A Natural Language Processing Pipeline of Chinese Free-text Radiology Reports for Liver Cancer Diagnosis

Honglei Liu¹ ², Yan Xu³, Zhiqiang Zhang¹ ², Ni Wang¹ ², Yanqun Huang¹ ², Zhenghan Yang³, Rui Jiang⁴§, Hui Chen¹ ²§

¹ School of Biomedical Engineering, Capital Medical University, Beijing 100069, China
² Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical Application, Capital Medical University, Beijing 100069, China
³ Department of Radiology, Beijing Friendship Hospital, Capital Medical University, Beijing, 100050, China
⁴ Ministry of Education Key Laboratory of Bioinformatics; Bioinformatics Division, Beijing National Research Center for Information Science and Technology; Department of Automation, Tsinghua University, Beijing 100084, China.

§ Corresponding author

Email addresses:

Rui Jiang: ruijiang@tsinghua.edu.cn
Hui Chen: chenhui@ccmu.edu.cn
A Natural Language Processing Pipeline of Chinese Free-text Radiology Reports for Liver Cancer Diagnosis

Abstract

Background
Despite the rapid development of natural language processing (NLP) implementation in electronic medical records (EMRs), Chinese EMRs processing remains challenging due to the limited corpus and specific grammatical characteristics, especially for radiology reports. Based on a manually constructed lexicon, this study sought to design an NLP pipeline for the direct extraction of clinically relevant features from Chinese radiology reports, which is the first key step in computer-aided radiologic diagnosis.

Methods
We implemented the NLP pipeline on abdominal computed tomography (CT) radiology reports written in Chinese. The pipeline was comprised of word segmentation, entity annotation, coreference resolution, and relationship extraction to finally derive the symptom features composed of one or more terms. The whole pipeline was based on a lexicon that was constructed manually according to Chinese grammatical characteristics. Least absolute shrinkage and selection operator (LASSO) and machine learning methods (support vector machine (SVM), random forest, decision tree, and logistic regression) were used to build the classifiers for liver cancer prediction. Random forest model was also used to calculate the Gini impurity for identifying the most important features in liver cancer diagnosis.

Results
The lexicon finally contained 831 words, 48 lists of synonyms and five entity types (Location, Morphology, Density, Enhancement, and Modifier). The features extracted by the NLP pipeline conformed to the original meaning of the radiology reports. SVM had a higher predictive performance in liver cancer diagnosis (F1 score 90.23%, precision 92.51%, and recall 88.05%). The regular state of size, shape, and contour were the top features extracted from the radiological reports for the liver cancer diagnosis.

Conclusions
Our study was a comprehensive NLP study focusing on Chinese radiology reports and the application of NLP in cancer risk prediction. The proposed method for the radiological feature extraction could be easily implemented in other kinds of Chinese clinical texts and other disease predictive tasks.

Keywords: Natural language processing, Radiology reports, Information extraction, Computer-aided diagnosis, Machine learning
Background
Massive electronic medical records (EMRs) are potentially valuable clinical sources for research for improving clinical care and support [1 2]. In the current digital age, artificial intelligence (AI)-based algorithms play a powerful role in data mining, which is useful in applications such as clinical decision-making, disease computer-aided diagnosis, and management [3 4].

As an important EMRs component, the radiology report is a primary method of communication between radiologists who interpret the image and physicians who make the final diagnosis. Radiological diagnosis is frequently formulated by relying on physicians’ experience, which may lead to limited accuracy and efficiency [5]. With the rapid growth of clinical big data, applying AI methods to process medical texts becomes executable. Extracting clinically relevant information from radiology reports has great importance in terms of advancing radiological research and clinical practice [6], although significant challenges still exist, mainly due to the free form of most reports [7]. Natural language processing (NLP) is a multistep process comprised of statistical and linguistic methods that can mine information from unstructured texts, which are then formed into a standardized structured format (i.e., a fixed collection of text features). NLP-based feature extraction has advantages for massive text processing compared with time-consuming manual extraction flow. Hence, NLP-based feature extraction has been effectively used in radiology for diagnostic surveillance, cohort building, quality assessment, and clinical support services [8-11]. Nevertheless, previous NLP studies of radiology reports primarily focused on documents written in English. With the rapid growth of clinical data in China, information extraction from vast amounts of Chinese radiology reports has become a meaningful task that has both theoretical and practical significance. Due to the limitations of the related corpus, NLP on Chinese clinical texts remains challenging [12 13].

