

# 中文医学量表内容实体及关系标注语料库构建与应用

陈振丽 孙海霞 郝洁 钱庆林 燕蒲思樾

(中国医学科学院/北京协和医学院医学信息研究所/图书馆 北京 100020)

**〔摘要〕** **目的/意义** 构建中文医学量表核心内容要素语料库, 为相关知识实体与关系抽取任务提供数据基础。**方法/过程** 设计涵盖量表名称、测量概念、测量条目及其编码等 5 类实体, 以及 4 类语义关系的标注体系, 制定统一标注规范; 采用双人背靠背人工标注方式, 对 1 491 篇中文核心期刊文献富语义段落进行标注, 并通过一致性计算和下游抽取任务实验评估语料库质量, 最终形成 CMedScale 语料库。**结果/结论** 各模型在引入语料库示例后, 实体识别的 Micro-F1 分数提升 2.95~13.89 个百分点, 关系抽取的 Micro-F1 分数提升 16.93~33.33 个百分点。CMedScale 语料库为中文医学量表知识抽取及相关下游任务研究提供了高质量数据支撑。

**〔关键词〕** 医学量表; 语料库构建; 大语言模型; 实体识别; 关系抽取

**〔中图分类号〕** R-058 **〔文献标识码〕** A **〔DOI〕** 10.3969/j.issn.1673-6036.2025.11.004

## Construction and Application of a Chinese Medical Scale Corpus with Entity and Relation Annotations

CHEN Zhenli, SUN Haixia, HAO Jie, QIAN Qing, LIN Yan, PU Siyue

Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

**〔Abstract〕** **Purpose/Significance** To construct a Chinese medical scale corpus of core content elements, so as to provide a data basis for the task of extracting related knowledge entities and relations. **Method/Process** An annotation schema covering five types of scale-related entities such as scale names, measurement concepts, measurement items, and their corresponding codes, as well as four types of semantic relations is designed, and a unified annotation standard is formulated. A double-blind manual annotation approach is adopted to annotate semantically rich paragraphs from 1 491 Chinese core journal articles. The corpus quality is further evaluated through inter-annotator agreement and downstream task experiments, ultimately resulting in the CMedScale corpus. **Result/Conclusion** After introducing the corpus examples, the Micro-F1 scores of entity recognition in each model increased by 2.95 to 13.89 percentage points, and those for relation extraction improved by 16.93 to 33.33 percentage points. The CMedScale corpus provides high-quality data support for Chinese medical scale knowledge extraction and related downstream research tasks.

**〔Keywords〕** medical scale; corpus construction; large language model; entity recognition; relation extraction

**〔修回日期〕** 2025-10-09

**〔作者简介〕** 陈振丽, 硕士研究生, 发表论文 3 篇; 通信作者: 孙海霞, 研究员, 硕士生导师。

**〔基金项目〕** 国家社会科学基金项目 (项目编号: 21BTQ069)。

## 1 引言

医学量表是临床诊疗、护理、心理评估及康复干预中广泛应用的标准化测量工具<sup>[1-2]</sup>。随着循证医学和精准医疗的发展,医学量表在科学研究与临床实践中的作用愈加突出<sup>[3]</sup>。当前,中文医学量表多以数据库、工具书或研究文献附件等形式呈现,整体以文档为最小发布单元<sup>[4]</sup>,缺乏对其内容要素的细粒度建模与语义解析,限制了医学量表资源在量表智能推荐、知识图谱扩展、辅助问答系统等任务中的应用。以大语言模型为代表的人工智能技术为医学量表资源的知识抽取与结构化提供解决方案,但其实现需要高质量领域数据集作支撑<sup>[5]</sup>。

