



在线全文

癫痫临床文本信息的结构化标注与抽取研究*

金 灵^{1, 2, 3}, 何孟桥^{1, 2}, 巴桑斯甲^{1, 2}, 冯端宇⁴, 雷文强⁴, 陈 蕾^{1, 2, Δ}

1. 四川大学华西医院 神经内科(成都 610041); 2. 四川省脑机调控工程中心及四川省神经调控工程技术中心(成都 610041);
3. 四川大学华西医院 疾病分子网络前沿科学中心(成都 610041); 4. 四川大学计算机学院(成都 610041)

【摘要】目的 面向癫痫专科临床及科研需求,构建一套细粒度、高覆盖的中文临床文本命名实体标注体系,并验证其在命名实体识别(named entity recognition, NER)任务中的有效性。**方法** 从疾病、病程时间、临床表现、医疗检查、非药物治疗、用药及影响因素等7大维度设计3级标注体系,共制定25类标签,明确标签边界与特殊表达处理规则。以四川大学华西医院2009~2023年癫痫住院患者脱敏病历为数据来源,由3名具有癫痫临床背景的标注员完成804例高质量语料标注,采用双人复标、专家仲裁与实体级标注者一致性(inter-annotator agreement, IAA)评估保障标注质量。以五种中文医学预训练语言模型(Base-BERT、chinese-bert、chinese-Roberta、MC-BERT、MedBERT)分别结合BiLSTM-CRF与GlobalPointer两种序列标注框架,共10种组合方案进行NER性能验证,并在外部文献病例集上开展跨域泛化评估。**结果** 最终语料库涵盖25类癫痫相关实体,标注实例804条,实体总数28 400个;三位标注员IAA为0.86~0.88,标注一致性达到计算语言学通行标准。NER验证中,GlobalPointer框架表现优于BiLSTM-CRF,整体Micro-F1为0.906、Macro-F1为0.760;高频核心实体(如发作症状、药物名称、时间信息)F1值均超过0.90;跨域文献病例验证中高频标签F1值维持在0.80以上,低频标签(如fac-inc-amb、trt、dru-adv)F1值为0.31~0.57,主要受样本量有限与表达多样性影响。**结论** 本研究构建的癫痫专科中文临床标注体系覆盖范围广、粒度细、一致性高,可有效支持NER模型的训练与评估,为癫痫病历结构化分析及下游智能诊疗工具的开发提供了可复用的语料基础。

【关键词】 癫痫 电子健康档案 自然语言处理 命名实体识别 语料库

Structured Annotation and Information Extraction of Epilepsy Clinical Texts

Jin Ling^{1, 2, 3}, He Mengqiao^{1, 2}, Basangsijia^{1, 2}, Feng Duanyu⁴, Lei Wenqiang⁴, Chen Lei^{1, 2, Δ}.

1. Department of Neurology, West China Hospital, Sichuan University, Chengdu 610041, China; 2. Sichuan Provincial Engineering Research Center of Brain-machine Interactive Neurodulation, Chengdu 610041, China; 3. Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu 610041, China; 4. College of Computer Science, Sichuan University, Chengdu 610041, China

Δ Corresponding author

This study was supported by the Major Science and Technology Special Project of Sichuan Province (No. 2024ZDZX0018).

[Abstract] Objective To develop a fine-grained, highly comprehensive Chinese clinical texts named entity annotation schema tailored to the needs of epilepsy specialty clinical practice and research, and to validate its effectiveness in named entity recognition (NER) tasks. **Methods** A three-level annotation schema covering 25 entity types was designed across seven major dimensions, including disease, disease course timeline, clinical manifestations, medical examinations, non-pharmacological treatments, medication, and influencing factors, with explicit label boundary definitions and rules for handling special expressions. De-identified inpatient records of epilepsy patients admitted to West China Hospital, Sichuan University from 2009 to 2023 served as the data source. Three annotators with epilepsy clinical backgrounds completed high-quality annotation of 804 cases, with annotation quality ensured through double annotation, expert arbitration, and entity-level inter-annotator agreement (IAA) evaluation. NER performance was validated using 10 model combinations comprising five Chinese medical pre-trained language models (Base-BERT, chinese-bert, chinese-Roberta, MC-BERT, MedBERT) paired with two sequence labeling frameworks (BiLSTM-CRF and GlobalPointer), with additional cross-domain generalization evaluation on an external case dataset established on the basis of published literature. **Results** The final corpus contains 25 categories of epilepsy-related entities, 804 annotated cases, and a total of 28 400 entities. The IAA among the three annotators ranged from 0.86 to 0.88, indicating annotation

