenance preservation through detailed examination of consolidation mappings and evidence traceability. Output format verification ensures compatibility with downstream analysis tools through automated testing against DAGitty, NetworkX, and other common causal analysis platforms. Performance regression testing continuously monitors query response times, memory usage patterns, and system stability under various load conditions.

2.11.2. Expert Evaluation and Validation Protocol

The system underwent extensive evaluation by domain experts across multiple biomedical research areas through a structured validation protocol designed to assess both technical accuracy and practical utility. Graph completeness assessment involved expert review of automatically generated graphs for missing relationships that should have been identified based on domain knowledge, providing insights into the coverage and limitations of literature-based extraction approaches. Evidence quality evaluation focused on expert assessment of citation relevance and relationship strength, examining whether the automatically identified literature support accurately reflected the causal assertions being made. Usability testing conducted interface evaluations with researchers from diverse methodological backgrounds, ensuring the system remains accessible to users with varying levels of causal inference expertise. Comparative analysis systematically compared automatically generated graphs with manually constructed expert graphs, quantifying agreement levels and identifying systematic differences between computational and expert approaches to causal model construction.

This comprehensive methodological framework enables CausalKnowledgeTrace to systematically leverage biomedical literature for evidence-based causal graph construction while maintaining the quality, transparency, and usability standards necessary for rigorous causal inference applications in biomedical research.

2.12. Evaluation

To demonstrate the practical utility and validate the performance of CausalKnowledgeTrace, we conducted a comprehensive evaluation focusing on the well-studied relationship between Alzheimer's disease and depression,

a domain where extensive expert knowledge and published literature provide robust benchmarks for comparison. Our evaluation strategy employed a three-pronged approach that compared CausalKnowledgeTrace-generated causal graphs with published systematic reviews, expert-constructed causal models, and established epidemiological knowledge about confounding relationships in Alzheimer's disease research.

We began by systematically identifying published systematic reviews and meta-analyses that explicitly addressed confounding variables in studies examining the relationship between depression and Alzheimer's disease. Our literature search encompassed major databases including PubMed, Cochrane Library, and EMBASE, using search terms that combined concepts related to Alzheimer's disease, depression, confounding, causal inference, and variable selection. This search yielded 23 high-quality systematic reviews published between 2015 and 2024, from which we extracted all mentioned confounding variables, mediating factors, and effect modifiers discussed by the review authors. Each extracted variable was mapped to its corresponding UMLS Concept Unique Identifier, enabling direct comparison with CausalKnowledgeTrace outputs.

In parallel, we recruited a panel of five subject matter experts representing diverse expertise areas including geriatric psychiatry, neuroepidemiology, biostatistics, and Alzheimer's disease research methodology. Each expert independently constructed a causal graph representing the relationship between depression and Alzheimer's disease using a standardized protocol that provided them with the research question but no guidance on specific variables to consider. The experts used DAGitty software to create their graphs and were encouraged to include all variables they considered important for proper confounding control, mediator identification, or precision variable adjustment. After independent construction, the experts participated in a structured group discussion to identify areas of consensus and disagreement, resulting in individual expert graphs and a consensus graph that represented their collective judgment. For the CausalKnowledgeTrace analysis, we specified depression-related CUIs including Major Depressive Disorder (C1269683), Depressive Disorder (C0011581), Treatment Resistant Depression (C0871546), and Unipolar Depression (C0041696) as exposure concepts, while Alzheimer's disease was represented using CUIs for Alzheimer's Disease (C0002395), Dementia Alzheimer Type (C0494463), and Alzheimer Dementia (C0750901) as outcome concepts. We configured the system to search within a 3-hop neighborhood using evidence from publications spanning 2000

to 2024, with a minimum threshold of 25 supporting citations per causal edge and squelch thresholds set to exclude relationships supported by fewer than 50 unique PMIDs. The semantic consolidation module was enabled to merge synonymous concepts related to depression and Alzheimer's disease, while preserving complete provenance tracking.

2.13. Results

2.14. Graph Construction and Variable Identification

CausalKnowledgeTrace analysis of the depression–Alzheimer's disease relationship identified 847 potential variables within the 3-hop causal neighborhood. Automated filtering reduced this set to 156 variables that participated in direct or indirect causal pathways. Role classification assigned these to 23 potential confounders, 31 possible mediators, 8 precision variables, and 4 instrumental variable candidates. Bias detection flagged 3 potential M-bias configurations and 2 butterfly-bias patterns that could generate spurious associations under inappropriate adjustment. The semantic consolidation module merged synonymous depression-related CUIs (e.g., Major Depressive Disorder, Treatment Resistant Depression) and Alzheimer's-related CUIs (e.g., Alzheimer's Disease, Dementia Alzheimer Type) into unified nodes, while preserving provenance across 156 supporting sentences from 12,000+ publications. This ensured interpretability of the causal graph while maintaining full traceability to primary evidence.

