Modeling and Simulation 建模与仿真, 2025, 14(10), 28-38
Published Online October 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.1410603

Hans汉斯

基于多源异构信息融合的肺癌专病知识图谱构建研究

雒增月，尹　裴

上海理工大学管理学院，上海

收稿日期：2025年8月30日；录用日期：2025年9月22日；发布日期：2025年9月30日

摘　要

肺癌是全球重大公共卫生挑战。医学知识图谱(MKG)可为智能诊疗提供关键支持，但现有图谱常面临信息源单一、覆盖不全、缺乏真实案例等问题。为此，本研究融合MIMIC-IV电子病历、DrugBank、PubMed、ICD-10等多源异构数据，构建肺癌专病知识图谱。创新性地采用模块化子图融合方法：先构建患者、疾病、药物三个子图，再通过实体对齐融合为总图谱。实验验证：1) 基于微调BioBERT的医疗实体识别模型性能优于基线；2) 利用TransE/TransH生成的图谱嵌入在药物/手术预测任务中，Top-3和Top-5命中率均≥92%。该图谱为肺癌临床决策提供了可靠知识支撑，其构建框架为多源医学数据融合与知识图谱构建提供了可复用的参考方案。

关键词

肺癌，知识图谱构建，多源信息融合，BioBERT模型

Research on the Construction of a Lung Cancer-Specific Knowledge Graph Based on Multi-Source Heterogeneous Information Fusion

Zengyue Luo, Pei Yin

Business School, University of Shanghai for Science and Technology, Shanghai

Received: August 30, 2025; accepted: September 22, 2025; published: September 30, 2025

Abstract

Lung cancer is a major global public health challenge. Medical Knowledge Graphs (MKG) can provide

文章引用: 雒增月, 尹裴. 基于多源异构信息融合的肺癌专病知识图谱构建研究[J]. 建模与仿真, 2025, 14(10): 28-38.
DOI: 10.12677/mos.2025.1410603

crucial support for intelligent diagnosis and treatment, but existing graphs often face issues such as single information sources, incomplete coverage, and a lack of real cases. To address these problems, this study constructs a lung cancer-specific knowledge graph by integrating multi-source heterogeneous data, including MIMIC-IV electronic medical records, DrugBank, PubMed, and ICD-10. It innovatively adopts a modular subgraph fusion approach: first constructing three subgraphs for patients, diseases, and drugs, then fusing them into an overall graph through entity alignment. Experimental verification shows that: 1) The medical entity recognition model based on fine-tuned BioBERT outperforms the baseline; 2) The graph embeddings generated using TransE/TransH achieve a hit rate of ≥92% for both Top-3 and Top-5 in drug/surgery prediction tasks. This graph provides reliable knowledge support for clinical decision-making in lung cancer, and its construction framework offers a reusable reference scheme for multi-source medical data fusion and knowledge graph construction.

Keywords

Lung Cancer, Knowledge Graph Construction, Multi-Source Information Fusion, BioBERT Model

1. 引言

肺癌是全球导致死亡的主要癌症类型之一，且伴随多种并发症，当下基于人工智能的方法在各医疗场景中的应用迅速增加，有助于肺癌的诊断和治疗[1]。然而医疗数据记录形式多样，缺乏组织、关系不清晰等问题阻碍了智能医疗的发展。医学知识图谱可以有效地提取、整合、表示、存储医疗知识，能有效解决以上问题。因此，研究构建有效高质量的医学知识图谱有重要意义。

现有知识图谱构建技术主要基于深度学习的方法。例如，Huang 等人[2]提出了常用的 Bi-LSTM-CRF 模型，广泛用于命名实体识别任务。通过长短期记忆网络(LSTM)和门循环单元(GRU)等各种神经模型从嵌入向量序列中捕获上下文，最后解码标签序列。但是使用深度学习方法训练模型时通常需要大量的标注数据与人工干预，这些都会极大程度影响所构建的知识图谱质量。

在大量通用数据上进行预训练的大语言模型具有强大的性能，GraphRAG 结合了图神经网络(GNNs)和生成模型通过图结构来建模和组织实体及其关系，从而实现知识图谱的自动构建[3]。然而所使用的通用领域预训练大模型，如 Llama 3 和 Qwen 2，缺乏领域专有知识，容易受到一词多义现象的影响，导致实体识别错误和关系抽取缺失等问题，影响知识图谱的准确性和完整性；而且在处理海量数据或使用强推理大模型时，计算成本剧增。

已有学者致力于医学知识图谱构建的研究[4] [5]，但是用于医疗场景的知识图谱数量有限，相关构建研究也较为匮乏并且面临多重挑战：1) 医疗知识分散于多源异构的医疗资源中，难以有效整合；2) 较少关注患者电子病历，使得医学知识图谱的应用难以满足患者个性化需求；3) 大多数研究缺乏对医学知识不确定性的量化；4) 现有大多数医学知识图谱中的实体关系过于繁琐，缺乏模块化结构，不利于图谱的修改、更新和补全。

为应对上述挑战，本文基于 MIMIC-IV 数据库中的电子病历，并结合 DrugBank、PubMed、ICD-10 等数据库搭建从多源异构医学文本中构建肺癌专病知识图谱的框架。研究贡献如下：

