BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Construction of a knowledge graph for breast cancer diagnosis based on Chinese electronic medical records: development and usability study

Xiaolong Li[1,2], Shuifa Sun[1,3], Tinglong Tang[1,3], Ji Lu[4], Lijuan Zhang[5], Jie Yin[5], Qian Geng[6] and Yirong Wu[6*]

Abstract

Background Electronic medical records (EMRs) contain a wealth of information related to breast cancer diagnosis and treatment. Extracting relevant features from these medical records and constructing a knowledge graph can significantly contribute to an efficient data analysis and decision support system for breast cancer diagnosis.

Methods An approach was proposed to develop a workflow for effectively extracting breast cancer-related features from Chinese breast cancer mammography reports and constructing a knowledge graph for breast cancer diagnosis. Firstly, the concept layer of the knowledge graph for breast cancer diagnosis was constructed based on breast cancer diagnosis and treatment guidelines, along with insights from clinical experts. .Next, a BiLSTM-Highway-CRF model was designed to extract the mammography features, which formed the data layer of the knowledge graph. Finally, the knowledge graph was constructed by combining the concept layer and the data layer in a Neo4j graph data platform, and then applied in visualization analysis, semantic query and computer assisted diagnosis.

Results Mammographic features were extracted from a total of 1171 mammography examination reports. The overall extraction performance of the model achieved an accuracy rate of 97.16%, a recall rate of 98.06%, and a F1 score of 97.61%. Additionally, 47,660 relationships between entities were identified based on the four different types of relationships defined in the concept layer. The knowledge graph for breast cancer diagnosis was constructed after inputting mammographic features and relationships into the Neo4j graph data platform. The model was assessed from the concept layer, data layer, and application layer perspectives, and showed promising results.

Conclusions The proposed workflow is applicable for constructing knowledge graphs for breast cancer diagnosis based on Chinese EMRs. This study serves as a reference for the rapid design, construction, and application of knowledge graphs for diagnosis and treatment of other diseases. Furthermore, it offers a potential solution to address the issues of limited data sharing and format inconsistencies present in Chinese EMR data.

Keywords Chinese electronic medical records, Breast cancer, Knowledge graph, Mammography, Computer assisted diagnosis

*Correspondence:
Yirong Wu
yrwu@bnu.edu.cn

Full list of author information is available at the end of the article

Background

Breast cancer poses a significant threat to women's health worldwide and is one of the most prevalent malignant tumors among women. Currently, the diagnosis and treatment of breast cancer heavily rely on experienced clinical experts. However, the scarcity of clinical experts and their overwhelming workload severely impact timely diagnosis and treatment for patients. Although certain deep learning algorithms can be used to detect breast cancer to some extent, their lack of interpretability remains a major technical barrier, preventing their practical application in clinical settings [1]. Advances in science and technology have significantly improved the accuracy of breast cancer diagnosis and treatments. Consequently, a vast amount of breast cancer examination data has been generated and stored in the Electronic Medical Record (EMR) system, containing valuable information about the disease. Extracted variables from EMRs can contribute to patients' condition management [2] and aid in breast cancer risk prediction [3]. Moreover, these variables can serve as knowledge-driven tools, enhancing diagnosis efficiency and interpretability, ultimately reducing errors and relieving clinician workloads [4]. Therefore, there is an urgent need to design an effective data processing and management system to help collect, manage, and use variables for breast cancer diagnosis and treatments.

Knowledge graph (KG) is one of the visualization technologies in the knowledge domain. It is composed of entities, semantic types, properties, and relationships [5]. In medical domains, research on the construction of knowledge graph has achieved many successes [6, 7]. However, most of the existing knowledge graphs in medical domains are based on data from medical literature, online community resources, or various open databases that are publicly available. Although this type of data is more convenient to obtain, it is not real-world EMR data [8]. Additionally, the quality of knowledge graphs does not meet the needs of disease diagnosis or treatment. Particularly, the Chinese EMR data has poor sharing characteristics, diverse storage approaches, different formats and standards, etc., making it difficult to build a knowledge graph [9]. Therefore, it is necessary to develop a workflow for breast cancer KG construction using Chinese EMR data.

For studies on breast cancer knowledge graph construction, An [10] proposed a method to extract features about breast cancer from EMRs to construct the knowledge graph for breast cancer diagnosis. The design of the concept layer and the data layer was specifically introduced. However, a description of the procedure used for knowledge graph construction is missing, which makes it impossible to check the effect of the constructed knowledge graph. Hasan et al. [11, 12] proposed a method to construct a knowledge graph from cancer registration data and developed a prototype of the knowledge graph for breast cancer patient management. However, information contained in the knowledge graph is very limited for breast cancer diagnosis. Jin et al. [13] proposed a method to construct knowledge graphs for breast cancer diagnosis and treatments. Nevertheless, a description of the construction procedure is missing and the construction results are not reported, affecting the evaluation of the construction method. There are some issues with existing methods for constructing breast cancer knowledge graphs from Chinese EMRs. Firstly, existing methods exhibit limited coverage in building breast cancer knowledge graphs, failing to encompass a comprehensive range of knowledge. Secondly, while some existing methods can be used to construct knowledge graphs, most of them lack sufficient detail and clarity when the construction process and methodology are described. Thirdly, these methods have not been validated through clinical applications.

Commonly used examination methods for breast cancer include clinical palpation, mammography, ultrasound, and nuclear magnetic resonance. Mammography is currently one of the most widely used diagnostic method for breast cancer. Using breast cancer mammography examination reports as research samples, case studies can be conducted on knowledge extraction, knowledge graph construction, and applications, which can help obtain standardized and high-quality examination results to facilitate effective disease management. Additionally, breast cancer risk prediction can be performed to assist in breast cancer diagnostic decision-making. However, there is still a scarcity of research in this domain. Therefore, it is necessary to develop a comprehensive workflow, from knowledge extraction from Chinese EMRs to knowledge graph construction, and even diagnosis applications.

Our study intends to utilize Named Entity Recognition (NER) technology to effectively identify and extract features from Chinese breast cancer mammography reports and construct a knowledge graph for breast cancer diagnosis. The contributions of this study are as follows:

- A workflow covering the design of the concept layer, feature extraction from Chinese EMRs, the construction of a knowledge graph, and a demonstration of its applications is proposed. Based on the workflow, a top-down knowledge graph for breast cancer diagnosis is constructed.
- A deep learning model is developed, the BiLSTM-Highway-CRF network, which achieves higher extraction performance compared to traditional models for feature extraction from Chinese EMRs.
- The constructed knowledge graph is utilized to realize visual analysis, semantic query, and computer-assisted diagnosis, effectively

demonstrating its usefulness and practicality in clinical applications.