NLP and Its Clinical Applications
Compared with structured text, free text is more natural and expressive in the record of the clinical events. To facilitate the application of clinical texts, information-mining research using NLP, which could automatically extract entities, events, and relations, is necessary. During the NLP workflow, such as semantic analysis and syntactic analysis, a lexicon of words with definitions and synonyms is useful. Several tools and systems could provide such support. For example, the Unified Medical Language System (UMLS) Metathesaurus [15] includes synonymous terms and specific semantic roles for each concept and relationships between concepts. Other useful lexicons and ontologies include RadLex® [16], which is a specialized radiological lexicon including imaging techniques. Algorithms for information extraction can be divided into two categories, rule-based methods and machine learning methods [1]. For clinical information, the most frequently used tool is cTAKES [17]. In cTAKES, lists of individual concepts identified from medical terms can be produced. In addition, MetaMap [18] and MedLEE [19] are also widely used for information extraction.

After information extraction to obtain the structured features, NLP can be further implemented on clinical tasks, such as disease studies [20-22], drug-related studies [23 24], and clinical workflow optimization [25]. Computer-aided diagnosis is an important research field in disease study, which aims to use computer algorithms to provide physicians a reference for disease diagnosis. Studies have investigated many diseases to date, such as hepatocellular cancer [20], colorectal cancer [26], pancreatic cancer [22], and celiac disease [27]. Wu et al. developed Med3R using a deep learning model that successfully provided a comprehensive aided clinical diagnosis service on EMRs [28].

For radiology reports, NLP has been utilized for identifying biomedical concepts [29], extracting recommendations [30], extracting actionable findings of appendicitis [31], determining the change level of clinical findings, and so on [32]. Machine learning methods are widely used today for other clinical applications. For example, Bahl et al. developed a random forest method to predict high-risk breast lesions using textual features [33]. Using IBM Watson’s NLP algorithm, Trivedi et al. developed a classifier to automatically assign the intravenous contrast use based on magnetic resonance imaging reports [34].

NLP in Chinese Clinical Information Extraction
Due to the limitation of Chinese EMRs corpus, NLP systems in clinical information extraction and application are challenging, which probably leads to a poor performance based on the general corpus. Therefore, corpus annotating and lexicon building are necessary for NLP in specific clinical applications. For example, in the research on word segmentation, which was the initial processing step of NLP, He et al. [35] and Xu et al. [36] annotated the comprehensive corpus for different kinds of clinical texts in their research fields to improve NLP performance. Recently, there are increasing numbers of studies on broader NLP element tasks in Chinese EMRs, such as named entity recognition (NER) [37 38] and speculation detection [39]. In information extraction, CMedTEX is a rule-based system for extracting and normalizing
temporal expressions [40]. Information extraction studies for Chinese clinical texts also include temporal expression extraction and normalization [12], entity relation extraction [13], and so on. In an NLP application, Miao et al. extracted BI-RADS findings to support clinical operation and breast cancer research in China [40]. Liang et al. applied an automatic NLP system using machine learning to extract information from Chinese EMRs and demonstrated high diagnostic accuracy in childhood disease diagnosis [41].

Although there are some studies based on Chinese clinical texts in NLP fundamental tasks, higher-level tasks and applications are limited, especially for research on radiology reports. Building a comprehensive NLP pipeline for information extraction from Chinese radiology reports has great importance for further NLP research. In this study, we designed an NLP pipeline that could extract clinically relevant features from abdominal computed tomography (CT) radiology reports written in Chinese. In consideration of the language characteristics of Chinese, we manually collected a lexicon, containing words, synonym lists, and entity types. Typically, patients with liver cancer are likely to be diagnosed with symptoms of advanced disease. Moreover, the diagnosis of liver cancer via early examination, such as radiological examination, is necessary [14]. Therefore, in terms of implementation, we applied different machine learning algorithms to liver cancer prediction using the features extracted by the NLP pipeline.

Methods

Dataset
Abdominal CT radiology reports were collected from a tertiary hospital in Beijing, China, between 2012 and 2018. All identifying information was removed to protect patient privacy. All radiology reports were unstructured and written in Chinese. According to the content, the radiology report included the Type of examination, Clinical history, Comparison, Technique, Findings, and Impressions. In the Findings section, a radiologist listed the observations regarding each area of the body examined. Whether and how the area was normal, abnormal or potentially abnormal was recorded. The Impressions section contained a diagnosis indicated by a radiologist when combining the radiological findings and clinical history. The NLP pipeline in this study was applied to the section of radiological findings.

As shown in Figure 1, of all the patients, 519 were diagnosed with liver cancer based on both the section of impressions and annotations by experienced radiologists. We further randomly selected 654 reports of 654 patients with the diagnosis of liver cirrhosis, liver cysts, hepatic or hemangioma.

