

Creating a Biomedical Causal Knowledge Graph via Semi-Supervised Learning

Jim Dorsey

jimdorsey4521@gmail.com

Miami University of Science and Technology

Research Article

Keywords: biomedical knowledge graph, causal inference, semi-supervised learning, self-training, relation extraction, drug safety, precision medicine

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11364075/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Creating a Biomedical Causal Knowledge Graph via Semi-Supervised Learning

Submitted By:

Jim Dorsey

Affiliation:

Miami University of Science and Technology

Abstract

Biomedical knowledge graphs have emerged as powerful tools for organizing heterogeneous biomedical data and supporting interpretable clinical reasoning. However, constructing causal knowledge graphs—which encode directed cause-effect relationships rather than mere statistical associations—remains challenging due to the scarcity of large-scale, expert-annotated corpora for causal relation extraction. This paper examines the role of semi-supervised learning as a methodological solution to this labeling bottleneck. We review the core semi-supervised paradigms—self-training, co-training, and active learning—and analyze how they have been applied to biomedical relation extraction and knowledge graph construction. Through case studies of recent causal knowledge graph frameworks, including Drug-Disease CKGs and ontology-enriched causal graphs, we demonstrate that semi-supervised approaches enable the effective use of unlabeled biomedical literature while preserving the causal specificity required for clinical validity. We conclude by discussing persistent challenges, including recall limitations and the need for better integration of large language models, and outline future directions for scaling causal knowledge graph construction in precision medicine.

Keywords: biomedical knowledge graph; causal inference; semi-supervised learning; self-training; relation extraction; drug safety; precision medicine

1. Introduction

The rapid expansion of biomedical literature has created an unprecedented challenge for knowledge synthesis. PubMed alone indexes more than 36 million citations, with thousands of new articles added daily. For researchers, clinicians, and computational systems alike, the task of extracting meaningful relationships from this deluge of text has become increasingly difficult to manage through manual curation alone (Mihăilă and Ananiadou, 2014). Knowledge graphs have emerged as a promising solution to this problem, offering structured

representations of entities and their relationships that can support automated reasoning, information retrieval, and clinical decision support (Zhang et al., 2026).

Within the broader landscape of biomedical knowledge graphs, causal knowledge graphs occupy a special position. Whereas conventional knowledge graphs encode associations, co-occurrences, and statistical correlations, causal knowledge graphs represent directed cause-effect relationships that carry stronger semantic commitments. The distinction matters enormously in biomedical contexts. A knowledge graph that records an association between a drug and a disease provides useful information for retrieval purposes, but a causal knowledge graph that specifies whether the drug causes, treats, or prevents the disease enables reasoning about interventions, adverse effects, and treatment pathways (Toonsi et al., 2026). This causal specificity is essential for applications ranging from drug safety surveillance to precision medicine, where understanding not merely what is correlated but what is causally efficacious determines clinical action.

The construction of causal knowledge graphs from biomedical text, however, faces a fundamental obstacle: the scarcity of labeled data. Supervised machine learning approaches to relation extraction, which have achieved impressive results on benchmark datasets, depend on large quantities of expert-annotated text. In the general domain, resources such as SemEval provide thousands of labeled examples for training relation extraction models (Zhang et al., 2026). In the biomedical domain, the situation is considerably more constrained. Annotating causal relations requires not only linguistic expertise but also domain knowledge to distinguish genuine causal claims from speculative associations, conditional relationships, and correlational statements. The cost of such annotation is prohibitive at the scale required for modern deep learning models.

Semi-supervised learning offers a principled response to this data bottleneck. By combining a small set of labeled examples with a much larger corpus of unlabeled text, semi-supervised methods can leverage the vast quantities of unannotated biomedical literature that already exist. The core insight is that unlabeled data, while lacking explicit supervision signals, contains distributional information that can improve model performance when appropriately incorporated into training. Self-training approaches iteratively expand the labeled set by adding high-confidence predictions on unlabeled instances. Co-training exploits multiple views of the data to train complementary models that teach each other. Active learning strategically selects the most informative unlabeled instances for human annotation, maximizing the value of limited expert time.

This paper investigates the construction of biomedical causal knowledge graphs through semi-supervised learning methods. The central research question concerns how semi-supervised techniques can be effectively deployed to overcome the annotation bottleneck in causal relation extraction while preserving the precision and causal specificity required for biomedical applications. Addressing this question requires engagement with several sub-problems: identifying which semi-supervised paradigms are most suitable for causal relation extraction, understanding how these methods interact with the distinctive characteristics of biomedical

text, and evaluating the quality of the resulting causal knowledge graphs against established benchmarks and clinical use cases.

The significance of this research extends across multiple domains. For biomedical informatics, improved methods for causal knowledge graph construction enable more comprehensive integration of literature-derived knowledge with electronic health record data and clinical guidelines. For machine learning, the biomedical causal extraction task provides a demanding testbed for semi-supervised methods operating under conditions of severe label scarcity and domain-specific linguistic complexity. For clinical practice, causal knowledge graphs constructed through these methods can support drug safety monitoring, adverse effect detection, and treatment recommendation systems that require causal rather than merely associative reasoning.

The paper proceeds as follows. Section 2 reviews the relevant literature on biomedical knowledge graphs, causal relation extraction, and semi-supervised learning methods. Section 3 describes the methodology employed in this study, including data sources, the semi-supervised learning framework, and evaluation procedures. Section 4 presents the results of experiments on causal relation extraction and knowledge graph construction. Section 5 discusses the findings in relation to existing work and considers limitations and implications. Section 6 concludes with a summary of contributions and directions for future research.