Methods

The approval was obtained from the hospital's ethics review committee for this study. There are 2989 mammography examination reports collected from the Radiology department of a Three-A hospital located in Yichang province of China, spanning from December 2018 to July 2019. The data were then subjected to de-identification processing. Each report was prepared by one doctor and verified by another doctor. A total of nine doctors participated in this work. After duplicated reports were removed and reports with incomplete or incorrect data were deleted, the final dataset consists of 1171 reports used to construct the knowledge graph for breast cancer diagnosis.

The knowledge graph construction process is shown in Fig. 1, which mainly includes four steps: (1) the design of the concept layer. (2) the development of the data layer, which is composed of data annotation, feature extraction, and knowledge fusion. (3) knowledge graph construction. (4) knowledge graph applications. The details of each step are described in the following subsections.

Design of the concept layer

A knowledge graph is composed of a concept layer (or schema layer) and a data layer (or instance layer) [14], as shown in Fig. 2. The concept layer is the core of the knowledge graph, including a hierarchical structure of entities and their attributes, which can be used to constrain data storage in the data layer [15].

In the design of the concept layer, we leveraged the Breast Imaging Reporting and Data System (BI-RADS) lexicon of the American College of Radiology [16], and the Breast Cancer Diagnosis and Treatment Guidelines issued by the Chinese Anti-Cancer Association [17]. Additionally, we engaged four mammography radiologists from different hospitals to participate in ontology definition and concept layer framework design. Moreover, we extensively referenced domestic and international literature, as well as mammography examination norms and standards, to revise the concept layer .

The concept layer is designed as a 3-level hierarchical structure with 15 types of mammography features

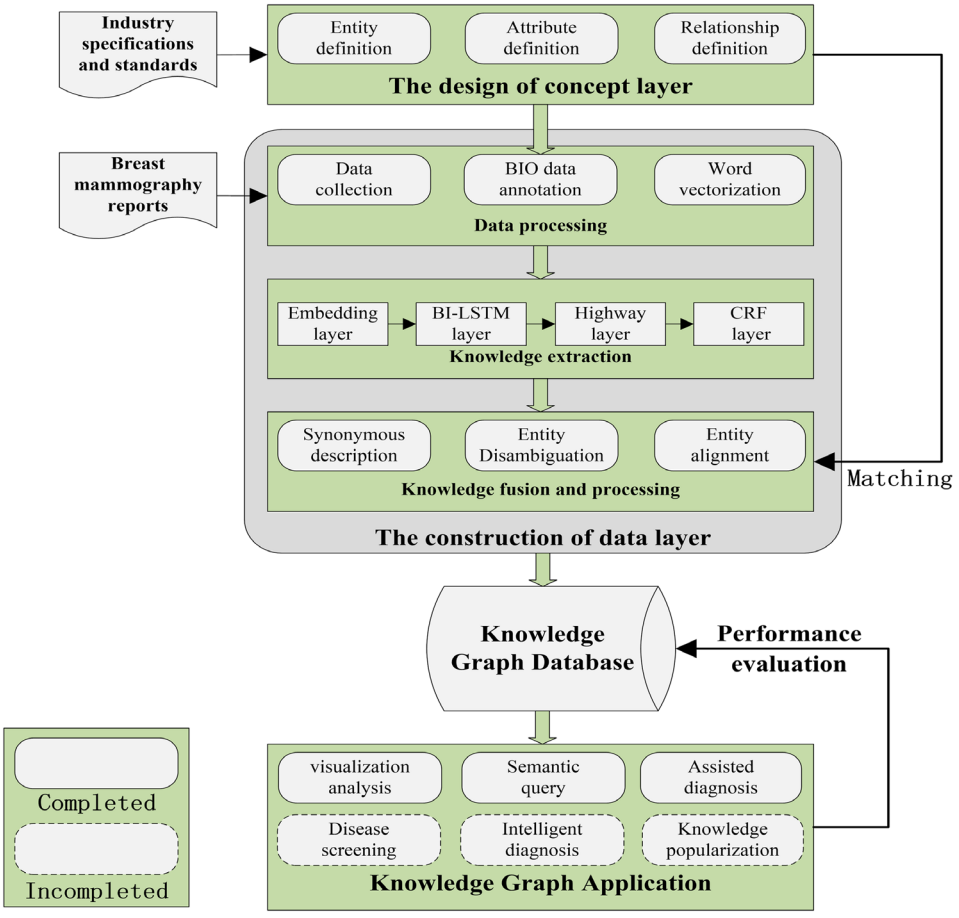

Fig. 1 The workflow of constructing a knowledge graph for breast cancer diagnosis based on mammography examination reports

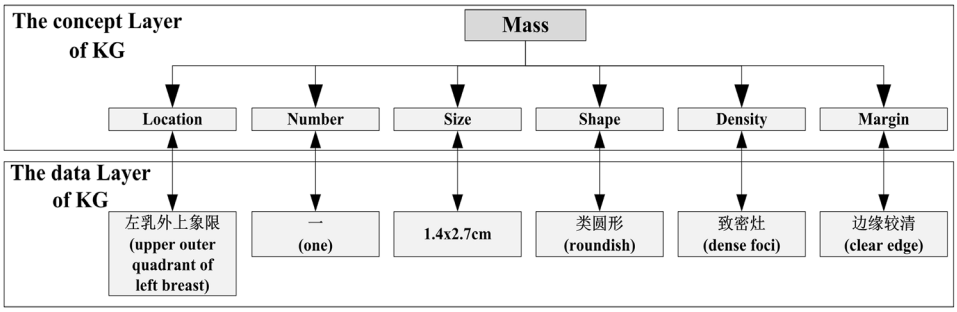

Fig. 2 The relationship between the concept layer and the data layer of a knowledge graph

Table 1 The entities and attributes related to mammography features in the concept layer

Project	First-level entity	Second-level entity	Entity attributes
Breast cancer mammography	Common signs	Mass	Location, Margin, Number, Shape, Type, Size, Density, Negation,
		Calcification	Location, Number, Shape, Type, Size, Density, Negation, Distribute
		Structure	Describe
	Special signs	Duct change	Describe
		Intramammary lymph nodes	Location, Number, Shape, Negation, Describe
		Asymmetrical	Location, Describe
	Merge signs	Vascular change	Location, Shape, Negation,
		Axillary lymph nodes	Location, Number, Shape, Negation, Describe
		Association anomaly	Location, Number, Negation, Describe
	Category descriptions		Location, Describe

primarily(Calcification, Density, Distribute, Location, Mass, Lymph Node, Margin, Merge, Number, Shape, Size, Special, Structure, Category descriptions, Negation). There are 4 entities such as "common signs" at the first level, there are 9 entities such as "Mass" at the second level, and related entity attributes are at the third level (Table 1). According to the three-element principle of knowledge graph construction [18], it is necessary to clarify the three elements of entity-relationship-entity attribute or entity-relationship-entity. Guided by clinicians, we established relationships among entities and attributes. This study not only defined a set of entities and their attributes, but also established their hierarchical relationships, Table 2 shows the three elements of subject, predicate and object among different entities and attributes in the concept layer. Once entities, entity attributes, and relationships were specified, the design of the concept layer of the knowledge graph for breast cancer diagnosis was completed. In the following context, the term "entity" refers to mammography examination variable such as "mass" and "calcification", while "entity attribute" refers to attribute variables such as "size" and "shape" owned by "entity". Both examination variables and attribute variables belong to BI-RADS variables.