![Figure 1. Data selection flow chart](image)

The NLP pipeline
Figure 2 shows the overview of the computer-aided diagnosis framework that consisted of an NLP section and a disease classifier section. NLP was performed to extract radiological features with terms from the radiology reports. Features in training reports were reduced to a smaller subset by the least absolute shrinkage and selection operator (Lasso, also LASSO) method and then were input into machine learning models.
Lexicon Building
The whole framework for feature selection and extraction was initialized with lexicon construction. In reports with and without a liver cancer diagnosis, a small number of reports (approximately 3% of overall data) were sampled randomly for generating the lexicon by manual reading. Another subset of radiology reports (approximately 1% of overall data) was sampled randomly to further manually integrate the lexicon. The specialized lexicon containing clinical terms, entity types, and lists of synonyms was built based on prior clinical knowledge and Chinese grammatical characteristics. The entity types in radiology reports included [Location] (e.g., 肝脏 (liver)), [Morphology] (e.g., 轮廓规整 (regular contour)), [Density] (e.g., 密度不均匀 (nonhomogeneous)), [Enhancement] (e.g., 动脉期 (arterial phase)), and [Modifier] (e.g., 结节状强化 (nodular enhancement)). Synonyms involved different locations of the liver, different presentations of items, such as “low density”, “irregular” and so on.

Figure 2. Overview of the natural language processing pipeline for clinically relevant features extraction and liver cancer diagnosis.

Word segmentation
As Figure 3 shows, the NLP pipeline consisted of a sequence of steps that could generate the structured radiological features from unstructured radiology reports. With the input of reports and the lexicon previously constructed, the first step was word segmentation by the forward maximum matching (FMM) method focusing on the imaging finding section of the radiological reports. The FMM is a basic word segmentation algorithm for Chinese text, which uses a greedy method preferring to match from the longest word (with the most Chinese characters) according to the provided lexicon [42]. Simultaneously, entity types were annotated based on the lexicon. Thus, the whole text was segmented into individual words. In the word-level coreference resolution step, words with the same meaning were unified into a single word according to the synonym lists generated previously.

Radiological feature extraction
We extracted symptom information among the single words towards the computer-aided diagnosis. Each report was divided into a series of sentences by a full stop. The sentences were further divided into several parts if more than one entity [Location] occurred. As described in Table 1, several rule-based patterns were then designed to extract relations according to semantic comprehension, syntactic structure, and knowledge-based characteristics. All patterns started with an entity [Location], followed by the combination of other entities in each sentence or part. Entity [Morphology] only represented the morphology description for different locations. Entity [Modifier] referred to the modifier of entity [Enhancement] or [Density]. If the entities [Enhancement] and [Density] occurred at the same time, the feature ([Location]+[Enhancement]+[Density]+Others (either [Modifier] or [Morphology])) would be recognized into several features all starting with the same [Location] ([Location]+[Enhancement]+Others...).
or [Location]+[Density]+Others or [Location] +Others). We scanned all the satisfactory patterns according to the above rules. A feature extraction example is shown in Figure 3.

The final results of all previous steps in the NLP pipeline converted each unstructured report into several itemized features with terms that could be easily manipulated. These lists of radiological features were subsequently used to build prediction models for liver cancer.

**A sentence from the imaging findings of a radiology report:**
肝脏形态大小正常，轮廓规整，肝实质密度不均匀，肝右叶可见巨大不规则低密度灶。
(The liver is normal in size and shape, and the contour is regular. The liver parenchyma is nonhomogeneous. Right lobe of liver has irregular low density area.)

**A. Word segmentation**

| Location + Density | 肝脏+低密度影 (liver + low density) |
|--------------------|-----------------------------------|
| Location + Enhancement | 肝脏+增强扫描未见强化 (liver + enhancement scan showed no enhancement) |
| Location + Enhancement + Modifier | 肝门+动脉期+结节状强化 (porta hepatis + arterial phase + nodular enhancement) |
| Location + Density + Modifier | 肝脏+低密度灶+边界清晰 (liver + low density area + clear boundary) |
| Location + Morphology | 肝脏+形态大小正常 (liver + normal in size and shape) |

**B. Entity annotation, coreference resolution and relation extraction**

**C. Radiological features**

| Pattern | Example |
|---------|---------|
| Location + Density | 肝脏+低密度影 (liver + low density) |
| Location + Enhancement | 肝脏+增强扫描未见强化 (liver + enhancement scan showed no enhancement) |
| Location + Enhancement + Modifier | 肝门+动脉期+结节状强化 (porta hepatis + arterial phase + nodular enhancement) |
| Location + Density + Modifier | 肝脏+低密度灶+边界清晰 (liver + low density area + clear boundary) |
| Location + Morphology | 肝脏+形态大小正常 (liver + normal in size and shape) |

**Figure 3.** An example of feature extraction in a sentence from a radiology report. Texts in parentheses are the corresponding English translations.