命名实体识别(named entity recognition, NER)和关系抽取(relation extraction, RE)是医学知识抽取的基本任务,也是实现医学知识结构化表示与组织的关键。为促进各类医学实体和关系抽取技术发展,医学领域已经基于科技文献、临床电子病历、药品说明书等不同类型数据源构建了含有疾病、症状、药物等通用医学实体及其相互间关系的语料库<sup>[6-9]</sup>。关于医学量表内容类实体和关系,目前相关标注语料较少,主要为笔者团队围绕“量表-概念-条目”内容结构构建的量表实体语料库 CMedS-NER<sup>[10]</sup>和关系语料库 CMedS-RE<sup>[11]</sup>,但其规模有限,不利于在更广范围内进行医学量表内容知识抽取技术测评。因此,本研究进一步遴选医学量表开发与应用评估类中文期刊文献,开展更大规模中文医学量表内容实体与关系标注语料库构建,并探究其对中文大语言模型结合提示工程技术进行信息抽取性能的提升作用。

## 2 相关工作

既有医学语料库在数据来源、实体类型及关系类型设计上已呈现多样化特征。数据来源涉及电子病历、医学文献、医学教材等。如跨语言信息检索(cross-language evaluation forum, CLEF)语料库<sup>[12]</sup>基于 20 000 份癌症患者病历开展语义标注;

美国国家生物技术信息中心的疾病语料库<sup>[13]</sup>对 793 篇 PubMed 摘要进行提及和概念标注; Xu Y 等<sup>[14]</sup>构建包含 336 份出院摘要的标注语料库;中文医学命名实体识别数据集(Chinese medical entity extraction dataset, CMeEE)<sup>[15]</sup>对儿科类医学教材进行了标注。在实体类型方面,医学语料库普遍涵盖疾病、症状、药物、治疗、检查、身体部位等通用医学实体,部分语料进一步扩展至基因、蛋白、表型等生物医学实体。CLEF<sup>[12]</sup>包含健康状况、干预、检查、药物等临床概念; CMeEE<sup>[15]</sup>覆盖疾病、临床表现、药物、微生物等类别; Xu Y 等<sup>[14]</sup>聚焦医学问题、检查、治疗和药物; Gao Y 等<sup>[16]</sup>在此基础上增加了身体部位、时间等类型; 杨锦峰等<sup>[17]</sup>研究涵盖疾病、诊断分类、症状、检查与治疗等类别。PGxCorpus<sup>[18]</sup>进一步引入蛋白质、变异、表型等基因组学相关实体。在关系类型方面,既有语料库已基本覆盖通用关系类型。CLEF<sup>[12]</sup>设定 5 类关系,用于刻画临床事件之间的基本联系; Gao Y 等<sup>[16]</sup>设定 4 类关系,主要聚焦疾病与检查、疾病与治疗等基础临床联系; Legrand J 等<sup>[18]</sup>和 Rosario B 等<sup>[19]</sup>定义 7 类关系,更细致地描述疾病、基因和药物之间的多样化相互作用; 杨锦峰等<sup>[17]</sup>设定 15 类关系,覆盖诊断、检查、治疗等更广泛的语义关联。但目前中文医学量表语料的构建和研究相对较少。

## 3 语料库标注与评估

### 3.1 标注体系设计

结合中文医学量表内容与结构特点,参考中文医学标注语料设计经验<sup>[20]</sup>,兼顾医学量表场景下知识结构特征和实际应用需求<sup>[21]</sup>,设计面向实体与关系的标注体系。

**3.1.1 实体类型定义** 设计 5 类核心实体类型:量表名称、测量概念、测量条目、概念编码和条目编码,见表 1。量表名称指向现实世界中某一特定量表的名称,而非某类量表的统称,包含“评分表”“量表”“问卷”等称谓。测量概念为量表中用于衡量某一心理、生理或功能维度的理论构件或变量,通常为条目的测量对象,有“维度”“领域”

“方面”“次领域”“子领域”“变量”“因子”等  
称谓<sup>[21]</sup>。测量条目是量表中获取评估对象主客观信  
息的具体问题或陈述，作为量表最小的测量单元，

通常配有评分选项（本研究未标注评分项）。概念  
编码与条目编码分别为对应测量概念、测量条目的  
标准化编码。

表 1 实体类型定义及标注示例

实体类型（标签）	实体定义	示例
量表名称（scale）	一种测量工具，由多个条目构成，旨在揭示不宜用直 接方法测量的变量的水平	简易精神状态量表；mini - mental state examination； MMSE
测量概念（concept）	量表中的理论构件，如领域与维度	躯体功能；心理功能；抑郁
测量条目（item）	量表中的基本测量单元，通常是一个问题或陈述句	您对家庭的生活方式满意吗
概念编码（concept - code）	特定量表中测量概念的编码	因子 2
条目编码（item - code）	特定量表中测量条目的编码	条目 1

