Research on Physical Examination Item Recommendation System Based on Large Model and Knowledge Graph

Jiangtao Zhang[1], Hu Lin[1], Yiteng Meng[2], Xuemei Liang[1*]

[1]The 305th Hospital of PLA, [2]College of Intelligence and Computing,Tianjin University

No. A13 Wenjin Street, Xicheng District, Beijing, China, 100017

Abstract

The complexity and redundancy of electronic medical records (EMRs) pose significant challenges for personalized examination recommendations in clinical practice. To address these limitations, we present a novel knowledge-grounded framework that synergizes large language models (LLMs) with a comprehensive Chinese medical knowledge graph (KG). Our KG, constructed from vertical medical resources, comprises 44,111 entities (including 8,807 diseases and 3,353 examination items) and 294,149 clinically validated relationships, establishing explicit connections between diseases, symptoms, and diagnostic procedures. The framework operates through three phases: 1) Multi-source EMR information extraction and structuring, transforming raw EMR data into a unified structured medical history text , 2) Context-aware knowledge retrieval leveraging disease-examination relationships from the KG, and 3) Recommendation generation via Qwen-7B enhanced with structured clinical prompts. Evaluated on real-world inpatient cases from a hospital, our system achieves 91.6% recommendation accuracy, reducing redundant tests by 4.2% compared to LLM-only approaches . Notably, the KG enables interpretable reasoning paths (e.g., Diabetes → Polyuria → HbA1c Test). This work provides a practical paradigm for integrating static medical knowledge with adaptive patient contexts, significantly advancing precision medicine in resource-constrained clinical settings.

1 Introduction

With the rapid accumulation of medical data, effectively utilize these data to provide personalized health management and physical examination recommendations for patients has become an important topic in current medical research. The selection of physical examination items usually depends on multiple factors such as the patient's medical history, diagnosis, treatment experience, and potential disease risks. In this context, Electronic Medical Records (EMRs), as digital repositories of patient health information, play a pivotal role. However, the inherent complexity, multi-source heterogeneity, and information redundancy of EMRs pose

* Corresponding author: Xuemei Liang (Email: 59525209@qq.com);

significant challenges. Traditional recommendation methods based on EMRs often rely on static rules and clinical experience, which lack the flexibility and precision required for true personalization.

To overcome these limitations, the field has witnessed a technological evolution in medical recommendation systems. Early medical recommendation systems were predominantly based on rule engines, constructing static rules through expert experience or health guidelines to match patients' risk factors and recommend corresponding examinations or screening items. To overcome these limitations, the field has witnessed a technological evolution in medical recommendation systems. Early systems were predominantly based on rule engines, constructing static rules from expert knowledge or clinical guidelines to match patient risk factors with corresponding examinations. While straightforward, these approaches struggle to cover all potential scenarios and lack generalization capabilities for complex or novel cases. In recent years, deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been applied to automatically extract features from EMRs for diagnosis prediction and recommendation [1, 2, 3]. Although deep learning offers significant advantages in feature extraction, its "black-box" nature often results in poor interpretability, making it difficult for clinicians to trust and verify the reasoning process, which hinders its practical adoption in high-stakes clinical decision-making.

More recently, Large Language Models (LLMs) have demonstrated remarkable potential across various medical fields, bringing a new paradigm to recommendation systems with their powerful natural language understanding and generation capabilities [4, 5, 6]. However, the direct application of general-purpose LLMs in clinical settings is fraught with challenges, most notably the risk of generating factually incorrect information or "hallucinations" due to their lack of specific, validated medical knowledge [7]. Furthermore, training or fine-tuning specialized medical LLMs requires vast amounts of high-quality data and computational resources, and updating their internal knowledge base is often a time-consuming and expensive process [8, 9].

To address these challenges, combining LLMs with Knowledge Graphs (KGs) has emerged as a promising and mainstream research direction [10]. As a structured knowledge base, a KG can provide explicit, verifiable, and interpretable knowledge, serving as a reliable external source to ground the reasoning of LLMs. This synergy allows the system to leverage the LLM's language capabilities while ensuring the factual accuracy and logical coherence of the output by anchoring it to the structured facts within the KG. Existing studies have

successfully integrated KGs with recommendation algorithms for applications such as drug recommendation [11], diagnostic assistance [12], and treatment planning [13].

Despite these advances, most current research focuses on these areas, while work on recommending specific medical examination items (including health screenings and diagnostic tests) remains limited. This gap is particularly prominent in the context of Chinese EMRs, where unique linguistic characteristics, terminologies, and healthcare system workflows necessitate localized solutions. Few studies have combined large-scale LLMs with a comprehensive, domain-specific Chinese medical KG for this task. Therefore, there is an urgent need for a hybrid framework that can harness the power of LLMs while being grounded by a localized medical KG to provide interpretable and trustworthy examination recommendations.

To this end, this study proposes a novel knowledge-grounded framework that synergizes a large language model with a comprehensive Chinese medical knowledge graph to deliver personalized physical examination item recommendations. The objective is to automatically extract potential risks from patients' historical data and generate clinically sound, cost-effective, and explainable examination suggestions.

2 Related Work

2.1 Clinical Diagnosis and Prediction of Electronic Medical Records

Electronic Medical Record (EMR) data provide detailed medical information about patients, including symptoms, medical history, examination results, and treatment records, and are widely used in patient care, clinical diagnosis, and treatment [14]. Previous studies mainly focused on designing deep learning models for EMR data to solve downstream tasks such as disease diagnosis and risk assessment [15,16,17].

LLMs have shown impressive performance in multiple medical tasks, including disease diagnosis and prediction in EMRs. Generally speaking, the applications of large models in the medical field usually include doctor-patient dialogue diagnosis, semantic segmentation, medical image analysis, or drug treatment with large models having medical knowledge. For example, products include BenTsao [18] and Med-PaLM [19] for assisting treatment and diagnosis, PanGu drug model and HelixFold-Single for drug design, DSI-Net and MedLSAM for medical image segmentation, and PubMed GPT [20], ChatDoctor [21], etc. for doctor-patient communication. This trend is also prominent in the industrial sector, where

companies have developed specialized medical LLMs by pre-training or fine-tuning on vast corpora of medical texts. For instance, Unisound's "Shanhai" and iFLYTEK's "Spark" medical large models are designed to enhance their capabilities in clinical scenarios like medical record generation and doctor-patient dialogues. Similarly, open-source contributors like Baichuan Inc. have released models such as Baichuan-M2, which achieve state-of-the-art performance on medical reasoning benchmarks through advanced training techniques like phased reinforcement learning [22]. These models integrate the latest medical knowledge, provide personalized treatment plans, accelerate the drug research and development process, improve the accuracy of image analysis, and enhance doctor-patient communication. The application situations of some medical large models are shown in Table 1:

