Tumor Event Extraction based on Semantic Features and Its Application in Auxiliary Diagnosis and Treatment

Xiaoxue Li*, Bairu Jia, Yao Liu
School of Information Engineering, Shaanxi University of International Trade & Commerce, Xi'an, Shaanxi, China
** Corresponding Author*

Abstract: To address the challenge of utilizing unstructured text in tumor electronic medical records for clinical decision-making, this study proposes a medical event extraction method based on fused semantic features and designs an oncology-assisted diagnosis and treatment visualization system. First, a pre-trained medical event extraction model is employed to automatically extract key event information from tumor records, including tumor type, lesion size, location, grade, and metastasis. Second, an auxiliary diagnosis and treatment system is developed, incorporating functions such as tumor grading, metastasis detection, personalized treatment plan formulation, prognosis evaluation, and data statistical analysis. Finally, a user interface is designed using Qt Designer to visualize the workflow of record input, model invocation, diagnostic recommendation generation, and evaluation. Practical applications demonstrate that this system effectively assists clinicians in rapidly extracting critical information from records and generating reasonable treatment plans, enhancing the efficiency and interpretability of tumor diagnosis and treatment while validating the feasibility of medical event extraction technology in clinical oncology assistance.

Keywords: Tumor Event Extraction; Semantic Feature Fusion; Auxiliary Diagnosis and Treatment; Electronic Medical Records; Visualization System.

1. Introduction

Tumors are one of the major diseases that threaten human health worldwide. According to the 2022 Global Cancer Statistics Report released by the International Agency for Research on Cancer (IARC) of the World Health Organization, there were nearly 20 million new cancer cases and approximately 9.7 million cancer deaths worldwide in 2022 [1]. In the comprehensive diagnosis and treatment of tumors, electronic medical record systems are widely used, generating massive amounts of clinical text data. These electronic medical records record important information such as demographic information, medical history, imaging examination results, pathological diagnosis, treatment plan, and follow-up status of patients, and are the core data source for tumor clinical decision-making and scientific research analysis.

However, a large amount of key information in electronic medical records exists in the form of unstructured natural language text. Traditionally, doctors need to manually read the entire medical record and extract key diagnostic and treatment information such as tumor type, staging, grading, lesion size, and metastasis. This process not only consumes a lot of time and human resources, but also easily affects the accuracy of subsequent diagnosis and treatment decisions due to information omissions or misunderstandings. Therefore, how to efficiently and accurately extract tumor event information that has guiding significance for clinical practice from unstructured medical record texts has become an important technical issue in the informationization of tumor diagnosis and treatment.

The event extraction for Chinese electronic tumor medical records faces the following special challenges. Firstly, there is a lack of annotated corpus, and the Chinese medical event extraction field lacks large-scale high-quality annotated datasets. Existing medical event extraction models often struggle to achieve ideal extraction performance due to insufficient training data. Secondly, there are a large number of professional medical terms and complex sentence structures in Chinese electronic medical records, and existing models often struggle to fully capture the impact of contextual semantic information on event recognition. In addition, the time information involved in tumor medical records (such as treatment start and end, disease progression time, etc.) is of great significance for assisting diagnosis and treatment decisions. However, research on tumor event extraction that focuses on events and considers temporal relationships is still relatively weak. Finally, most current research on medical event extraction remains at the methodological level, and the exploration of end-to-end applications from event extraction results to clinical auxiliary diagnosis and treatment systems is not yet sufficient.

2. Related Work

In recent years, scholars both domestically and internationally have conducted extensive research in the field of medical event extraction. In the field of biomedical text mining, researchers continue to explore event extraction methods to address the problem of information overload in academic literature and clinical texts [2,3]. In terms of clinical event extraction, pre trained language models have shown good performance, and models based on Transformer architectures such as BERT are significantly better than traditional Bi LSTM models in medical text classification and event extraction tasks [4].

In the field of medical information extraction for Chinese electronic medical records, there has been a review system that summarizes the research progress in named entity recognition entity relationship extraction, and other aspects [5]. Specialized research in the field of cancer is also gradually advancing. Some studies have proposed a medical event extraction framework for Chinese electronic medical records, and conducted extraction experiments on event elements such as tumor primary site, lesion size, and metastasis site, achieving good results [6]. Another study proposed the CACER corpus, which contains fine-grained annotations of over 48000 medical issues and drug events. The BERT model fine tuned based on this corpus achieved an F1 score of 88.2 in tumor event extraction tasks, which is comparable to manual annotation and validates the effectiveness of the Transformer based model in cancer event extraction [7,8]. In terms of extracting information on tumor progression, there have been studies based on NLP pipeline to automatically extract structured and time anchored clinical events from tumor EMR, achieving an F1 score of 79.7% in the extraction of tumor progression concepts, verifying the practical feasibility of NLP technology in tumor EMR structuring [9].