**Table 1.** Patterns summarized to extract features according to the word’s entity type

| Entity pattern | Example |
|----------------|---------|
| Location + Density | 肝脏+低密度影 (liver + low density) |
| Location + Enhancement | 肝脏+增强扫描未见强化 (liver + enhancement scan showed no enhancement) |
| Location + Enhancement + Modifier | 肝门+动脉期+结节状强化 (porta hepatis + arterial phase + nodular enhancement) |
| Location + Density + Modifier | 肝脏+低密度灶+边界清晰 (liver + low density area + clear boundary) |
| Location + Morphology | 肝脏+形态大小正常 (liver + normal in size and shape) |

**Predictive models**

Itemized features derived from the previous steps were binary, representing the absence or presence of a certain feature (Figure 1). They served as the input of the classifier for liver cancer prediction. The classifier output was also binary, indicating whether the patient was diagnosed with liver cancer or not.

Lasso, a regression analysis method, was introduced for the feature selection to improve the prediction performance and interpretability on training reports. Features selected by Lasso were further used by the test reports. Imposing L1 penalty on the feature vector, the Lasso method encouraged to use only a subset of the overall features rather than all of them [43]. We used the binomial distribution for Lasso logistic regression due to the binary response (whether diagnosed with liver cancer or not) in this study.
Machine learning-based classifiers, including the decision tree, random forest, support vector machine (SVM) and logistic regression were then built for the liver cancer prediction. With good interpretability, logistic regression is usually used to explain the relationship between the independent variables and the binary dependent variable. A decision tree can be considered as a set of if-then rules, which describes the process of instances classification based on trees. The prediction model and results generated by a decision tree are easy to understand [44]. Random forest is an ensemble learning method constructed with a multitude of decision trees, which usually gets higher performance than a single decision tree [45]. Based on the structural risk minimization principle, SVM is a robust model for prediction problems by maximizing the margin. Different types of kernels can be chosen to solve both linear and non-linear problems [46], while a linear kernel was used in this study.

Fivefold cross-validation was employed when assessing and comparing the predictive models. Performance measures used in this classification study included recall (also called sensitivity), precision and F1 score. To rank the radiological features associated with the diagnosis of liver cancer, Gini impurity was computed by the random forest method [47]. Gini impurity is a measurement of the probability that a sample is classified incorrectly in tree-based models without a specific feature. The greater the Gini impurity is, the more important the radiological feature is.

Results

We finally collected 831 words and 48 lists of synonyms (Supplementary Table 1) in the lexicon. All words were classified into one of the five entity types. Entity types [Location], [Morphology], [Density], [Enhancement], and [Modifier] were annotated among 226, 295, 125, 131, and 54 words, respectively. After feature extraction, only features that occurred more than twice were retained. We finally got 398 features to formulate the feature vectors. According to the presence or absence of each feature, every radiology report was represented by a 0-1 vector in the feature vector space. The statistics of the extracted radiological features were shown in Table 2. There were 6 features with a proportion higher than 30% of all the reports. The top 5 features with high proportion were all associated with liver morphology, which were usually required to be recorded in every radiology report.

| Radiological Features | Count | Proportion in all the reports | Proportion in all the features |
|-----------------------|-------|-----------------------------|-------------------------------|
| 肝脏/形态大小正常 | 703 | 59.93% | 6.50% |
| (liver / normal in size and shape) | | | |
| 肝脏/轮廓规则 | 617 | 52.60% | 5.70% |
| (liver / contour is regular) | | | |
| 肝固 /未见异常 | 597 | 50.90% | 5.52% |
| (porta heptatis / regular) | | | |
| 肝叶/比例异常 | 522 | 44.50% | 4.83% |
| (liver lobe / normal proportion) | | | |
| 肝裂/无增宽 | 503 | 42.88% | 4.65% |
| (hepatic fissures / no broadening) | | | |
| 肝脏/低密度影 | 373 | 31.80% | 3.45% |
| (liver / low density) | | | |

All the models got a relatively high performance (Table 3, Supplementary Figure 1), and Lasso worked efficiently in performance improvement. All four classifiers with Lasso-based feature reduction got a higher F1 score compared with classifiers without such feature dimension reduction. The highest F1 score of 90.23% was seen in the SVM model, whose precision was also the highest (92.51%). Compared with SVM, the random forest got a lower precision but a higher recall. All the evaluation indicators of random forest were higher than those of a decision tree. After Lasso being applied, the performance of SVM and logistic regression improved greatly. F1 score increased by 7.58% for logistic regression and 3.82% for SVM. However, the decision tree and random forest were not sensitive to the reduced input features, only increasing by 1.02% and 0.53% respectively in the F1 score (Table 3).