1) 通过微调医学领域的预训练语言模型，识别肺癌患者出院记录中的癌细胞组织类型和症状实体，

提升实体识别的全面性与准确性。

2) 在知识图谱构建中计算疾病相关三元组的概率属性，以此量化医学知识的不确定性。

3) 采用相似度和大语言模型结合的实体对齐方法，将 Drugbank、PubMed 等外部知识与 MIMIC-IV 数据库中的医学实体进行精确对齐，保证了多源数据的一致性和准确性。

4) 首先构建病人、疾病、药物三个模块化子图，再通过实体对齐构建总图，提升了知识图谱的可扩展性和灵活性。

5) 基于 TransE 和 TransH 嵌入进行肺癌相关的药物和手术预测任务，验证了所构建的知识图谱有效性，并为临床决策提供了潜在的辅助支持。

2. 数据集构建

2.1. 数据获取

本研究通过整合多源异构医疗数据构建肺癌专病知识图谱，数据来源包括：重症监护医学数据库 MIMIC-IV 提供的临床诊疗数据；PubMed 生物医学文献数据库收录的研究证据；DrugBank 药物数据库包含的药品信息；国际疾病分类(ICD-10)标准术语体系。上述多源数据的系统整合为知识图谱构建提供了全面的数据基础。

2.2. 数据预处理

本研究基于 ICD-10 编码 C34.x 提取出 1460 份患者的医疗记录用于筛选潜在肺癌病例，通过遍历每一条患者记录，生成其诊断代码与用药、症状和手术之间的所有可能组合，构建疾病－药物对、疾病－症状对、疾病－手术对。最终筛选提取出 676 名肺癌患者多次入院的医疗记录，患者组织类型和分期分布情况如表 1 所示。

Table 1. The number of lung cancer patients by histological type and stage
表 1. 肺癌患者组织类型和分期数量表

特征	分类	N = 676
Gender (性别)	Male/Female	676
Histology (组织类型)	Adenocarcinoma	223
	Squamous Cell Carcinoma	113
	Small Cell Carcinoma	106
	The Others	78
Stage (分期)	Stage I NSCLC	20
	Stage II NSCLC	32
	Stage III NSCLC	41
	Stage IV NSCLC	166
	Limited Stage SCLC	16
	Extensive Stage SCLC	30

3. 研究方法

基于上述预处理数据，本研究提出模块化融合的肺癌专病知识图谱构建方法如图 1 所示，首先在知

识抽取阶段，采用微调 BioBERT 对 MIMIC-IV 临床数据库的出院记录进行症状和组织类型两类医疗实体的识别。进而基于模块化构建策略，分别建立疾病、药物和患者三个核心子图。基于相似度与大语言模型结合的实体对齐方法将子图融合为总图。最后将构建的知识图谱总图使用基于距离约束的知识表示方法(TransE 和 TransH)嵌入到低维向量空间中，并使用药物预测任务和手术预测任务来验证评估所构建图谱的有效性。

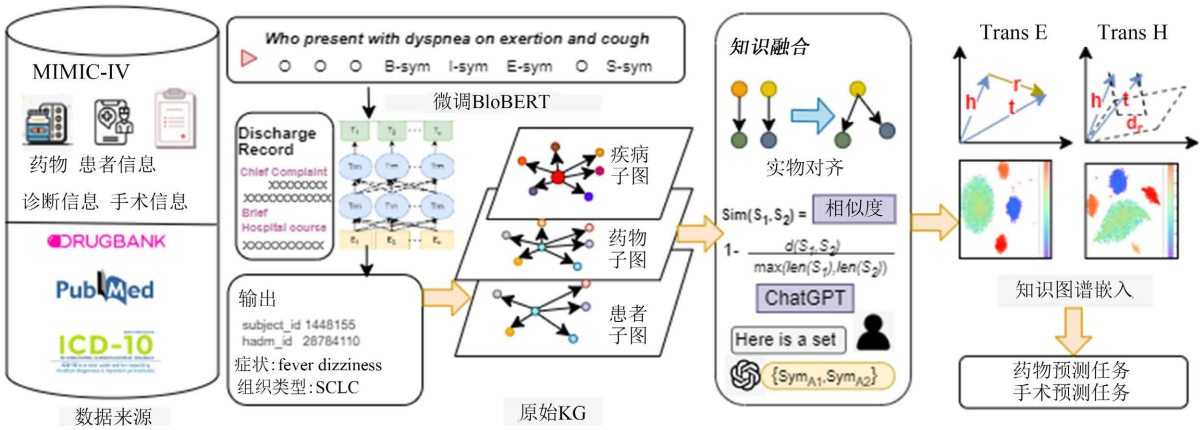

Figure 1. Construction flowchart of a lung cancer-specific knowledge graph based on multi-source heterogeneous information
图 1. 基于多源异构信息源的肺癌专病知识图谱构建流程图

3.1. 基于预训练语言模型的医疗命名实体识别模型构建

基于以上已经构建的数据集进行实体提取，针对结构化数据所包含的药物、诊断结果、手术等医学实体，根据 MIMIC-IV 数据库中的字段直接进行系统化的提取。针对非结构化的患者出院记录数据，使用微调后的 BioBERT 模型从患者信息、现病史、主诉、医嘱、住院经历概述等医学文本描述中提取癌细胞组织类型和症状两类医学实体。