Table 1. Applications of Medical Large Models

Large Model Name	Release Unit	Application Scenario	Data Type
Pangu Drug Molecule Large Model	Huawei Cloud Computing Technology Health Intelligence Laboratory	Drug Research and Development	Multimodal
BioMedLM	Stanford Center for Foundation Models	Medical Question Answering	Text
HealthGPT	Dingdang Health Technology Group Co., Ltd.	Drug Consultation, Nutrition Guidance, Health Advice	Text
HuaTuo GPT	Shenzhen Big Data Research Institute	Health Consultation, Medical Guidance, Emotional Companionship	Multimodal
Med-PaLM	Google	Medical Question Answering	Text
ChatDoctor	Hangzhou Dianzi University	Doctor-Patient Dialogue under Information Retrieval	Text

Investigations have found that the accuracy of physical examination item recommendations based on large models mostly depends on the performance or accuracy ability of the models. Most studies focus on English EMR datasets such as MIMIC-III [23], which mainly contains ICU data and may not be sufficient to model mild cases, rehabilitation, or routine treatments.

Research on Chinese EMR datasets is still limited.

2.2 General Knowledge Graph-Enhanced LLMs

Knowledge graphs have advantages in dynamic and explicit structured knowledge representation and storage, and are easy to add, delete, modify, and query [10], which has aroused great interest among researchers in combining knowledge graphs with large language models. A typical paradigm is to incorporate knowledge graph triples into the training data during the training stage and obtain their embedding representations through graph neural network modules [24,25,26]. However, LLMs usually have large-scale requirements for pre-training corpora, and it is both difficult and expensive to find or create knowledge graphs of matching scale [27].

In recent studies, researchers have tried to combine knowledge graphs with LLMs through prompts [27,28,29]. They usually identify entities in the input text and locate the corresponding triples or subgraphs in the knowledge graph. These triples or subgraphs will be converted into natural language [27], entity sets [28], or reorganized triples [29], etc., and then combined with the input prompts to provide additional knowledge for LLMs. Another method is to use an iterative strategy, in which the LLM acts as an agent and gradually reasons on the knowledge graph until enough knowledge is obtained or the maximum number of iterations is reached [14,30]. However, this method is more suitable for shorter questions. In scenarios with longer contexts, larger knowledge graph scales, and more complex structures, it may lead to excessive interaction of the LLM and failure to find the correct path.

In this field, researchers have explored various methods: Jiang et al. (2023a) [31] used LLMs and biomedical knowledge graphs to construct patient-specific knowledge graphs and adopted a Bidirectional Attention Augmented Graph Neural Network (BAT GNN); RAM-HER [14] converted multiple knowledge sources into text format and used retrieval enhancement and consistency regularization for co-training; DR.KNOWS [32] combined the knowledge graph constructed using the Unified Medical Language System (UMLS) and the graph model based on clinical diagnosis reasoning to improve the accuracy and interpretability of diagnosis; REALM [33] integrated clinical notes and multivariate time series data, adopted LLMs and RAG technology, and used an adaptive multimodal fusion network.

2.3 Knowledge Graph-Enhanced Medical Recommendation Systems

In medical recommendation, relying solely on LLMs may lead to "hallucinatory" suggestions due to the lack of domain constraints; while pure KG methods struggle to process unstructured EMR texts. The integration of the two can balance empirical knowledge and

language understanding capabilities. For instance, some studies have leveraged medical knowledge graphs (such as drug-disease-gene association networks) and combined graph embedding with sequence models for precise drug recommendation [11]. Other studies have integrated entity relationships in knowledge graphs with attention mechanisms to provide auxiliary explanations for clinical imaging diagnosis [12]. In addition, with the breakthroughs of large language models in natural language understanding and generation, some studies have begun to explore the fusion of knowledge graphs and large language models to improve the knowledge accuracy and interpretability of medical question-answering and recommendation systems. For example, Zhao et al. proposed a knowledge graph-enhanced prompting framework, which injects structured medical concepts into large language models to improve the accuracy and credibility of treatment plan recommendations [13].This hybrid approach is also being actively pursued in industrial applications. For example, Yidu Tech has developed a proprietary medical LLM trained on a massive, multi-dimensional knowledge graph constructed from real-world medical records. This demonstrates a practical application of using KGs to ground LLM-based systems for tasks like clinical decision support and intelligent inquiry.

Currently, the vast majority of KG-LLM applications focus on diagnosis, question-answering, and treatment plans, with few systems specifically designed for physical examination item recommendation. On one hand, physical examination recommendation needs to extract potential risks from multi-source EMRs; on the other hand, it also needs to combine the disease-examination mapping relationships in the KG to generate interpretable examination suggestions. Existing studies have not yet constructed a medical KG covering complete physical examination items in the Chinese EMR scenario and linked it with LLMs to generate recommendations, nor have they demonstrated its interpretable reasoning chains through real cases.

3 System Design and Methods

3.1 Overall System Architecture

This study proposes a medical examination recommendation system based on collaborative reasoning between large language models (LLMs) and knowledge graphs (KGs). As shown in Figure 1, the framework follows the logical flow of "information acquisition-knowledge enhancement-intelligent recommendation-evaluation feedback," operating primarily through the following four stages:

Multi-source EMR Information Extraction and Structuring (A): The system first performs

multi-source information extraction (such as structured tabular data and semi-structured text) and preliminary structuring on the patient's Electronic Medical Record (EMR) data. At this stage, fragmented raw EMR data are integrated into a unified "structured medical history text".

Direct Diagnosis Prediction (B - Auxiliary or Comparative Pathway):As an auxiliary validation or comparative baseline for model performance, this pathway allows the LLM to directly attempt diagnosis prediction from the preliminarily structured EMR information.

Entity Recognition and Knowledge Graph-Enhanced Reasoning (C.1 & C.2):Based on the structured medical history text, the system uses the LLM for medical entity recognition and keyword extraction (C.1). These keywords are then used for context-aware knowledge retrieval to obtain associations between diseases and medical examination items from the constructed Chinese medical knowledge graph, followed by deep reasoning combined with the LLM (C.2).

Recommendation Generation and Calibration (D):Patient information extracted from the EMR is combined with knowledge graph-enhanced reasoning results and input into the large language model to generate personalized medical examination recommendations and their rationales through structured clinical prompts. Meanwhile, the system introduces a manual calibration and feedback mechanism to continuously optimize the recommendation strategy. The following sections will elaborate on the design and implementation of each module.

