BMC Medical Informatics and Decision Making

RESEARCH

Open Access

A study on large-scale disease causality discovery from biomedical literature

Shirui Yu[1,2], Peng Dong[3], Junlian Li[3], Xiaoli Tang[3*] and Xiaoying Li[3*]

Abstract

Background Biomedical semantic relationship extraction could reveal important biomedical entities and the semantic relationships between them, providing a crucial foundation for the biomedical knowledge discovery, clinical decision making and other artificial intelligence applications. Identifying the causal relationships between diseases is a significant research field, since it expedites the identification of underlying disease pathogenesis mechanisms and promote better disease prevention and treatment. SemRep is an effective tool for semantic relationship extraction in the biomedical field, but it is not accurate enough for disease causality extraction, bringing challenges for downstream tasks. In this study, we proposed an optimization strategy for SemRep to enhance its accuracy in disease causality extraction.

Methods This study aims to optimize disease causality extraction of SemRep tool by constructing a semantic predicate vocabulary that precisely expresses disease causality to support the automatic extraction of disease causality knowledge from biomedical literature. The proposed method invloves the following four steps: Firstly, we obtained a collection of semantic feature words expressing disease causality based on current causality predicate studies and the disease causality pairs extracted from SemMedDB. Then, we constructed a disease causality semantic predicate vocabulary by filtering and evaluating the clue words using quantitative comparisons. Following that, we extracted disease causality pairs from the biomedical literature using 36 semantic predicates with an accuracy greater than 80% for more meaningful knowledge discovery. Finally, we conducted knowledge discovery based on the extracted disease causality triples, which primarily includes unidirectional disease causality, bidirectional disease causality, as well as two specific types of disease causality: primary disease causality and rare disease causality.

Results We obtained a disease causality semantic predicate vocabulary containing 50 textual predicates with an accuracy of above 40%. 36 semantic predicates from the 60% accuracy group were used for disease causality extraction, yielding 259,434 disease causality pairs for subsequent knowledge discovery. Among them, 92,557 types with 176,010 unidirectional disease causality triples, and 6084 types with 83,424 bidirectional disease causality triples were found eventually. Two other types of disease causality, primary disease causality and rare disease causality, were also discovered.

*Correspondence:
Xiaoli Tang
tang.xiaoli@imicams.ac.cn
Xiaoying Li
lixiaoying@imicams.ac.cn
Full list of author information is available at the end of the article

Conclusions The novelty of this research is that the proposed method enhanced the disease causality extraction of SemRep tool, resulting a more accurate and comprehensive disease causality extraction. It also facilitates an automatic disease causality extraction from large-scale biomedical literature. Additionally, a customized extraction of disease causality for its accuracy and comprehensiveness can be made possible by leveraging the quantified causality predicate vocabulary, allowing for flexible extraction of disease causality according to the actual circumstance.

Keywords Disease causality, Knowledge discovery, SemRep, Semantic predicates

Introduction

As the scientific and technical literature continues to expand, while the knowledge structures, knowledge sequences and knowledge associations become more and more complex, it's increasingly difficult for researchers to stay abreast of the latest development in their fields and discover further connections between knowledge [1]. Luckily, Natural Language Processing (NLP) technology offers a solution to this problem by extracting biomedical concepts and their relationships from free text, which can then be transformed into computable semantic representations using contextual features. Therefore, NLP supports biomedical knowledge discovery, clinical decision making and other valuable applications, proving to be a flexible and effective method [2].

Among them, the Semantic Knowledge Representation (SemRep) tool has gained much popularity in natural language processing and semantic relation extraction of biomedical literature. This tool is a rule-based semantic relation extractor from biomedical text, which could extract a total of 58 types of semantic relationships including "CAUSES". A significant advantage of SemRep is, it could align the various biomedical entities names to UMLS concepts and then efficiently fulfill their normalization. For example, Table 1 lists some of the extracted disease causal knowledge in a structured format, namely Subject – Predicate – Object (SPO) triple. However, there are a few limitations of SemRep tool. No public documents are available about the rules used to predefine the 58 semantic types. Thus, it is not obvious which clue words were incorporated into to the "CAUSES" and whether they could be adequate to express the causality from biomedical text [3]. Based on numerous evaluations, it has been revealed that SemRep's semantic relation extraction results produce some errors with accuracy rates ranging from 53 to 83%, based on many task-based evaluations. It has been found that the lack of accurate semantic predicate recognition is a major cause of errors. Thus, the main focus of this research is to improve the accuracy of SemRep's semantic relation extraction through semantic predicate optimization.

Mining disease causality in the field of biomedicine is critical to uncover the essential associations between diseases, thereby enabling the identification of risk factors for disease, which in turn can understand the root causes of disease [4]. Understanding disease causality will also facilitate the study of disease phenotypes, and researchers can develop attribution studies that can rapidly reveal the underlying mechanisms of disease. This can provide new insights into the molecular mechanisms of disease, leading to better disease prevention and treatment. As a result, it is of great value in real-world application [5].

Biomedical semantic relation extraction methods are primarily categorized into three types: rule-based methods, traditional machine learning methods, and deep learning methods. Rule-based methods leverage biomedical knowledge resources, combined with co-occurrence analysis and manually formulated semantic relation rule templates, to extract semantic relations. Semantic relations are represented with triplet patterns. For instance, Lee et al. [6] utilized manually created semantic relation templates to extract PPIs triples from texts, achieving a high accuracy of 97% on the AIMed dataset [7]. Traditional machine learning methods extract semantic relations by automatically extracting semantic predicates in the corpus that can express accurate semantic relations between entities, and supervised learning is the predominant methods. For example, in the SemMedDB [8] based biomedical literature data mining, Zhang [9] employed a machine

Table 1 Examples of SemRep's disease causalities in SPO format

No	Biomedical text *w*	Subject – Predicate – Object (SPO)
1	Gallic acid (GA) plays a significant role in **cardiovascular disorders resulted from diabetes**	**Diabetes – CAUSES – Cardiovascular Diseases**
2	Endothelial dysfunction, activation, inflammation, and endothelial barrier leakage are key factors contributing to vascular complications in diabetes, plus the development of **diabetes-induced cardiovascular diseases**	
3	**Diabetes** is a leading **cause of cardiovascular disease** and its associated morbidity	

learning approach to construct a semantic predicate filter. This filter aimed to extract semantic predicates with high accuracy in expressing semantic relations. Consequently, this method resulted in an enhancement of the accuracy in the extraction of gene-drug semantic relations from 58 to 69%. Deep learning methods can automatically learn underlying features from large datasets and form more abstract high-level representations of attribute categories or features. For instance, Lai et al. [10] employed a LC-CNN method to extract semantic relations between diseases. This method achieved an accuracy of 82%, a recall of 85%, and an F-score of 84% on the DDAE dataset based on biomedical literature data in PubMed.