Table 2 Triple-structure of the subject, predicate, and object in the concept layer

	subject	predicate	object
1	Entity (Basic information)	select	Attributes (Name, Sex, etc.)
2	Entity (Report)	has_a	Entity (Basic information, Mammography information, BI-RADS information)
3	Entity (Mammography information)	instance_of	First-level entity (Common signs, Merge signs, etc.)
4	First-level entity (Common signs, Merge signs, etc.)	part_of	Second-level entity (Mass, Calcification, Structure, etc.)
5	Second-level entity (Structure, Mass, etc.)	select	Attributes (Location, Size, etc.)

Development of the data layer

Data annotation

In this study, entity type annotation was performed on mammography examination reports for developing entity recognition models. The annotation step was carried out using the web-based labeling tool BRAT [19]. Then, the annotation text was prepared in the BIO format

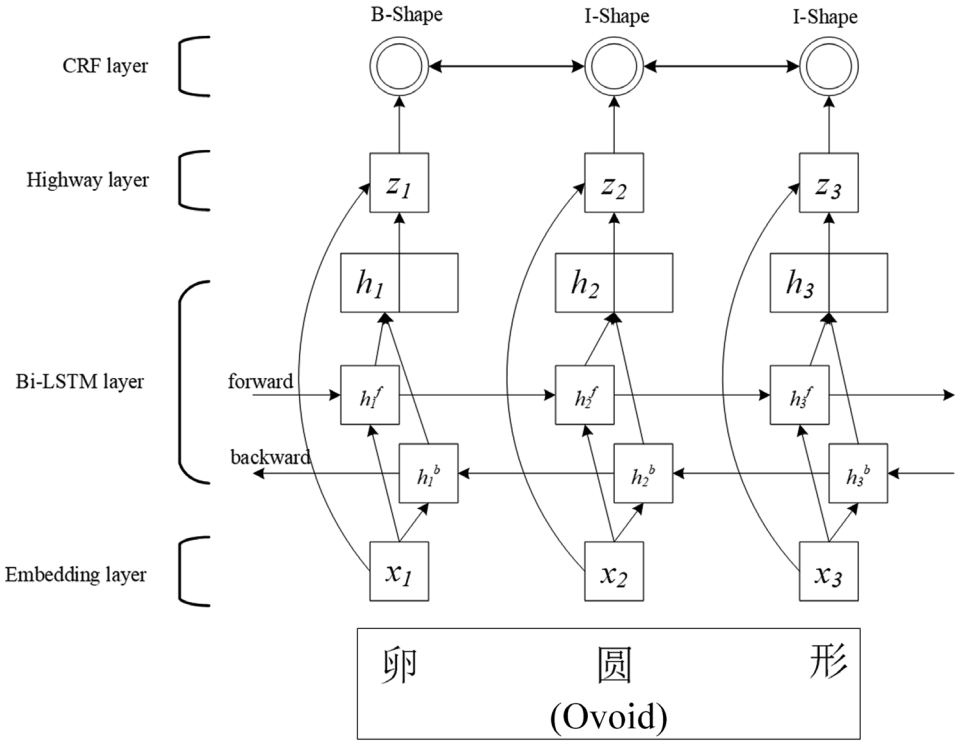

Fig. 3 BiLSTM-Highway-CRF model

("B" indicates the start tag of an entity, "I" indicates the tag inside an entity, and "O" indicates the tag outside an entity) [20], commonly used in NER tasks. Finally, the Word2Vec model was used for word vectorization [21].

Feature extraction

There are three types of NER methods that can be used to extract BI-RADS variables from EMRs: rule-based methods [22, 23], machine learning-based methods [24–26], and deep learning-based methods [27–29]. The first type of methods pays more attention to the doctor's clinical experience, while the second and third type of methods rely more on data to train the models. Rule-based NER methods identify entities based primarily on hand-crafted semantic and syntactic rules, and have high development costs in terms of time and efforts from domain experts. Therefore, there exist the disadvantages such as poor portability, the problems of expanding to other entity types or datasets, and the difficulty of migrating to other fields [30]. Although machine learning-based methods have overcome some of these shortcomings, with the increase of data volume, efficiency and effectiveness may become issues. On the other hand, deep learning-based methods are suitable for nonlinear transformation, and they can handle the linguistic and structural variability of free text more effectively, which can be used to build a more complex network [22].

The BiLSTM-CRF model is currently one of the most popular NER models [31]. In this study, a BiLSTM-Highway-CRF model was constructed, comprising a BiLSTM module, a Highway network module, and a CRF module. The Highway network defines two gate structures, transfer gate $\mathrm{T}(\mathrm{W_T}, \mathrm{x})$ and carry gate $\mathrm{C}(\mathrm{W_C}, \mathrm{x})$, to control information flow. The transfer gate controls the amount of information transmitted by the feed forward network, while the carry gate controls the amount of information transmitted by the input x. Its mathematical expression is defined in Eq. (1), the highway network also sets the relationship between the transfer gate T and the carry gate C, that is, the above mathematical expression can be simplified as Eq. (2). The design of the relationship between switching gates makes the Highway network more flexible.

$$y = \mathrm{H}\left(\mathrm{W_H}, \mathrm{x}\right) * \mathrm{T}\left(\mathrm{W_T}, \mathrm{x}\right) + \mathrm{x}^{*}\mathrm{C}\left(\mathrm{W_C}, \mathrm{x}\right) \quad (1)$$

$$y = \mathrm{H}\left(\mathrm{W_H}, \mathrm{x}\right) * \mathrm{T}\left(\mathrm{W_T}, \mathrm{x}\right) + \mathrm{x}^{*}\left(1 - \mathrm{T}\left(\mathrm{W_T}, \mathrm{x}\right)\right) \quad (2)$$

Based on the idea of the Highway network [32], through comparing the feature information obtained by the training of the BiLSTM network with the original word vector information, the retention ratio of the trained feature information and the word vector information can be optimized, so as to reduce the degradation impacts of the deep neural network on recognition performance.

Figure 3 illustrates the structure of the BiLSTM-Highway-CRF model. "卵圆形"(Ovoid) is an input Chinese word. B-Shape, I-Shape, and I-Shape correspond to their recognition results.

In this study, the recognition performance of the model was evaluated using commonly used metrics in NER tasks, which include precision rate P, recall rate R, and $F1$ score. P represents the proportion of real positive samples among the entities recognized by the model to be positive. R represents the proportion of all positive samples that are correctly recognized. The $F1$ score is the harmonic average of precision and recall. The score of $F1$ is 1 (perfect precision and recall) to reach the best value, and 0 to reach the worst value.