2. Literature Review

2.1 Biomedical Knowledge Graphs

Knowledge graphs have become a foundational technology for organizing and accessing biomedical information. The general architecture of a knowledge graph consists of entities represented as nodes and relationships represented as edges, with both entities and relations typically drawn from controlled vocabularies or ontologies (Noy and McGuinness, 2001). In the biomedical domain, this structure aligns naturally with the relational nature of biological knowledge: genes interact with proteins, drugs target receptors, diseases affect anatomical systems, and so forth.

Early biomedical knowledge graphs were constructed primarily through manual curation. Resources such as the Gene Ontology (Ashburner et al., 2000), the Unified Medical Language System, and SNOMED CT (Willett et al., 2018) provided the ontological backbone for representing biomedical entities and their hierarchical relationships. These resources remain essential, but their manual construction limits their coverage and currency. As the volume of biomedical literature has grown, automated and semi-automated methods for knowledge graph construction have become necessary.

Recent work has explored increasingly sophisticated approaches to biomedical knowledge graph construction. The CREOG framework (Causal Reasoning over Ontology-Enriched Graphs) introduced by Thuy et al. (2026) demonstrates how ontology constraints can guide causal structure learning while providing human-readable explanations for discovered causal links. Their approach transforms OWL-based ontologies into causal knowledge graphs

enriched with semantic and quantitative causal-strength annotations, constructing a diabetes-specific graph with over 21,000 entities and 107,000 causal or associative edges. The framework integrates edge-level confidence scores into the embedding learning process, enabling the model to prioritize high-certainty causal links.

The Drug-Disease Causal Knowledge Graph (DD-CKG) developed by Toonsi et al. (2026) represents another significant advance. This framework integrates disease progression pathways, drug indications, side effects, and hierarchical disease classification to enable automated large-scale mediation analysis. Applied to UK Biobank and MIMIC-IV cohorts, the DD-CKG successfully reproduced known adverse drug reactions while identifying previously undocumented candidate effects. This work exemplifies the potential of causal knowledge graphs to support clinical inference at scale.

2.2 Causal Relation Extraction in Biomedical Text

Extracting causal relationships from biomedical text presents distinctive challenges. Causality in scientific discourse is expressed through a diverse and often subtle range of linguistic constructions. Explicit causal triggers such as "causes," "leads to," and "results in" provide clear signals, but many causal relationships are expressed through semi-explicit constructions like "is associated with" or through entirely implicit means that require inference from context (Mihăilă and Ananiadou, 2014).

The BioCause corpus, developed by Mihăilă et al., provides one of the few gold-standard resources for biomedical causality annotation. Containing 850 causal relationships across 19 full-text articles in the infectious diseases domain, BioCause distinguishes between causal triggers, cause arguments, and effect arguments (Mihăilă and Ananiadou, 2014). The corpus reveals the diversity of causal expression in biomedical text: 381 unique explicit triggers are used on average only 2.10 times each, indicating that causal language is highly varied and context-dependent.

The annotation of causal relations requires careful attention to the distinction between correlation and causation. In biomedical research, observational studies frequently report associations without establishing causality, and the language used to describe findings reflects this epistemic caution. A sentence stating that a gene is "associated with" a disease does not assert that the gene causes the disease, yet this distinction is crucial for the downstream applications of causal knowledge graphs. This annotation challenge contributes to the scarcity of labeled causal data, as expert annotators must make fine-grained judgments about the strength and nature of causal claims.

2.3 Semi-Supervised Learning for Relation Extraction

Semi-supervised learning addresses the fundamental tension between the need for large training sets and the cost of annotation. The core paradigm involves training a model on a small labeled dataset and then using that model to make predictions on unlabeled data, which are in turn used to improve the model. This iterative process can substantially expand the effective training set without additional annotation cost.

Self-training, also known as bootstrapping, represents the most straightforward semi-supervised approach. A classifier is trained on the labeled data, applied to the unlabeled data, and the most confident predictions are added to the training set. The process repeats until convergence or until no new predictions exceed the confidence threshold. Mihăilă and Ananiadou (2014) applied self-training to causal trigger detection in biomedical text, achieving an F-score of 83.47% for causal triggers, a substantial improvement over supervised baselines. Their results demonstrate that the large quantities of unlabeled biomedical text can be effectively exploited to improve causal relation extraction.

The choice of confidence threshold in self-training is a critical design decision. A threshold that is too high excludes potentially useful training examples, while a threshold that is too low introduces noise that can degrade model performance. FABLE, a medication extraction system described in research on the i2b2 shared task, found that conditional thresholds outperformed fixed thresholds of 0.95 and 0.90 (FABLE, 2018). By calculating confidence rates separately for each entity category and adjusting thresholds accordingly, the system achieved significant improvements in F1-measure across medication, dosage, mode, frequency, duration, and reason categories. This finding suggests that the optimal threshold varies by relation type and that adaptive strategies may outperform uniform confidence criteria.

Co-training offers an alternative to self-training by exploiting multiple views of the data. If the feature space can be naturally partitioned into conditionally independent subsets, separate classifiers can be trained on each view and their predictions used to augment each other's training sets. In biomedical relation extraction, potential views include character-level features, word-level features, syntactic features, and semantic features. Zhang et al. (2026) introduce a semi-supervised relation extraction approach that models relationships between labels and instances using a unique label distribution for model training supervision. Their method achieves an average F1 score increase of 2.64% on the general-domain SemEval dataset and notably higher gains on the biomedical ChemProt dataset, demonstrating the effectiveness of semi-supervised approaches in noisy supervision settings.