The discovery of disease causality can support clinical diagnosis and treatment, which is an important research topic in the study of disease. Currently, we mainly constructed disease causality networks to promote the disease causality relations discovery and support the hypothesis generation of disease progression pathways and disease causality. For example, Bang et al. [11] utilized various biomedical data, including genes, proteins, clinical information, etc., to determine the causality between diseases and constructed a disease causality network. Zhou et al. [12] proposed a method to determine the association between diseases based on disease phenotypes. Lee [13] constructed a disease causality network based on biomedical literature text mining, and proposed the construction of a disease causality network. causal network construction and further identified the disease progression pathways in the network. As a crucial knowledge source in the field of medicine, biomedical literature provided valuable semantic information for the identification of specific disease causality patterns. This facilitated the construction of intricate disease causality networks, thus enhancing the research of disease causality discovery.

The available relevant disease-disease association datasets can facilitate the in-depth exploration of disease causality, and provide a basis for the development and evaluation of text mining methods, which plays an important role. For example, Lai et al. [14] used 521 PubMed abstracts to formulate a disease-disease association extraction (DDAE) dataset, consisting of disease mentions, Medical Subject Heading IDs, and relation annotations. A neural network model for extracting disease-disease associations from the literature were developed on the basis of the dataset. The construction of the dataset is helpful for the research and optimization of the model performance. Xu et al. [15] proposed a semi-supervised iterative pattern-learning approach to learn disease-disease association patterns from PubMed abstracts. Based on this approach, they constructed a disease-disease risk relationship knowledge base (dRiskKB) consisting of 34,000 unique disease pairs. Nicia et al. [16] proposed SicknessMiner, a methodology that encompasses Named Entity Recognition (NER) and Named Entity Normalization (NEN), and used DisGeNET to evaluate the testing results.

In this study, a method of expanding the text semantic predicates in SemRep and then evaluating the extracted text semantic predicates was proposed to achieve the optimization of semantic predicates. This process improved the accuracy of SemRep in automatically extracting disease causality and thus contributing to a better performance in finding disease causality from biomedical literature. The overall research path of our study is outlined in Fig. 1.

Methods

Data preprocessing

The complete data up to December 2021 were obtained from the SemMedDB database. Then the data were filtered and processed to construct a base set required for this research and a small test set specifically designed for evaluation purpose. In constructing the base set, the disease relation pairs were extracted by cleaning and screening according to the subject and object semantic types, as well as the predicates. Generalized and misidentified data were further eliminated from the dataset. All in all, a total of 1,268,284 disease relation pairs were obtained, serving as the base set for this study. In addition, the SemRep-processed relation pair data were automatically subsumed for semantic predicates, and the corresponding text predicates could not be obtained. As a result, we resorted to the text sentences and the textual predicates based on the start and end positions of characters to extract the text predicates.

A small test set was constructed to evaluate the effectiveness of the SemRep's automatic disease causality extraction and to explore its shortcomings so as to develop an optimization strategy. Specifically, 500 documents were randomly selected from the base set yielding a total of 741 disease relation pairs. Two experts manually identified and labeled the data on their own by analyzing whether the relation pairs belonged to disease causality according to text sentences, resulting in 304 labeled disease causality pairs. The semantic predicates in the 304 disease causality pairs were further examined and evaluated, and 28 semantic predicates expressing disease causality were obtained. The indices