Knowledge fusion and processing

During the process of extracting entities, we observed that the difference between clinician's reporting habits and hospital operation guidelines results in different qualities of mammography reports. The same entity may be inconsistent in the context of different mammography reports. As a result, the extracted entities will be different. Based on breast cancer diagnosis and treatment guidelines, regulations, and industry standards [16, 17], we performed text similarity analysis [33] on the extracted mammography features of the same type to achieve the effects of synonymous descriptions. Additionally, we performed standard semantic replacements for words having semantic errors and synonymous descriptions to attain the goal of entity disambiguation. In the process of coreference resolution, we supplemented and aligned features to ensure the consistency and accuracy of the extracted mammography features.

Knowledge graph construction and application

There are two main approaches for knowledge graph construction: bottom-up and top-down [34]. The bottom-up approach involves obtaining entities and frameworks directly from the data. However, due to its automated nature, the knowledge obtained through this approach may lack completeness and its quality may not reach a practical level. On the other hand, the top-down approach is mainly used to construct domain-specific knowledge graphs. For this approach, the construction process begins with domain experts and clinical professionals who define the entities and frameworks at the concept layer based on the needs of breast cancer diagnosis. The data is then filled into the knowledge graph from heterogeneous data sources.

To ensure that the knowledge graph is effectively utilized for clinical applications, especially in assisting diagnosis, this study adopted the top-down approach for knowledge graph construction. At first, the entities and frameworks at the concept layer were specified, and then data annotation was carried out with the assistance of clinical experts and doctors. Next, deep learning models were used for knowledge extraction from the annotated data. The extracted knowledge was then subjected to knowledge fusion, which involved standardizing the entities and entity attributes in the data layer. Afterward, the fused data layer was processed and matched with the concept layer, establishing relationships between entities and between entities and attributes to develop a large-scale hierarchical graph system. Finally, the knowledge graph construction was completed.

Due to the top-down method used to construct knowledge graph, all entities, entity attributes and relationships were defined in the concept layer [34]. Relying on the structure of the concept layer, the entities and entity attributes extracted from mammography reports were matched with the entities and attributes of the concept layer, so as to construct a complete knowledge graph. In order to facilitate the top-down application research, this study used Neo4j graph data platform (hereinafter referred to as Neo4j) to store knowledge in the form of graphic database. The high quality of knowledge graphs is required in the medical domain [35]. In order to verify the practicability and rationality of the knowledge graph, this study carried out experiments such as visual analysis, semantic query and computer-aided diagnosis using the knowledge graph.

Results

Annotation results

After the 1171 mammography examination reports were annotated with those 15 types of mammography features, the number of entities in each type was obtained (Table 3). Four breast cancer mammography examination doctors were invited to participate in and guide data annotation. After data annotation was completed, clinical doctors conducted multiple rounds of sampling and corrected any incorrectly annotated data. In the final round of sampling, 951 labeled data points from 50 data samples were inspected, and the accuracy of the annotated data reached 99.26%.

The results of feature extraction

For 1171 text reports, we used the featured extracted from 820 reports in the training set for model training, the features from 116 reports in the validation set for model parameter adjustment, and the features from 235 reports in the test set for evaluating the performance of the model.

The results of the BiLSTM-Highway-CRF model for each mammography feature type are described in Table 4. To compare the model performance, we also developed three widely used NER models: Hidden Markov Model (HMM), Conditional Random Field model (CRF), and

Table 3 The type and number of its entities annotated in this study

Entity type	Definition	Number of entities
Calcification	Calcification	1824
Density	Mass density, Tissue density	630
Distribute	Calcification distribution	144
Location	Description area, location information, like: upper limit on the left	6038
Mass	Mass information	1169
Lymph Node	Lymph Node information	1163
Margin	Mass boundary information	507
Merge	Merge signs. Sunken skin, Syndromes, such as thickening	1154
Number	The number of mass or calcification	1082
Shape	The shape of a mass or calcification	1270
Size	The size of mass or calcification	562
Special	Special signs, like: Vascular thickening	1217
Structure	Normal or distorted structure	1934
Category descriptions	The typing features about breast densities	1162
Negation	Negative Words	3288

Table 4 Performance of BiLSTM-Highway-CRF model for each entity type(unit: %)

Entity Type	Precision	Recall	F1
Calcifications	99.71	100.00	99.86
Density	91.59	93.33	92.45
Distribute	93.02	93.02	93.02
Location	97.42	97.89	97.65
Mass	98.71	99.57	99.13
Lymph Node	98.30	98.72	98.51
Margin	89.57	94.81	92.11
Merge	94.14	96.98	95.54
Number	97.96	97.56	97.76
Shape	93.57	96.04	94.79
Size	97.56	98.77	98.16
Special	97.47	97.88	97.67
Structure	98.16	98.42	98.29
Category descriptions	98.28	99.57	98.92
Negation	99.84	100.00	99.92

Table 5 Overall feature extraction performance(unit: %)

Model	Precision	Recall	F1
HMM	93.42	94.91	94.17
CRF	97.00	95.84	96.42
BiLSTM-CRF	96.71	97.20	96.96
BiLSTM-Highway-CRF	**97.16**	**98.06**	**97.61**

BiLSTM-CRF. The results of the precision, recall, and F1 of those three NER models and the BiLSTM-Highway-CRF model are shown in Table 5. For the BiLSTM-Highway-CRF model, the precision rate is 97.16%, the recall rate is 98.06%, and F1 is 97.61. The experimental results show that the performance of the CRF model is higher than that of the HMM model in NER task, indicating that considering the proximity label information is essential for predicting the current label when extracting BI-RADS features from mammography examination reports. Additionally, the BiLSTM-Highway-CRF model demonstrates a higher performance than the CRF and BiLSTM-CRF models, which indicates that the introduction of the highway network mechanism is beneficial for the model to learn features in mammography examination reports.

The results of relationship matching

There are a total of 47,660 relationships between entities and between entities and attributes in 1,171 reports. Among them, there are 3,513 "*has_a*" relationships, 4,684 "*instance_of*" relationships, 17,565 "*part_of*" relationships and 21,898 "*select*" relationships.

The applications of knowledge graph

Visualization analysis

The knowledge graph for breast cancer diagnosis is constructed by combining the concept layer and the data layer. Additionally, the demographic patient information and BI-RADS categories representing the results of breast cancer diagnosis are integrated into the knowledge graph (Fig. 4). In the knowledge graph, nodes represent entities or entity attributes, and edges represents the relationships between entities or between entities and their attributes. This visualization makes it convenient for patients and clinicians to view and analyze patient situations.