3.1.2 关系类型定义 设计 4 类关系类型，见表  
2。一是包含测量概念，用于表示量表所涵盖的测量  
概念或编码。二是包含测量条目，用于表示量表名  
称、测量概念、概念编码 3 类实体，与各自所包含的

测量条目或其编码之间的关联。三是有编码，用于建  
立测量概念或测量条目，与对应编码之间的映射关  
系。四是有等同名称，用于关联同一类型的实体在  
不同文本语境中出现的同义称谓或表达变体。

表 2 关系类型定义及标注示例

关系类型（标签）	主体	客体	示例
有等同名称（hasSynonymousName）	量表名称	量表名称	（Conners 家长量表，hasSynonymousName，CPRS）
包含测量概念（containConcept）	量表名称	测量概念	（CPRS，containConcept，品行问题）
		概念编码	（CPRS，containConcept，I 因子）
包含测量条目（containItem）	量表名称	测量条目	（CPRS，containItem，对大人粗鲁无礼）
		条目编码	（CPRS，containItem，条目 2）
	测量概念	测量条目	（品行问题，containItem，对大人粗鲁无礼）
		条目编码	（品行问题，containItem，条目 2）
	概念编码	测量条目	（I 因子，containItem，对大人粗鲁无礼）
有编码（hasCode）	测量概念	概念编码	（品行问题，hasCode，I 因子）
		条目编码	（对大人粗鲁无礼，hasCode，条目 2）

3.2 标注规范制定

实体标注准则如下。一是提及层面标注，标注所有 5 类实体的具体提及。二是出现即标注，暂不考虑前置修饰的研究方法词（如“删除”“翻译”“修订”）。三是语种规定，量表实体标注其中英文提及；概念与条目实体仅标注中文形式。四是最长标注原则，优先标注语义上最具专指性的完整表达。五是不嵌套标注，一个实体不能在另一实体内

部出现。六是不重复标注，每个实体提及只能归属一种实体类型。七是不连续实体不标注，如“中文版的 DQoL”不作为整体标注。八是指示代词不标注，如“该量表”“其”“该因子”等。

关系标注准则如下。一是仅标注句内关系。二是支持一对多关系，即同一实体可与多个实体建立关联。三是每对实体仅标注唯一关系类型。四是当同一实体多次出现时，逐一标注其与其他实体提及的关系。

3.3 数据筛选与预处理

以中文医学核心期刊论文为数据来源，聚焦以医学量表开发为核心主题的原始研究文献。首先，在中国知网设置检索范围为“学术期刊”，来源类别选择“北大核心”“CSSCI”“CSCD”，以“篇关摘=量表 AND 篇关摘=开发 AND 中图分类号=%‘R’”为检索条件，检索时间截至 2023 年 10 月。其次，选取发文量最多的 3 个学科领域（肿瘤学、护理学和精神病学）的论文，对摘要进行人工筛选，剔除综述、经验介绍等非原始研究论文，仅保留与量表构建相关的研究型论文，并进一步获取其结构化格式（XML）的全文内容。最后，纳入 1 491 篇论文，覆盖 257 本核心期刊，年代跨度为 1998—2023 年。

为提高标注效率和语料的语义丰富度，随机抽取 12 篇文献进行人工全文阅读和预标注，发现研究所需的实体和关系主要集中在论文的“方法”

“结果”与“讨论”3 部分，尤其是测量概念和测量条目。因此，首先抽取上述 3 部分作为候选段落。再依据总结的关键线索词集 {维度、领域、变量、概念、因子、项目、条目} 进行段落过滤，仅保留富含目标实体的富语义段落作为标注语料。