The Gini impurity of all radiological features derived from random forest and the top ten features associated with the liver cancer diagnosis were shown in Figure 4. The top two features both had a Gini impurity greater than 0.1, which presented the regular state of the liver in contour, size, and shape.

| Predictive model | Precision (%) | Recall (%) | F1 Score (%) |
|------------------|---------------|------------|--------------|
|                  |               |            |              |
### Table 1: Performance of Different Models

| Model                | Without Lasso | With Lasso |
|----------------------|---------------|------------|
| Logistic Regression  | 82.14         | 87.70      |
| Decision Tree        | 85.41         | 85.58      |
| Random Forest        | **88.80**     | 88.91      |
| Support Vector Machine| 87.90        | **92.51**  |

**Figure 4.** Top ten radiological features associated with liver cancer diagnosis ranked by Gini impurity.

### Discussion

Liver cancer is a substantial economic burden for both patients and the government in China. Limited by the diagnostic technology, many patients are diagnosed at the stage of terminal liver cancer, resulting in a much poorer prognosis in China compared with that of developed countries [48]. Therefore, the early diagnosis of liver cancer by the benefit of informative examination, such as radiological examination, has great significance [14 49]. With the development of AI technology in recent eras, computer-aided early diagnosis for cancer built on massive clinical data became feasible. Feature extraction from these clinical data was an important step. For clinical texts, NLP was widely used for information extraction, and it was implemented in disease study areas, especially for the category of neoplasms [1].

Since the insufficient corpus and lexicon made it harder to process clinical texts in Chinese than those in English, some studies focused on the corpus construction and other fundamental NLP studies focused on Chinese clinical texts [36 37]. In this work, we constructed a lexicon from a small proportion of radiology reports randomly sampled in the overall dataset. This radiological lexicon, rather than a general dictionary, was then used in the subsequent study. The lexicon was manually collected and annotated by some experienced radiologists based on their prior clinical knowledge and Chinese grammatical rules. Different from English, Chinese has its own specific semantic characteristics and grammatical rules, especially in the medical domain. Chinese text has more flexibility in word combinations. For example, the word 肝脏 (English: liver), belonging to entity [Location], could also be written as a specific segment of the liver in radiology reports, such as “肝S8”, “肝S3”, or just one character, “肝”. Therefore, the constructed lexicon included a list of synonyms to unify different presentations and different sections of 肝脏 into a single word. The synonyms also contained other Chinese expressions, such as negative words. The lexicon only took clinically relevant words into consideration. As a result, other words remained...
unique characters and would be ignored during information extraction.

In the consideration of the characteristics of radiology reports, we annotated five entity types and designed five patterns (i.e., entity combinations) for the feature extraction. The extracted itemized features could present the meaning of corresponding sentences. Although the listed patterns and entity annotation could restrict the number of word combinations, the features still had a high dimension. Screened by word count and the Lasso method, the extracted features decreased to a tiny amount, which was a relatively limited number compared with the free-text. As presented in Table 2, the normal morphology of different locations had the highest counts. The main reason for this may be that the morphology of some locations, such as liver, liver lobe, and porta hepatis, should be recorded in every radiology report no matter whether the patient had liver disease or not.

With the derived radiological features containing terms, all four machine learning models had good performance with an F1 score higher than 85% with Lasso (Table 3), where SVM achieved the highest F1 score. All the evaluation indicators of random forest were higher than the decision tree since the random forest was an ensemble method constructed by a large number of decision trees. For the two classification models with an overall high performance, SVM achieved a higher precision but lower recall than the random forest, meaning that the SVM-based classification model had a higher positive predictive value and was more likely to provide a false negative prediction. We could see that different performance occurred with the same features using different classifiers. In clinical application, lower recall represented a higher under-diagnosed rate, which was not beneficial for disease screening. In contrast, lower precision showed a lower prediction reliability, leading to a lower clinical application value. Therefore, the four machine learning methods adapted to various application requests. SVM had the highest reliability and random forest had the highest completeness in liver cancer prediction. We could conclude that the structured features extracted by the NLP pipeline could obtain effective information from the original reports, and we would expect a wide array of clinical applications using the structured features instead of the free-text reports.