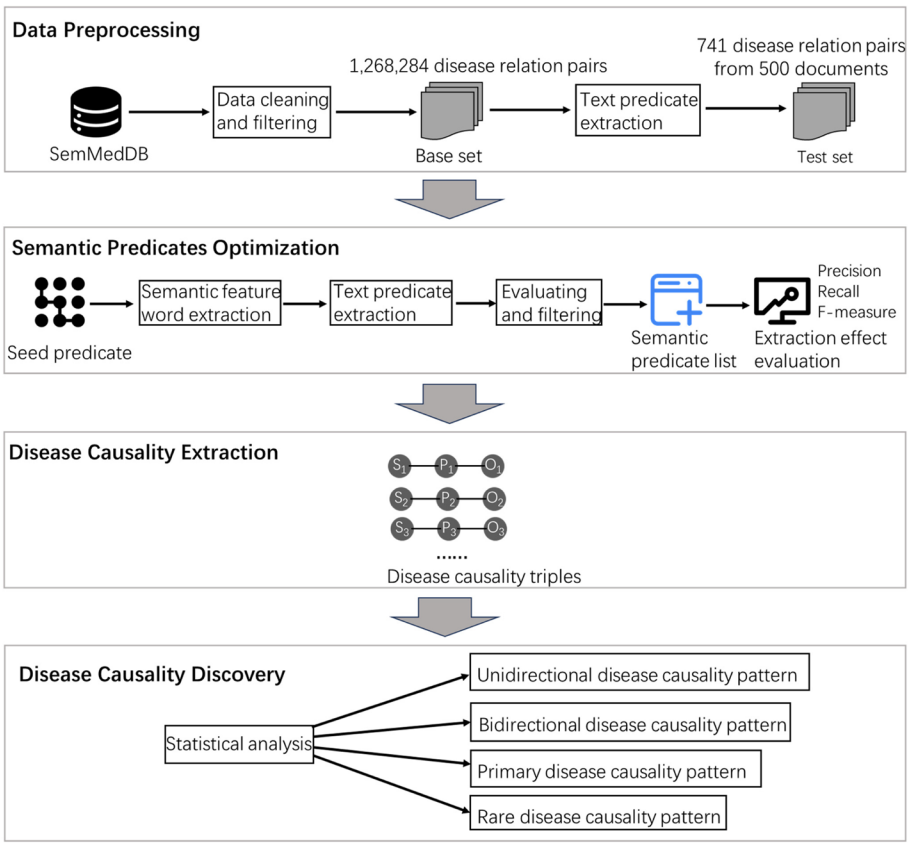

Fig. 1 Research map

of accuracy, recall, and *F*-value were used to evaluate SemRep 's disease causality extraction, with the result of 85.34%, 53.62%, and 65.86%, respectively. The analysis showed that SemRep's disease causality extraction had the problem of incomplete and inaccurate subsumption of semantic predicates. Therefore, the optimization of semantic predicates by means of semantic predicate supplementation and screening was proposed to improve the performance of SemRep for automatic disease causality extraction.

Semantic predicates optimization

In this study, the textual semantic predicates were extended by filtering the semantic feature words extracted from two sources, SemRep's parsing and identification results from literature and existing research results. The semantic predicates were screened according to the accuracy of causality between diseases. Finally, a semantic predicate vocabulary for disease causality was constructed to achieve the optimization of semantic predicates.

To achieve this, we first extracted semantic feature words expressing disease causality from a published research paper by Xu [15] which identified 26 semantic predicates of disease causality with high accuracy and 28 semantic predicates obtained from the test set. These two parts formed the basis of the reference predicates. By supplementing the semantic feature words, the incomplete coverage of semantic feature words due to different tenses or lexical properties can be addressed. After removing duplicates, a total of 22 semantic feature words and complementary semantic feature words were extracted (Table 2).

Then, more forms of predicates were filtered from the base set based on the extracted semantic

Table 2 A vocabulary of disease causality semantic feature words

Semantic Feature Words			
Cause	induce	responsible	complicating
Result	lead	risk	etiologies
Attribute	led	secondary	inducing
Complicate	owing	trigger	producing
Due	pathogenesis	causing	
Etiology	produce	causative	

feature words. All textual predicates containing the feature predicates in the base set were extracted. The extracted predicates were independently reviewed by two experts in terms of two aspects, "whether these predicate forms are reasonable" and "whether they can convey the meaning of disease causality". Furthermore, the predicate forms with too many characters or obviously unreasonable were removed. Similarly, those that did not indicate disease causality or were wrongly extracted were also removed. Finally, we obtained 56 semantic predicates that could express disease causality.

In order to quantitatively reveal the accuracy of each semantic predicate expressing disease causality, 50 disease causality pairs were randomly selected for each semantic predicate in the base set (the total number of relation pairs for some predicates may be less than 50). The accuracy of each semantic predicate was evaluated by manual audit as described above, in which 36 predicates had an accuracy rate of not less than 80%, 42 predicates not less than 60%, and 50 predicates not less than 40%. A semantic predicate vocabulary of disease causality was constructed using 50 textual predicates with an accuracy rate of not less than 40% (Table 3).

Finally, the effect of the automatic extraction of disease causality based on the semantic predicate vocabulary was tested on the test set. The predicates in the semantic predicate vocabulary were divided into 80% accuracy group, 60% accuracy group and 40% accuracy group according to the accuracy rate, and 36, 42 and 50 semantic predicates were used for each group. The results (Table 4) showed that the accuracy of disease causality extraction tended to decrease as the accuracy of the semantic predicates decreased, while the recall and F-score tended to increase, which was in accordance with the general rule. Compared with SemRep, which did not optimize semantic predicates and whose performance in disease causality pair extraction was 85.34%, the accuracy rate of our method increased by 13.63%, 12.74%, and 8.31%, respectively. The results indicated that the optimization of semantic predicates could improve the accuracy of SemRep's automatic disease causality extraction, verifying the feasibility of this method.

Table 4 The evaluation of disease causality semantic predicate extraction

Number of Predicates	Precision	Recall	F-score	Accuracy Improvement
36	96.97%	63.16%	76.50%	13.63%
42	96.21%	66.78%	78.84%	12.74%
50	92.43%	76.32%	83.60%	8.31%

Disease causality extraction

In this study, the disease causality semantic predicate vocabulary was used to extract disease causality pairs from the base set. Since more attention is paid to the accuracy of disease causality in medical diagnosis and treatment decision-making and other related applications, we used the predicates with higher accuracy and smaller number to automatically extract disease causality pairs for disease causality discovery to avoid the introduction of noise. We selected 36 textual predicates with 80% and above accuracy in semantic predicates for disease causality extraction, and a total number of 259,434 disease causality triples were found eventually.

Disease causality discovery

This study focused on disease causality discovery from the perspective of fine-grained knowledge units based on the extracted triples. By performing frequency statistics and relation matching on the extracted disease causality pairs, we could reveal the specific type or characteristic of different disease causality.

Table 3 A vocabulary of disease causality semantic predicates

Semantic Predicates				
Caused	resulting	precipitate	triggering	as the result
Play important causative roles	result	lead	triggered	play an etiological role
As a results	results	causes	leading	etiology
As the end result	producing	causing	triggers	etiologies
As the results	risk factor	resulted	trigger	aetiology
Due to	produce	induces	led	produced
Secondary	induce	leads	responsible	causative
Cause	attributable	as a result	risk factors	pathogenesis
Precipitated	owing to	produces	as result	risk
Induced	due	inducing	play causative roles	as a direct result

The accuracy rate is ranked from high to low

Results

Disease causality discovery results

In this study, a total of 259,434 disease causality pairs were extracted. Among them, various forms of disease causality structures were identified, primarily unidirectional disease causality, bidirectional disease causality, and two specific types: primary disease causality and rare disease causality.

(1) *Unidirectional disease causality*: Among these disease causality pairs, a total of 92,557 types, or 176,010 unidirectional disease causality triples were obtained. We visualized 32 disease causality pairs with a frequency of not less than 100 times for analysis (Fig. 2).