3.4 数据标注

使用 Label Studio 标注工具进行实体标注。流程方面，先完成实体标注，形成实体语料库的金标准，再开展实体间关系标注。整体标注流程分为预标注与正式标注两个阶段，见图 1。在预标注阶段，标注团队接受为期两周的系统培训，以熟悉标注规范并掌握标注工具的具体使用方法。正式标注阶段采用多轮迭代标注机制，每期期刊文献由两名标注人员分别独立完成背靠背标注，由第 3 名研究人员对两者标注结果进行一致性审核，记录标注差异及存疑部分，并组织医学领域专家展开讨论，最终确定一致的修订方案。

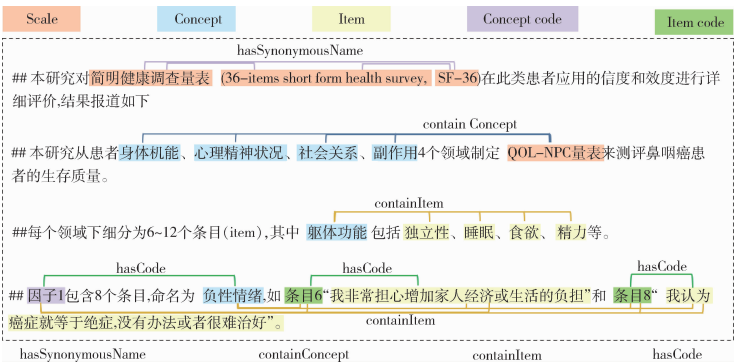


图 1 实体和关系标注示例

3.5 一致性评估

标注一致性是指两名独立标注人员标注结果达成一致的程[22]。在手工标注语料库研究中，通常使用 Kappa 值来衡量一致性，数值越大表明数据集标注质量越高。选用 Cohen’s Kappa 系数[23]作为一致性评价指标，评估两名标注人员对 N 条数据做多分类（定性）任务时的一致性： $k = (p_0 - p_e) / (1 - p_e)$ 。其中  $p_0$  是观察到的比例一致性，即评估者在所有样本上达成一致的比例， $p_e$  是预期的随机一致性比例，即在假设评估者独立且随机进行分类的情况下，其达成一致的[22]。

实体一致性界定原则：两个实体标注的文本、类型标签和文本起止位置 3 项均完全一致时，认为该实体标注一致。关系三元组一致性界定原则：两个关系标注的主体文本、类型标签与客体文本 3 项均一致时，认为该关系标注一致。经一致性评估，实体 Cohen’s Kappa 为 0.92，关系 Cohen’s Kappa 为 0.84，表明标注结果可信[22]。

4 语料库统计与分析

4.1 语料统计

本研究构建的中文医学量表知识语料库命名为 CMedScale，其语料基本统计，见图 2。

该语料库共包含 1 491 篇中文医学核心期刊论文，共提取 16 036 个段落与 62 219 个句子，标注 68 745 个实体和 28 720 个关系三元组。

4.2 实体字符特征分析

不同实体类型的字符长度分布，见图 3。整体来看，量表名称的长度多集中在 3 ~ 14 字；测量条目普遍较短，主要集中在 2 ~ 6 字，说明条目设计倾向于简洁明了，便于受试者理解与作答；测量概念多为 2 ~ 6 字的短语，尤以 4 字居多；概念编码和条目编码则具有高度规范化特征，字符长度集中且统一，其中概念编码多为 3 ~ 4 字，而条目编码则主要为 1 ~ 4 字。