At the level of clinical decision support systems, a systematic review has evaluated the potential application of artificial intelligence and NLP technology in electronic medical records, pointing out that NLP engines can effectively improve the accuracy of clinical decision support systems, and a two-stage system combining AI algorithms and manual review may optimize clinical workflows [10,11]. There are also reviews that provide a systematic overview of deep learning and large-scale language models in EHR modeling, providing a roadmap for AI driven clinical decision support research [12-15]. However, most existing tumor assisted decision-making systems rely on structured data input, and the systematic exploration of directly integrating event extraction results from unstructured medical records into the diagnosis and treatment decision-making process is still in its infancy.

In response to the above issues and shortcomings, this article adopts a medical event extraction model that integrates semantic features based on previous work, applies it to the analysis and processing of tumor electronic medical records, and designs and implements a visualization system for tumor assisted diagnosis and treatment. The main tasks include the following aspects:

(1) Using a pre trained fusion semantic feature medical event extraction model, automatically identify key event information in tumor medical records, including tumor primary site, lesion size, tumor grading, metastatic site, and patient basic attributes.

(2) Based on the results of event extraction, a tumor assisted diagnosis and treatment system was designed, which includes core functional modules such as tumor grading recommendation, metastasis site detection, personalized treatment plan formulation, prognosis evaluation, and case data statistical analysis.

(3) A visual user interaction interface was built based on Qt Designer, which supports end-to-end automated processes such as medical record text input, event extraction model calling, diagnosis and treatment suggestion generation and evaluation. The usability of the system was verified through practical application examples.

3. Application of Tumor Event Extraction

3.1 Tumor Adjuvant Diagnosis and Treatment

Tumor event extraction technology can provide key information support for tumor assisted diagnosis and treatment. Tumor assisted diagnosis and treatment refers to the comprehensive use of modern technologies such as artificial intelligence, machine learning, and big data to systematically analyze and evaluate the patient's condition, in order to assist doctors in completing diagnosis, treatment, and prognosis judgment. This model focuses on personalization and aims to improve the accuracy and effectiveness of diagnosis and treatment, ultimately improving the quality of life of cancer patients.

In practical diagnosis and treatment, doctors need to collect multidimensional clinical data such as imaging, pathology, physiology, and genetics, and use intelligent technology for analysis to determine the optimal diagnosis, treatment, and prognosis plan. Specifically, it includes: analyzing clinical and imaging data to determine the presence, type, location, size, and depth of invasion of tumors; Develop personalized treatment plans based on tumor type, staging, histological characteristics, and genetic variations; Preoperative use of chemotherapy, radiation therapy and other auxiliary methods to reduce the size of the lesion, lower surgical risks and postoperative complications; Continue adjuvant therapy after surgery to reduce recurrence and metastasis; And conduct long-term follow-up and prognosis evaluation of patients, timely detect and treat recurrence or metastasis. The above methods enable doctors to dynamically adjust individualized plans, thereby reducing treatment side effects and improving patients' quality of life.

The specific methods of adjuvant therapy vary depending on the patient's disease type, lesion location, and severity of the condition. Common types include chemotherapy (drugs kill tumor cells), radiotherapy (ionizing radiation), immunotherapy (mobilizes the immune system), hormone therapy (hormone inhibits tumor growth), and blood or stem cell transplantation. The dosage and duration of various treatments need to be determined individualized. Based on the electronic medical record of tumors, the following plan can be planned: surgical resection is determined according to the primary site and lesion size; Select radiotherapy based on the primary site and the presence or absence of metastasis; Develop chemotherapy plans based on the primary site and metastatic status; And targeted therapy targeting the primary site and metastatic characteristics to inhibit tumor cell growth and spread.