基于 BioBERT 微调的医疗实体识别模型

BioBERT 是利用 Transformer 模型搭建的多层双向编码网络[6]，在大量的医学语料库中进行了预训练，具有丰富的医学术语和概念知识，因此能够较好地应用于各生物医学自然语言处理任务中。然而，其在特定临床场景下的表现仍存在两方面局限：1) 专病领域知识覆盖不足，特别是针对诊疗相关的细粒度实体(如组织学亚型分类)；2) 对电子健康记录(EHR)特有的叙述性文本结构适应性有限[7]。

BioBERT 模型支持小样本微调训练，可提升对自定义类别实体的识别效果。因此本研究采用 MIMIC-IV 临床数据库的出院记录进行领域自适应训练。从 Note 模块 Discharge 表格随机抽取 500 例肺癌患者的出院记录构建微调语料，并基于 BIOES 标注体系对组织类型和症状两类实体进行人工标注。

在 BioBERT 上下文编码层，给定医疗文本序列 $X=\{x_1,x_2,\cdots,x_n\}$ 通过 BioBERT 的 Transformer 编码器生成上下文感知的隐藏状态 h_t：

$$H = BioBERT(X) = [h_1, h_2, \cdots, h_n] \in R^{n\times h} \tag{1}$$

其中每个 h_t 的计算包含多头注意力机制，$Q,K,V \in R^{n\times d_k}$ 分别为查询、键、值矩阵。

$$Attention(Q,K,V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

在标签预测层模型需学习条件概率分布：

$$P\left(Y\left|X \rightleftarrows;\ \theta\right.\right)=\Pi_{t=1}^{n}P\left(y_t\left|h_t \rightleftarrows;\ \theta\right.\right)\ y_t\in\Gamma \tag{3}$$

$$\Gamma=\left\{B-sym,I-sym,E-sym,B-hist,E-hist,S-hist,O\right\} \tag{4}$$

为优化模型对医疗实体的捕捉能力，标注后的语料通过图 2 所示的流程输入 BioBERT 进行微调训练。

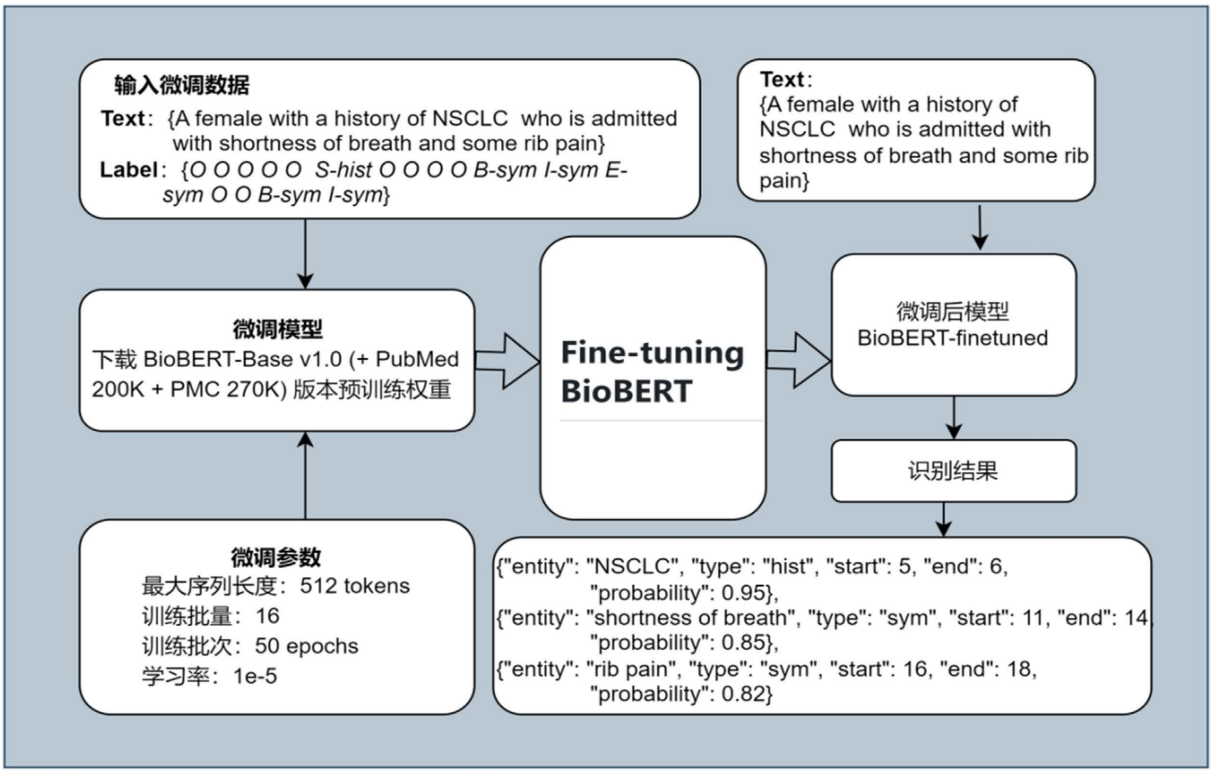

Figure 2. BioBERT fine-tuning flowchart
图 2. 微调 BioBERT 流程图

3.2. 基于模块化子图融合的医学知识图谱构建

3.2.1. 基于共现概率模型的疾病子图构建

医学领域中，疾病与临床表现、药物、手术等实体构成的三元组关系存在不确定性[8] [9]，例如考虑(pneumonia, disease_to_symptom, cough)这一知识三元组，由经验知识可得咳嗽是肺炎患者的常见症状，但并不是每一个诊断中都有“pneumonia”，即(ICD10: J18.901)的患者都表现出咳嗽的症状。因此，本研究在构建以疾病为中心的知识图谱时对三元组中关系的不确定性进行量化以增强图谱的准确性。