Through the analysis of misjudgment samples, we identified two main reasons for misclassification. Some patients who were diagnosed with liver cirrhosis were easily classified as liver cancer since some radiological features of liver cirrhosis were close to liver cancer. Patients with liver cirrhosis had the potential to progress to cancer [14 49]. Therefore, our results could be an early warning for these patients. Another reason for the incorrect classification may be the radiological features omitted during NLP extraction. Due to the size of the dataset, the missing of clinical terms during lexicon construction was inevitable. Especially due to Chinese grammatical characteristics, the long term could retain the same meaning after the emendation of several characters in unstructured form. Thus, extending the lexicon to cover as much term as possible in free-text radiology reports was a challenging task. We could collect more data for information extraction or more samples for lexicon construction to decrease this kind of error in future studies.

The prediction of cancer and other diseases is an important and significant application of medical language processing. Extracted features from EMRs could be part or all of the features for classifier input. Studies of cancer prediction using administrative data and EMRs have been published [20]. In recent years, there were also several studies of disease evaluation using NLP on Chinese clinical data [28]. In contrast with these studies of NLP applications, our work is the first study on liver cancer assisted diagnosis based on Chinese radiology reports. The NLP pipeline in this work focused on features extracted only from texts, which could represent the whole free text in further applications. Furthermore, this study was focused on radiology reports, which are a valuable resource for the detection of some kind of diseases.

To get the radiological features strongly associated with the liver cancer diagnosis, we ranked the features by Gini impurity derived from the random forest method. The regular state of size, shape, and contour had the highest Gini impurity which coincided with clinical knowledge. These features were important and basic risk factors in liver disease diagnosis. Broadening of hepatic fissures was also characteristic features that occurred in a patient with liver cirrhosis and liver cancer progressed from liver cirrhosis [14 49]. Except for the radiological features presented in Figure 4, the top features also included “liver / arterial phase / enhancement”, and this imaging finding of hyperechoic enhancement in the arterial phase was one of the significant features in the liver cancer diagnosis [50].

Our pipeline took both Chinese grammar and radiological characteristics into consideration during the NLP pipeline design and provided an executable framework in automatic clinical feature extraction. Therefore, our overall study could be easily applied to the processing of clinical documents in Chinese other than radiology reports. The application in this study was to help physicians with the radiological diagnosis. Since medical health services are very unbalanced in China, the potential application of our work is in assisting physicians in small and medium-sized urban and rural areas. In these regions, qualified doctors are insufficient, which leads to the risk of inaccurate disease diagnosis.
Although our pipeline has been shown to have high performance in liver cancer diagnosis, there are still some limitations. The lexicon construction was based on limited annotation resources from one hospital. Hence, some clinical key terms had a risk of omission, and the performance of some NLP procedures might weaken across different hospitals. In addition, in the current lexicon version, some words for detailed descriptions were ignored, such as different position sizes and density area shapes. Deriving a common lexicon in radiology based on massive resources and more details could further improve NLP performance.

Conclusions
This study described an NLP pipeline of Chinese free-text radiology reports for liver cancer diagnosis. To improve the accuracy of feature extraction, we constructed a lexicon containing clinical terms, entity types, and synonyms, instead of using a common corpus. Itemized features composed of terms were extracted based on rules and a lexicon, which finally made the reports structured. Our model achieved a high performance in the application of liver cancer prediction. Our study was a comprehensive study of a liver cancer computer-aided diagnosis model using the NLP method based on Chinese radiology reports. The NLP pipeline proposed in this project could be generalized to the lexicon construction of other diseases and other kinds of clinical texts in Chinese. Furthermore, the radiological feature extraction method could be an important step towards the international use of massive Chinese clinical data for health research.

Abbreviations
NLP: natural language processing; CT: computed tomography; LASSO: Least absolute shrinkage and selection operator; SVM: support vector machine; EMRs: electronic medical records

Declarations
Authors’ contributions
Honglei Liu proposed and designed the whole pipeline, analyzed the results and wrote the paper. Yan Xu and Zhenghan Yang collected the original data, annotated the radiology reports and provided clinical knowledge guidance in lexicon construction. Zhiqiang Zhang, Ni Wang and Yanqun Huang cleaned the data. Rui Jiang and Hui Chen provided theoretical guidance and the revision of this paper.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The datasets used and analyzed during the current study are not available since the privacy of residents is included.

Competing interests
The authors declare that they have no competing interests.