3.2 Design of Tumor Diagnosis and Treatment Functions

Based on the characteristics of tumor electronic medical records, this article designs an auxiliary diagnosis and treatment system, whose core functions include tumor grading, metastasis site detection, treatment plan formulation, prognosis evaluation, and data analysis and statistics. Specifically, the system classifies tumors based on their primary location and lesion size, providing reference for treatment; By analyzing imaging data, determine whether the tumor has metastasis and the specific location and quantity of metastasis; Generate personalized treatment recommendations based on the tumor type, grading, metastasis status, and patient's physical condition; Based on the transfer status and physical condition, evaluate the prognosis and recommend appropriate treatment strategies; At the same time, mining and statistical analysis of the accumulated data within the system can generate valuable conclusions to assist clinical doctors in more effectively carrying out tumor treatment. Medical event extraction technology provides the system with fast and accurate judgment criteria, supporting precise auxiliary treatment decisions. The process of adjuvant therapy for tumors is shown in Figure 1.

In the initial stage of adjuvant therapy for tumors, the first step is to complete the input of electronic medical record information. After receiving medical record input, the system will automatically call the medical event extraction model to extract various attribute information related to tumors and perform corresponding processing. Based on these attributes, the system combines multiple factors such as tumor grading, metastasis detection, and whether the lesion size meets the surgical resection criteria to conduct a comprehensive analysis and preliminarily develop a corresponding treatment plan.

In the process of formulating the plan, the system will comprehensively evaluate the applicability of various treatment methods such as radiotherapy, chemotherapy, and targeted therapy, and simultaneously carry out prognosis evaluation to determine the optimal treatment strategy. During the implementation of treatment, the system continuously monitors the patient's treatment response and conducts statistical analysis on relevant data to support dynamic adjustment of treatment plans and provide reference for doctors' clinical decision-making. Overall, the system aims to provide patients with a better diagnosis and treatment experience and achieve optimal treatment outcomes through intelligent auxiliary means.

3.3 Implementation of Tumor Assisted Diagnosis and Treatment

The tumor assisted diagnosis and treatment system is supported by a medical event extraction model that integrates semantic features. By analyzing and processing patient tumor medical record data, it provides decision-making assistance for doctors, aiming to improve the accuracy and efficiency of diagnosis and treatment. The system is based on the medical event extraction model that integrates semantic features proposed in Chapter 3 of this study. It can automatically call the trained model for information parsing and processing based on user input of tumor patient medical records. The system architecture shown in Figure 2 allows users to quickly obtain the required tumor event information and corresponding diagnosis and treatment recommendations by performing simple input operations.