2.4 Causal Knowledge Graph Construction

The construction of causal knowledge graphs from semi-supervised relation extraction involves additional considerations beyond those of standard relation extraction. Not all extracted relations should be treated as equally causal, and the graph structure must encode the strength and reliability of causal claims.

The CREOG framework addresses this by incorporating edge weights that reflect causal strength into the embedding learning process (Thuy et al., 2026). Their CausE-S embedding model integrates edge-level confidence scores, enabling the model to prioritize high-certainty causal links during training. This approach recognizes that a knowledge graph derived from automated extraction contains relations of varying reliability, and that treating all edges equally ignores this heterogeneity.

The DD-CKG framework takes a complementary approach by integrating multiple data sources with different reliability characteristics (Toonsi et al., 2026). Disease progression pathways

derived from text-mined literature are combined with drug indications from the MEDI-C dataset and side effects from OnSIDES and OFFSIDES. The framework supports deconfounding via explicitly marked causal edges and facilitates hypothesis formulation aligned with both encoded and entailed background knowledge. Applied to adverse drug reaction detection, the approach successfully identified known adverse effects while discovering novel candidates that were validated through side effect similarity analysis.

2.5 Gaps in the Literature

Despite significant progress, several gaps remain in the literature on semi-supervised causal knowledge graph construction. First, the integration of semi-supervised relation extraction with causal graph construction has received limited attention. Most work treats these as separate problems: one body of research addresses relation extraction, while another addresses graph construction from pre-extracted relations. The feedback loop between extraction quality and graph structure, and the potential for graph-level constraints to improve extraction, remains underexplored.

Second, the evaluation of causal knowledge graphs presents methodological challenges that have not been fully addressed. Standard relation extraction metrics such as precision, recall, and F1-measure evaluate extraction performance at the relation level, but they do not capture the quality of the resulting graph for downstream reasoning tasks. Metrics that assess causal graph utility, such as the ability to support accurate causal inference or to identify known adverse effects, are needed to complement traditional extraction metrics.

Third, the interaction between semi-supervised learning and the distinctive features of biomedical causal language deserves further investigation. Biomedical text exhibits high lexical diversity in causal expression, domain-specific syntactic constructions, and an epistemic caution that frequently expresses causal claims tentatively. How semi-supervised methods should be adapted to these characteristics is not well understood.

3. Methodology

3.1 Research Design

This study employs a mixed-methods approach combining quantitative experiments on causal relation extraction with qualitative analysis of the resulting knowledge graphs. The design is structured around three phases: (1) construction of a causal relation extraction model using semi-supervised learning, (2) application of the extraction model to construct a biomedical causal knowledge graph, and (3) evaluation of the knowledge graph through comparison with established benchmarks and clinical use cases.

The rationale for this design reflects the multi-faceted nature of the research question. Evaluating semi-supervised learning for causal knowledge graph construction requires not only measuring extraction accuracy but also assessing whether the extracted causal relations form a coherent graph structure that supports meaningful inference. The mixed-methods approach enables both fine-grained analysis of extraction performance and holistic assessment of graph quality.

3.2 Data Sources

The study draws on multiple data sources for labeled and unlabeled text. For labeled causal relation data, the BioCause corpus serves as the primary resource (Mihăilă and Ananiadou, 2014). BioCause contains 19 full-text open-access journal articles from the infectious diseases domain, manually annotated with 850 causal relationships. Each relation is composed of three annotations: a trigger, a cause argument, and an effect argument. Of the 850 relations, 800 have explicit triggers, while 50 have implicit triggers that are excluded from this study due to their small number and the difficulty of their identification.

For unlabeled data, a corpus of 50 full-text open-access journal articles from the same infectious diseases subdomain was collected. These articles are of similar age to those in BioCause and are drawn from the same subject area, ensuring that the unlabeled data is in-distribution with respect to the labeled data. The unlabeled articles contain no gold-standard annotations; all features are derived from automatic parses.

For evaluation of the constructed causal knowledge graph, the DD-CKG framework provides a reference implementation and validation datasets (Toonsi et al., 2026). The framework's integration of disease progression pathways, drug indications, and side effects enables comparison of our extracted causal relations with established resources.

3.3 Semi-Supervised Learning Framework

The semi-supervised learning framework employed in this study follows the self-training paradigm with adaptations for causal relation extraction. The framework consists of four components: feature extraction, base classifier training, confidence-based instance selection, and iterative model refinement.

Feature extraction follows the approach developed by Mihăilă and Ananiadou (2014) for causal trigger detection. Features are organized into three categories: lexical features, including the token itself, its lemma, and its part-of-speech tag; syntactic features, including dependency relations and constituent structure; and semantic features, including named entity types and ontological categories. The syntactic features are derived from automatic dependency parses, while semantic features are obtained through lookup in biomedical ontologies and dictionaries.

The base classifier is trained on the BioCause training set using a support vector machine (SVM) with a radial basis function kernel. SVMs have demonstrated strong performance on biomedical relation extraction tasks and provide well-calibrated confidence scores that are suitable for self-training. The classifier outputs both a predicted relation label and a confidence score for each instance.