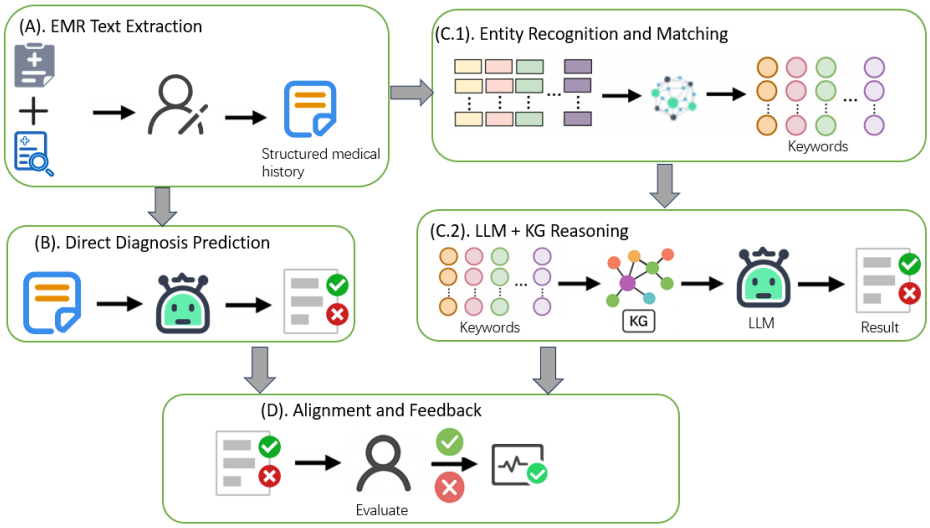


Figure 1. Overall Workflow

3.2 Information Extraction from Electronic Medical Records and Direct Diagnosis by LLMs

Considering that EMRs contain a large amount of redundant information, which will interfere with the diagnosis process if used directly, we first designed a series of rules to extract key information from electronic medical records. The data sources mainly include the following parts: Basic hospitalization information: patients' gender, age, inpatient department, number of hospitalizations, etc. Diagnostic information: patients' disease history, including primary diagnosis and related diagnoses. Surgical information: patients' surgical history and related records. Examination reports: various examinations (such as blood routine, imaging examinations, etc.) conducted during the patient's hospitalization and their results. External data: consultation reports after patients' outpatient visits. Table 2 provides a summary of these data sources and their main features.

Table 2. Data Sources

Electronic Spreadsheet Name	Description	Main Features
00-Sample Serial Number.xlsx	Patient Identifier and Metadata	Admission Number, Unique Patient Identifier, ETL Batch ID
01-MED_INP_INFO.xlsx	Basic Hospitalization Information	Age, Gender, Admission/Discharge Date, Department
02-MED_INP_ORDER.xlsx	Inpatient Doctor's Order Information	Order Item Name
03-MED_HP_BASE.xlsx	Basic Information of the Front Page of Inpatient Medical Records	Medical Record Type
04 MED_HP_DIAG.xlsx	Diagnosis Information on the Front Page of Inpatient Medical Records	Detailed Patient Medical History, Diagnosis Code, Diagnosis Description, Diagnosis Date
05-MED_HP_SURG.xlsx	Surgical Information on the Front Page of Inpatient Medical Records	Surgical Name, Surgical Duration, Surgical Date
07-MED_EMR_FILE_INFO.xlsx	Information about Electronic Medical Record (EMR) Files	File Type, Creation Date, Content Summary

Electronic Spreadsheet Name	Description	Main Features
11-MED_EXAM.xlsx	Imaging Examinations and Other Diagnosis Reports	Examination Type, Result, Interpretation
12-MED_LAB_TEST.xlsx	Main Items of Laboratory Test Records/Reports	Test Name
13-MED_LAB_RESULT.xlsx	Results of Laboratory Sub-items ((Microbiological Tests)	Test Result, Unit, Reference Range, Date and Time of the Result

All this information needs to be extracted from structured spreadsheets, converted into natural language text, and combined with patients' health records to form input data. A complete sample of the input data is detailed in Appendix A.

After obtaining the above preliminary structured text, we use a large language model (LLM) for direct reasoning and prediction (corresponding to Stage B in Figure 1). Examples of LLM prompt words can be found in Appendix B.1.

3.3 Collaborative Reasoning between Knowledge Graph and LLM

To achieve personalized medical examination recommendations, we designed a collaborative reasoning module based on a Knowledge Graph (KG)-enhanced Large Language Model (LLM). This module aims to deeply integrate patients' structured medical history information with comprehensive medical knowledge graphs and provide accurate and interpretable recommendations through the powerful reasoning and generation capabilities of LLMs.

3.3.1 Knowledge Graph Construction

To support querying and reasoning in the Retrieval-Augmented Generation (RAG) model, we first constructed a comprehensive Chinese medical Knowledge Graph (KG). This knowledge graph structurally stores a large number of medical concepts and their associations in the form of entities (nodes) and relationships (edges). The core schema of our KG is illustrated in Figure 2.

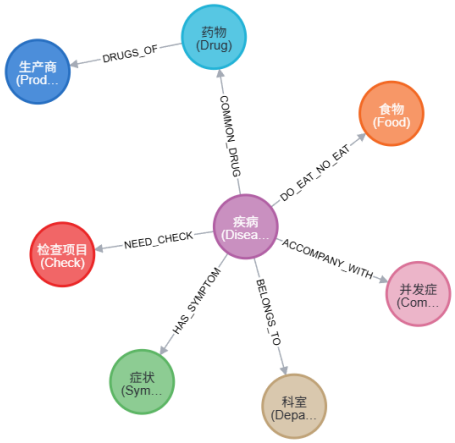


Figure 2. The schema of our Chinese Medical Knowledge Graph

The graph illustrates the core entity types (nodes) and the semantic relationships (edges) connecting them. The Disease entity serves as the central hub, linked to related symptoms, recommended examinations, drugs, and potential complications.

This knowledge graph is constructed based on existing medical data and standards, as well as professional medical websites such as "Xunyiwenyao" (a well-known Chinese medical consultation platform). We invited 2 physicians with more than 5 years of clinical experience to review the schema of the KG and 200 randomly sampled relational triples of the 'need_check' type, aiming to verify their clinical relevance and accuracy. Based on the experts' feedback, only 3 of these relationships were revised and supplemented.

3.3.2 Collaborative Reasoning Workflow

The core recommendation generation process of this system (corresponding to stages C.1 and C.2 in Figure 1) is a multi-step collaborative reasoning workflow. It aims to generate medical examination recommendations by integrating individual patient information with structured knowledge from the knowledge graph.