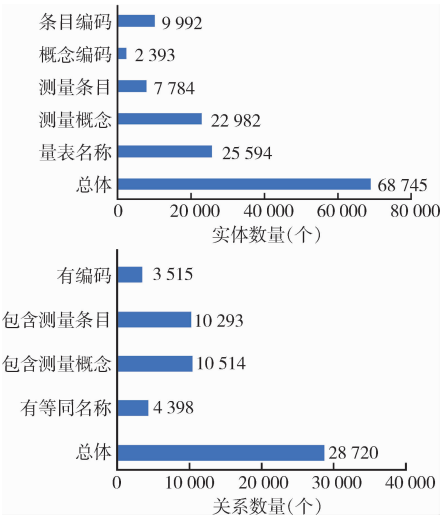


图 2 CMedScale 语料基本统计

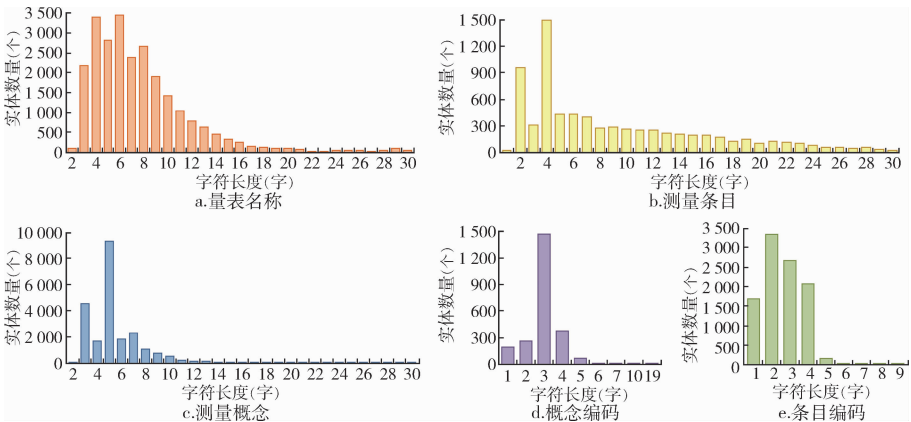


图 3 实体字符长度统计

各实体类型的字符类型组成，见图 4。量表名称的构成最为复杂，纯英文（29.00%）、纯中文（19.41%）、中英结合（20.53%）比例接近，且符号使用频繁，符号与其他字符的组合占 29.61%。测量概念与测量条目以纯中文为主，分别占比 96.04% 和 76.84%，其中测量条目中符号比例（17.68%）相对较高，说明条目常包含提示或附加说明。概念编码以“中文 + 数字”组合为主，占比 60.11%。条目编码以纯数字（45.58%）和“中文 + 数字”组合（41.09%）为主。

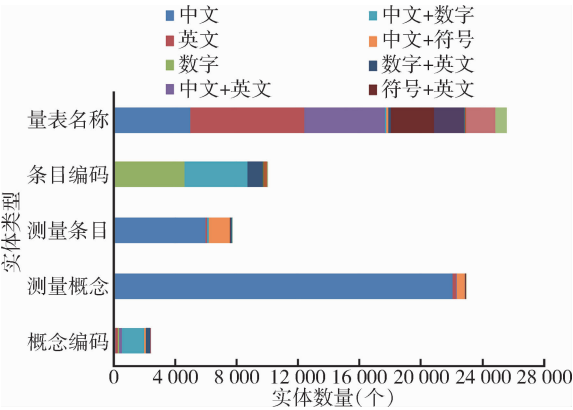


图 4 实体字符特征统计

不同实体类型在字符长度与字符类型上存在明显差异，反映了医学量表文本在命名、条目设计与编码规范上的多层次需求。量表名称和测量条目字符长度较长且混合中英文及符号，体现跨语言和符号的多样性；测量概念字符短小、类型单一，体现领域规范性；编码以数字和英文为主。

4.3 实体和关系密度分析

4.3.1 实体密度分析 为进一步分析实体的分布特征，对 5 类实体在不同文本单元（文献、段落和句子）中的平均密度进行统计，见表 3。从文献级别来看，每篇论文平均包含 46.11 个实体提及，其中量表名称与测量概念占实体总数的主要部分，显示出量表开发研究中对测量工具名称和测量维度的频繁引用。概念编码实体出现频次最低，说明编码信息在论文中更为稀疏。

表 3 各实体类型在文献、段落和句子级别的密度统计（个）

数据类型	平均每篇文献	平均每个段落	平均每个句子
量表名称	17.17	1.60	0.41
测量概念	15.41	1.43	0.37
测量条目	5.22	0.49	0.13
概念编码	1.60	0.15	0.04
条目编码	6.70	0.62	0.16
总体	46.11	4.29	1.10