一名患者一次特定访问所产生的所有医疗记录中包括多个医疗实体即疾病、症状、药物、手术的具体实例。所有肺癌患者入院所产生的 EHR 数据可以被视为医疗访问记录的集合。在疾病知识子图中，三元组的头实体表示诊断的具体实例，疾病知识子图中每一个三元组$\left(h_i,r,t_{ij}\right)$概率的计算方法如公式(5)：

$$P\left(h_i,r,t_{ij}\right)=P_r\left(t_{ij}\left|h\right.\right)=\frac{N_{\text{co}}\left(h_i,r,t_{ij}\right)}{N\left(h_i\right)} \tag{5}$$

$N(h_t)$表示诊断结果中含有h_t的就诊次数，$N_{co}(h_i,r,t_{ij})$表示在所有访问记录中诊断结果h_i与疾病相关的医疗实体t_{ij}，在关系 r 下的共现次数。

肺癌患者通常同时患有各种共病，本研究通过在疾病子图中融入肺癌共病关系三元组，以识别疾病之间的潜在联系，帮助管理多疾病共存的复杂病例。以“comorbid lung cancer”为检索关键词借助 PubTator Central (PTC)生物医学文本挖掘平台，系统获取 PubMed 文献资源，包含 150 篇全文与 500 篇摘要构成的初始语料库。

由于文献中的疾病术语较为规范，因此我们直接将 PTC 标注结果输入 MetaMap，将它们与 UMLS 中的标准化概念对齐从而获得对应的 ICD-10 代码。基于共现频率筛选的高关联疾病 ICD 编码(阈值：共现频次 ≥ 15)构建肺癌核心共病知识子图，形成如(C34, Comorbidity, ICD_x)的三元组集合。

3.2.2. 基于结构化病例数据的患者子图构建

本研究基于结构化的表格病例构建以患者为中心的知识子图，患者的入院 id (Hadm_id)、年龄(Age)、性别(Gender)、APR (All Patient Refined)作为患者实体节点的属性。APR 考虑了患者的主要诊断、治疗、严重程度以及并发症等情况，因此能够更准确地反映出患者的病情。

3.2.3. 基于药物作用机制的药物子图构建

本研究基于从 DrugBank 药物数据库中提取的药物关系构建药物子图。在具有药物相互作用机制的两个药物实体之间，添加“Drug-Interaction”关系标签。并将经过语义识别系统处理的关系描述作为药物知识图谱中“Drug-Interaction”关系的属性进一步丰富图谱信息，为后续的图谱扩展、查询和推理提供了基础。

3.3. 基于语义相似度与大语言模型的实体对齐方法

在将子图整合为总图的过程中为了消除不同来源之间的冗余实体，需要对药物、疾病、症状三类实体进行对齐以改进提取的知识。

① 对于疾病实体，我们直接通过 MIMIC-IV 结构化的诊断表格来获取患者的疾病 ICD 编码。

② 对于药物实体，我们借助 MIMIC-IV prescription.csv 中药物给定的 NDC 索引编码与 DrugBank 数据库进行检索匹配。

③ 对于症状实体，由于临床医疗文本中对同一症状的表示方式不尽相同，因此我们采用语义相似度与大语言模型相结合的实体对齐方法。

1) 基于语义相似度的实体对齐策略

从 MIMIC-IV 数据库中病人的现病史、主诉、医疗过程简述等文段提取出的症状实体中，我们发现相同含义的实体概念的名称十分类似，因此首先，我们基于 Levenshtein 距离度量症状实体间的相似性，给定两个症状实体字符串 S_1、S_2，则它们的相似度定义为：

$$Similarity(S_1 \rightleftarrows, S_2) = 1 - \frac{d(S_1,S_2)}{\max(len(S_1),len(S_2))} \tag{6}$$

对于每个症状实体，创建一个对应的图节点。如果两个症状实体之间的相似度大于或等于 0.85 时，则在它们之间添加一条无向边，形成无向图。通过连通分量算法，将每个连通分量视为一个症状实体簇C_i；将其中的症状实体归为同一个标准化症状实体(Symptom Cluster)，并赋予统一的标签S_i作为该簇的代表，将词形相似的症状实体归为一个标准化的症状群组，从而在后续分析中避免了症状实体冗余和重复表达的问题，基于相似度的症状对齐结果见表 2。

Table 2. Similarity-based symptom entity alignment results
表 2. 基于相似度的症状实体对齐结果

Symptom-clusters
Difficulty Ambulating　Difficulty with Ambulation　Difficulty Ambulating
Hypotension　Hypotensive　Hypotension
Unsteady　Unbalanced
Pneumonia　Pneumonia　Penumaonitis
Lightheadedness　Light-Headedness