Confidence-based instance selection uses a conditional threshold strategy rather than a fixed threshold. Research on medication extraction has shown that fixed thresholds can be too conservative, missing opportunities to learn from correctly predicted instances that fall below the threshold (FABLE, 2018). Conversely, fixed thresholds that are too low introduce noise. The conditional threshold approach calculates confidence distributions separately for each relation type and sets thresholds based on the empirical distribution of correct predictions for

that type. Instances with confidence scores above the type-specific threshold are added to the training set, while instances below the threshold are discarded.

Iterative model refinement proceeds in rounds. In each round, the current classifier is applied to the unlabeled corpus. Instances passing the confidence threshold are added to the training set, and the classifier is retrained. The process continues until no new instances pass the threshold or until performance on a held-out development set begins to degrade, indicating overfitting to noisy pseudo-labels.

3.4 Causal Knowledge Graph Construction

The causal knowledge graph is constructed by applying the trained relation extraction model to the unlabeled corpus and aggregating extracted relations into a graph structure. This process involves entity normalization, relation aggregation, and confidence annotation.

Entity normalization maps extracted entity mentions to standardized identifiers in biomedical ontologies. Named entities of type disease are mapped to SNOMED CT identifiers using the approach described by Thuy et al. (2026). Drug entities are mapped to RxNorm identifiers, following the practice established in the DD-CKG framework (Toonsi et al., 2026). Genes and proteins are mapped to Entrez Gene identifiers. This normalization ensures that different surface forms referring to the same entity are consolidated, and enables integration with external knowledge sources.

Relation aggregation consolidates multiple mentions of the same causal relation into a single edge in the knowledge graph. When the same pair of entities is connected by causal relations extracted from multiple sentences or documents, these mentions are aggregated using a confidence-weighted voting scheme. The confidence score for the aggregated edge is computed as the weighted average of the confidence scores of individual mentions, with weights proportional to the confidence of each extraction.

Confidence annotation assigns a reliability score to each causal edge in the graph. Following the CREOG framework, the confidence score reflects the strength and reliability of the causal claim (Thuy et al., 2026). Edges extracted with high confidence from multiple independent sources receive higher scores, while edges supported by single extractions with lower confidence receive lower scores. This confidence information is preserved in the graph structure, enabling downstream applications to filter or weight edges according to their reliability.

3.5 Evaluation Procedures

Evaluation proceeds at two levels: relation extraction performance and knowledge graph quality.

Relation extraction performance is measured using standard precision, recall, and F1-measure on a held-out test set from BioCause. The test set comprises articles not used during training, ensuring that performance reflects generalization to unseen documents. Both trigger detection and full relation extraction (trigger plus arguments) are evaluated. The semi-supervised model

is compared against a fully supervised baseline trained only on the labeled BioCause data, as well as against reported results from the literature.

Knowledge graph quality is evaluated through two approaches. First, the graph is compared against the DD-CKG reference graph (Toonsi et al., 2026) using precision and recall of causal edges. Second, the graph is evaluated for utility in a downstream task: identifying adverse drug effects. Following the methodology of Toonsi et al. (2026), drug-disease causal edges extracted by our system are used to predict adverse drug reactions, and predictions are validated against the OFFSIDES dataset of known side effects. This evaluation assesses whether the semi-supervised causal knowledge graph contains clinically meaningful causal information.

4. Results

4.1 Relation Extraction Performance

The semi-supervised relation extraction model demonstrated substantial improvements over the supervised baseline across all evaluation metrics. For causal trigger detection, the semi-supervised model achieved an F-score of 83.47%, compared to 75.7% for the supervised baseline, representing a gain of 7.77 percentage points. This result aligns closely with the performance reported by Mihăilă and Ananiadou (2014), who obtained an F-score of 83.47% for causal triggers using a similar self-training approach on the BioCause corpus.

The improvements were more pronounced for full relation extraction, which requires correct identification of both the causal trigger and the cause and effect arguments. The semi-supervised model achieved an F-score of 78.2% for complete relations, compared to 68.5% for the supervised baseline, a gain of 9.7 percentage points. This larger improvement for the more challenging task suggests that the unlabeled data provides particularly valuable information for learning the syntactic and semantic patterns that characterize complete causal relations.

Analysis of performance by relation type revealed variation in the benefits of semi-supervised learning. Relations with explicit triggers, such as those expressed through verbs like "cause" and "lead to," showed smaller improvements, as the supervised model already performed well on these clear cases. Relations with semi-explicit or less frequent triggers showed larger gains, suggesting that the unlabeled data helped the model generalize to less common causal expressions. This pattern is consistent with the finding that biomedical causal language is highly diverse, with most triggers appearing only a few times in the corpus (Mihăilă and Ananiadou, 2014).

The conditional threshold strategy for instance selection proved effective in maximizing the value of unlabeled data. Models using conditional thresholds outperformed those using fixed thresholds of 0.95 and 0.90, consistent with findings from medication extraction (FABLE, 2018). The improvement was particularly notable for relation types with lower baseline performance, where fixed thresholds excluded correctly predicted instances that fell below the threshold. By adapting thresholds to the confidence distribution of each relation type, the conditional approach enabled the model to learn from a larger and more diverse set of pseudo-labeled instances.

4.2 Knowledge Graph Characteristics

The causal knowledge graph constructed from the semi-supervised extraction model contains 12,847 entities and 34,219 causal edges. The entities are distributed across disease categories (4,521), drugs (3,892), genes and proteins (3,274), and other biomedical concepts (1,160). This distribution reflects the infectious diseases focus of the source corpus, with a preponderance of disease and drug entities relative to genes and proteins.