Semantic query

To facilitate clinicians in accessing and analyzing data, we have customized several advanced query statements using the Neo4j graph data platform. For example, based on common queries required by clinicians, we have predefined template statements with parameters. By entering specific queries, such as "patient information between 46 and 50 and BI-RADS level 3," clinicians can retrieve the corresponding patient information (Fig. 5). Relying on the knowledge graph, clinicians can complete simple query, combined query, conditional query, path query and even deep relationship query of patient information without a need of learning a lot of non-medical professional knowledge.

Computer assisted diagnosis

Breast cancer diagnosis involves numerous mammography features, making it essential to identify the most critical risk factors used for accurate diagnosis. Through analysis of the knowledge graph, we have discovered that for some patients who have the mass that is an irregular shape with spiculated margins, their examination result indicate BI-RADS category of 4 C (Fig. 6). A BI-RADS

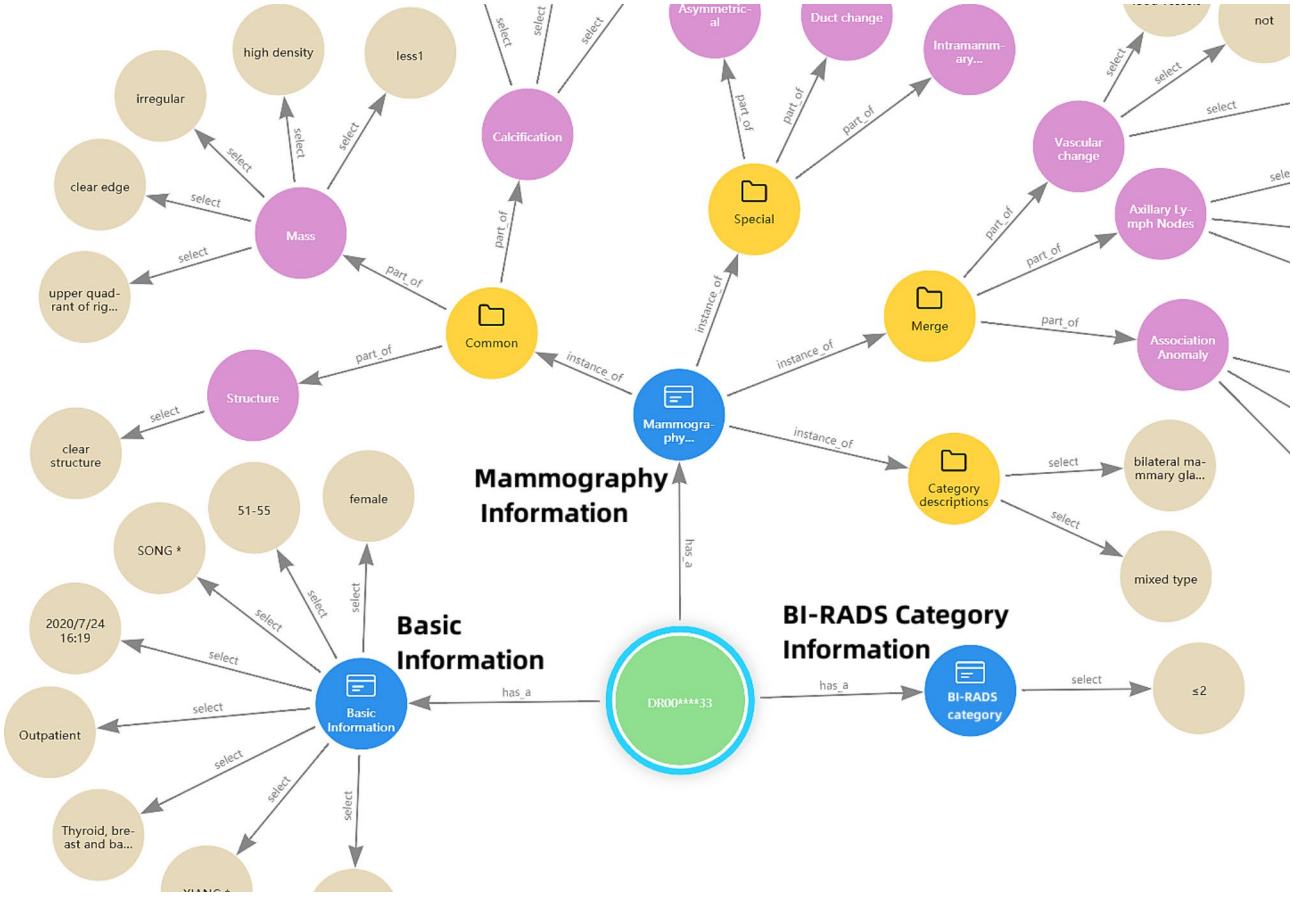

Fig. 4 An example of the knowledge graph(part)

category of 4 C indicates a highly suspicion of malignancy, and a blurred margin is considered one of the highest risk factors. Clinical evaluation has confirmed the accuracy of this conclusion. By extension, we can assess the impact of individual other features or combinations of features on the BI-RADS classification.

Furthermore, the knowledge graph enables us to assess the impact of individual features or combinations of features on the BI-RADS classification. By leveraging the information from the knowledge graph, clinicians can gain valuable insights into the diagnostic process and utilize it to assist in the diagnosis of breast cancer more effectively. Overall, the knowledge graph constructed in this study has the potential to play a vital role in improving breast cancer diagnosis.

Quality evaluation

To assess the quality of experimental data, model performance, and application effectiveness, we invited eight clinical experts and natural language processing experts to conduct testing, analysis, and data quality assessment at the concept layer, data layer, and application layer (as shown in Table 6). The specific evaluations are as follows:

In the concept layer design, the ontology and concept layer design of the breast cancer mammography diagnosis knowledge graph were completed (as shown in Tables 1 and 2). The design includes 4 first-level entities, 9 s-level entities, and 39 entity attributes, encompassing common BI-RADS variables used in breast cancer mammography examinations from both domestic and international sources. This demonstrates the accuracy, rationality, completeness, and standardization of the knowledge graph framework. However, the top-down construction approach places high demands on the design of the concept layer. It is recommended to further standardize and refine the knowledge graph concept layer framework to ensure its generality and effectiveness, thus ensuring that the knowledge graph complies with clinical standards and meets clinical application requirements.