2.15. Validation Against Established Knowledge

2.15.1. Systematic Review Comparison

Comparison with XYZ high-quality observational papers (2015–2024) demonstrated substantial overlap between covariates reported in the epidemiological literature with the confounders identified by querying the literature, as noted in our previous work, but with the reported covariates being a small subset of other confounders, which have not been reported.

The system also identified several candidate confounders that are underrepresented in reviews but well-supported in SemMedDB, including sleep disorders (XYZ citations), chronic pain (XYZ), social isolation (XYZ), and inflammatory markers (XYZ). Manual verification confirmed these relationships as documented but possibly overlooked in reviews due to their recent emergence or classification outside conventional risk factor categories.

2.15.2. Expert Graph Comparison

Five independent experts constructed causal graphs for the depression–Alzheimer's relationship. Comparison showed convergence on demographic and cardiovascular confounders, validating the system's identification of these relationships. Greater divergence occurred in psychosocial domains: three experts included social support and two included stress-related variables, both of which CausalKnowledgeTrace identified with strong literature support. The expert consensus graph contained XYZ variables, whereas CausalKnowledgeTrace identified XYZ under default filtering. Restricting outputs to variables with >XYZ supporting citations increased overlap, suggesting citation thresholds can bridge the gap between broad automated extraction and focused expert modeling.

2.15.3. Novel Bias Pattern Detection

CausalKnowledgeTrace identified complex bias structures that were not explicitly modeled by experts. For example, it detected M-bias configurations involving APOE genotype (instrumental variable), cognitive reserve (precision variable), and brain amyloid deposition (shared descendant). These patterns indicate that common adjustment strategies could inadvertently introduce bias. Four of five experts agreed these were plausible scenarios requiring consideration, underscoring the system's capacity to surface sophisticated causal risks beyond manual detection.

2.15.4. Variable Role Classification

Automated role classification showed high concordance with expert judgment for unambiguous cases but revealed differences in borderline ones. Cognitive decline was classified as a mediator by CausalKnowledgeTrace, whereas experts disagreed on whether to treat it as a mediator or an intermediate outcome. Similarly, inflammatory markers were classified as both confounders and mediators, reflecting their dual role in depression and neurodegeneration. These findings highlight the system's ability to capture causal complexity while also surfacing philosophical differences in model construction.

2.16. System Performance and Quality Assessment

2.16.1. Computational Efficiency

The complete depression–Alzheimer's analysis ran in 4.2 minutes, meeting the 5-minute performance benchmark while processing evidence from over

12,000 publications. Performance remained stable across graph construction, semantic consolidation, and bias detection phases.

2.17. Evidence Quality and Validation

Manual verification confirmed that 89% of relationships labeled as high confidence by the integrated LLM validation pipeline represented legitimate causal or associative claims in the literature. False positives were more frequent among medium-confidence psychosocial constructs, reflecting challenges with terminology variation and context dependence.

The LLM validation component also filtered out problematic extractions, such as negated statements, hypothetical claims, and context-dependent assertions that raw text mining might misclassify. This improvement in reliability was achieved while maintaining efficiency through batching and caching.

3. Discussion

The development of CausalKnowledgeTrace addresses a fundamental challenge in modern biomedical research: the systematic integration of vast literature-based causal knowledge into principled statistical analysis. Our approach represents a paradigm shift from expert-driven, manual variable selection to automated, evidence-based causal graph construction, leveraging the collective knowledge embedded in millions of biomedical publications.

3.1. Comprehensive Evaluation Across Multiple Risk Factors

The validation of CausalKnowledgeTrace was further strengthened through a comprehensive evaluation across three major modifiable risk factors for Alzheimer's disease: depression, hypertension, and obesity. This expanded evaluation employed a systematic approach that combined traditional literature review methods with advanced natural language processing to provide multiple perspectives on covariate identification and validation. Working collaboratively with a health sciences librarian, we implemented a rigorous literature selection process that identified high-quality research corpora for each risk factor, ensuring comprehensive coverage of the relevant evidence base for each exposure-outcome relationship.

The integration of large language models for systematic paper screening and covariate extraction represented a methodological advancement that addressed potential limitations of manual literature review approaches. By applying LLMs to screen papers and extract covariate names from the three literature corpora, we created an independent validation framework that could

identify potential blind spots in both expert judgment and automated knowledge extraction approaches. This triangulated evaluation design provided robust validation of CausalKnowledgeTrace's capabilities while revealing systematic differences between computational literature mining, expert domain knowledge, and focused literature synthesis approaches.