2) 基于提示词工程引导分组的实体对齐策略

然而基于相似度进行的对齐主要依赖于文本的结构和语义信息，缺乏医疗知识和文本上下文信息。大语言模型经过全面的预训练，可以为医疗实体对齐任务提供相关医疗知识，并且不需要额外数据。经随机选择样本的测试结果表明，ChatGPT 在医疗实体对齐任务中表现良好。因此本研究采取 ChatGPT 4.0 版本来进一步对症状实体进行细粒度的对齐。

为了提高利用效率即最小化提问的数量和提高响应的一致性，我们采用了特定的提示工程技术。具体来说，我们需要 ChatGPT 将我们提供的一组低召回阈值的术语进行分组，并提供一个示例作为输出规范。较低的相似度阈值说明粗粒度对齐未能将它们进行有效对齐，因此提供了更多可能指代同一实体的术语供 GPT 过滤。

自上而下的语义聚类分组模式有效地减少了总问题数量，并且有效提升了对齐的效率并节约了成本，我们提供了一个示例作为规范便于输出答案的一致性。图 3 为 ChatGPT 4.0 版本为症状术语分组的一个例子。

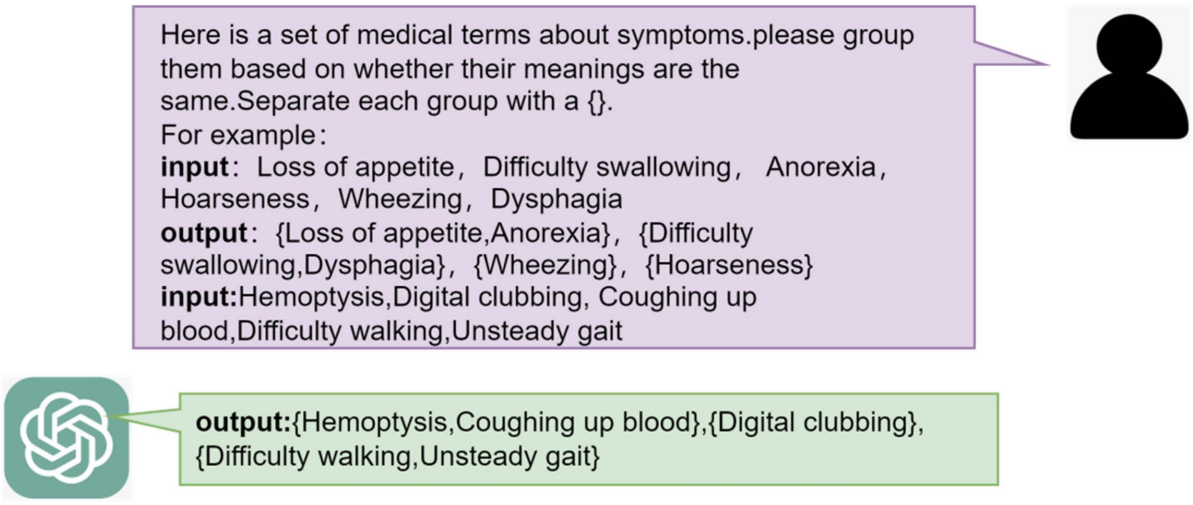

Figure 3. Fine-grained symptom entity alignment using ChatGPT 4.0
图 3. 利用 ChatGPT 4.0 进行细粒度症状实体对齐

通过症状实体对齐归一化，我们构建了一个包含肺癌患者相关症状的词表字典，保证了数据的一致性并且丰富和增强了图谱的可扩展性。

4. 实验设置与评价

4.1. 实验数据收集与处理流程

4.1.1. 微调语料标注方案

基于 MIMIC-IV 数据集 Note 模块中 600 条肺癌患者出院记录(154,080 tokens)，采用“人工主导 + GPT 辅助”的混合标注策略，在 Label Studio 平台上使用 BIOES 体系标注症状与组织类型实体。500 条记录用于微调 BioBERT 模型，100 条双重标注记录构成评估集，用于横向对比 NER 模型性能，经临床医学生交叉审核(Kappa = 0.93)确保标注可靠性。

4.1.2. 药物与手术预测实验训练集划分策略

基于构建的肺癌专病医学知识图谱，将包含三元组形式实体关系的数据集科学划分为训练集、验证集和测试集。实验过程划分训练集、测试集以及验证集的标准如下：

1) 对于任何给定的 h 和 r，所有的尾实体要么都在训练数据中要么都在测试数据；

2) 与癌症分期和转移以及疾病相关症状、药物、手术关系三元组作为背景知识全被划分在训练集；

3) 除了背景知识外的三元组 70%作为训练集，10%作为验证集，20%作为测试集。

4.2. 实验参数设置

4.2.1. 微调 BioBERT 实验参数设置

本研究选择 BioBERT-Base v1.0 版本参数，其中网络中隐藏层 12 层，每层 768 节点，12 个注意力头数，共约 110 M 参数。此版本预训练参数训练数据集包括英文维基百科、BooksCorpus 和 PubMed 摘要数据，能够更好地适应医学文本。