The figure indicates that the disease causality related to blind vision is of the highest frequency. The major causes of blindness come from trachoma and diabetic macular edema and blindness is often a serious consequence of these diseases. The figure shows that hyperhomocysteinemia is a risk factor for both cardiovascular diseases and atherosclerosis. In addition, the figure also clearly demonstrates that the diseases trigger acute kidney failure, chronic kidney failure and end stage renal failure.

(2) *Bidirectional disease causality*: In this category of disease causality pairs, we found 6,084 types, or 83,424 bidirectional disease causality triples. The association between obesity, diabetes and hypertensive disease is widely acknowledged to the public, so in this

Fig. 2 A visualization of bidirectional disease causality. The direction of the connecting line is from "cause" to "effect", representing the direction of disease causality. The weight on the line indicates the frequency of disease causality

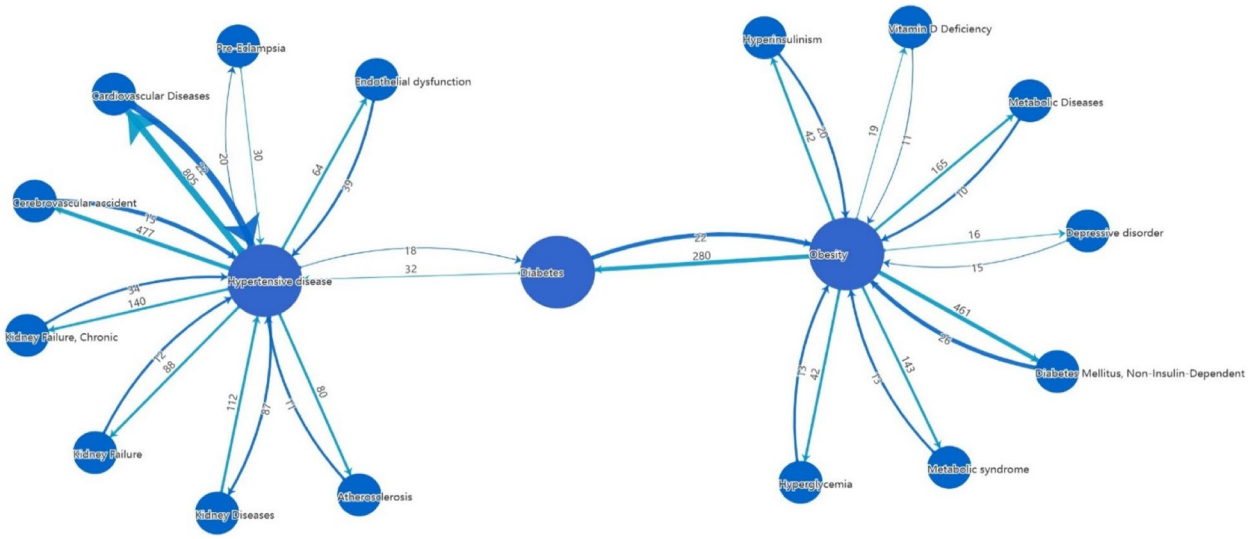

Fig. 3 Weighted bidirectional disease causality graph between obesity and hypertension. The weight is determined by the frequency of disease causality pairs in the literature, which is the sum of the forward and reverse frequencies. The direction of the connecting line represents the direction of disease causality. The thickness of the connecting line represents the weight

study we selected these three diseases as examples of bidirectional disease causality for detailed analysis. A weighted bidirectional disease causality diagram with obesity and hypertension at its core is shown below (Fig. 3). The diagram depicts a simple disease causality network involving 18 diseases, where obesity and hypertension are connected together through the intermediary node of diabetes.

In recent years, the global prevalence of obesity has been increasing year by year. Obesity has become one of the major concerns in human health because it increases the risk of developing a variety of other diseases. In the relationship diagram, eight diseases exhibit high-frequency bidirectional causality with obesity. Notably, one of the most strongly correlated diseases is non-insulin-dependent diabetes mellitus, commonly referred to as type 2 diabetes. Individuals with obesity often exhibit insulin resistance in their adipose tissue, which induces increased insulin secretion from pancreatic β-cells. However, this overcompensation for insulin secretion could lead to pancreatic β-cell failure and inadequate insulin secretion, ultimately resulting in non-insulin-dependent diabetes mellitus [17]. The underlying pathogenesis of this correlation between obesity and non-insulin-dependent diabetes mellitus has been studied extensively. Additionally, genetic studies provide insights suggesting that non-insulin-dependent diabetes mellitus causes obesity [18].

This research has found that nine diseases exhibit bidirectional causality with hypertensive disease. It is widely known that hypertensive disease is closely related to cardiovascular disease and cerebrovascular accident. Hypertension is a significant risk factor for cardiovascular disease, and cardiovascular disease could also cause hypertension, which means that there is a bidirectional causality between two diseases. Evidence suggests that the renin–angiotensin–aldosterone system (RAAS), the role of natriuretic peptides and the endothelium, the sympathetic nervous system (SNS) and the immune system in a comprehensive neurohumoral system play an important part in the regulation of blood pressure levels. In which the dysfunction involving blood pressure control factors may contribute to hypertension, and induce damage to target organs over time. A number of other syndromes, including coronary artery disease, stroke and cardiovascular diseases, are manifestations of this dysfunction [19]. By understanding the underlying mechanisms, it is possible to reduce the risk of the disease and provide useful reference for diagnosis and treatment of the disease.

It was also found that there was a remarkable difference in frequency between the forward and reverse disease causality pairs, which is referred to as bidirectional disease causality with high frequency difference in this study. To quantify this difference, we used the following formula:

$$D = (F - N)/N \tag{1}$$

(D is the value of the frequency difference, F denotes the forward frequency and N denotes the reverse frequency.)

In this study, the high-frequency differential bidirectional disease causality with reverse frequency less than 10 and frequency difference greater than 10 were selected for analysis. The erroneous data were further eliminated according to literature (Table 5). This type of disease causality is particularly valuable because it indicates a high degree of uncertainty in the causality between two diseases, which can provide important clues for further scientific research and clinical investigation.