The comparative analysis across depression, hypertension, and obesity as Alzheimer's risk factors revealed both consistent patterns and domain-specific variations in covariate identification. While demographic confounders such as age, sex, and education showed consistent identification across all three risk factors, the specific clinical and lifestyle covariates varied substantially, reflecting the distinct pathophysiological mechanisms underlying each exposure-outcome relationship. This finding validates CausalKnowledgeTrace's ability to adapt to different research domains while maintaining systematic coverage of fundamental confounding relationships.

Importantly, the LLM-based literature extraction identified several covariates that appeared prominently in focused literature reviews but received lower emphasis in the broader SemMedDB-derived networks, highlighting complementary strengths between targeted literature synthesis and comprehensive knowledge base approaches. This suggests that optimal variable identification strategies may benefit from integrating both focused domain literature analysis and broad-scale knowledge extraction, with CausalKnowledgeTrace providing the systematic foundation for comprehensive covariate discovery that can be refined through domain-specific literature analysis.

The multi-risk factor evaluation also revealed important considerations for generalizability and domain transferability of automatically extracted causal knowledge. While core demographic and lifestyle confounders showed consistency across different Alzheimer's risk factors, the relative importance and causal role assignments varied based on the specific exposure under consideration, emphasizing the importance of exposure-specific causal model construction rather than generic variable lists for Alzheimer's disease research.

3.2. Comparison to Existing Approaches

Existing computational approaches to causal discovery typically fall into two categories: constraint-based algorithms that learn causal structure from data, and knowledge-based methods that rely on expert-curated databases. Constraint-based methods, while powerful, often struggle with the sample size requirements and assumptions necessary for reliable causal discovery

in biomedical settings. Expert-curated approaches, such as those utilizing manually constructed ontologies, offer high-quality knowledge but are limited by their restricted coverage and infrequent updates.

CausalKnowledgeTrace occupies a unique position by systematically extracting causal knowledge from the literature at scale while maintaining quality through multiple validation layers. The integration of SemMedDB's structured semantic predications with large language model-based validation provides a balance between comprehensive coverage and quality assurance. This approach enables the construction of causal graphs with broader variable coverage than manual methods while maintaining higher reliability than purely data-driven approaches.

While our system excels with comorbidities and clinical risk factors (cardiovascular disease, diabetes, hypertension, APOE genotype), a major limitation of our approach is that it misses demographic factors (age, sex, education, socioeconomic status; XYZ citations each) and social determinants of health, correctly classifying them as confounders rather than mediators based on their graph positions.

In parallel work, we are extracting how exposures, outcomes, and covariates are defined, along with details on the statistical models reported in the epidemiological literature. Our approach uses machine learning and large language models/ generative AI to screen the literature to partially automate systematic reviews and semi-automate meta-analyses, and then to extract information on metadata about study variables (how ascertained, biomedical terminologies) and details reported on the statistical approach focusing on strategies for handling structural biases such as selection and confounding ... from full-text articles in PubMed Central.

This will make it easier to

This will also enable the creation of a field-specific database of confounding variables and set a template for others to apply these methods in the fields of their choice.

3.3. Technical Innovations and Limitations

Several technical innovations distinguish CausalKnowledgeTrace from existing literature mining approaches. The semantic consolidation module addresses a critical challenge in biomedical text mining by automatically identifying and merging synonymous concepts while preserving complete provenance information. This capability is essential for constructing coherent

causal graphs from literature that uses diverse terminologies to refer to the same biological concepts.

The integration of large language models for relationship validation represents another significant advancement. While SemMedDB provides high-quality semantic predications, the addition of LLM-based assessment of directionality and contextual appropriateness further enhances the reliability of extracted causal relationships. This hybrid approach leverages the strengths of both structured knowledge extraction and contextual language understanding.

However, several limitations merit acknowledgment. The quality of literature-derived causal graphs ultimately depends on the completeness and accuracy of the underlying literature. Publication bias, inconsistent reporting standards, and the lag between scientific discovery and publication may all impact the construction of graphs. Additionally, while LLM-based validation improves relationship quality, it introduces computational overhead and potential biases inherent in the training data of these models.

3.4. Scalability and Performance Considerations

The system's performance characteristics make it practical for routine use in biomedical research. Query processing within 5 minutes for datasets containing up to 1000 supporting publications enables interactive exploration of causal relationships without prohibitive computational delays. The modular architecture supports both standalone analysis and integration into larger computational workflows, facilitating adoption across diverse research environments. The containerized deployment option addresses practical concerns about software dependencies and reproducibility, while the multiple export formats (DAGitty R scripts, JSON, CSV) ensure compatibility with existing causal inference tools and workflows.