Funding
The work is supported by grants from National Natural Science Foundation of China (No. 81701792 and No. 81971707), and National Key Research and Development Program of China (No. 2018YFC0910404)

References
1. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S et al: Clinical information extraction applications: A literature review. J Biomed Inform 2018, 77:34-49.
2. Ellsworth MA, Dziadzko M, O'Horo JC, Farrell AM, Zhang J, Herrasevich V: An appraisal of published usability evaluations of electronic health records via systematic review. *J Am Med Inform Assoc* 2017, 24(1):218-226.

3. Jensen PB, Jensen LJ, Brunak S: Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012, 13(6):395-405.

4. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI: Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015, 13:8-17.

5. Cai T, Giannopoulou AA, Yu S, Kelil T, Ripley B, Kumamaru KK, Rybicki FJ, Mitsouras D: Natural Language Processing Technologies in Radiology Research and Clinical Applications. *Radiographics* 2016, 36(1):176-191.

6. Pons E, Braun LM, Hunink MG, Kors JA: Natural Language Processing in Radiology: A Systematic Review. *Radiology* 2016, 279(2):329-343.

7. Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K, Lindsetmo RO, Kouskoumvekaki I, Girolami M, Olav Skovseth S, Augustad KM: Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep* 2017, 7:46226.

8. Goft DJ, Loehfelm TW: Automated Radiology Report Summarization Using an Open-Source Natural Language Processing Pipeline. *J Digit Imaging* 2018, 31(2):185-192.

9. Huhdanpaa HT, Tan WK, Rundell SD, Suri P, Chokshi FH, Comstock BA, Heagerty PJ, James KT, Avins AL, Nedeljkovic SS et al: Using Natural Language Processing of Free-Text Radiology Reports to Identify Type 1 Modic Endplate Changes. *J Digit Imaging* 2018, 31(1):84-90.

10. Kang SK, Garry K, Chung R, Moore WH, Iitтратe E, Swartz JL, Kim DC, Horwitz LI, Blecker S: Natural Language Processing for Identification of Incidental Pulmonary Nodules in Radiology Reports. *J Am Coll Radiol* 2019.

11. Kehl KL, Elmarakeby H, Nishino M, Van Allen EM, Lepisto EM, Hassett MJ, Johnson BE, Schrag D: Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. *JAMA Oncol* 2019.

12. Hao T, Pan X, Gu Z, Qu Y, Weng H: A pattern learning-based method for temporal expression extraction and normalization from multi-lingual heterogeneous clinical texts. *BMCMedical Informatics and Decision Making* 2018, 18(S1).

13. Zhang Z, Zhou T, Zhang Y, Pang Y: Attention-based deep residual learning network for entity relation extraction in Chinese EMRs. *BMCMed Inform Decis Mak* 2019, 19(Suppl 2):55.

14. Kudo M, Trevisani F, Abou-Alfa GK, Rimassa L: Hepatocellular Carcinoma: Therapeutic Guidelines and Medical Treatment. *Liver Cancer* 2016, 6(1):16-26.

15. (UMLS) UMLS! U.S. National Library of Medicine. [http://www.nlm.nih.gov/research/umls/](http://www.nlm.nih.gov/research/umls/).

16. RadLex: Radiological Society of North America. [http://www.rsna.org/RadLex.aspx](http://www.rsna.org/RadLex.aspx).

17. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010, 17(5):507-513.

18. Aronson AR, Lang FM: An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010, 17(3):229-236.

19. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB: A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994, 1(2):161-174.

20. Sada Y, Hou J, Richardson P, El-Serag H, Davila J: Validation of Case Finding Algorithms for Hepatocellular Cancer From Administrative Data and Electronic Health Records Using Natural Language Processing. *Med Care* 2018, 31(2):185-195.

21. Bhattacharjee A, Wei F, Zhu Y, Sun Y, Wu Q, Clarke D, et al: Comparison of machine learning algorithms for cancer diagnosis and prognosis using electronic health records. *Med Care* 2015, 53(12):1134-1141.

22. Roch AM, Mehrabi S, Krishnan A, Schmidt HE, Kesterson J, Beesley C, Dexter PR, Palakal M, Schmidt CM: Automated pancreatic cyst screening using natural language processing: a new tool in the early detection of pancreatic cancer. *HPB (Oxford)* 2015, 17(5):447-453.

23. Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, Dai Q, Levy M, Shah A, Han X, Ruan X et al: Validating drug repurposing signals using electronic health records: A case study of metformin associated with reduced cancer mortality. *J Am Med Inform Assoc* 2015, 22(1):179-191.

24. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC: MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010, 17(1):19-24.

25. Tamang S, Patel MI, Blayney DW, Kuznetsov J, Finlayson SG, Vetteth Y, Shah N: Detecting unplanned care from clinician notes in electronic health records. *J Oncol Pract* 2015, 11(3):e313-319.