In stage C.1, the structured text from multi-source EMRs and natural language descriptions are used as the medical history summary P_{EMR}. Using $\mathrm{LLM}_{\mathrm{ER}}$, medical entities such as diseases and symptoms are identified in P_{EMR}, then deduplicated and aggregated into a keyword set:$K = \{k_1, k_2, \ldots, k_m\}$.

After obtaining the keyword set K, we perform context-aware knowledge retrieval from the KG using these keywords to acquire medical facts most relevant to the patient's current condition. In stage C.2, the keyword set K and the complete medical knowledge graph $G = (E, R)$ (where E is the entity set and R is the relationship set) are utilized.For each $k_i \in K$, a subgraph retrieval is executed in the graph database to obtain triples directly connected to k_i and related to physical examination items:

$$T = \bigcup_{k_i \in K} \{(h, r, t) \in G \mid h = k_i \wedge r = suitable\_exam\}$$

These triples are then filtered by expert rules to retain the top 5 most diagnostically valuable paths.

The output of this step is $T_{\text{retrieved}}$, representing the set of relevant triples retrieved from the KG. These triples typically include associations between diseases and examination items, symptoms and diagnostic procedures, etc. At this point, the retrieved triple set is $T_{\text{retrieved}} = \{(k_i, r_j, e_j)\}$.

In the final recommendation generation stage, the system integrates the patient's structured medical history P_{EMR}, the knowledge retrieved from the graph $T_{\text{retrieved}}$, and the guiding prompt $\text{Prompt}_{\text{Rec}}$ (see Appendix B.2), then inputs them into the LLM_{Gen} (Generation) module. LLM_{Gen} synthesizes this information to generate personalized examination recommendations and detailed justifications R. This process can be formalized as:

$$R = \text{LLM}_{\text{Gen}}(P_{\text{EMR}}, T_{\text{retrieved}}, \text{Prompt}_{\text{Rec}}),$$

where R includes the list of recommended physical examination items and their rationales.

This corresponds to the second half of stage (C.2) in Figure 1, where the large language model combines KG-derived information to generate the final output. Through this formalized process, the system ensures that the generated examination recommendations are not only based on the patient's actual health status but also on reasoning results derived from the knowledge graph [34]. This significantly enhances the accuracy, interpretability, and credibility of the recommendations.

3.3.3 System Evaluation and Optimization

To evaluate the effect of the system, we design the following evaluation indicators:

1.Recommendation Accuracy: Whether the physical examination items recommended by the system are in line with the patient's potential disease risks.

2.Generation Quality: Whether the recommendation reasons generated by the large model have sufficient medical basis and can provide useful decision support for doctors.

3.User Feedback: Through cooperation with doctors, collect feedback on the recommendation results and further optimize the recommendation strategy.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

MMEMR Dataset: This dataset is constructed from real patients' electronic medical record data collected by a hospital. It contains historical hospitalization records and consultation reports of 500 patients who were treated in the hospital during February and March 2023. Each patient's historical medical records, diagnostic information, surgical records, and examination reports are used as input data. Data quality is ensured through manual inspection, screening, and exclusion of records with issues or missing key information.
Figure 2 displays the key statistical distributions of the integrated MMEMR dataset, including the age distribution of patients, the top 10 most frequent disease diagnoses, the 10 most commonly performed medical examination items, and the 10 most common abnormal high-value laboratory indicators.

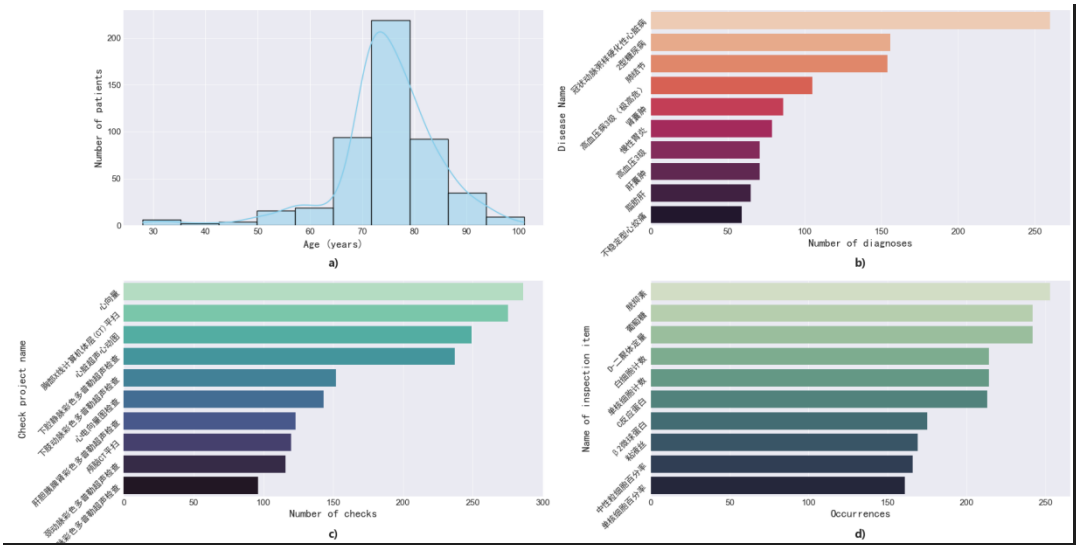


Figure 3. Statistical Information of the MMEMR Dataset

This figure displays four key statistical dimensions of the integrated MMEMR dataset. (a) A histogram of patient age distribution shows the age group distribution of patients. (b) A bar chart of the top 10 most common disease diagnoses reflects the primary disease types in the dataset. (c) A bar chart of the 10 most frequently performed medical examination items reveals patients' main examination needs. (d) A bar chart of the 10 most common abnormal high-value laboratory indicators indicates the prevalent physiological abnormalities in the patient population. These statistical insights provide a critical foundation for data understanding and model design in this study.

Additional statistical analysis of this dataset is included in Appendix C.

To ensure patient data privacy, this study implemented strict desensitization and de-identification procedures for all EMR data.

A total of 42 items of information that could directly or indirectly identify patients were located and marked in the original electronic spreadsheets, including name, ID card number, date of birth, home address, contact phone number, names and phone numbers of accompanying persons, as well as information about attending physicians, recorders, and

examination performers.

Each of the above sensitive fields was replaced with "desensitized" fields to ensure that no real personal information remained. Scripts were used to automatically scan all spreadsheets to verify that all sensitive fields had been processed; meanwhile, two independent auditors conducted manual reviews of sample data.