4.2.2. 知识图谱嵌入与预测实验参数设置

本研究采用 TransE 和 TransH 算法生成构建知识图谱的嵌入表示。具体参数设置如下：嵌入维度设为 50，学习率采用 0.001，训练轮数设置为 20 次，批量大小设定为 256，优化计算效率与训练稳定性。负样本生成策略采用随机替换同类型头或者尾实体的方式，增强模型对错误关系的识别能力。

5. 实验结果与分析

5.1. 描述性统计分析结果

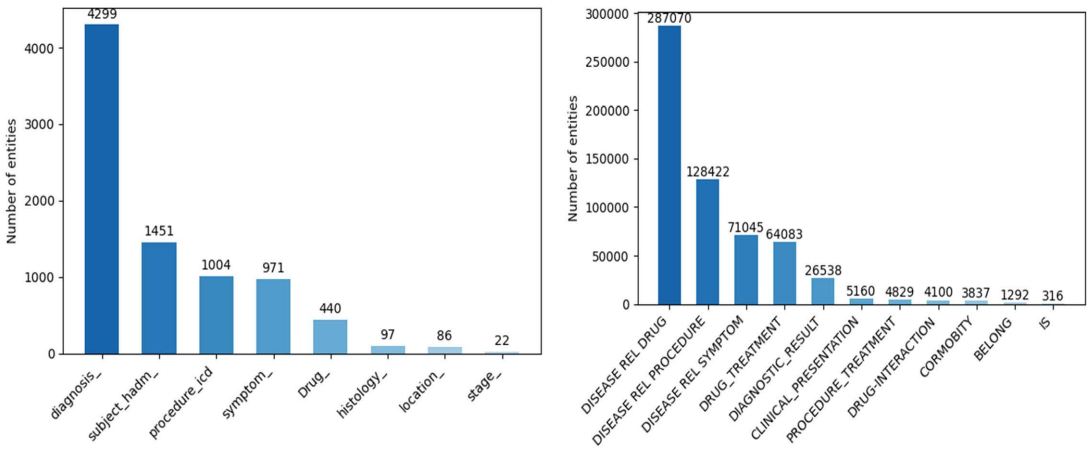

Figure 4. Distribution diagram of entities and relationship types in the lung cancer knowledge graph
图 4. 肺癌知识图谱实体和关系类型分布图

我们从知识图谱实体、关系类型及数量等维度进行统计分析，如图 4 所示。相较于传统医学知识图谱，其不仅覆盖疾病、症状、药品等常见实体类型，更创新性地纳入癌细胞组织类型等与临床诊疗紧密相关的特异性实体，同时通过计算概率特征对医疗知识的不确定性进行量化，有效增强了知识图谱对复杂临床知识的表征能力。

5.2. 知识表示及可视化分析

本研究采用知识图嵌入技术，将图数据有效地映射到真实的低维向量空间中，使智能算法能够更好地挖掘隐藏在图数据中的信息[10]。选用基于距离约束的知识表示方法(TransE 和 TransH)以生成本研究所构建的肺癌医学知识图谱嵌入。

采用 t-SNE 降维方法，将原始的 768 维向量降低到二维平面。嵌入结果，如图 5 所示。我们对图嵌入进行可视化分析能够发现明显的聚类现象，这表明图嵌入能够将语义差异转化为向量空间中的距离差异，这种分布规律与实际医学知识中的语义关系相契合，该可视化结果为后续实体分类、关系预测等任务提供了直观的理论依据。

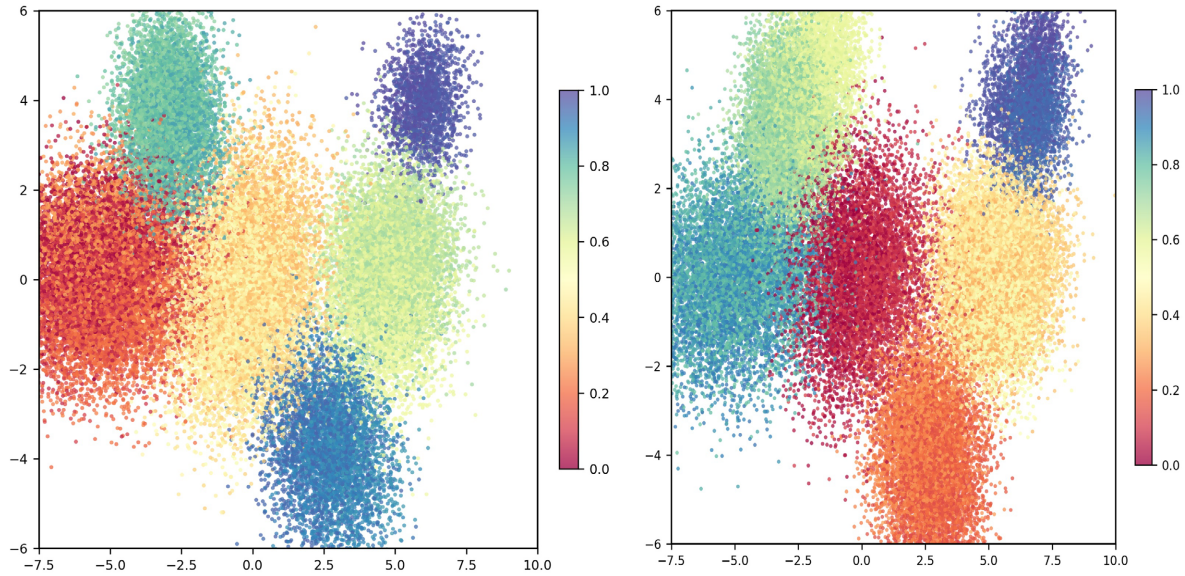

Figure 5. Embedding representation utilizing TransE (left) and TransH (right)
图 5. 利用 TransE (左)和 TransH (右)嵌入表示可视化