26. Xu H, Fu Z, Shah A, Chen Y, Peterson NB, Chen Q, Mani S, Levy MA, Dai Q, Denny JC: Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA*
Ludvigsson JF, Pathak J, Murphy S, Durski M, Kirsch PS, Chute CG, Ryu E, Murray JA: Use of computerized algorithm to identify individuals in need of testing for celiac disease. *J Am Med Inform Assoc* 2013, 20(2):e306-310.

Wu J, Liu X, Zhang X, He Z, Lv P: Master clinical medical knowledge at certificated-doctor-level with deep learning model. *Nat Commun* 2018, 9(1):4352.

Flynn RW, Macdonald TM, Schembri N, Murray GD, Doney AS: Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes. *Pharmacoepidemiol Drug Saf* 2010, 19(8):843-847.

Yetisen-Yildiz M, Gunn ML, Xia F, Payne TH: A text processing pipeline to extract recommendations from radiology reports. *J Biomed Inform* 2013, 46(2):354-362.

Rink B, Roberts K, Harabagiu S, Scheuermann RH, Toomay S, Browning T, Bosler T, Peshock R: Extracting actionable findings of appendicitis from radiology reports using natural language processing. *AMIA Annu Symp Proc* 2013, 2013:221.

Hassanpour S, Bay G, Langlotz CP: Characterization of Change and Significance for Clinical Findings in Radiology Reports Through Natural Language Processing. *J Digit Imaging* 2017, 30(3):314-322.

Bahl M, Barzilay R, Yedidia AB, Locascio NJ, Yu L, Lehman CD: High-Risk Breast Lesions: A Machine Learning Model to Predict Pathologic Upgrade and Reduce Unnecessary Surgical Excision. *Radiology* 2018, 286(3):810-818.

Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH: Automatic Determination of the Need for Intravenous Contrast in Musculoskeletal MRI Examinations Using IBM Watson’s Natural Language Processing Algorithm. *Journal of Digital Imaging* 2017, 31(2):245-251.

He B, Dong B, Guan Y, Yang J, Jiang Z, Yu Q, Cheng J, Qu C: Building a comprehensive syntactic and semantic corpus of Chinese clinical texts. *Journal of Biomedical Informatics* 2017, 69:203-217.

Xu Y, Wang Y, Liu T, Liu J, Fan Y, Qian Y, Tsuji J, Chang EI: Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries. *J Am Med Inform Assoc* 2014, 21(1):e84-92.

Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H: A comprehensive study of named entity recognition in Chinese clinical text. *J Am Med Inform Assoc* 2014, 21(5):808-814.

Chen Y, Lasko TA, Mei Q, Denny JC, Xu H: A study of active learning methods for named entity recognition in clinical text. *J Biomed Inform* 2015, 58:11-18.

Zhang S, Kang T, Zhang X, Wen D, Elhadad N, Lei J: Speculation detection for Chinese clinical notes: Impacts of word segmentation and embedding models. *J Biomed Inform* 2016, 60:334-341.

Miao S, Xu T, Wu Y, Xie H, Wang J, Jing S, Zhang Y, Zhang X, Yang Y, Zhang X et al: Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches. *International Journal of Medical Informatics* 2018, 119:17-21.

Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, Cai W, Kermany DS, Sun X, Chen J et al: Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med* 2019, 25(3):433-438.

Chen K, Liu S: Word identification for Mandarin Chinese sentences. *Proceedings of the Fourteenth International Conference on Computational Linguistics (Association for Computational Linguistics, Stroudsburg, PA)* 1992, 1:101-107.

Tibshirani R: The lasso method for variable selection in the Cox model. *Stat Med* 1997, 16(4):385-395.

Hastie T, Tibshirani R, Friedman J: The Elements of Statistical Learning; 2008.

Ho TK: Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal* 1995:278-282.

Vapnik VN, Lerner AY: Recognition of patterns with help of generalized portraits. *Avtomat i Telemekh* 1963, 24(6):774-780.

Savin I, Ershova K, Kurdyumova N, et al. Healthcare-associated ventriculitis and meningitis in a neuro-ICU: Incidence and risk factors selected by machine learning approach. *Journal of Critical Care*, 2018, 45:95-104.

WHO: World Cancer Report. 2014.

Nagtegaal ID, Odze RD, Klimstra D, Paradis V, Rugge M, Schirmacher P, Washington MK, Carneiro F, Cree IA: The 2019 WHO classification of tumours of the digestive system. *Histopathology* 2019.

Guidelines of Chinese Society of Clinical Oncology (CSCO): Hepatocellular Carcinoma. 2018.