All data processing procedures comply with relevant medical data privacy regulations and ethical guidelines. Data access is strictly controlled and limited to authorized researchers only, and all analyses are conducted in a securely isolated environment to maximize the protection of patients' personal information security.

4.1.2 Benchmark Methods

This experiment is compared with the following benchmark methods:

- LLM-only: A way of reasoning relying only on the knowledge of Large Language Models (LLMs) themselves without using external knowledge, including methods such as CoT, ToT, and Sc-CoT.
- LLM ⊕ KG: A way of reasoning by combining LLM and Knowledge Graph (KG), including methods such as MindMap, ICP, HyKGE, and KG-rank [34].

4.1.3 Evaluation Indicators

Doctor Feedback: Through cooperation with doctors, collect feedback on the recommendation results, and calculate the indicators of the correct rate, deletion rate, and addition rate. The calculation methods of the three are as follows:

$$Correct\ Rate\ = \frac{Total\ Number\ of\ Correctly\ Recommended\ Items}{Total\ Number\ of\ Recommended\ Examination\ Items}$$

$$Deletion\ Rate\ = \frac{Total\ Number\ of\ Redundant\ Recommended\ Items}{Total\ Number\ of\ Recommended\ Examination\ Items}$$

$$Addition\ Rate\ = \frac{Total\ Number\ of\ Examination\ Items\ Needing\ to\ be\ Added}{Total\ Number\ of\ Recommended\ Examination\ Item}$$

4.2 Experimental Results

4.2.1 Overall Performance

The performance on this dataset is significantly better than the benchmark method (using the

LLM-only paradigm), proving its effectiveness in the EMR physical examination item recommendation task. We tested our method on the dataset, and the experimental results are shown in Table 3.

Table 3. Experimental Results

Methods	Correct Rate	Deletion Rate	Addition Rate
LLM-only	87.42%	12.58%	9.2%
LLM⊕KG	91.6%	8.39%	7.44%

4.2.2 In-depth Analysis

Impact of Knowledge Graph-Enhanced Prompts: By comparing different knowledge graph-enhanced prompt templates, it is found that the effect of using only prompts of related entities is poor because the relationship information in the knowledge graph is not utilized. Using prompts with complex structures such as reasoning chains and mindmaps may lead to an overly large prompt context, especially for small-scale models, which is likely to affect the reasoning effect.

4.3 Case Study Analysis

To more intuitively demonstrate the application effect and interpretability advantages of our proposed LLM⊕KG framework in actual clinical scenarios, this section will elaborate on the complete workflow of the system through a real, desensitized patient case.

We selected a complex case involving a patient with multiple chronic diseases. After preliminary information extraction and structuring (as described in Section 3.3), the system generated the patient's summary text (P_structured) as follows:

Partial summary text of the patient (P_structured) (the complete sample text is in Appendix A, but different from this case):

"Patient, male, 68 years old, hospitalized. Main diagnoses include: type 2 diabetes mellitus, grade 3 hypertension (very high risk), pulmonary nodules. Past medical history shows that the patient has been taking hypoglycemic and antihypertensive drugs for a long time, and blood glucose control is acceptable. Recent chest CT examination indicates an 8mm

ground-glass nodule in the upper lobe of the right lung. Laboratory test reports show a high level of C-reactive protein (CRP)."

After receiving the above patient summary, the system conducts collaborative reasoning in the following steps:

Step 1: Keyword Extraction

The LLM_ER model analyzes the patient summary, extracts key medical entities, and generates the keyword set K:

{ ["type 2 diabetes mellitus", "hypertension", "pulmonary nodules", "high C-reactive protein"] }

Step 2: Knowledge Graph Retrieval

The system uses the keyword set K to retrieve in our Chinese medical knowledge graph (KG) and obtain recommended examination items related to these diseases and abnormal indicators. The retrieved key knowledge paths T_retrieved are as follows:

(Hypertension) --[acompany_with]--> (Left ventricular hypertrophy) --[need_check]--> (Echocardiography)

(Hypertension) --[acompany_with]--> (Hypertensive nephropathy) --[need_check]--> (Urinary microalbumin test)

(Type 2 diabetes mellitus) --[acompany_with]--> (Diabetic retinopathy) --[need_check]--> (Fundus examination)

(Type 2 diabetes mellitus) --[acompany_with]--> (Diabetic nephropathy) --[need_check]--> (Urinary microalbumin test)

(Pulmonary nodules) --[need_check]--> (Chest CT)

Step 3: Recommendation Generation

The LLM_Gen model integrates the patient summary P_structured and the retrieved knowledge paths T_retrieved to generate the final physical examination item recommendations R, and provides interpretable reasons based on the knowledge graph for each recommendation.

Final recommended output of the system (R):

Recommendation 1: Echocardiography

Reason for recommendation: Based on the patient's diagnosis of "grade 3 hypertension (very high risk)", the knowledge graph indicates that hypertension is a major risk factor for "left ventricular hypertrophy". The echocardiography is recommended to evaluate the cardiac structure, especially whether the left ventricular wall is thickened, so as to clarify the degree of target organ damage to the heart caused by hypertension.

Recommendation 2: Urinary microalbumin test

Reason for recommendation: The patient suffers from both "hypertension" and "type 2 diabetes mellitus". The knowledge graph clearly points out that both diseases may lead to kidney damage (hypertensive nephropathy, diabetic nephropathy). Urinary microalbumin is a sensitive indicator of early renal injury, so this test is recommended to screen for early renal lesions.

Recommendation 3: Fundus examination

Reason for recommendation: According to the knowledge graph, "diabetic retinopathy" is a common and serious complication of "type 2 diabetes mellitus". To assess the impact of diabetes on fundus blood vessels and prevent visual impairment, an annual fundus examination is recommended.

Recommendation 4: Chest CT

Reason for recommendation: The patient has been found to have "pulmonary nodules". The knowledge graph recommends "chest high-resolution CT" as a key examination for qualitative diagnosis and follow-up. This recommendation aims to conduct a more accurate morphological evaluation of the nodules and compare with previous images to determine their dynamic changes.

Recommendation 5: Tumor marker screening (such as CEA, CYFRA21-1)

Reason for recommendation: Considering that the patient has "pulmonary nodules" and "elevated C-reactive protein" (which may be related to inflammation or tumors), tumor marker screening is recommended as an auxiliary means to assess the nature of pulmonary nodules.

5 Result Analysis

The experimental results show that the recommendation system combining knowledge graphs and large models can generate targeted physical examination item recommendations

according to the patient's medical history and potential disease risks. In particular, the system can generate recommendation reasons with clear medical basis according to the relationship between diseases and examination items.