* 四川省重大科技专项项目(No. 2024ZDZX0018)资助

Δ 通信作者

出版日期: 2026-07-20

consistency that met the accepted standards in computational linguistics. In NER validation, the GlobalPointer framework outperformed BiLSTM-CRF, achieving an overall Micro-F1 score of 0.906 and Macro-F1 score of 0.760. High-frequency core entities (e.g., seizure symptoms, drug names, and temporal information) all yielded F1 scores exceeding 0.90. In cross-domain validation on literature-based cases, high-frequency entity F1 scores remained above 0.80, while low-frequency entities (e.g., factors with incomplete/ambiguous induction [fac-inc-amb], treatment information [trt], and adverse drug reactions [dru-adv]) achieved F1 scores of 0.31-0.57, primarily attributable to limited sample size and high linguistic variability. **Conclusion** The epilepsy-specific Chinese clinical annotation schema developed in this study demonstrates broad coverage, fine granularity, and high inter-annotator consistency. It effectively supports the training and evaluation of NER models and provides a reusable corpus foundation for the structured analysis of epilepsy medical records and the development of downstream intelligent diagnostic and therapeutic tools.

[Key words] Epilepsy Electronic health records Natural language processing Named entity recognition Corpus

癫痫是最常见的神经系统慢性疾病之一,全球约有 7 000 万例癫痫患者^[1],其中我国癫痫患者总数达 900 万,年新增病例约 40 万^[2]。该疾病的病理特征为大脑神经元异常放电导致的反复性、发作性症状,其临床表现具有高度异质性^[3]。准确的发作症状描述(包括症状特征、发作频率等)及发作环境信息(如时间、场所等)是癫痫诊断分型的关键依据^[4]。电子健康记录中积累了海量癫痫患者的临床诊疗文本,这些真实世界的医疗记录为癫痫疾病研究提供了宝贵的数据资源。然而,由于不同医师的表述习惯和临床经验存在差异,癫痫病历文本中症状描述存在显著的术语不一致、表达非标准化等问题,为数据结构化提取带来巨大挑战。

本研究针对以上问题,制定了癫痫临床文本的细粒度实体标注规则体系,涵盖症状、诱因、治疗等关键医学实体,并在医院脱敏病历数据上通过命名实体识别任务(named entity recognition, NER)验证了标注方案的有效性。该标注规范可为癫痫健康咨询、智能辅助诊断等应用提供标准化信息提取框架,并为癫痫文本语料库的批量化建设奠定方法论基础。

1 资料与方法

为系统推进癫痫临床文本的结构化信息抽取研究,本研究设计并实施了包含三个阶段的整体工作流程,如图 1 所示。首先基于国际标准与临床实际制定细粒度标注体系与规范;其次通过人工标注与规范调试构建高质量语料库;最后以多种主流命名实体识别模型进行性能验证,以评估语料的可行性与模型适配性。

1.1 标签体系及规则制定

本研究基于国际抗癫痫联盟(ILAE)的最新分类标准^[5]和国内《临床诊疗指南·癫痫病分册》的临床规范^[6],同时参考通用医学命名实体体系(UMLS^[7]、SNOMED CT^[8])

和中文电子病历标注体系^[9],针对临床实际需求^[10]构建了一套针对癫痫专科临床文本的细粒度标注体系。

在标注规则设计上严格遵循简单性原则,通过详细定义各类医学实体的子类划分,并辅以大量典型示例说明,使标注者无需具备深厚医学背景即可准确理解标注要求。同时,本研究在标注过程中尽量采用非嵌套原则,即对包含多个子实体的复合表述优先标注最外层完整实体,减少实体间交叉重叠带来的复杂性。但在部分特定标签中,如发作频率(clm-frq)和药物剂量(dru-dos),存在必要的嵌套情况,以确保信息完整性和剂量/频率细节的准确表达。此外,贯彻整体性原则,规定除临床表现外的其他医学实体(如疾病名称、药物等)均作为一个完整单元进行标注,不进行内部切分。

通过与神经内科专家团队开展多轮深度研讨,重点解决了癫痫专科文本中实体标签的界定与标准化问题。系统明确了各类癫痫相关临床表述应归属的具体标签类别,并严格界定了各标签的标注边界范围。通过典型病例讨论和标注实例分析建立详细的标注示例库,并制定特殊情况处理规则以及歧义鉴别说明,确立了统一的标注标准,确保不同标注人员对“标什么”和“怎么标”形成一致理解,有效解决了标签适用范围的模糊性问题,为后续标注工作奠定了规范基础。