注：各实体类型数值为该类型实体总数除以相应文本单元数；“总体”并非各类型均值的简单相加，而是所有实体类型总数除以对应单元数（如平均每篇文献总体实体数 =  $68\,745 \div 1\,491 = 46.11$ ）。

4.3.2 关系密度分析 4 类关系在不同文本单元中的平均密度，见表 4。

表 4 各关系类型在文献、段落和句子级别的密度统计（个）

数据类型	平均每篇文献	平均每个段落	平均每个句子
有等同名称	2.95	0.27	0.07
包含测量概念	7.05	0.66	0.17
包含测量条目	6.90	0.64	0.17
有编码	2.36	0.22	0.06
总体	19.26	1.79	0.46

注：各关系类型数值为该类型关系总数除以相应文本单元数；“总体”并非各类型均值的简单相加，而是所有关系类型总数除以对应单元数（如平均每篇文献总体关系数 =  $28\,720 \div 1\,491 = 19.26$ ）。

语料在实体与关系文献层面的密度均维持在较高水平，平均密度分别为 46.11 个实体和 19.26 个关系，为大语言模型的提示学习提供了丰富上下文，有助于模型更有效地捕捉语义结构<sup>[24]</sup>。总体而言，这些分布规律揭示了医学量表语料的结构性特征，不仅可为后续模型训练与提示设计提供参考，也提示后续抽取模型在优化目标时可根据类型密度权重进行差异化设计<sup>[18,25]</sup>。

5 语料库应用

为验证 CMedScale 对模型性能的提升作用，分别在有语料支持和无语料支持下，开展命名实体识别与关系抽取实验，整体采用 Micro-F1 分数作为评估指标，以全面衡量模型表现。

5.1 实验设置

将 CMedScale 语料库按照 9:1 划分为训练集和测试集。考虑时间和算力限制，从测试集中随机抽取 20 篇文献（共计 841 个句子）作为测试数据。训练集用于构建示例池，支持检索增强生成（retrieval augmented generation, RAG）策略。

选用中文大语言模型 GLM-4<sup>[26]</sup>、Moonshot-v1、Qwen-turbo<sup>[27]</sup> 和 DeepSeek-v3 作为基础模型。参数设置方面，采样方式采用温度采样（temperature = 0.02），max\_tokens 设为 2 048，其余超参数保持默认配置。为保证结果可比性，每轮仅保留模型输出的首条结果。

为验证语料对模型表现的价值，在 NER 和 RE 任务中均设计两类条件。无语料条件（0-shot），不提供任何标注示例，仅依赖模型的通用能力完成任务。有语料条件（k-shot），设置 1-shot、2-shot、4-shot、8-shot 共 4 种配置，通过 RAG 策略引导模型识别与抽取<sup>[28]</sup>，以观察模型在不同示例支持下的性能变化。RAG 策略通过 acge\_text\_embedding 模型<sup>[29]</sup> 计算输入句与训练样本的句向量，基于余弦相似度选取最相近的 k 个示例，按统一模板整合至提示中，用于引导实体识别和关系抽取。

5.2 实体识别结果分析

在 NER 任务中，引入 CMedScale 语料库示例后，各模型整体性能均较 0-shot 条件有不同程度提升，见表 5，验证了语料在支持医学量表实体识别中的价值。整体来看，模型的 Micro-F1 分数提升幅度为 2.95~13.89 个百分点。具体而言，GLM-4 在仅提供 1 个示例时性能由 51.13% 上升至 54.56%，DeepSeek-v3 同样在 1-shot 时达到最佳性能，显示模型对有限示例较为敏感。Qwen-turbo 的性能提升趋势更为明显，从 16.33% 上升至 30.22% ( $k=2$ )。Moonshot-v1 在  $k=8$  时达到最高性能，相比 0-shot 提升 6.3 个百分点。

表 5 各模型在不同 k-shot 下实体识别 Micro-F1 分数 (%)

k-shot	GLM-4	Moonshot-v1	Qwen-turbo	DeepSeek-v3
0	51.13	27.03	16.33	47.24
1	54.56	29.33	26.16	50.19
2	53.85	31.29	30.22	49.04
4	53.56	30.84	25.76	49.84
8	51.93	33.33	22.88	47.01