5.1 System Performance

Recommendation Accuracy: The recommended physical examination items have a high degree of agreement with the actual recommendations of doctors, which can effectively improve the decision-making efficiency of doctors.

Processing Speed: The query speed of the system is relatively fast, and it can generate personalized physical examination recommendations within a few seconds.

5.2 Discussion

5.2.1 Analysis of System Performance

Our experimental results demonstrate that the proposed LLM ⊕ KG framework significantly outperforms the LLM-only baseline, achieving a higher recommendation accuracy while reducing the rate of redundant suggestions. The primary advantage stems from the integration of the structured Chinese medical knowledge graph. By grounding the LLM's reasoning process in a verifiable knowledge base, our system can generate recommendations that are not only accurate but also clinically coherent and explainable. As illustrated in the case study, the KG provides explicit reasoning paths (e.g., disease-complication-examination links) that enable the model to recommend proactive and targeted examinations, a capability often lacking in standalone LLMs that rely solely on patterns learned from text corpora. This knowledge-driven approach ensures that each recommendation is supported by a clear medical basis, which is crucial for building trust with clinicians.

However, the system's performance is also subject to certain inherent dependencies. The quality and coverage of the constructed knowledge graph are paramount; omissions or inaccuracies in the KG could directly impact the relevance and correctness of the recommendations. Furthermore, the final recommendations are highly dependent on the quality of the knowledge retrieved from the graph. If the initial keyword extraction is imprecise or the retrieved subgraphs are irrelevant, the LLM's generation quality will be compromised, highlighting the importance of the synergy between all components of the framework.

5.2.2 Limitations and Future Work

While our study validates the effectiveness of the proposed framework, we acknowledge several limitations that also present opportunities for future research.

First, our experimental evaluation was primarily conducted using Qwen-7B as the base large language model. Although Qwen-7B is a representative and high-performing open-source model, and our results have successfully demonstrated the feasibility of our knowledge-enhancement approach, testing on a broader range of LLMs is necessary to fully assess the generalizability of the framework. Due to significant computational resource constraints and the substantial expert involvement required for evaluation—each new model's output would necessitate a new round of review by physicians—we were unable to include additional base models (such as Baichuan or ChatGLM) in the current study. Future work will focus on extending our experiments to these and other models to investigate the framework's performance across different architectures and scales.

Second, a direct comparison with state-of-the-art (SOTA) methods was challenging. The lack of publicly available baseline models and standardized benchmarks specifically for the task of personalized physical examination item recommendation from Chinese EMRs prevented a like-for-like SOTA comparison. Our future work aims to contribute to the establishment of such benchmarks to facilitate more direct and comprehensive evaluations within the research community.

Finally, while our knowledge graph is comprehensive, the dynamic and ever-evolving nature of medical knowledge necessitates continuous updates. Future efforts will be directed towards establishing a semi-automated pipeline for KG maintenance and expansion, incorporating the latest clinical guidelines and research findings to ensure its long-term relevance and accuracy.

6 Conclusions

This study proposes a physical examination item recommendation system based on large models and knowledge graphs. By transforming complex, multi-source EMR data into a structured format and leveraging the KG for context-aware, interpretable reasoning, our system achieves high recommendation accuracy and reduces redundancy compared to LLM-only approaches. The presented case study further highlighted the framework's ability to generate clinically relevant and explainable recommendations, demonstrating its potential to serve as a valuable tool for assisting physicians in clinical decision-making.

This work provides a practical paradigm for integrating static, structured medical knowledge

with adaptive patient contexts, representing a significant step towards advancing precision medicine, particularly in resource-constrained clinical settings. Future research will focus on addressing the current limitations, including expanding the evaluation to more diverse LLMs, contributing to benchmark development, and enhancing the KG's maintenance mechanisms to further improve the system's robustness, reliability, and clinical utility.

References

[1] Noshad, M., Jankovic, I. and Chen, J.H., 2020. Clinical recommender system: predicting medical specialty diagnostic choices with neural network ensembles. arXiv preprint arXiv:2007.12161.

[2] Chenquan Dai, Xiaobin Zhuang, and Jiaxin Cai. 2023. Chinese Electronic Medical Record Named Entity Recognition Based on Bi-RNN-LSTM-RNN-CRF. In Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition (ICCPR '22). Association for Computing Machinery, New York, NY, USA, 577 – 583. https://doi.org/10.1145/3581807.3581892

[3] Xu, Jiabao, Xuefeng Xi, Jie Chen, Victor S. Sheng, Jieming Ma, and Zhiming Cui. 2022. "A Survey of Deep Learning for Electronic Health Records" Applied Sciences 12, no. 22: 11709. https://doi.org/10.3390/app122211709

[4] Hyunsu Lee. 2023. The rise of chatgpt: Exploring its potential in medical education. Anatomical sciences education.

[5] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to pa tient questions posted to a public social media forum. JAMA internal medicine, 183(6):589–596.

[6] Ashwin Nayak, Matthew S Alkaitis, Kristen Nayak, Margaret Nikolov, Kevin P Weinfurt, and Kevin Schulman. 2023. Comparison of history of present illness summaries generated by a chatbot and senior internal medicine residents. JAMA Internal Medicine, 183(9):1026–1027.

[7] Isaac A Bernstein, Youchen Victor Zhang, Devendra Govil, Iyad Majid, Robert T Chang, Yang Sun, Ann Shue, Jonathan C Chou, Emily Schehlein, Karen L Christopher, et al. 2023. Comparison of ophthalmol ogist and large language model chatbot responses to online patient eye care questions. JAMA Network Open, 6(8):e2330320–e2330320.

[8] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

[9] Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C Park, and Sung Ju Hwang. 2023b. Knowledge augmented language model verification. arXiv preprint arXiv:2310.12836.

[10] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Ji apu Wang, and Xindong Wu. 2024a. Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engi neering.

[11]Fan Gong, Meng Wang, Haofen Wang, Sen Wang, Mengyue Liu. 2021. SMR: Medical Knowledge Graph Embedding for Safe Medicine Recommendation. In Big Data Research, pages 100174.

[12]Xie, Y., Yang, B., Guan, Q., Zhang, J., Wu, Q. and Xia, Y., 2023. Attention mechanisms in medical image segmentation: A survey. arXiv preprint arXiv:2305.17937.

[13]Zhao, X., Liu, S., Yang, S.Y. and Miao, C., 2025, April. MedRAG: Enhancing Retrieval-augmented Generation with Knowledge Graph-Elicited Reasoning for Healthcare Copilot. In Proceedings of the ACM on Web Conference 2025 (pp. 4442-4457).

[14] Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May D Wang, Joyce C Ho, and Carl Yang. 2024. Ram-ehr: Retrieval augmentation meets clinical pre dictions on electronic health records. arXiv preprint arXiv:2403.00815.