In the data layer design, the BiLSTM-Highway-CRF deep learning model for knowledge extraction significantly improved the precision, recall, and F1 score compared to the HMM, CRF, and BiLSTM-CRF models, achieving 97.16%, 98.06%, and 97.61%, respectively. This demonstrates the accuracy and precision of data extraction. However, since the model data all originate from the same hospital, its accuracy on data from other sources

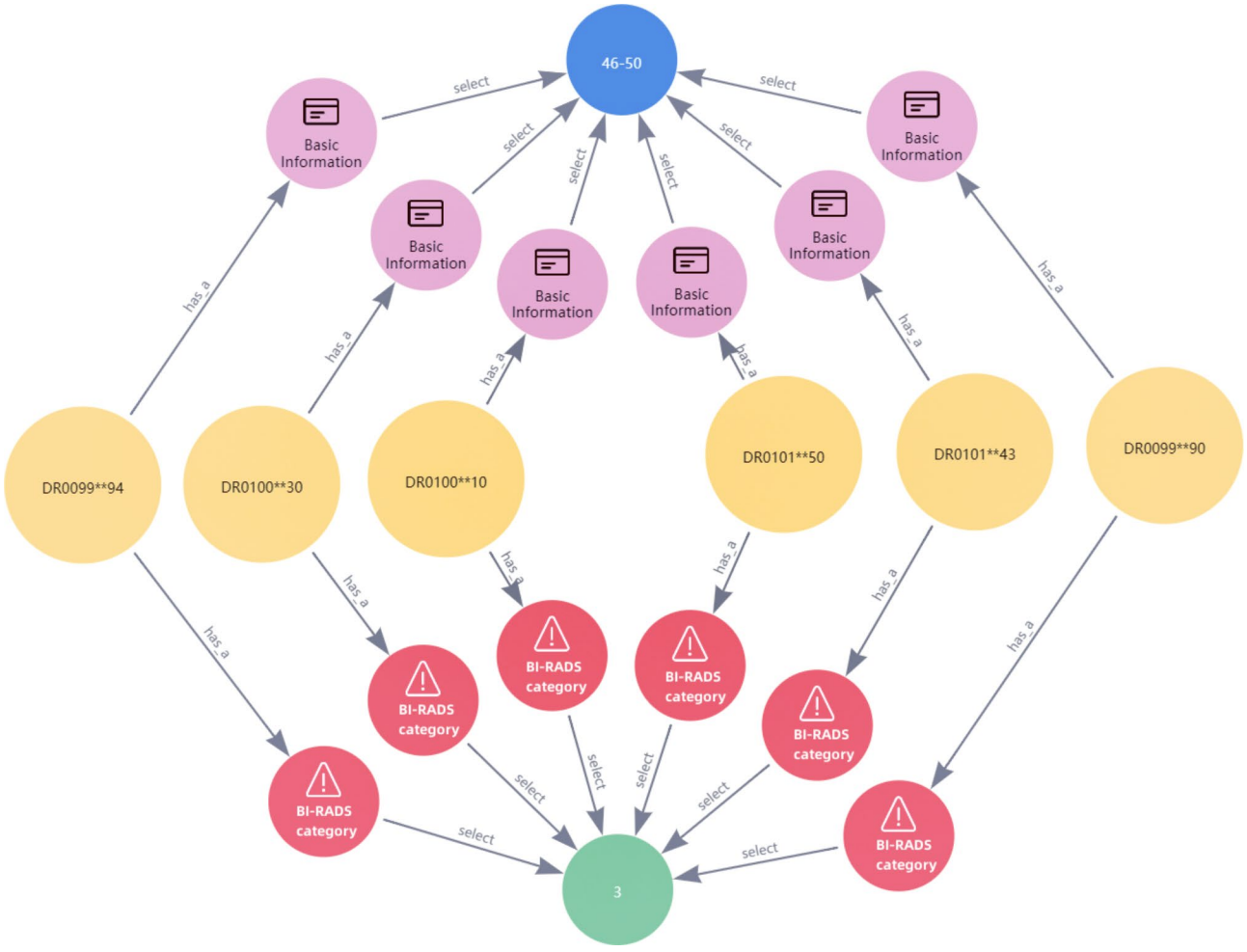

Fig. 5 One example of customizing the query statements

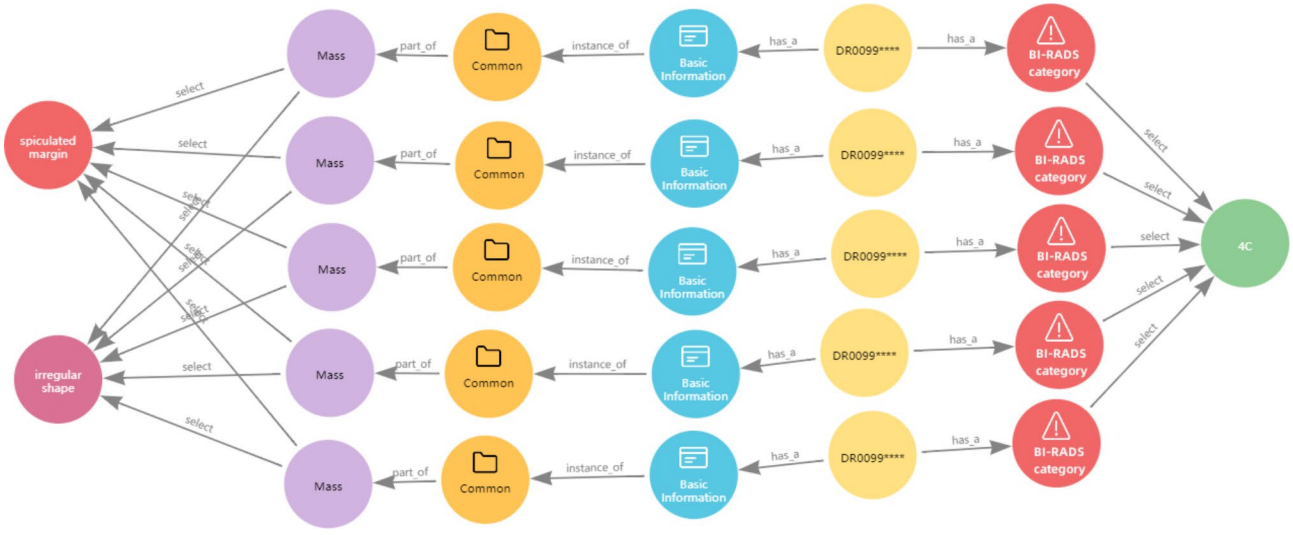

Fig. 6 The relationship between the feature "irregular shape and spiculated margins" and BI-RADS category "4 C" in the knowledge graph

Table 6 The indexes and results of effective evaluation

Knowledge graph	Test and Analysis Content	Data Quality Assurance and Evaluation	Suggestion
the concept layer	Ontology and attribute definition, relationship construction, concept layer framework	**Positive aspects**: Accuracy, rationality, completeness, standardization **Areas for improvement**: Generality, effectiveness	Further standardize and refine the concept layer framework of the knowledge graph to ensure its generality and effectiveness.
the data layer	Data preprocessing, knowledge extraction, knowledge fusion, knowledge processing	**Positive aspects**: Precision, accuracy, recall, F1 **Areas for improvement**: Robustness, portability	Conduct additional experiments on multi-source data to optimize the model and enhance its robustness and generalization capabilities.
the application layer	Visualization analysis, data queries, assisted diagnosis	**Positive aspects**: Visualization, structured presentation, accuracy, usability, user-friendliness **Areas for improvement**: Completeness, operability, aesthetics, practicality	Integrate other feature knowledge graphs (such as more detailed basic information, other breast cancer examination data, breast cancer malignancy information), and suggest utilizing machine learning or deep learning for knowledge inference to improve the clinical application of the model.

cannot be guaranteed. Further experiments on multi-source data are recommended to optimize the model and enhance its robustness and generalization capabilities.