5.3. 面向医疗实体识别的对比实验与分析

Table 3. Model's named entity recognition results
表 3. 模型命名实体识别结果

模型	Categories Symptom			Caterories Histology		
	Precision	Recall	F1-score	Precision	Recall	F1-score
微调后 BioBERT	**86.7**	**84.5**	**85.5**	**74.6**	**76.4**	**75.6**
未微调 BioBERT	79.3	75.9	79.6	61.1	67.2	63.9
Bilstm-CFR	64.6	65.2	64.89	54.2	56.8	55.5
GraphRAG	82	87.5	84.6	45	65.8	53.3

本实验基于标注的 100 条医疗记录，系统对比了微调后的 BioBERT 模型、未微调的 BioBERT 模型、BiLSTM-CRF 和 GraphRAG 模型在症状和癌细胞组织类型实体识别任务中的性能表现，实验结果见表 3。微调后的 BioBERT 展现出显著优势，在症状实体的识别中准确率、召回率和 F1-score 分别达到 86.7%、84.5%和 85.5%，表现出良好的识别能力。结果表明自适应微调策略能有效注入肺癌特定领域知识，并且能够增强对电子病历文本的适应能力从而增强医学命名实体识别能力。

同时对比发现各个模型对症状实体的识别指标均高于癌细胞组织类型实体。这表明模型在识别较短且相对简单的症状实体时表现出较高的准确性和召回能力，由于癌细胞组织类型实体表述较长、表述方式不固定、实体被分隔、复杂性较高等原因导致各个模型在癌细胞组织类型实体的识别性能较低。此外，由于预训练以及微调过程中与癌细胞组织类型实体相关标注数据的稀缺性，也导致信息的稀疏性加大，进一步降低了识别的准确性，使识别效果与实际情况存在差异。尽管存在性能差异，微调后的 BioBERT 模型结果准确率较高，经人工验证后，其输出结果仍可作为构建医学知识图谱的可靠数据源。

5.4. 面向肺癌专病知识图谱应用的实验与分析

本研究利用 TransE 生成实体和关系的嵌入向量，然后将这些嵌入向量输入一个包含三个全连接层的神经网络模型进行训练。在测试阶段，输入测试集中的特定患者头实体向量及关系向量(关系为 DRUG_TREATMENT)，然后从测试集中提取所有药物作为候选药物列表。通过计算每个候选药物实体的概率，选取概率前 5 高的药物代码，并与实际使用药物进行对比，计算模型评估指标。

在手术预测任务中，方法与药物预测相似。我们通过相同的 TransE 嵌入模型生成实体和关系的嵌入向量，然后将其输入神经网络进行训练。在测试阶段，输入患者的头实体向量和相应的手术关系向量(关系为 PROCEDURE_TREATMENT)，并从测试集中提取所有手术作为候选手术列表。根据头实体和关系的嵌入向量，计算每个候选手术的概率，最终推荐前 5 个概率最高的手术，并与实际发生的手术进行对比，评估模型的效果。药物及手术预测评估结果见图 6。