[15] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. 2020. Stagenet: Stage-aware neural networks for health risk prediction. In Pro ceedings of The Web Conference 2020, pages 530 540.

[16] Ran Xu, Yue Yu, Chao Zhang, Mohammed K Ali, Joyce C Ho, and Carl Yang. 2022. Counterfac tual and factual reasoning over hypergraphs for in terpretable clinical predictions on ehr. In Machine Learning for Health, pages 259–278. PMLR.

[17] Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. 2023b. Hierarchical pretraining on multimodal electronic health records. In Proceedings of the 2023 Conference on Empirical Methods in Natural Lan guage Processing, pages 2839–2852.

[18] Yang, S., Zhao, H., Zhu, S., Zhou, G., Xu, H., Jia, Y., & Zan, H. (2024). Zhongjing: Enhancing the Chinese Medical Capabilities of Large

[19] Singhal, K., Tu, T., Gottweis, J. *et al.* Toward expert-level medical question answering with large language models. *Nat Med* (2025).

[20] Waisberg, E., Ong, J., Masalkhi, M. *et al.* GPT-4: a new era of artificial intelligence in medicine. *Ir J Med Sci* **192**, 3197–3200 (2023).

[21] Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus. 2023 Jun 24;15(6):e40895. doi: 10.7759/cureus.40895. PMID: 37492832; PMCID: PMC10364849.

[22] Dou, C., Liu, C., Yang, F., Li, F., Jia, J., Chen, M., Ju, Q., Wang, S., Dang, S., Li, T. and Zeng, X., 2025. Baichuan-M2: Scaling Medical Capability with Large Verifier System. arXiv preprint arXiv:2509.02208.

[23] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessi ble critical care database. Scientific data, 3(1):1–9.

[24] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129.

[25] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for lan guage understanding and generation. arXiv preprint arXiv:2107.02137.

[26] Rikui Huang, Wei Wei, Xiaoye Qu, Wenfeng Xie, Xi anling Mao, and Dangyang Chen. 2024. Joint multi facts reasoning network for complex temporal ques tion answering over knowledge graph. arXiv preprint arXiv:2401.02212.

[27] Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. arXiv preprint arXiv:2308.09729.

[28] Jiageng Wu, Xian Wu, and Jie Yang. 2024. Guiding clinical reasoning with large language models via knowledge seeds. arXiv preprint arXiv:2403.06609.

[29] Rui Yang, Haoran Liu, Qingcheng Zeng, Yu He Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and Irene Li. 2024. Kg rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. arXiv preprint arXiv:2403.05881.

[30] Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Suhang Wang, Yu Meng, and Jiawei Han. 2024. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. arXiv preprint arXiv:2404.07103.

[31] Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2023a. Graphcare: Enhancing healthcare pre dictions with open-world personalized knowledge graphs. arXiv preprint arXiv:2305.12788.

[32] Yanjun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, and Majid Af shar. 2023. Leveraging a medical knowledge graph into large language models for diagnosis prediction. arXiv preprint arXiv:2308.14321.

[33] Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, et al. 2024. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models.

[34] Jiang, Pengcheng, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha A. Kass-Hout, Jimeng Sun and Jiawei Han. “Reasoning-Enhanced Healthcare Predictions with Knowledge Graph Community Retrieval.” *ArXiv* abs/2410.04585 (2024): n. pag.

Appendix

Appendix A. Input data sample

性别男，年龄 81 岁，住院次数为 5，住院科室是东院保健五(呼吸与危重症医学)科。住院病案首页的诊断信息：肺动脉高压、呼吸困难、肺间质纤维化、心脏瓣膜病、重症肺炎、I 型呼吸衰竭、三尖瓣关闭不全、心功能 III 级（NYHA 分级）、阵发性心房颤动。住院病案曾做过手术有：无创呼吸机辅助通气(双水平气道正压[BiPAP])。

住院期间的全部检查记录有（示例为：检查项目名称：对应检查报告对应事实结果：对应报告医生的诊断信息）：

名称：下腔静脉彩色多普勒超声检查。结果：下腔静脉可探及处内透声好，血流充盈可，内径约 16mm，呼吸塌陷率 > 50%。提示右房压正常。。诊断建议：下腔静脉未见明显异常。

名称：床旁心脏超声心动图。结果：声窗差，结果仅供参考。1. M 超+二维超声+彩色多普勒+频谱多普勒： 心脏各房室腔内径正常，室壁厚度正常。 主动脉瓣呈三叶瓣，瓣膜增厚钙化，舒张期见微量返流，收缩期前向血流速度 1.0m/s。 二尖瓣结构正常，收缩期见微量返流，舒张期 E 峰 0.6m/s，A 峰 0.7m/s。 三尖瓣结构正常，收缩期见少量返流，返流 束最大压差 43mmHg，估测右房压 3mmHg，估测肺动脉收缩压 46mmHg。 肺动脉显示不清。2.组织多普勒：二尖瓣瓣环间隔 e'6.9cm/s，侧壁 e'7.2cm/s。3.彩色室壁运动分析：可探及处室壁运动未见明显异常。4.左室收缩功能评估：EDV：89ml， FS：30%，EF：58%, SV、CO、CI 均位于正常范围。 右室收缩功能评估: TAPSE: 18.9mm, 右室侧壁 TDI-PW 测 S'峰 9.2cm/s。5.左室舒张功能评估：左室舒张功能正常。。诊断建议：主动脉瓣钙化肺动脉高压（轻度）。

名称：胸部正位。结果：胸廓形态可，气管、纵隔未见明显偏移。双肺纹理增多，双肺见多发斑片影，局部网格状改变，双侧肺门未见增大。心影大小、形态未见异常。双侧膈面光滑，肋膈角锐利。。诊断建议：双肺间质纤维灶，不除外合并炎症，较 2023.02.02 日片未见明显变化，请结合 CT 检查。。

名称：心电向量图检查。结果：p:100ms QRS:86ms P-R:138ms QT/QTc:282/423ms QRS 电轴:71° Rv5/Sv1:0.88/0.52mv 心率:133bpm。诊断建议：1、窦性心律 2、窦性心动过速。

名称：胸部正位。结果：胸廓对称，气管、纵隔居中。双肺纹理增多，双肺野示弥漫性网格影。双侧肺门未见增大。心影增大。双侧膈面光滑，肋膈角锐利。。诊断建议：1、双肺弥漫性网格影，间质性纤维化？请结合临床 2、心影增大。