5.3 关系抽取结果分析

在 RE 任务中，引入 CMedScale 语料库示例后，各模型整体性能同样显著优于 0-shot 条件，见表 6。整体来看，各模型的 Micro-F1 分数提升幅度为 16.93~33.33 个百分点，充分体现了语料对模型抽取复杂语义关系的支撑作用。

表 6 各模型在不同 k-shot 下关系抽取 Micro-F1 分数 (%)

k-shot	GLM-4	Moonshot-v1	Qwen-turbo	DeepSeek-v3
0	57.99	39.57	48.01	47.25
1	64.05	52.48	55.26	72.71
2	74.92	56.61	60.17	80.13
4	72.92	55.06	66.19	73.89
8	70.70	59.93	71.66	80.58

6 结语

本研究构建了面向中文科技文献全文的医学量表内容实体与关系标注语料库 CMedScale，包含

1 491 篇中文医学期刊论文，共标注 68 745 个实体和 28 720 个关系三元组，为医学量表知识的抽取与应用提供了坚实的数据基础。初步实验结果表明，无论在实体识别还是关系抽取任务中，提供语料示例均能有效提升模型性能，其中关系抽取任务的提升幅度更为显著，说明语料在帮助模型捕捉复杂语义关系方面发挥了关键作用。

未来，CMedScale 语料库可从以下 3 方面拓展与深化。一是在数据层面，扩充学科覆盖范围，增加文献类型（如图书、量表手册），并引入不同数据模态（如表格、图像），完善跨句关系标注体系。二是在模型评估方面，引入更多大语言模型和传统基线模型进行系统对比，从不同任务维度全面评估语料的适用性，并在实际应用场景中进一步验证其价值，以确保其在中文医学量表知识抽取中的实用性与推广性。三是可结合当前提示工程的发展趋势，探索提示调优、动态提示生成、自我验证等技术在医学量表相关知识抽取中的应用潜力，助力推动高质量医学语料与技术的协同发展。

**作者贡献：**陈振丽负责标注规范制定、数据收集、数据标注与分析、实验实施、论文撰写；孙海霞负责提出选题、标注体系与规范设计、论文修订；郝洁参与数据标注、实验实施；钱庆负责提供指导、论文修订；林燕、蒲思樾参与数据标注。

**利益声明：**所有作者均声明不存在利益冲突。

参考文献

- 1 陶立元, 张华, 赵一鸣. 临床研究中量表研制的基本思路与方法 [J]. 中华儿科杂志, 2019, 57 (5): 400.
- 2 DEVELLIS R F. Scale development: theory and applications [M]. Thousand Oaks: Sage Publications, Inc., 1991.
- 3 CROSSNOHERE N L, SCHUSTER A L R, BRUCKEL J, et al. Patient-reported outcome measures add value as clinical trial endpoints [J]. Nature medicine, 2025 (8): 1-4.
- 4 孙海霞, 郝洁, 郭臻, 等. 基于知识元的细粒度医学量表文档知识表示框架构建 [J]. 数字图书馆论坛, 2023 (12): 86-98.
- 5 余辉, 岳少丹. 浅析语料库在自然语言处理中的重要性和应用 [J]. 语言文字教学与研究, 2024, 2 (2): 48-50.
- 6 GUAN T, ZAN H, ZHOU X, et al. CMeIE: construction