1.2 人工标注及规则调整

本研究的数据来源于 2009—2023 年期间四川大学华西医院住院并被诊断为癫痫的患者,所有病例均根据国际疾病分类第十版(ICD-10)中的编码(G40或G40.x)确诊癫痫。数据使用经过四川大学华西医院伦理委员会审查批准,批准号 2023 年审(1208)号,研究豁免知情同意并对数据实施严格的隐私保护措施,包括删除患者姓名、登记号等标识信息。为匹配 NER 任务训练所需,研究以现病史文本作为主要分析对象,针对超过 500 字符的长文本进

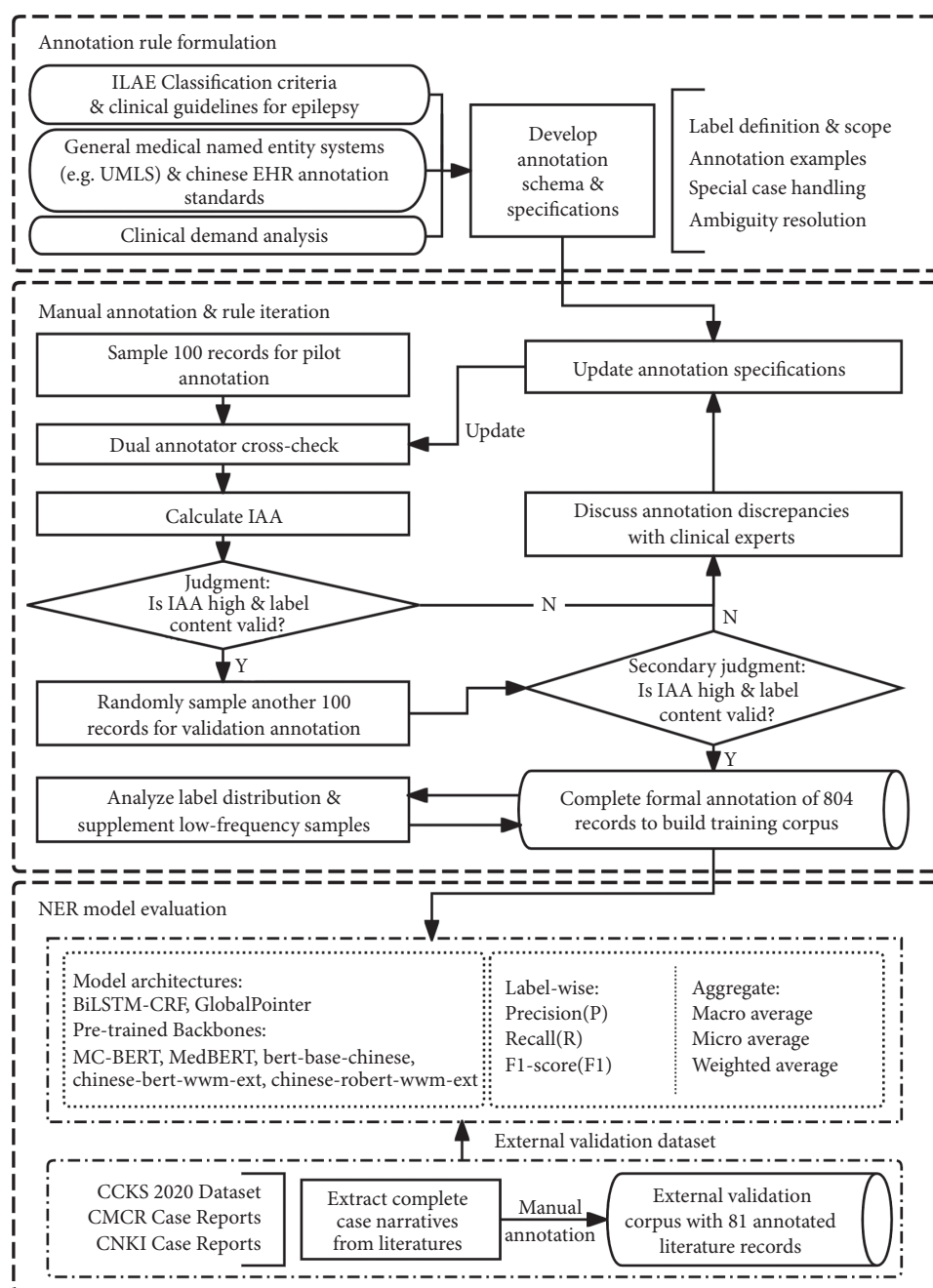


图1 研究总体流程

Fig 1 Overall research workflow

ILAE: International League Against Epilepsy; UMLS: Unified Medical Language System; EHR: electronic health record; IAA: inter-annotator agreement; NER: named entity recognition; P: precision; R: recall; F1: F1-score; CCKS: China Conference on Knowledge Graph and Semantic Computing; CMCR: Chinese Medical Case Reports; CNKI: China National Knowledge Infrastructure.