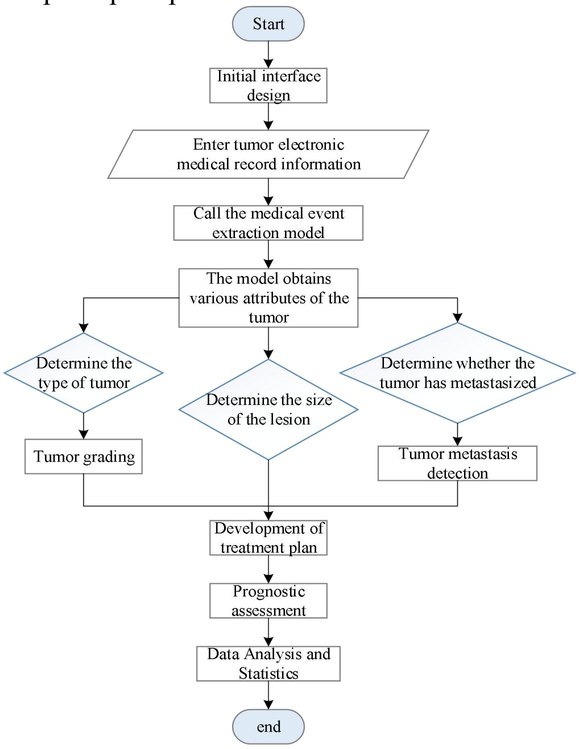


Figure 1. Tumor Adjuvant Diagnosis and Treatment Process

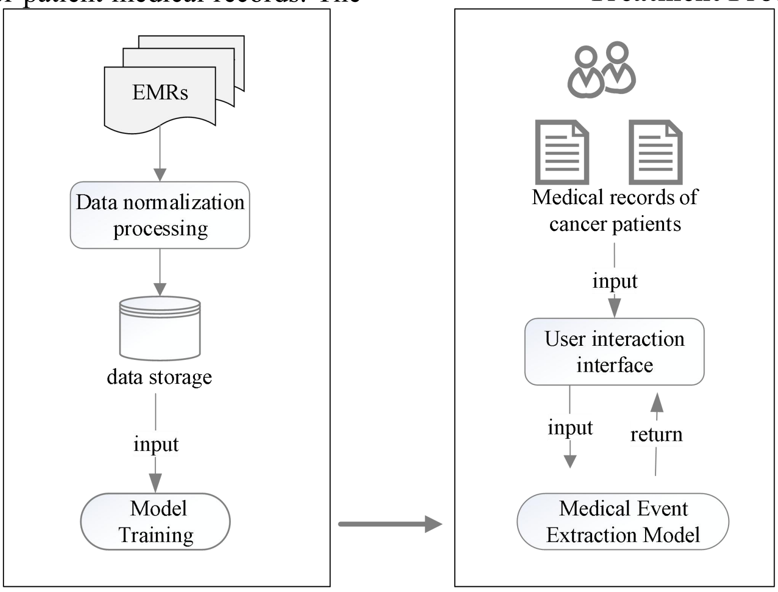


Figure 2. Architecture of Auxiliary Diagnosis and Treatment System

The system consists of three major modules: preprocessing, event extraction, and decision support. The preprocessing module is responsible for cleaning, annotating, and standardizing tumor medical record data to improve the accuracy and efficiency of subsequent event extraction. The event extraction module serves as the core to conduct in-depth analysis of Chinese tumor medical record data, extracting key information such as tumor type, size, location, grading, and metastasis status. The decision support module is based on the extraction results, combined with clinical practice and the latest research results, to provide doctors with complete diagnosis and treatment recommendations, and supports human-computer interaction for flexible adjustment of plans. The application of this system can significantly improve the effectiveness of tumor diagnosis and treatment, alleviate patient pain, and provide strong support for clinical practice and research.

In addition, the system is divided into two parts in terms of structure: model training and auxiliary diagnosis and treatment. The model training part is based on the PyCharm environment, which normalizes the electronic medical record text and stores it in the database. It is then used for training and parameter optimization of the medical event extraction model. The auxiliary diagnosis and treatment part is reflected in the user interaction interface: after the user inputs medical record information, the system automatically selects the optimal model for processing, outputs prediction results and diagnosis and treatment suggestions. This model has been pre trained with a large number of tumor medical records and corresponding diagnostic results.

4. User Interaction Interface

The visual interface of the auxiliary diagnosis and treatment system consists of three main modules: initialization module, model training module, and auxiliary diagnosis and treatment module. The initialization module is used to display the basic functions and usage instructions of the system, helping users quickly understand and become proficient in operating the system; The model training module automatically extracts various attributes of tumors by calling the medical event extraction model; The auxiliary diagnosis and treatment module covers the functions of developing diagnosis and treatment plans, evaluating plans, and analyzing and statistically analyzing tumor data. For each module's role in the diagnosis and treatment process, the system has carried out targeted visual design to fully meet the actual needs of tumor adjuvant treatment.

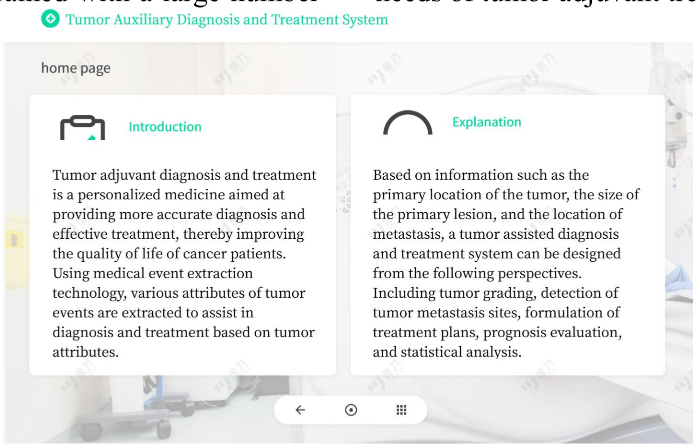
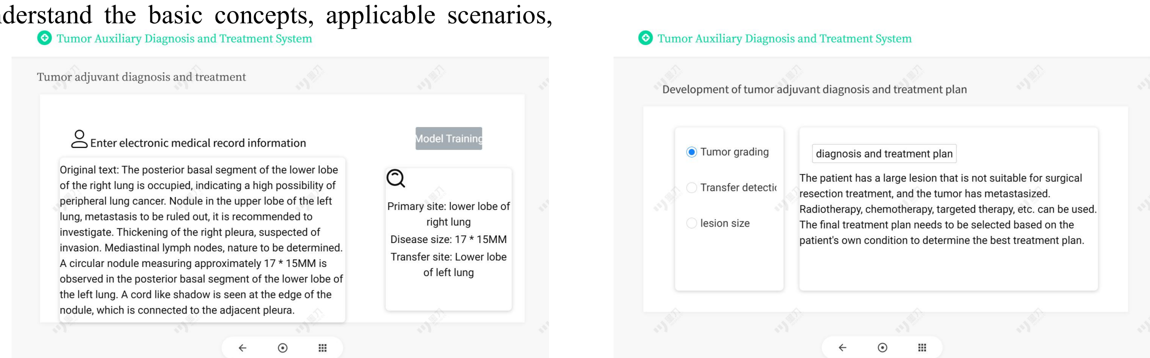