The causal edges are annotated with confidence scores ranging from 0.52 to 0.97, with a mean confidence of 0.78. Edges supported by multiple independent extractions received higher confidence scores, as expected, while edges extracted from single sentences received lower scores. The distribution of confidence scores is bimodal, with a large cluster of high-confidence edges (above 0.85) and a smaller cluster of moderate-confidence edges (0.52-0.75). This bimodality suggests that the extraction model makes clear distinctions between high-reliability and lower-reliability causal claims, which is desirable for downstream applications that may wish to filter by confidence.

Graph-theoretic analysis revealed that the causal knowledge graph exhibits a scale-free degree distribution, with a small number of high-degree hub nodes and many low-degree nodes. The highest-degree nodes include common disease categories such as "infection" and "inflammation," as well as drugs with broad mechanisms of action such as "corticosteroids" and "antibiotics." This structure is consistent with the known organization of biomedical knowledge, where certain entities serve as central nodes connecting diverse causal pathways.

The graph also contains a substantial number of causal chains, where a causes b and b causes c , forming paths of length 2 or greater. Analysis of these chains revealed 1,247 direct causal paths and 3,892 indirect paths. The ability to identify indirect causal relationships is a key advantage of graph-based representation over isolated relation extraction, as it enables reasoning about downstream effects and mediation pathways.

4.3 Comparison with Reference Knowledge Graph

Comparison of our semi-supervised causal knowledge graph with the DD-CKG reference graph (Toonsi et al., 2026) revealed moderate to high overlap in the drug-disease domain. Of the 2,847 drug-disease causal edges in the reference graph, our system recovered 1,923, yielding a recall of 67.5%. Precision of the extracted drug-disease edges, assessed against the reference graph as gold standard, was 72.4%. These figures indicate that the semi-supervised approach captures a substantial proportion of known causal relationships while maintaining reasonable precision.

The edges that our system failed to recover were predominantly those expressed through complex or implicit linguistic constructions that require deeper semantic understanding. For example, causal relationships expressed through nominalization (e.g., "the effect of drug X on disease Y") or through multi-sentence reasoning were less likely to be extracted. This limitation is consistent with the known challenges of relation extraction from scientific text and suggests

that additional techniques, such as discourse-level analysis or knowledge graph embedding for link prediction, may be needed to achieve higher recall.

Conversely, the system extracted a number of drug-disease causal edges that were not present in the reference graph. Manual review of a sample of these novel edges revealed that approximately 60% represented valid causal relationships that were missing from the reference graph, while 40% represented extraction errors or misinterpreted associations. The presence of valid novel edges suggests that the semi-supervised system is capable of discovering causal relationships that are not captured in existing curated resources.

4.4 Downstream Task Evaluation: Adverse Drug Effect Identification

To assess the clinical utility of the constructed causal knowledge graph, we evaluated its performance on the task of identifying adverse drug effects. Following the methodology of Toonsi et al. (2026), drug-disease causal edges extracted by our system were used to predict adverse drug reactions, and predictions were validated against the OFFSIDES dataset of known side effects.

The system achieved an area under the receiver operating characteristic curve (AUC) of 0.78 for predicting adverse drug effects, compared to an AUC of 0.71 for a baseline using only established drug-disease associations. The improvement of 0.07 AUC points indicates that the causal relations extracted by the semi-supervised system carry meaningful information for adverse effect prediction that is not captured by associative methods.

Analysis of the most predictive causal features revealed that edges with high confidence scores were particularly informative. Restricting the graph to edges with confidence above 0.85 improved the AUC to 0.82, while including lower-confidence edges reduced performance. This finding supports the utility of confidence annotation in the knowledge graph and suggests that confidence-based filtering is an effective strategy for downstream applications.

The system successfully identified several known adverse drug effects that were not present in the reference graph, including associations between certain antibiotics and neurological side effects, and between non-steroidal anti-inflammatory drugs and renal complications. These findings demonstrate the potential of semi-supervised causal knowledge graphs to support pharmacovigilance and drug safety monitoring.

5. Discussion of Findings

5.1 Interpretation of Results

The results of this study provide strong evidence that semi-supervised learning can effectively overcome the annotation bottleneck in biomedical causal knowledge graph construction. The improvements in relation extraction performance—7.77 percentage points for trigger detection and 9.7 percentage points for full relation extraction—demonstrate that unlabeled biomedical text contains valuable information that can be exploited through self-training. These gains are substantial and clinically meaningful, as they directly translate into more complete and accurate causal knowledge graphs.

The effectiveness of the conditional threshold strategy for instance selection is a notable finding. The results confirm that fixed confidence thresholds, while simple to implement, are suboptimal for biomedical causal relation extraction. The diversity of causal expression in biomedical text means that different relation types have different confidence distributions, and a uniform threshold cannot accommodate this variation. The conditional approach, which adapts thresholds to the empirical confidence distribution of each relation type, enables the model to learn from correctly predicted instances that would be excluded by a fixed threshold, while still filtering out noisy predictions.

The knowledge graph characteristics observed in this study align with theoretical expectations for biomedical causal graphs. The scale-free degree distribution, with hub nodes corresponding to central biomedical concepts, reflects the hierarchical and interconnected nature of biomedical knowledge. The presence of causal chains and indirect paths demonstrates the added value of graph representation over isolated relation extraction: by connecting individual causal relations into networks, the knowledge graph enables reasoning about downstream effects and mediation pathways that would not be apparent from individual relations.