An analysis of the source literature on disease causality revealed that obstructive sleep apnea (OSA) is a common sleep disorder and that obesity has been identified as a major risk factor for OSA. However, in studies of OSA leading to obesity, the literature published in 2013 concluded that it is uncertain whether OSA causes obesity [20]. And a more recent study published in 2020 explained that "OSAS itself leads to obesity by causing both leptin and insulin resistance as a consequence of activation of the sympathetic nervous system." (PMID:33264533). In addition, no other relevant studies were found, suggesting that researchers and clinicians may consider paying more attention to this pair of disease relationships or conducting in-depth studies.

(3) *Other Disease Causality*: The analysis also identified two other special disease causality patterns. These two patterns are defined as "primary disease causality" and "rare disease causality" in this study. Overall, the understanding of these unique disease causality patterns is more noteworthy than the regular ones that have been studied previously because of their distinct characteristics.

(a) Primary disease causality: *"*Primary disease causality*"* is defined as having at least two disease causality pairs extracted from a single sentence in a way that can be represented as a chained or bidirectional disease causality structure. Examples are shown in the table below (Tables 6 and 7).

Primary disease causality was extracted from the same sentence and these diseases were closely related and had been scientifically verified. Therefore, the primary disease causality pattern is more scientific and credible.

(b) *Rare disease causality*: *"*Rare disease causality*"* refers to a disease that rarely causes another disease. In this study, the disease causality pairs with low frequency (10 *times or less) were defined as "*rare disease causality*"*. The specific examples are shown in the following table (Table 8).

The study of rare disease causality is of great significance for two reasons. Firstly, it suggests that the risk factors and implications for the pathogenesis of the disease are infrequent, and therefore require serious consideration. Second, because the disease is not commonly studied, it holds promise for further exploration and provides insight into uncovering the underlying mechanism.

Discussion

In this study, we proposed a disease causality extraction method based on the semantic predicate optimization of SemRep. By screening textual predicates in SemRep

Table 5 Bidirectional disease causality pairs with high-frequency difference

Disease Causality Pairs			D
Number	***Cause***	***Effect***	
1	Coronary Artery Vasospasm	Myocardial Infarction	31.33
2	Hypertension, Pulmonary	Right ventricular failure	30.75
3	Obesity	Fatty Liver	25.25
4	Ischemia	Ventricular Fibrillation	21.57
5	Hypertensive disease	End stage renal failure	21.5
6	Subarachnoid Hemorrhage	Cerebral Vasospasm	20.71
7	Obesity	Asthma	20.57
8	Obesity	Sleep Apnea, Obstructive	18
9	Bacterial Infections	Septicemia	12.8
10	Strabismus	Amblyopia	11.6

Table 6 Example of primary chain disease causality

Source of Sentence	**PMID:24847674**
Text Sentence	Vitamin D deficiency due to malabsorption syndromes (e.g., Crohn's disease, ulcerative colitis, celiac disease, and jejuno-ileal bypass for obesity) may cause osteomalacia
Disease Causality Pairs	Malabsorption Syndromes – due to – Vitamin D Deficiency – cause – Osteomalacia

Table 7 Example of primary bidirectional disease causality

Source of Sentence	**PMID:30017041**
Text Sentence	CONCLUSIONS: A possible bidirectional relationship of psoriasis and major depression exists; i.e. the depression leads to psoriasis, and psoriasis leads to depression
Disease Causality Pairs	Depressive disorder – leads – Psoriasis Psoriasis – causes – Depressive disorder

Table 8 Example of rare disease causality

Source of Sentence	PMID:30665352
Text Sentence	Periodontal disease, including periodontitis, has been reported to be a rare cause of septic pulmonary embolism (SPE)
Disease Causality Pairs	Periodontitis –rare cause – Pulmonary Embolism

and supplementing semantic feature words from other sources, a vocabulary list containing 50 disease causality semantic predicates was constructed to optimize the semantic predicates. The accuracy of these predicates was not less than 40%, including 36 textual predicates with an accuracy of not less than 80% and 42 textual predicates with an accuracy of not less than 60%. The accuracy of using the SemRep tool to extract the disease causality pairs was 85.34%, and after semantic predicate optimization, the accuracy of the three experiments was 96.97%, 96.21%, and 92.43%, which was an improvement of 13.63%, 12.74%, and 8.31%, respectively. It shows that this study can improve the performance of automatic disease causality extraction by screening and optimizing semantic predicates, and achieve a more accurate and comprehensive disease causality extraction. We further used 36 semantic predicates with an accuracy of not less than 80% in the semantic predicate vocabulary list of disease causality to automatically extract 259,434 disease causality pairs from biomedical literature. It was discovered that disease causality pairs had different types, including high-frequency unidirectional disease causality, high-frequency bidirectional disease causality, primary disease causality and rare disease causality.

The method for discovering disease causality from large-scale biomedical literature optimized disease causality semantic predicates through semantic analysis and textual predicate filtering. Compared with the SemRep tool, firstly, it removed and supplemented the normalized disease causality predicates in SemRep according to the accuracy of the predicates, which enhanced the accuracy and comprehensiveness of disease causality extraction to a certain extent. Secondly, the disease causality semantic predicates vocabulary list provided the accuracy of each textual predicate in the evaluation experiment, which enabled the flexibility of semantic predicate selection in terms of the actual demand, and allows for the comprehensive consideration of extracting more accurate or more comprehensive disease causality semantic relations, so as to obtain more desirable results. By discovering highly accurate disease causality pairs from biomedical literature, it is possible to realize the fine-grained revelation of knowledge in biomedical literature, thereby improving the efficiency of researchers in utilizing biomedical literature and facilitating the study of disease attribution.