In the application layer, the knowledge graph is completed, which can be used to effectively present patient information, as indicated by visualization analysis and data queries. (as shown in Fig. 4). The retrieval results (as shown in Fig. 5) demonstrate accuracy, comprehensiveness, and usability of the knowledge graph. However, there is room for improvement in terms of user-interface. Additionally, the knowledge graph can provide a theoretical basis and practical foundation for assisted diagnosis of breast tumor malignancy (as shown in Fig. 6). It is recommended to integrate other knowledge graphs as well as machine learning or deep learning algorithms for knowledge inference to expedite the clinical application of the knowledge graph.

Discussion

Currently, the majority of widely used knowledge graphs in the medical domains are constructed using the bottom-up approach [36], where the data is generally obtained from medical literature, online community resources, or various open databases. Although this kind of data is easier to obtain, it does not represent real-world EMR data. Additionally, the quality of knowledge graphs constructed using this approach may not meet the requirements for disease diagnosis or treatment [37]. To address these challenges, we have not only designed the workflow which utilizes the top-down approach for knowledge graph construction, incorporating real-world EMR data, specific domain knowledge, and clinician's experience, but we have also successfully constructed the knowledge graph for breast cancer diagnosis following this approach.

In this study, we have developed the BiLSTM-Highway-CRF model to effectively extract mammography features from mammography reports. The performance of the BiLSTM-Highway-CRF model outperforms that of both the CRF model and the BiLSTM-CRF model, indicating that the incorporation of the highway network mechanism is advantageous for the model to learn features in examination reports more accurately and comprehensively. The extracted BI-RADS variables from breast cancer mammography reports are then integrated and organized within the knowledge graph specifically designed for breast cancer diagnosis. By leveraging the graphic structure of the knowledge graph, we have established an efficient data visualization and management approach. Overall, the successful development and application of the BiLSTM-Highway-CRF model and the knowledge graph demonstrate the potential of these techniques in advancing breast cancer diagnosis and management.

Conclusions

The main contribution of this study lies in the development of a comprehensive workflow that encompasses the extraction of Chinese EMR variables and the application of knowledge graphs, providing guidance for the construction and utilizing medical knowledge graphs in disease diagnosis and treatment. In this study, we design the concept layer of the knowledge graph with reference to BI-RADS standards and several breast cancer diagnosis and treatment guidelines. Based on the deep learning NLP methods, mammographic features are extracted from examination reports and imported into the Neo4j graph data platform. Leveraging the design of the concept layer, we develop the knowledge graph for breast cancer diagnosis. Through the evaluation of the design of the concept layer, the construction of the data layer, and the functions of the application layer, the rationality, effectiveness, and practicability of the knowledge graph are demonstrated.

In conclusion, this study provides a valuable workflow that serves as a guide for designing, constructing, and

applying knowledge graphs in the diagnosis and treatment of breast cancer. Moreover, it offers insights for designing and constructing knowledge graphs for other disease diagnosis and treatment scenarios. To a certain extent, it contributes to addressing issues related to poor data sharing and format inconsistencies in Chinese EMR data. For future research, we aim to emphasize the construction of the concept layer in the knowledge graph to enhance its effectiveness and generalizability. Additionally, we will utilize larger datasets from multiple hospitals to further enrich and develop the knowledge graph, thereby improving the model's robustness. Lastly, ethical considerations will be considered when the knowledge graph is implemented in clinical disease diagnosis applications.

Abbreviations

NLP	Natural Language Processing
EMR	Electronic Medical Record
KG	Knowledge Graph
BI-RADS	Breast Imaging-Reporting And Data System
CRF	Conditional Random Fields
BiLSTM	Bi-directional Long Short-Term Memory
RDF	Resource Description Framework

Acknowledgements
Not applicable.

Authors' contributions
WYR and SSF conceived the study and developed algorithm. LXL and WYR designed experimental and result analysis. LXL, LJ and ZLJ collected and preprocessed the data. SSF, TTL, LJ and GQ supervised the work and contributed to the study design and to the redaction of the manuscript. LXL, SSF and WYR contributed to the modification of the manuscript. All authors have read and approved the final manuscript.

Funding
This work was supported by the National Social Science Foundation of China (20BTQ066). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Social Science Foundation of China. The funding body did not play any role in the design of the study, collection, analysis, and interpretation of the data, or writing the manuscript.

Data availability
This research made use of the data of EMR from Yichang Central People's Hospital. The data can be made available upon reasonable request from the corresponding author. The data are not publicly available due to privacy restrictions. The source codes of this research are available at, https://github.com/hbxiaolong/NER-KG.

Declarations

Ethics approval and consent to participate
The study was approved by the Medical Science Research Ethics Committee of Yichang Central People's Hospital (Serial No. 2022-082-01). Informed consent was obtained from all subjects and their legal guardian(s). All methods were performed in accordance with the relevant guidelines and regulation.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Yichang Key Laboratory of Intelligent Medicine, Yichang, Hubei, China
[2]College of Economics and Management, China Three Gorges University, Yichang, Hubei, China
[3]College of Computer and Information Technology, China Three Gorges University, Yichang, Hubei, China
[4]Yichang Central People's Hospital, Yichang, Hubei, China
[5]The Fourth Affiliated Hospital of Nanjing Medical University (Pukou Hospital), Nanjing, Jiangsu, China
[6]Institute of Advanced Studies in Humanities and Social Sciences, Beijing Normal University, No18 Jinfeng Road, Tangjiawan, Xiangzhou District, Zhuhai 519087, Guangdong, China

Received: 17 February 2023 / Accepted: 3 October 2023
Published online: 10 October 2023