住院期间的验血或验便检验结果：其中体内含量低的成分有：血液中 O2 总浓度、白蛋白、阴离子间隙、钾、钙、PT 活动度、嗜酸性粒细胞百分率、球蛋白、淋巴细胞计数、血液渗透压、单核细胞百分率、a/AO2、氧分压、氧合血红蛋白、血小板计数、离子钙、氯、肺泡动脉氧分压、淋巴细胞百分率、肺泡动脉氧分压比、胆红素、肌酐。体内含量高的成分有：乳酸脱氢酶、血糖、可溶性细胞角蛋白 19 片段、还原血红蛋白、尿素氮、鳞癌抗原 SCC、缓冲碱、活化部分凝血活酶时间、神经元特异性烯醇化酶、RBC 体积分布宽度、二氧化碳分压、间接胆红素、中性粒细胞计数、热休克蛋白 90α、APTT 比值、直接胆红素、酵母菌计数、碱剩余、PT 国际标准化比值、中性粒细胞百分率、C 反应蛋白、粘液丝、D-二聚体定量、癌胚抗原、白细胞计数、PH 值 、透明管型、细菌计数、凝血酶原时间、肺泡动脉氧分压比、碳酸氢根、单核细胞计数。

Appendix B.1. Prompts Directly Recommended

You are a professional medical consultant, capable of recommending appropriate imaging examination items based on a patient's disease risks. Please adhere to the following principles:

* **Targeted:** Select highly specific examinations that directly address the patient's diseases.
* **Efficient:** Strictly avoid redundant or unnecessary examinations.
* **Cost-effective:** Prioritize economic and effective examination methods when diagnostic value is comparable.
* **Priority:** Order examination items based on their diagnostic value and the severity of the patient's condition.
* **Standardized Terminology:** Use standardized medical terminology.

Please recommend imaging examination items for the patient based on the following information:

Patient's Disease Risks: {diseases} (separated by semicolons)

Please reply in the following JSON format:

```json
{
  "diseases": [
    "Disease 1",
    "Disease 2",
    ...
  ],
  "Imaging Tests": [
    {
      "Examination Item": "Examination Item 1",
      "Purpose": "Purpose 1"
    },
    {
      "Examination Item": "Examination Item 2",
      "Purpose": "Purpose 2"
    },
    ...
  ]
```

Appendix B.2.Prompts for collaborative reasoning

You are a professional medical consultant. Based on the patient's disease risks, and **by fully utilizing the relevant medical knowledge retrieved from the knowledge graph provided below**, recommend the most appropriate imaging examination items for the patient. Your recommendations must be accurate, targeted, and explainable.

Please strictly adhere to the following recommendation principles:

* **Targeted:** Select highly specific examinations that directly address the patient's diseases and their known associations from the provided knowledge.
* **Efficient:** Strictly avoid redundant or unnecessary examinations, ensuring a concise and effective recommendation plan.
* **Cost-effective:** Prioritize economic and less invasive examination methods when diagnostic value is comparable.
* **Priority:** Order examination items based on their diagnostic value and importance for patient risk assessment.
* **Standardized Terminology:** Use standardized medical terminology.

Relevant Medical Knowledge Retrieved from Knowledge Graph:

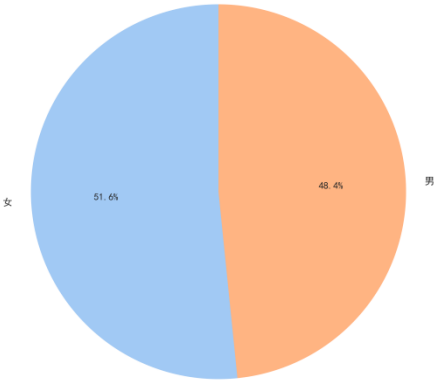
- **For "Hypertension"**: The knowledge graph suggests imaging examinations such as:
  - Cardiac Ultrasound: To assess heart structure and function, especially left ventricular hypertrophy.
  - Carotid Artery Ultrasound: To evaluate carotid atherosclerosis plaques and stenosis, understanding cerebral blood supply.
- **For "Diabetes"**: The knowledge graph indicates a potential for "Diabetic Nephropathy" as a complication. Suitable examinations for this complication include:
  - Kidney Ultrasound: To evaluate kidney size, morphology, and detect lesions.
- **For "Pulmonary Nodule" (Lung Nodule)**: The knowledge graph advises a **Chest CT (High-Resolution)** for characterization.
- **For "Elevated C-Reactive Protein (CRP)"**: The knowledge graph indicates a possible underlying inflammation or infection. Imaging tests may help locate the source of infection.

Patient's Disease Risks: {diseases} (separated by semicolons, e.g., Hypertension; Diabetes; Pulmonary Nodule)

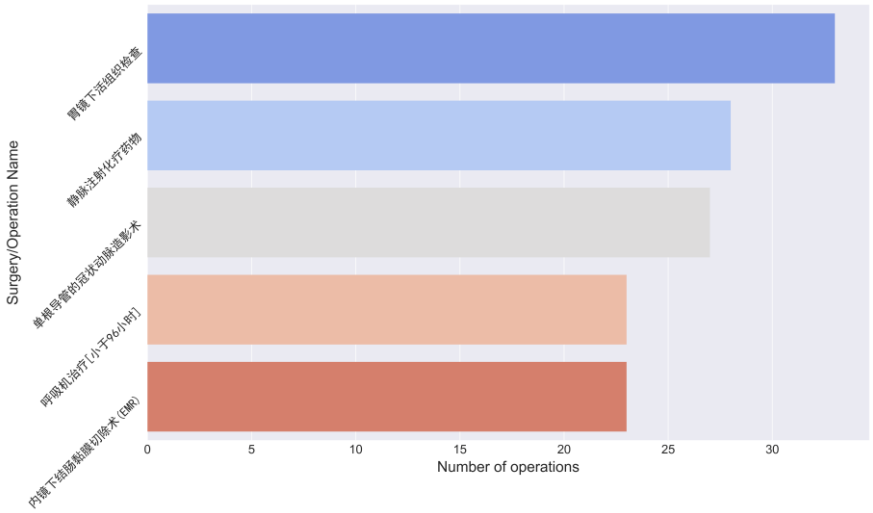
Please combine the above "Relevant Medical Knowledge Retrieved from Knowledge Graph" and the "Patient's Disease Risks" to recommend imaging examination items for the patient. Your response must be in the following JSON format, and clearly state the basis for each recommendation in the "Reason for Recommendation" field, specifically referencing how the knowledge graph information was used:

```json

Appendix C.1.Gender distribution of patients
```



Appendix C.2.Top 5 types of surgeries/procedures



Appendix C.3.Top 10 abnormally low value test indicators