Figure 3. Initialization Interfacer

The main function of this interface is to introduce the functional features of the tumor assisted diagnosis and treatment system to users, and provide corresponding operation instructions. With the help of this interface, users can comprehensively understand the basic concepts, applicable scenarios, and advantages of the system, clarify how the system can help doctors improve the accuracy and work efficiency of diagnosis and treatment, and thus bring better medical services and nursing support to patients.

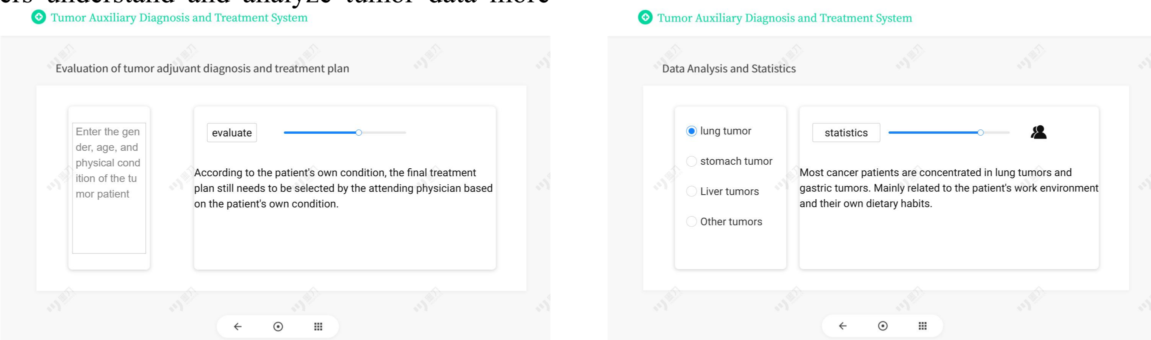


(a)Auxiliary Diagnosis and Treatment Interface (b)Development of Diagnosis and Treatment Plan

Figure 4. Auxiliary Diagnosis and Treatment and Plan Formulation

Figure 4 shows the interface of the system's auxiliary diagnosis and treatment, as well as the diagnosis and treatment plan. In the auxiliary diagnosis and treatment model training interface shown in Figure 4 (a), the user needs to manually input the patient's electronic medical record information, and then the system starts training the medical event extraction model to extract various attributes of the tumor. The design of this interface needs to balance efficiency and accuracy, while providing corresponding data visualization tools to help users understand and analyze tumor data more intuitively.

In the auxiliary diagnosis and treatment plan development interface shown in Figure 4 (b), the system can integrate multi-dimensional information such as tumor grading, metastasis detection results, and lesion size to provide the optimal diagnosis and treatment plan. The design of this section should embody the concept of personalized treatment and be equipped with visual aids to enable users to develop more scientific and reasonable treatment plans based on understanding tumor data.



(a)Diagnosis and Treatment Plan Evaluation **(b)Data Analysis and Statistics**

Figure 5. Diagnosis and Treatment Plan Evaluation, Data Analysis, and Statistical Interface

Figure 5 shows the diagnosis and treatment plan evaluation interface as well as the data analysis and statistics interface of the system. In the diagnosis and treatment plan evaluation interface shown in Figure 5 (a), users can input the patient's gender, age, physical condition, and other information to evaluate the current diagnosis and treatment plan and obtain more suitable plan recommendations. The design of this interface needs to meet the needs of personalized diagnosis and treatment, while providing corresponding visualization tools to assist users in understanding tumor data and formulating optimal treatment strategies.

In the data analysis and statistics interface shown in Figure 5 (b), the system can perform statistical analysis and mining on accumulated diagnosis and treatment data, generate valuable analysis conclusions, and help clinical doctors improve the effectiveness of tumor treatment. The design of this section should focus on the visualization of data and the convenience of analysis operations, enabling users to easily explore data and infer conclusions. At the same time, it is necessary to attach great importance to the security and confidentiality of data, and ensure the accuracy and reliability of data while safeguarding patient privacy.