References

1. Banerjee I, Bozkurt S, Alkim E, et al. Automatic inference of BI-RADS final assessment categories from narrative mammography report findings. J Biomed Inform. 2019;92:103137. https://doi.org/10.1016/j.jbi.2019.103137.
2. Savova GK, Danciu I, Alamudun F, et al. Use of Natural Language Processing to Extract Clinical Cancer phenotypes from Electronic Medical Records Natural Language Processing for Cancer phenotypes from EMRs. Cancer Res. 2019;79(21):5463–70. https://doi.org/10.1158/0008-5472.CAN-19-0579.
3. Esmaeili M, Ayyoubzadeh SM, Ahmadinejad N, et al. A decision support system for mammography reports interpretation. Health Inform Sci Syst. 2020;8:1–8. https://doi.org/10.1007/s13755-020-00109-5.
4. Pereira JW, Ribeiro MX. Semantic annotation and classification of mammography images using ontologies//2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2021: 378–383. https://doi.org/10.1109/CBMS52027.2021.00043.
5. Rossanez A, Dos Reis JC, Torres RS, et al. KGen: a knowledge graph generator from biomedical scientific literature. BMC Med Inf Decis Mak. 2020;20(4):1–24. https://doi.org/10.1186/s12911-020-01341-5.
6. Zhao J, Liu Z, Cui M, et al. Design and construction of knowledge graph of Electronic Medical Record based on BiLSTM-CRF. Proc 4th Int Conf Big Data Technol. 2021;72–8. https://doi.org/10.1145/3490322.3490334.
7. Li N, Yang Z, Luo L, et al. KGHC: a knowledge graph for hepatocellular carcinoma. BMC Med Inf Decis Mak. 2020;20(3):1–11. https://doi.org/10.1186/s12911-020-1112-5.
8. Seneviratne O, Rashid SM, Chari S et al. Knowledge integration for disease characterization: A breast cancer example//The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17. Springer International Publishing, 2018: 223–238. https://doi.org/10.1007/978-3-030-00668-6_14.
9. Gong F, Wang M, Wang H, et al. SMR: medical knowledge graph embedding for safe medicine recommendation. Big Data Research. 2021;23:100174. https://doi.org/10.1016/j.bdr.2020.100174.
10. An B. Construction and application of chinese breast cancer knowledge graph based on multi-source heterogeneous data. Math Biosci Eng. 2023;20(4):6776–99. https://doi.org/10.3934/mbe.2023292.
11. Hasan SMS, Rivera D, Wu XC et al. A knowledge graph approach for the secondary use of cancer registry data//2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, 2019: 1–4. https://doi.org/10.1109/BHI.2019.8834538.
12. Hasan SMS, Rivera D, Wu XC, et al. Knowledge graph-enabled cancer data analytics. IEEE J Biomedical Health Inf. 2020;24(7):1952–67. https://doi.org/10.1109/JBHI.2020.2990797.
13. Park J, Cho Y, Lee H et al. Knowledge graph-based question answering with electronic health records//Machine Learning for Healthcare Conference. PMLR, 2021: 36–53. https://arxiv.org/abs/2010.09394.
14. Chen Z, Wang Y, Zhao B, et al. Knowledge graph completion: a review. Ieee Access. 2020;8:192435–56. https://doi.org/10.1109/ACCESS.2020.3030076.
15. Ji S, Pan S, Cambria E, et al. A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural Networks Learn Syst. 2021;33(2):494–514. https://doi.org/10.1109/TNNLS.2021.3070843.
16. Magny SJ, Shikhman R, Keppke AL. Breast imaging reporting and data system[M]//StatPearls [Internet]. StatPearls publishing, 2022.

17. Breast Cancer Committee of Chinese Anti-Cancer Association. Chinese guidelines for diagnosis and treatment of breast Cancer. China Oncol. 2021;31(10):609–80.
18. Li L, Wang P, Yan J, et al. Real-world data medical knowledge graph: construction and applications. Artif Intell Med. 2020;103:101817. https://doi.org/10.1016/j.artmed.2020.101817.
19. Stenetorp P, Pyysalo S, Topić G, et al. BRAT: a web-based tool for NLP-assisted text annotation. Proc Demonstrations 13th Conf Eur Chapter Association Comput Linguistics. 2012;102–7. https://doi.org/10.5555/2380921.2380942.
20. Mikolov T, Chen K, Corrado G et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013. http://arxiv.org/abs/1301.3781.
21. Lai S, Liu K, He S, et al. How to generate a good word embedding. IEEE Intell Syst. 2016;31(6):5–14. https://doi.org/10.48550/arXiv.1507.05523.
22. Gorinski PJ, Wu H, Grover C et al. Named entity recognition for electronic health records: a comparison of rule-based and machine learning approaches. arXiv preprint arXiv:1903.03985, 2019. http://arxiv.org/abs/1903.03985.
23. Tiftikci M, Özgür A, He Y, et al. Machine learning-based identification and rule-based normalization of adverse drug reactions in drug labels. BMC Bioinformatics. 2019;20(21):1–9. https://doi.org/10.1186/s12859-019-3195-5.
24. Trienes J, Trieschnigg D, Seifert C et al. Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records. arXiv preprint arXiv:2001.05714, 2020. https://arxiv.org/abs/2001.05714.
25. Ravikumar J, Kumar PR. Machine learning model for clinical named entity recognition. Int J Electr Comput Eng. 2021;11(2):1689–1677. https://doi.org/10.11591/ijece.v11i2.pp1689-1696.
26. Zhang Y, Wang X, Hou Z, et al. Clinical named entity recognition from chinese electronic health records via machine learning methods. JMIR Med Inf. 2018;6(4):e9965. https://doi.org/10.2196/medinform.9965.
27. Zhou M, Tang T, Lu J, et al. EXTRACTING BI-RADS FEATURES FROM MAMMOGRAPHY REPORTS IN CHINESE BASED ON MACHINE LEARNING. J Flow Visualization Image Process. 2021;28(2). https://doi.org/10.1615/JFlowVisImageProc.2020035208.
28. Qin Q, Zhao S, Liu C. A BERT-BiGRU-CRF model for entity recognition of chinese electronic medical records. Complexity. 2021;2021:1–11. https://doi.org/10.1155/2021/6631837.
29. Wu Y, Huang J, Xu C et al. Research on named entity recognition of electronic medical records based on roberta and radical-level feature. Wireless Communications and Mobile Computing, 2021, 2021: 1–10. https://doi.org/10.1155/2021/2489754.
30. Li M, Zhang Y, Huang M et al. Named entity recognition in Chinese electronic medical record using attention mechanism//2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE, 2019: 649–654. https://doi.org/10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00125.
31. Dai Z, Wang X, Ni P, et al. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records//2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei). IEEE. 2019;1–5. https://doi.org/10.1109/CISP-BMEI48845.2019.8965823.
32. Srivastava RK, Greff K, Schmidhuber J. Highway networks. arXiv preprint arXiv:1505.00387, 2015. http://arxiv.org/abs/1505.00387.
33. Qurashi AW, Holmes V, Johnson AP. Document processing: Methods for semantic text similarity analysis//2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). IEEE, 2020: 1–6. https://doi.org/10.1109/INISTA49547.2020.9194665.
34. Hogan A, Blomqvist E, Cochez M, et al. Knowledge graphs. ACM Comput Surv (CSUR). 2021;54(4):1–37. https://doi.org/10.1145/3418294.
35. Lin J, Zhao Y, Huang W, et al. Domain knowledge graph-based research progress of knowledge representation. Neural Comput Appl. 2021;33:681–90. https://doi.org/10.1007/s00521-020-05057-5.
36. Zhu X, Li Z, Wang X, et al. Multi-modal knowledge graph construction and application: a survey. IEEE Trans Knowl Data Eng. 2022. https://doi.org/10.1109/TKDE.2022.3224228.
37. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023;10(1):67. https://doi.org/10.1038/s41597-023-01960-3.

Publisher's Note