To achieve a deeper analysis of the performance of the method in this research, we compared our work with two other disease-disease association extraction methods, one is the method based on lexical semantics and document-clause frequency [13], the other is a deep learning method called LC-CNN [10]. Here, a comparative analysis of the three methods from a qualitative and quantitative point of view is carried out, based on their principles, strengths, weaknesses and accuracy, as is shown in the table below (Table 9). The method in this study mainly constructs a more accurate and comprehensive semantic predicate table expressing disease causality to achieve an accurate extraction of disease causality. Lexical semantics and document-clause frequency method defined the concepts of causality term strength based on lexical semantics and the causality frequency based on the number of biomedical publications, along with their strength and directions. LC-CNN is a large margin context-aware convolutional neural network architecture. In terms of the strength and weakness of these methods, our method achieves a more accurate disease causality extraction by optimizing semantic predicates, which provides textual predicates expressing disease causality and the corresponding accuracy, making the disease causality extraction more flexible and contributing to an in-depth understanding of the key features of semantic predicates. However, the proposed method has some limitations for the semantic predicates could hardly be fully enumerated and the following evaluation is time-consuming. Lexical semantics and document-clause frequency based method could determine more causalities, show higher correlation with associated diseases, and provide the strength of causality. However, this method uses only a small amount of literature from PubMed for study, so the results of the disease causality pairs are limited and cover only 195 diseases. The method also suffers from the difficulty of verifying experimental results. LC-CNN integrates context features and convolutional neural networks through the large margin function to achieve more accurate disease-disease association extraction. However, LC-CNN method easily suffers from the symptom/subclass errors (a disease is a symptom/subclass of another disease), negation errors (two diseases are negative relation) and co-occur errors (two diseases co-occur in the sentence but with no association), etc. In term of the method accuracy, our method could achieve an accuracy of 96.97% for disease causality extraction, which is a 13.63% improvement in accuracy over the pre-optimization tool. Lexical semantics and document-clause frequency based method shows higher correlation in disease causality with the

Table 9 Comparison with existing disease-disease association extraction methods

Name	Principle	Advantage	Disadvantage	Accuracy
Method in this study	To achieve an accurate extraction of disease causality by optimizing the disease causality semantic predicate list	Can achieve a more accurate and flexible disease causality extraction	Disease causality extraction is incomplete; the manual extraction of predicates is time-consuming and labour-intensive and the evaluation of predicates is subjective	Achieve an accuracy of 96.97% in disease causality extraction
Method based on lexical semantics and document-clause frequency	Identify disease causality based on the lexicon-based causality term strength and document and clause frequencies in the literature	Reflect the direction and strength of disease causality	The results of disease causality pairs are limited; only 195 diseases are covered and the results are difficult to verify	Show higher correlation with associated diseases with the Spearman's rank correlation coefficient of 0.83
LC-CNN	A neural network-based approach	Achieve more accurate DDA extraction by combining the hinge loss function of SVM with a convolutional neural network into a single neural network architecture	There are symptom/subclass errors, negation errors and co-occur errors	Achieve a Precision measure of 82.36%

Table 10 Comparison with existing disease-disease association extraction tools

Tool	Principle	Advantage	Disadvantage	Disease entity	Relation type	Performance
Method in this study	A disease causality extraction method through the optimization of the disease causality semantic predicate list	Can achieve a more accurate and flexible disease causality extraction	Disease causality extraction is incomplete; the manual extraction of predicates is time-consuming and labour-intensive and the evaluation of predicates is subjective	Obtain 14,335 standardized disease entities	Use 36 textual predicates and extract 58 types of semantic relationships	Achieve an accuracy of 96.97% in disease causality extraction
PubTator	A tool supporting biomedical entities and relations search in the biomedical literature, which involves disease association extraction	Improve the model's ability to generalize to unseen data, has enhanced entity normalization and relation extraction performance	The accuracy of entity annotation and relation extraction remains imperfect, and relations extraction is restricted to abstracts	Involve 12,850 disease entities	Cover 12 types of relations, and two of them express disease relations	F-score is 82%, and demonstrates an overall precision of 90.0%
SicknessMiner	A deep-learning-driven text-mining tool for disease-disease associations extraction	Provide an easy to use text mining pipeline to postulate new relevant DDAs	Limited disease ontologies hinder the comparison and integration of data from different sources	Retrieve 5,443 unique diseases	Retrieve 12,263 co-mentions	Can retrieve 92% of all associations of a DDAs benchmark and still contribute with 16% of new DDAs

Spearman's rank correlation coefficient of 83%, while LC-CNN achieves a precision of 82.36%. The disease causality extraction results of the three methods were further analyzed. In the lexical semantics and document-clause frequency based method, for example, from sentence "AIMS/HYPOTHESIS: Obesity is an independent risk factor for heart diseases but the underlying mechanism is not clear." (PMID: 16612592), it can extract that "Obesity CAUSE heart diseases", which pays more attention to the direction of disease causality, but ignores the predicates that express disease causality. "Risk factor" usually conveys stronger disease causality than "cause". However, our method could be able to extract "Obesity PREDISPOSES heart diseases", which could achieve a more accurate extraction of disease causality. In LC-CNN, for example "Other large-artery aneurysms, including carotid, subclavian, and iliac artery aneurysms, have also been associated with Marfan syndrome." (PMID:23891252), the carotid, subclavian, and iliac artery aneurysms are symptoms of Marfan syndrome, which is not included in the DDA definition. It fails to identify the relation between iliac artery aneurysms and the Marfan syndrome. Luckily, our method can identify the disease and syndrome entity, which recognizes the above two disease entities and their associations, thus achieving an accurate extraction of the disease-disease causality.