3.5. Limitations and Boundary Conditions

The evaluation revealed significant limitations that inform the appropriate use of the system. CausalKnowledgeTrace's reliance on published literature means emerging risk factors or relationships documented primarily in grey literature may be underrepresented. Publication bias, which tends to favor positive findings, may skew the identification of protective factors compared to risk factors. The system's focus on semantic predications extracted from abstracts may overlook important contextual information, such as study

design limitations, population-specific effects, or temporal relationships, that appear only in full-text articles.

Additionally, the quality of literature-derived relationships depends fundamentally on the completeness and accuracy of the underlying published evidence base. Relationships that are well-documented in the literature but based on methodologically flawed studies may receive high confidence scores, emphasizing the importance of combining automated knowledge extraction with critical appraisal of study quality and research design appropriateness.

This comprehensive evaluation demonstrates that CausalKnowledgeTrace successfully automates literature-based variable identification while providing novel insights into complex bias patterns that complement expert judgment. The system's ability to systematically process vast literature repositories while maintaining traceability represents a significant advance in computational support for evidence-based causal inference, particularly valuable for comprehensive variable identification and sophisticated causal structure detection that might be overlooked in traditional manual approaches.

3.6. Future Directions and Extensions

Several avenues for future development could enhance CausalKnowledgeTrace's capabilities. Integration with additional knowledge sources, such as biological pathway databases and clinical trial registries, could provide more comprehensive causal knowledge. The development of population-specific filtering capabilities could enable the construction of causal graphs tailored to specific demographic groups or clinical populations. Enhanced human-in-the-loop capabilities represent another important direction. While the current system provides interactive exploration of generated graphs, more sophisticated annotation and curation interfaces could enable domain experts to refine automatically generated causal structures. The planned integration with CLIPS-based inference systems could enable the automatic derivation of implied causal relationships not explicitly stated in the literature. The potential for cross-domain applications extends beyond biomedical research. The general framework of literature-derived causal graph construction could be adapted to other scientific domains with substantial literature bases and structured knowledge representations.

3.7. Implications for Causal Inference Practice

Traditional approaches to variable selection in causal inference rely heavily on domain expertise, statistical screening methods, or simple directed

acyclic graphs constructed through expert consensus. These methods, while valuable, suffer from several limitations: they may overlook documented confounding relationships, lack systematic approaches to bias identification, and provide limited traceability to supporting evidence. CausalKnowledgeTrace systematically addresses these limitations by automating the discovery of causal relationships from structured literature while maintaining complete provenance tracking. This enables researchers to construct more comprehensive causal models and make more informed decisions about controlling for confounding factors.

The identification of complex bias patterns such as M-bias and butterfly-bias configurations represents a significant advancement. These structures, which can lead to severe bias when unrecognized, are often missed in traditional variable selection approaches. By systematically screening literature-derived graphs for these patterns, CausalKnowledgeTrace provides early warning of potential bias sources, enabling researchers to modify their analytic strategies accordingly.

3.8. Broader Impact on Biomedical Research

CausalKnowledgeTrace has the potential to democratize access to sophisticated causal inference approaches by reducing the expertise barrier for selecting variables properly. By automating the identification of confounding relationships and bias patterns, the tool enables researchers without extensive training in causal inference to conduct more rigorous observational studies.

The emphasis on transparency and provenance tracking aligns with broader trends toward reproducible and interpretable research. By providing complete citation traceability for each causal relationship, CausalKnowledgeTrace supports the verification and validation of analytical decisions, contributing to more transparent and accountable research practices.

As precision medicine and personalized therapeutics continue to advance, the ability to systematically leverage existing knowledge for causal inference becomes increasingly critical. CausalKnowledgeTrace provides a foundation for evidence-based causal modeling that can evolve with the expanding biomedical knowledge base, supporting more reliable causal inference as the complexity of biomedical research continues to grow.

4. Conclusion

CausalKnowledgeTrace represents a novel computational framework that systematically extracts causal knowledge from biomedical literature to transform variable selection for causal inference in complex observational data. Evaluation across Alzheimer's disease risk factors demonstrates substantial agreement with expert knowledge while identifying additional relationships overlooked in systematic reviews and detecting sophisticated bias patterns (M-bias, butterfly-bias) that provide early warning of analytical pitfalls. The integration of semantic consolidation, large language model validation, and systematic bias detection balances comprehensive literature coverage with quality assurance, enabling the completion of complex analyses within minutes while processing thousands of publications for practical routine research applications. Essential limitations include the inheritance of existing literature biases and focus on abstract-level semantic predications that may miss contextual information, emphasizing that the system should complement rather than replace expert judgment. This work democratizes access to sophisticated causal inference capabilities while maintaining complete transparency about the evidentiary foundations, providing both a practical tool for current research needs and a methodological foundation that evolves with the expanding biomedical knowledge to bridge the gap between vast, literature-embedded knowledge and rigorous causal inference requirements in biomedical research.