- and evaluation of Chinese medical information extraction dataset [C]. Zhengzhou: Natural Language Processing and Chinese Computing; 9th CCF International Conference, 2020.
- 7 ZHANG N, CHEN M, BI Z, et al. CBLUE: a Chinese biomedical language understanding evaluation benchmark [EB/OL]. [2025-06-11]. <https://arxiv.org/abs/2106.08087>.
- 8 ROJAS-CARABALI W, AGRAWAL R, GUTIERREZ-SINISTERRA L, et al. Natural language processing in medicine and ophthalmology: a review for the 21st-century clinician [J]. *Asia-Pacific journal of ophthalmology*, 2024, 13 (4): 100084.
- 9 HAO T, RUSANOV A, BOLAND M R, et al. Clustering clinical trials with similar eligibility criteria features [J]. *Journal of biomedical informatics*, 2014, 52 (12): 112-120.
- 10 HAO J, CHEN Z, PENG Q, et al. Prompt framework for extracting scale-related knowledge entities from Chinese medical literature: development and evaluation study [J]. *Journal of medical internet research*, 2025, 27 (1): e67033.
- 11 CHEN Z, HAO J, SUN H, et al. MedScaleRE-PF: a prompt-based framework with retrieval-augmented generation, chain-of-thought, and self-verification for scale-specific relation extraction in Chinese medical literature [J]. *Information processing & management*, 2025, 62 (6): 104278.
- 12 ROBERTS A, GAIZAUSKAS R, HEPPLER M, et al. Building a semantically annotated corpus of clinical texts [J]. *Journal of biomedical informatics*, 2009, 42 (5): 950-966.
- 13 DOĞAN R I, LEAMAN R, LU Z. NCBI disease corpus: a resource for disease name recognition and concept normalization [J]. *Journal of biomedical informatics*, 2014, 47 (2): 1-10.
- 14 XU Y, WANG Y, LIU T, et al. Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries [J]. *Journal of the American medical informatics association*, 2014, 21 (1): e84-e92.
- 15 ZHANG H, ZONG Y, CHANG B, et al. Medical entity annotation standard for medical text processing [C]. Haikou: The 19th Chinese National Conference on Computational Linguistics, 2020.
- 16 GAO Y, GU L, WANG Y, et al. Constructing a Chinese electronic medical record corpus for named entity recognition on resident admit notes [J]. *BMC medical informatics and decision making*, 2019, 19 (2): 56.
- 17 杨锦锋, 关毅, 何彬, 等. 中文电子病历命名实体和实体关系语料库构建 [J]. *软件学报*, 2016, 27 (11): 2725-2746.
- 18 LEGRAND J, GOGDEMIR R, BOUSQUET C, et al. PGx-Corpus, a manually annotated corpus for pharmacogenomics [J]. *Scientific data*, 2020, 7 (1): 3.
- 19 ROSARIO B, HEARST M. Classifying semantic relations in bioscience texts [C]. Barcelona: The 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), 2004.
- 20 咎红英, 刘涛, 牛常勇, 等. 面向儿科疾病的命名实体及实体关系标注语料库构建及应用 [J]. *中文信息学报*, 2020, 34 (5): 19-26.
- 21 Food and Drug Administration. Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims [EB/OL]. [2024-09-15]. <https://www.who.int/tools/whoqol>.
- 22 ARTSTEIN R, POESIO M. Inter-coder agreement for computational linguistics [J]. *Comput linguist*, 2008, 34 (4): 555-596.
- 23 COHEN J. A coefficient of agreement for nominal scales [J]. *Educational and psychological measurement*, 1960, 20 (1): 37-46.
- 24 LU R S, LIN C C, TSAO H Y. Empowering large language models to leverage domain-specific knowledge in e-learning [J]. *Applied sciences*, 2024, 14 (12): 5264.
- 25 CUI Y, JIA M, LIN T Y, et al. Class-balanced loss based on effective number of samples [EB/OL]. [2025-05-09]. <https://arxiv.org/abs/1901.05555>.
- 26 GLM T, ZENG A, XU B, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools [EB/OL]. [2025-05-19]. <https://arxiv.org/abs/2406.12793>.
- 27 YANG A, YANG B, ZHANG B, et al. Qwen2.5 Technical Report [EB/OL]. [2025-05-19]. <https://arxiv.org/abs/2412.15115>.
- 28 ZHOU H, LI M, XIAO Y, et al. LEAP: LLM instruction-example adaptive prompting framework for biomedical relation extraction [J]. *Journal of the American medical informatics association*, 2024, 31 (9): 2010-2018.
- 29 IntSig TextIn. Acge\_text\_embedding. [EB/OL]. [2024-05-28]. [https://github.com/intsig-textin/acge\\_text\\_embedding](https://github.com/intsig-textin/acge_text_embedding).