We further compared our work with two disease-disease association extraction tools, which are PubTator [21] and SicknessMiner [16]. The comparisons were analyzed from the aspects of principle, advantage, disadvantage, disease entity, relation type and performance, with the details shown in Table 10. For the method in this study, we mainly used predicate optimization to achieve an accurate extraction of disease causality. PubTator is a tool that extracts entities and relationships from biomedical literature, including the extraction of disease relationships. And SicknessMiner is a deep-learning-driven disease association extraction tool. This pipeline encompasses Named Entity Recognition (NER) and Named Entity Normalization (NEN) steps, where SOTA models are used as BioBERT for NER and NormCo for NEN. Each of these tools has its own advantages. The presented method could achieve an accurate and flexible extraction of disease causality by using semantic predicate table. PubTator has an enhanced ability to generalize to unseen data., which makes improvement in entity normalization performance by converting both mentions and lexicon names into high-dimensional TF-IDF vectors and learning a mapping. Moreover, PubTator could achieve more accurate relation extraction by using a data-centric approach to construct a comprehensive, unified training dataset. SicknessMiner provides a comprehensive, highly upgradeable and customizable, easy to use TM pipeline to postulate new relevant DDAs. However, these tools have some limitations. In our method, the disease causality extraction is incomplete and mainly done manually, which is time-consuming and labor-intensive. Also, the evaluation of predicates is subjective. For PubTator, the accuracy of extraction remains imperfect and the extraction is restricted to abstracts. SicknessMiner only uses several eminent disease ontologies in the NEN step, which makes a direct correspondence or mapping impossible, and hinders the integration of data from different sources. Among these tools, our method obtains 14,335 standardized disease entities, extracts 58 types of semantic relationships by using 36 textual predicates. PubTator covers 12,850 disease entities and 12 types of relations, of which two relation types express disease associations ("associate" and "cause"). SicknessMiner retrieves 12,263 co-mentions between 5443 unique diseases. For the performance of these tools, our method achieves an accuracy of 96.97% in disease causality extraction. PubTator demonstrates the F1 score of 82% and an overall precision of 90.0%. SicknessMiner attains a precision of 0.87, recall of 0.89 and F1-score of 0.88 for the NER module, and a precision of 0.80, recall of 0.83 and F1-score of 0.81 for the NEN module. It can retrieve close to 92% of DDAs from a well-established benchmark and still contribute with 16% of new DDAs.

We also compared our work with two disease-disease association extraction datasets, namely dRiskKB [15] and the DDAE dataset extracted from literature [14]. We analyzed these datasets from the following aspects: feature, disease entity, relation pair and performance (Table 11). Regarding the feature of the datasets, our method facilitates the automatic identification of disease causalities from biomedical literature by constructing a disease causality semantic predicate list. dRiskKB uses 21,354,075 MEDLINE records as the text corpus and uses typical disease risk-specific syntactic patterns to automatically extract disease risk pairs. A publicly available DDAE dataset extracted from literature consists of 521 PubMed abstracts, containing positive, negative, and null DDAs, as well as DDA sentences with more complex expressions. In the process of DDAE dataset construction, dependency tree-based relation rules and DNorm are used to annotate disease mentions. Among three datasets, our dataset contains 14,335 standardized disease entities, 6,084 types of bidirectional relations (66,393 SPOs) and 92,557 types of unidirectional relations (17,608 SPOs). dRiskKB covers 12,981 diseasesand 34,448 unique disease relation pairs. A publicly available DDAE dataset extracted from literature contains 12,346 diseases and 3,322 disease-disease pairs. In terms of the performance of these datasets, our method achieves an accuracy of 96.97% in disease causality extraction. For

Table 11 Comparison with existing disease-disease association extraction datasets

Dataset	Feature	Disease entity	Relation pair	Performance
Method in this study	Construct a disease causality semantic predicate list to facilitate the automatic identification of disease causalities	Obtain 14,335 standardized disease entities	Include 6,084 types of bidirectional relations (66,393 SPOs) and 92,557 types of unidirectional relations (17,608 SPOs)	Achieve an accuracy of 96.97% in disease causality extraction
dRiskKB	21,354,075 MEDLINE records comprised the text corpus under study, and use disease risk-specific syntactic pattern to automatically extract disease risk pairs	Cover 12,981 diseases	Consist of 34,448 unique disease relation pairs	The identified patterns have an average precision of 0.99, the exactly matched pairs of 0.919 and the partially matched pairs of 0.988
A publicly available DDAE dataset extracted from literature [10]	Consisting of 521 PubMed abstracts, containing positive, negative, and null DDAs, and dependency tree-based relation rules and DNorm are used to annotate disease mentions	Contain 12,346 diseases	Consist of 3,322 disease-disease pairs	An annotated DDAE dataset with the final kappa value of 76%

dRiskKB, the identified patterns have an average precision of 99% in specifying the risk-specific relationships among diseases, and the precisions of extracted pairs are 91.9% for those that are exactly matched and 98.8% for those that are partially matched. The publicly available DDAE dataset extracted from literature was annotated by biomedical specialists with a final kappa value of 76%.

This method mainly contributes in the following aspects: (1) This study constructed a disease causality semantic predicate vocabulary by accurately filtering and expanding the causal semantic predicates. A more accurate and comprehensive disease causality extraction can be achieved through the optimization of the semantic predicates. (2) Furthermore, a deeper understanding of the predicate feature, as well as the quantified accuracy of each corresponding predicate was made possible by the construction of the semantic predicate vocabulary, which permits flexible disease causality extraction in accordance with practical needs. (3) In this study, we improved the disease causality extraction of SemRep, a popular tool for extracting semantic relationship from biomedical literature. By optimizing the performance of the tool, disease causality can be automatically extracted from extensive biomedical literature. This supports efficient biomedical knowledge discovery, clinical decision-making and other downstream applications.

However, there are several limitations in our study: (1) This study involved the manual filtering and extraction of disease causality semantic predicates, which is a time-consuming and tedious process. (2) Textual predicates in SemRep and existing semantic feature predicates were used as a guide when choosing the disease causality semantic predicates. It was still challenging to achieve a comprehensive extraction of disease causality pairs due to the inadequate extraction of disease causality semantic predicates. (3) The disease causality knowledge units in scientific literature were not always definite, but with a certain degree of uncertainty. Extracting disease causality patterns solely based on the disease causality predicates overlooks the certainty degree in the knowledge units, resulting in the inaccurate extraction of the disease causality semantic relations.