5. Conclusion

This article addresses the problem of unstructured text in tumor electronic medical records being difficult to directly use for clinical decision-making. Based on a medical event extraction model that integrates semantic features, a tumor assisted diagnosis and treatment visualization system is designed and implemented. The system can automatically extract key event information such as tumor type, lesion size, grading, and metastasis, providing personalized diagnosis and treatment plan recommendations, prognosis evaluation, and data statistics functions. Through a visual interactive interface, doctors can easily input medical records, call models, and adjust treatment plans, achieving an automated process from medical records to diagnosis and treatment recommendations. Experiments have shown that the system can effectively improve the accuracy and efficiency of tumor event extraction, assist doctors in quickly grasping the condition and optimizing decisions, and has good clinical practical value. The study verified the feasibility of medical event extraction technology in tumor assisted diagnosis and treatment, providing reference for the integration of natural language processing technology into clinical practice. Future work will expand high-quality annotated datasets to enhance model generalization ability, introduce knowledge graphs and large language models to improve the interpretability and personalization level of diagnosis and treatment recommendations, and explore integration deployment with existing hospital information platforms.

Acknowledgments

This study is supported by the Scientific Research Program of Shaanxi Provincial Department of Education (No. 24JK0333), the Special Research Project of Shaanxi Institute of International Trade & Commerce (No. SMXY202462), the University-level Research Project of Shaanxi Institute of International Trade & Commerce (No. SMXY25042), and the University-level Research Project of Shaanxi Institute of International Trade & Commerce (No. SMXY25043).

References

[1] Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 2024, 74(3): 229-263.

[2] Goyal N, Singh N .Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions. Neurocomputing, 2025, 618:129171.

[3] Bilal M, Hamza A, Malik N, et al. NLP for Analyzing Electronic Health Records and Clinical Notes in Cancer Research: A Review. Journal of Pain and Symptom Management, 2025.

[4] Kim M, Kim S, Park B .Leveraging BERT and Network-Based Attention Analysis for Identifying Treatment Milestones in EHRs. 2024 IEEE International Conference on Big Data (BigData), 2024:8707-8709.

[5] Ji X R, Wei D J, Zhang J Z, Zhang S, Cao H. Research progress on information extraction methods for Chinese electronic medical records. Computer Engineering and Science, 2024, 46(2): 325-337.

[6] An B, Zhang H, Ma L, Zhao Z. End-to-end Chinese clinical event extraction based on large language model. Scientific Reports, 2025, 15(1): 16943.

[7] Fu Y, Ramachandran G K, Halwani A, et al. CACER: Clinical Concept Annotations for Cancer Events and Relations. Journal of the American Medical Informatics Association, 2024: ocae231.

[8] Namasivayam G, et al. Use of large language models to extract cancer diagnosis, histology, grade, and staging from unstructured electronic health records. Journal of Clinical Oncology, 2025, 43(16_suppl): 11176.

[9] Yao J, Goldner E, Hochheiser H,et al. Systemic Anticancer Therapy Timelines Extraction From Electronic Medical Records Text: Algorithm Development and Validation. JMIR Bioinformatics & Biotechnology, 2025, 6(1).

[10]Gangopadhyay S, et al. Clinical Decision Support and Natural Language Processing in Medicine: Systematic Literature Review. JMIR Medical Informatics, 2024.

[11]Ren W, Zhu J, Liu Z, et al. A Comprehensive Survey of Electronic Health Record Modeling: From Deep Learning Approaches to Large Language Models. 2025.

[12]Fu Y, Ramachandran G K, Halwani A, et al. CACER: Clinical Concept Annotations for Cancer Events and Relations. Journal of the American Medical Informatics Association, 2024: ocae231.

[13]Yao J, Goldner E, Hochheiser H, et al. Systemic Anticancer Therapy Timelines Extraction From Electronic Medical Records Text: Algorithm Development and Validation. JMIR Bioinformatics & Biotechnology, 2025, 6(1).

[14]Delourme S, Redjdal A, Bouaud J, et al. Leveraging Guideline-Based Clinical Decision Support Systems with Large Language Models: A Case Study with Breast Cancer. Methods of Information in Medicine, 2024, 63(03/04): 085-096.

[15]Niu M, He Z, Xu X, et al. Liver cancer knowledge graph construction based on dynamic entity replacement and masking strategies RoBERTa-wwm-large-BiLSTM-CRF model with clinical Chinese EMRs. Frontiers in Artificial Intelligence, 2025, 8: 1663877.