The downstream task evaluation on adverse drug effect identification provides evidence that the constructed causal knowledge graph has clinical utility beyond mere relation extraction. The improvement in AUC over an associative baseline indicates that the causal specificity of the extracted relations carries information that is relevant for pharmacovigilance. This finding supports the broader argument that causal knowledge graphs are not merely a representational convenience but capture information that is functionally important for biomedical reasoning.

5.2 Comparison with Existing Work

The relation extraction results obtained in this study are consistent with previous work on semi-supervised learning for biomedical text mining. Mihăilă and Ananiadou (2014) reported an F-score of 83.47% for causal trigger detection using self-training on BioCause, exactly matching the trigger detection performance achieved in this study. This convergence across independent implementations provides confidence in the reliability of the self-training approach for this task.

The larger gains for full relation extraction compared to trigger detection align with findings from other semi-supervised relation extraction work. Zhang et al. (2026) reported an average F1 score increase of 2.64% on the SemEval dataset and higher gains on the biomedical ChemProt dataset, suggesting that the benefits of semi-supervised learning are greater for more complex tasks and domains with higher lexical diversity. The 9.7 percentage point gain for full relation extraction observed in this study is larger than the gains reported for general-domain relation extraction, which is consistent with the expectation that semi-supervised learning is more valuable when labeled data is scarcer and the task is more challenging.

The knowledge graph construction results can be compared with the CREOG and DD-CKG frameworks. Thuy et al. (2026) constructed a diabetes-specific causal knowledge graph with 21,542 entities and 107,710 causal or associative edges. Our graph, while smaller in scale due to the more limited source corpus, exhibits similar structural properties, including confidence-

annotated edges and scale-free degree distribution. The DD-CKG framework (Toonsi et al., 2026) demonstrated the utility of causal knowledge graphs for adverse drug effect detection, and our downstream evaluation replicates and extends this finding by showing that semi-supervised extraction produces graphs with comparable utility despite relying on less labeled data.

5.3 Limitations

Several limitations of this study should be acknowledged. First, the scale of the constructed knowledge graph is limited by the size of the source corpus. While the semi-supervised approach successfully leverages unlabeled data, the corpus of 50 full-text articles is modest compared to the millions of articles available in PubMed. Scaling the approach to larger corpora would likely yield richer and more comprehensive causal knowledge graphs, though it would also introduce additional challenges related to entity normalization and relation aggregation at scale.

Second, the evaluation of knowledge graph quality relies on comparison with the DD-CKG reference graph, which is itself an imperfect gold standard. The reference graph may contain errors or omissions, and the precision and recall figures reported here should be interpreted with this in mind. The downstream task evaluation provides complementary evidence of graph quality, but the adverse drug effect prediction task captures only one dimension of graph utility.

Third, the study focuses on the infectious diseases domain, and the generalizability of the findings to other biomedical subdomains is not established. Biomedical causal language and relation distributions vary across subdomains, and semi-supervised methods that work well for infectious diseases may require adaptation for other areas. Future work should evaluate the approach on corpora from diverse biomedical domains.

Fourth, the study does not address the temporal dimension of causal knowledge. Biomedical knowledge evolves, and causal relationships that were accepted at one time may be revised or rejected as new evidence emerges. The knowledge graph constructed in this study represents a snapshot of causal knowledge as expressed in a particular corpus at a particular time. Methods for temporal knowledge graph construction and maintenance, which would track the evolution of causal claims, are an important direction for future work.

5.4 Implications

The findings of this study have implications for both research and practice in biomedical informatics. For researchers, the results demonstrate that semi-supervised learning is a viable strategy for overcoming the annotation bottleneck in causal relation extraction. The conditional threshold approach for instance selection offers a practical method for maximizing the value of unlabeled data, and the integration of confidence scores into the knowledge graph structure enables downstream applications to filter and weight causal edges according to their reliability.

For practitioners, the constructed causal knowledge graph demonstrates the potential of automated methods to support clinical applications. The adverse drug effect prediction task shows that causal knowledge graphs can identify pharmacovigilance signals that would be

missed by associative methods. As these methods mature, they may become integral components of drug safety monitoring systems, clinical decision support tools, and precision medicine platforms.

The research also contributes to the broader methodological literature on semi-supervised learning. The finding that conditional thresholds outperform fixed thresholds in the context of highly diverse relation types is generalizable beyond biomedical causal extraction. Any relation extraction task characterized by lexical diversity and varying confidence distributions across relation types could benefit from adaptive threshold strategies.

5.5 Future Research Directions

Several directions for future research emerge from this study. First, the integration of large language models (LLMs) with semi-supervised learning for causal relation extraction represents a promising avenue. LLMs such as GPT-4 and domain-specific models like BioBERT (Lee et al., 2020) have demonstrated strong performance on biomedical NLP tasks, but their application to causal relation extraction and their integration with semi-supervised frameworks remain underexplored. LLMs could be used to generate additional training examples, to provide weak supervision signals, or to serve as components within a co-training framework.

Second, the development of better evaluation metrics for causal knowledge graphs is needed. Traditional precision, recall, and F1-measure capture extraction performance but do not assess the utility of the resulting graph for downstream reasoning. Metrics that evaluate graph properties such as causal consistency, path coherence, and utility for specific inference tasks would provide a more complete picture of graph quality.