The following enhancements could be adopted to address the aforementioned limitations: (1) To improve efficiency in disease causality semantic predicates extraction, the combination of manually extracted features and machine learning has proven effective. Deep learning is another promising area worthy of further exploration. (2) For a precise and comprehensive identification of disease causality semantic relations, efficient methods for detecting and evaluating disease causality semantic predicates ought to be investigated. Utilizing extensive medical knowledge and establishing standardized rules and criteria facilitates the extraction of disease causality semantic relations from vast biomedical literature. (3) For the accurate inference and discovery of disease causality from biomedical literature, merely focusing on knowledge units is insufficient. It is also imperative to consider the surrounding context in scientific texts. The underlying evidence and data should be correlated with the disease causality knowledge units, and the certainty level of the disease causality semantic relations should be examined. By doing so, the overall quality of disease causality extraction can be enhanced.

Conclusions

In this study, both the accuracy and recall of disease causality extraction have been enhanced by constructing a semantic predicate vocabulary, allowing for automatic disease causality extraction from large-scale biomedical literature. The precision and comprehensiveness of using SemRep tool for disease causality extraction have also been improved through the optimization of semantic predicates, and our approach has the flexible ability to modify the required disease causality strength according to particular needs. Additionally, this approach makes automatic disease causality extraction possible, thereby facilitating knowledge discovery. The idea of this method goes beyond the extraction of disease causality, it can be applied to the extraction of other types of relation pairs, such as treatment. Furthermore, our method helps to identify key features of predicates expressing disease causality, supporting the advancement of machine-learning based knowledge discovery algorithms. The automatic extraction of disease causality aids in uncovering valuable insights hidden in biomedical literature and improve the efficiency of literature exploitation, thus accelerating the process of knowledge transformation and discovery. It could also provide evidence-based data support for clinical diagnosis, disease prevention and control resulting in a far-reaching impact.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02893-0.

Supplementary Material 1.

Acknowledgements

The authors would like to gratefully acknowledge support from the Innovation Fund for Medical Sciences of Chinese Academy of Medical Sciences (grant: 2021-I2M-1-033).

Authors' contributions

S.Y. wrote the manuscript text. P.D. prepared Figs. 2 and 3. X.T. and X.L. provided research ideas. All authors reviewed the manuscript. The first two authors are the primary author. The last two authors are the corresponding author.

Funding
This work was supported by the Innovation Fund for Medical Sciences of Chinese Academy of Medical Sciences (grant: 2021-I2M-1–033).

Data availability
The related dataset are provided in the supplement file. The publicly available dataset of SPOs from biomedical literature using SemRep is accessible in "https://lhncbc.nlm.nih.gov/temp/SemRep_SemMedDB_SKR/SemMedDB_download.html".

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610041, China. [2]Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China. [3]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China.

Received: 29 January 2024 Accepted: 23 January 2025
Published online: 18 March 2025

References

1. Zhao S, Su C, Lu Z, et al. Recent advances in biomedical literature mining. Brief Bioinform. 2021;22(3):bbaa057.
2. Kilicoglu H, Rosemblat G, Fiszman M, et al. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics. 2020;21(1):188.
3. Du J, Li X. A Knowledge graph of combined drug therapies using semantic predications from biomedical literature: algorithm development. JMIR Med Inform. 2020;8(4):e18323.
4. Lee DG, Kim M, Shin H. Inference on chains of disease progression based on disease networks. PLoS ONE. 2019;14(6):e0218871.
5. An N, Xiao Y, Yuan J, et al. Extracting causal relations from the literature with word vector mapping. Comput Biol Med. 2019;115:103524.
6. Lee J, Kim S, Lee S, et al. On the efficacy of per-relation basis performance evaluation for PPI extraction and a high-precision rule-based approach. BMC Med Inform Decis Mak. 2013;13(Suppl 1):S7.
7. Bunescu R, Ge R, Kate RJ, et al. Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med. 2005;33(2):139–55.
8. Kilicoglu H, Shin D, Fiszman M, et al. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics. 2012;28(23):3158–60.
9. Zhang R, Adam TJ, Simon G, et al. Mining biomedical literature to explore interactions between cancer drugs and dietary supplements[A]. In: AMIA Joint Summits on Translational Science proceedings[C]. Bethesda: AMIA; 2015. p. 69–73.
10. Lai PT, Lu WL, Kuo TR, et al. Using a large margin context-aware convolutional neural network to automatically extract disease-disease association from literature: comparative analytic study. JMIR Med Inform. 2019;7(4):e14502.
11. Bang S, Kim JH, Shin H. Causality modeling for directed disease network. Bioinformatics. 2016;32(17):437–44.
12. Zhou XZ, Menche J, Barabási AL, et al. Human symptoms-disease network. Nat Commun. 2014;5:4212.
13. Lee DG, Shin H. Disease causality extraction based on lexical semantics and document-clause frequency from biomedical literature. BMC Med Inform Decis Mak. 2017;17(Suppl 1):53.
14. Kartheeswaran KP, Rayan AXA, Varrieth GT. Enhanced disease-disease association with information enriched disease representation. Math Biosci Eng. 2023;20(5):8892–932.
15. Xu R, Li L, Wang Q. dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text. BMC Bioinformatics. 2014;15(1):105.
16. Rosário-ferreira N, Guimarães V, Costa VS, et al. SicknessMiner: a deep-learning-driven text-mining tool to abridge disease-disease associations. BMC Bioinformatics. 2021;22(1):482.
17. Guan W, Li S, Sun W, et al. Endocrine characteristics and risk factors of type 2 diabetes complicated with gastrointestinal autonomic neuropathy: a single-center retrospective study. Medicine (Baltimore). 2023;102(15):e33467.
18. Gandhi GR, Stalin A, Balakrishna K, et al. Insulin sensitization via partial agonism of PPARγ and glucose uptake through translocation and activation of GLUT4 in PI3K/p-Akt signaling pathway by embelin in type 2 diabetic rats. Biochim Biophys Acta. 2013;1830(1):2243–55.
19. Oparil S, Acelajado MC, Bakris GL, et al. Hypertension. Nat Rev Dis Primers. 2018;4(1):18014.
20. Kabeloğlu V, Senel GB, Karadeniz D. Positive airway pressure normalizes glucose metabolism in obstructive sleep apnea independent of diabetes and obesity. Ideggyogy Sz. 2020;73(11–12):417–25.
21. Wei CH, Allot A, Lai PT, et al. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. ArXiv. 2024.

Publisher's Note