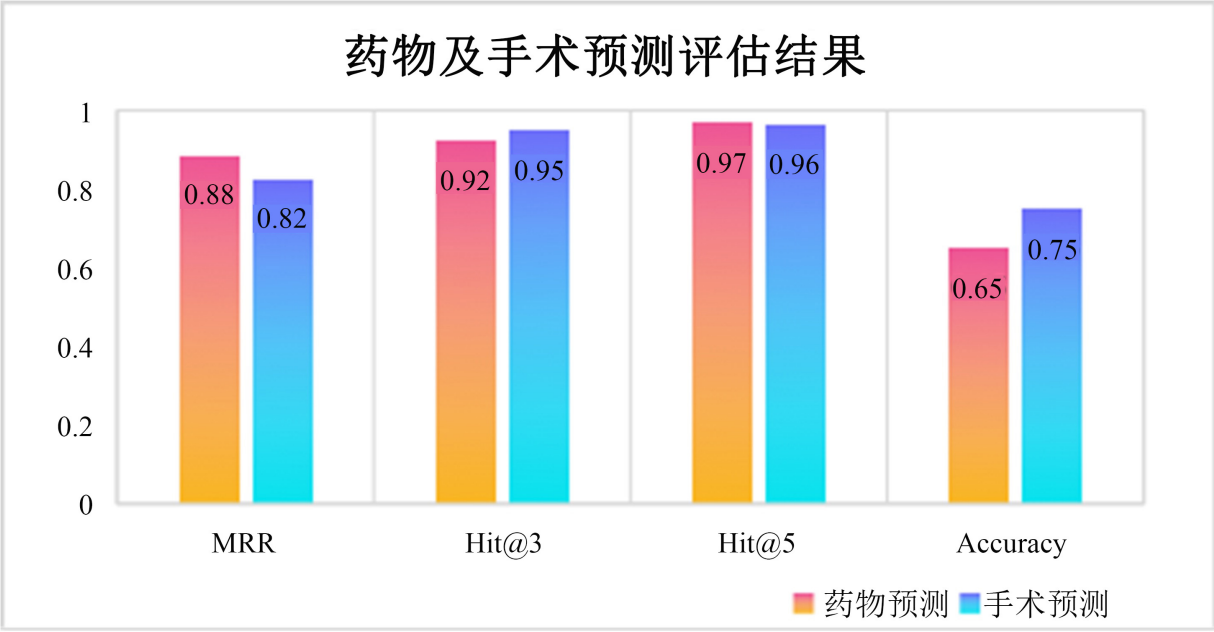

Figure 6. Experimental results of prediction using TransE embeddings
图 6. 利用 TransE 嵌入预测实验结果

在我们的实验中选取平均排名 MRR、Hit@K 以及准确率 Accuracy 作为实验的评估指标，药物预测结果如图 6 所示。在药物预测实验中平均排名(Mean Reciprocal Rank)为 0.88，表明模型在推荐药物时，正确药物的排名通常较高。这反映出知识图谱中的实体及其关系结构相对清晰，能够有效支持模型的学习过程。在前 3 个和前 5 个推荐中至少有一个正确药物的比例分别为 0.92 和 0.97。

在手术预测任务中，我们构建的肺癌知识图谱同样取得了优异的结果，MRR 为 0.82 表明模型在推荐手术时，正确手术的排名也较高；Hit@3 为 0.95、Hit@5 为 0.96 表示在前 3 个推荐中至少有一个正确手术的比例为 95%，在前 5 个推荐中至少有一个正确手术的比例为 96%。推荐的准确率为 0.75，表明知识图谱提供的手术相关知识能够帮助模型高效地推荐合适的手术方案。

6. 总结与讨论

医学知识图谱对于智慧医疗的发展有重要意义，本文基于 MIMIC-IV 数据库中的电子病历，并结合 DrugBank、PubMed、ICD-10 等数据库搭建融合多源异构医学信息的肺癌专病知识图谱构建框架。在命名实体识别环节，运用微调后的预训练语言模型 BioBERT，识别肺癌患者出院记录中的癌细胞组织类型和症状实体，能够让模型适应电子医疗记录文本特征，提升实体识别的准确性。同时本研究在疾病相关三元组中创新性地引入概率属性，以此量化医学知识的不确定性，并基于相似度计算与大语言模型融合的实体对齐方法，实现多源数据的有效融合。最终构建了一个多源、具有概率属性并融入患者信息的肺癌医学知识图谱，且该图谱在药物预测和手术预测任务中表现良好，能够为肺癌临床决策提供有效的知识支持。

参考文献

[1] Kim, M., Park, H., Kho, B., Park, C., Oh, I., Kim, Y., *et al*. (2020) Artificial Intelligence and Lung Cancer Treatment Decision: Agreement with Recommendation of Multidisciplinary Tumor Board. *Translational Lung Cancer Research*, **9**, 507-514. https://doi.org/10.21037/tlcr.2020.04.11

[2] Huang, Z., Xu, W. and Yu, K. (2015) Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv:1508.01991.

[3] Edge, D., Trinh, H., Cheng, N., *et al*. (2024) From Local to Global: A Graph Rag Approach to Query-Focused Summarization. arXiv:2404.16130.

[4] Yang, P., Wang, H., Huang, Y., Yang, S., Zhang, Y., Huang, L., *et al*. (2024) LMKG: A Large-Scale and Multi-Source Medical Knowledge Graph for Intelligent Medicine Applications. *Knowledge-Based Systems*, **284**, Article 111323. https://doi.org/10.1016/j.knosys.2023.111323

[5] 靳淑雁, 王爽, 黄琼, 邱五七, 林怿昊. 基于乳腺癌专病库的知识图谱构建研究[J]. 医学信息学杂志, 2023, 44(12): 65-70.

[6] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., *et al*. (2019) BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, **36**, 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

[7] 杨善林, 丁帅, 顾东晓, 等. 医疗健康大数据驱动的知识发现与知识服务方法[J]. 管理世界, 2022, 38(1): 219-229.

[8] Chandak, P., Huang, K. and Zitnik, M. (2023) Building a Knowledge Graph to Enable Precision Medicine. *Scientific Data*, **10**, Article No. 67. https://doi.org/10.1038/s41597-023-01960-3

[9] Li, L., Wang, P., Yan, J., *et al*. (2020) Real-World Data Medical Knowledge Graph: Construction and Applications. *Artificial Intelligence in Medicine*, **103**, Article 101817. https://doi.org/10.1016/j.artmed.2020.101817

[10] Yang, H. and Liu, J. (2021) Knowledge Graph Representation Learning as Groupoid: Unifying TransE, RotatE, QuatE, ComplEx. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, Virtual Event Queensland, 1-5 November 2021, 2311-2320. https://doi.org/10.1145/3459637.3482442