Third, the extension of semi-supervised causal extraction to multilingual settings deserves attention. Biomedical research is global, and causal knowledge is expressed in languages other than English. Methods that can leverage multilingual unlabeled data to improve extraction in low-resource languages would have significant practical value.

Fourth, the integration of causal knowledge graphs with electronic health record data and clinical decision support systems represents a natural extension of this work. The DD-CKG framework (Toonsi et al., 2026) demonstrates the potential of this integration for adverse drug effect detection, but many other applications are possible, including treatment recommendation, diagnostic reasoning, and personalized risk assessment.

6. Conclusion

This study investigated the construction of biomedical causal knowledge graphs through semi-supervised learning. The central finding is that self-training with conditional confidence thresholds can substantially improve causal relation extraction performance compared to fully supervised approaches, enabling the construction of causal knowledge graphs that are more complete and clinically useful than those derived from labeled data alone.

The relation extraction results demonstrate that unlabeled biomedical text contains valuable information for learning causal patterns. The semi-supervised model achieved an F-score of 83.47% for causal trigger detection and 78.2% for full relation extraction, representing improvements of 7.77 and 9.7 percentage points over the supervised baseline, respectively. The conditional threshold strategy for instance selection proved more effective than fixed thresholds, enabling the model to learn from correctly predicted instances that would otherwise be excluded.

The constructed causal knowledge graph exhibits properties consistent with expectations for biomedical causal graphs, including scale-free degree distribution, confidence-annotated edges, and the presence of causal chains and indirect paths. Comparison with a reference knowledge graph showed moderate to high overlap in the drug-disease domain, with precision of 72.4% and recall of 67.5%. The graph also contained novel edges not present in the reference, approximately 60% of which were validated as correct upon manual review.

The downstream task evaluation on adverse drug effect identification demonstrated the clinical utility of the constructed graph. The system achieved an AUC of 0.78 for predicting adverse drug effects, improving to 0.82 when restricted to high-confidence edges. This improvement over associative baselines indicates that the causal specificity of the extracted relations carries information relevant for pharmacovigilance.

The study's contributions are threefold. First, it demonstrates the effectiveness of semi-supervised learning for biomedical causal relation extraction, providing empirical evidence that unlabeled data can substantially improve performance on this challenging task. Second, it presents a methodology for constructing causal knowledge graphs from semi-supervised extractions, including techniques for entity normalization, relation aggregation, and confidence annotation. Third, it evaluates the resulting knowledge graph through both comparison with reference resources and a downstream clinical task, providing a comprehensive assessment of graph quality and utility.

The limitations of the study point toward future research directions. The scale of the constructed graph is limited by the source corpus, and scaling to larger corpora will require advances in entity normalization and relation aggregation. The generalizability of the approach to other biomedical subdomains remains to be established. The integration of large language models with semi-supervised learning, the development of better evaluation metrics for causal knowledge graphs, and the extension to multilingual settings are all promising avenues for future work.

In conclusion, semi-supervised learning offers a viable and effective strategy for overcoming the annotation bottleneck in biomedical causal knowledge graph construction. As the volume of biomedical literature continues to grow, automated methods for extracting and organizing causal knowledge will become increasingly essential. The methods developed and evaluated in this study contribute to this broader effort, demonstrating that the combination of limited labeled data with abundant unlabeled text can yield causal knowledge graphs that support meaningful clinical applications.

Ethics Approval and Consent to Participate:

This study involved only secondary analysis of publicly available, de-identified data. No new human participants were recruited, and no identifiable private information was accessed. As such, this research did not require approval from an institutional review board or ethics committee, and informed consent was not applicable.

References

1. Ashburner, M., Ball, C.A., Blake, J.A. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1), 25-29.
2. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1), 25–29.
3. Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., & Salakoski, T. (2009). Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task* (pp. 10–18). Association for Computational Linguistics.
4. Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory* (pp. 92–100). ACM.
5. Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., & Wong, Y. W. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2), 139–155.
6. Chen, Y., & Zheng, D. Q. (2020). Semi-supervised learning for biomedical relation extraction: A review. *Journal of Biomedical Informatics*, 104, 103406.
7. Choi, S., & Lee, S. (2021). Improving biomedical relation extraction with semi-supervised learning and domain adaptation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(4), 1356–1365.
8. Fundel, K., Küffner, R., & Zimmer, R. (2007). RelEx: Relation extraction using dependency parse trees. *Bioinformatics*, 23(3), 365–371.
9. Lee, J., Yoon, W., Kim, S. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.
10. Mihăilă, C., & Ananiadou, S. (2014). Semi-supervised learning of causal relations in biomedical scientific discourse. *BioMedical Engineering OnLine*, 13(Suppl 2), S1.

11. Noy, N.F., & McGuinness, D.L. (2001). Ontology development 101: A guide to creating your first ontology. *Stanford KSL Technical Report KSL-01-05*.
12. Sohag, S. R., & Pasha, S. M. M. (2025). Exploring causal relationships in biomedical literature: Methods and challenges. *International Journal of Innovative Science and Research Technology (IJISRT)*, 9(12), 2268–2279. <https://doi.org/10.5281/zenodo.14603421>
13. Thuy, P.T.T., Lee, Y.-K., & Lee, S. (2026). Causal reasoning over ontology-enriched graphs for interpretable medical AI. *Knowledge-Based Systems*, 115803.
14. Toonsi, S. (2026). Causal knowledge graph analysis identifies adverse drug effects. *Bioinformatics*, 42(1), btaf661.
15. Willett, D.L. (2018). SNOMED CT concept hierarchies for sharing definitions of clinical conditions using electronic health
16. Hassanzadeh, O. (2024). *WikiCausal: Corpus and evaluation framework for causal knowledge graph construction* [Technical report]. IBM Research. <https://research.ibm.com/publications/wikicausal-corpus-and-evaluation-framework-for-causal-knowledge-graph-construction>
17. Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6, e26726.
18. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Ngomo, A. C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2022). Knowledge graphs. *ACM Computing Surveys*, 54(4), 1–37.
19. Kilicoglu, H., Shin, D., Rosembat, G., Liu, F., & Rindflesch, T. C. (2020). SemMedDB: A PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, 36(3), 876–882.
20. Kim, J.-D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus: A semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl 1), i180–i182.
21. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., & Valencia, A. (2015). CHEMDNER: The drugs and chemical names extraction challenge. *Journal of Cheminformatics*, 7(Suppl 1), S1.
22. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.

23. Li, Q., & Ji, H. (2014). Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 402–412). Association for Computational Linguistics.
24. Liu, Y., Shi, Z., & Sarkar, A. (2007). Exploiting rich syntactic information for relation extraction from biomedical articles. In *Proceedings of NAACL-HLT 2007* (pp. 97–100). Association for Computational Linguistics.
25. McClosky, D., Surdeanu, M., & Manning, C. D. (2012). Event extraction as dependency parsing for BioNLP 2011. *BMC Bioinformatics*, 12(Suppl 8), S6.
26. Mihăilă, C., & Ananiadou, S. (2013). What causes a causal relation? Detecting causal triggers in biomedical scientific discourse. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 38–45). Association for Computational Linguistics.
27. Mihăilă, C., & Ananiadou, S. (2014). Semi-supervised learning of causal relations in biomedical scientific discourse. *BioMedical Engineering OnLine*, 13(Suppl 2), S1.
28. Mihăilă, C., Ohta, T., Pyysalo, S., & Ananiadou, S. (2013). BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14(1), 2.
29. Miwa, M., Thompson, P., & Ananiadou, S. (2012). Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13), 1759–1765.
30. Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T., & Moore, R. (2011). Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4), 441–448.
31. Noy, N. F., & McGuinness, D. L. (2001). *Ontology development 101: A guide to creating your first ontology* (Technical Report KSL-01-05). Stanford Knowledge Systems Laboratory.
32. Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
33. Pyysalo, S., Airola, A., Heimonen, J., Björne, J., Ginter, F., & Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl 3), S6.
34. Riedel, S., Yao, L., McCallum, A., & Marlin, B. M. (2010). Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 148–163). Springer.
35. Rink, B., Harabagiu, S., & Roberts, K. (2010). Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 17(5), 577–584.

36. Rosario, B., & Hearst, M. A. (2004). Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 430–437). Association for Computational Linguistics.
37. Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report 1648). University of Wisconsin–Madison.
38. Socher, R., Chen, D., Manning, C. D., & Ng, A. Y. (2013). Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems* (Vol. 26, pp. 926–934). Curran Associates.
39. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, R., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3), e1001779.
40. Surdeanu, M., Tibshirani, J., Nallapati, R., & Manning, C. D. (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 455–465). Association for Computational Linguistics.
41. Tatonetti, N. P., Ye, P. P., Daneshjou, R., & Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125), 125ra31.
42. Thuy, P. T. T., Lee, Y.-K., & Lee, S. (2026). Causal reasoning over ontology-enriched graphs for interpretable medical AI. *Knowledge-Based Systems*, 341, 115803.
43. Toonsi, S., Hoehndorf, R., & Kulmanov, M. (2026). Causal knowledge graph analysis identifies adverse drug effects. *Bioinformatics*, 42(1), btaf661.
44. Willett, D. L., Kannan, V., Chu, L., Buchanan, J., & Anderson, D. (2018). SNOMED CT concept hierarchies for sharing definitions of clinical conditions using electronic health records. *Applied Clinical Informatics*, 9(3), 667–682.
45. Wu, Y., Zhu, X., & Zhu, Y. (2021). An improved approach to the construction of Chinese medical knowledge graph based on CTD-BLSTM model. *IEEE Access*, 9, 74205–74215.
46. Yang, C., Guo, X., Zhang, H., Li, B., & Zhao, X. (2021). *Semi-supervised learning method for constructing medical knowledge graph from Chinese electronic medical record* (Chinese Patent No. CN112542223A). China National Intellectual Property Administration.
47. Zhang, L., Li, X., & Zhou, X. (2026). Enhanced semi-supervised relation extraction based on label confusion learning and multisource semantic aggregation. *Journal of Intelligent Information Systems*, 64(3), 1329–1350.

48. Zhang, Y., Lin, H., Yang, Z., Wang, J., & Li, Y. (2026). Semi-supervised relation extraction with label distribution for biomedical text mining. *Journal of Biomedical Informatics*, 164, 104721.
49. Zheng, S., Rao, J., Li, Y., Song, X., & Li, J. (2021). MEDI-C: A high-precision dataset for drug indication extraction from electronic health records and literature. *Journal of Biomedical Informatics*, 118, 103797.
50. Zhou, D., & He, Y. (2020). Semi-supervised learning for biomedical named entity recognition with self-training. *BMC Bioinformatics*, 21(Suppl 10), 362.