行分段处理,以距离500字符位置最近的句号为断点,将长文本切分为相对独立的片段,平衡语义连贯性与模型处理能力,为后续分析提供了高质量的数据基础。

基于制定的标签体系,人工标注过程通过严格的流程控制和质量评估保障数据可靠性。本研究由3名具有癫痫相关临床背景的研究者担任标注员,在正式标注前统一接受系统化培训和讨论,包括:(1)癫痫领域医学基

础知识与常见临床表述规范;(2)初版实体体系及标注规则的逐条讲解与示例演练;(3)Label Studio工具的操作培训(共约8 h);(4)试标阶段通过分批样本(20、50、100条)集中讨论,统一答疑并校准标注规则。同时培训过程本身也是标注体系进一步迭代与规则校准的过程,在试标阶段发现的问题均被整理并用于补充、细化标注指南。培训完成后,每位标注员独立完成试标并解决初步不一

致,随后进入正式标注阶段。整个标注及讨论复核过程累计耗时约1个月。项目期间每周开展一次规范复盘会议,对疑难样本进行讨论并持续对规则进行微调,确保标注标准在标注周期内保持一致。

为评估标注质量与标注者之间的一致性,本研究采用基于实体级匹配的一致性分析(inter-annotator agreement, IAA),即将两位标注员标注的结果视作系统输出及人工参考的关系计算精确率、召回率与F1值进行评估^[11]。该方法以实体边界与类型完全一致为匹配标准,适用于多类别NER任务中评估人-人一致性。

$$P = \frac{|A \cap B|}{|A|} \quad (1)$$

$$R = \frac{|A \cap B|}{|B|} \quad (2)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (3)$$

上式(1)~(3)中,集合A、B分别代表两名标注员标注得到的实体集合;|A|、|B|为对应集合内实体总数;|A ∩ B|为两者边界与类别完全匹配的重合实体数量;P为精确率,R为召回率,F1为综合一致性F1分数。对于分歧案例,标注员通过汇报形式与3名上级癫痫专家开展仲裁会议讨论,动态调整标注规则,例如明确发作频率和发作频率-模糊标签的子类范围以适配实际临床描述。最终完成的标注数据经过三轮交叉校验,当IAA达到0.8时,可认为语料库的一致性已满足计算语言学领域的通行标准^[12],为后续模型训练提供基础。

1.3 NER模型验证

在模型验证阶段,研究采用8:1:1的比例划分训练集、验证集和测试集,确保模型评估的可靠性。通过对五类预训练语言模型(Base-BERT、chinese-bert、chinese-Roberta、MC-BERT、MedBERT)与两种序列标注框架(BiLSTM-CRF和GlobalPointer)的10种组合方案进行系统性评估。BiLSTM-CRF模型通过双向LSTM网络捕捉上下文特征,结合CRF层优化标签序列预测^[13];GlobalPointer则利用全局指针网络有效处理嵌套实体识别^[14]。所有模型均重点调整了学习率、dropout、batch_size和权重衰减等关键参数。

在评估过程中,除标准的实体级别精确率、召回率和F1值外,研究特别关注总体的微平均F1值(Micro-F1)、宏平均F1值(Macro-F1)及加权平均F1值(Weighted-F1)三个维度的综合性能指标,以全面评估模型在不同类别上的表现均衡性。最终由神经内科专家对模型预测结果进行临床合理性盲审,重点验证“临床表现”子类的关键医

疗实体的识别准确性,确保模型输出既满足技术指标要求,又符合临床诊疗逻辑。

在自建语料库的验证基础上,研究进一步在文献病例数据上开展实验,以评估语料体系与模型在跨域文本中的适配性与泛化能力。文献病例主要来源于 CCKS 2020医疗命名实体识别数据集、中国临床案例成果数据库及中国知网(CNKI)。为避免特殊病例和误诊病例信息干扰,设定检索词“癫痫 AND NOT(标题:误诊)”进行检索与筛选,保留其中包含完整病例叙述的信息片段,构建外部验证集。与电子病历相比,文献病例在写作风格、结构化程度和实体表述方式上存在显著差异,因而能够更有效地检验模型在异构医学语料上的鲁棒性与泛化能力。

2 结果

2.1 标注规则及体系

本研究针对癫痫临床文本的结构化信息抽取需求,构建了一套面向任务的命名实体标注体系,共设计了25类标签,覆盖癫痫诊疗过程中的核心医疗信息。如表1所示,标注体系主要包括以下七大类:①疾病诊断信息,包括癫痫分型(dis-ept)及共患病与病史(dis-his),可支持对复杂诊断的精细表达;②时间信息(tim)用于统一标注癫痫患者共患病发生,发作时间,用药调整或治疗干预等关键节点上的时间相关描述,以明确疾病发展过程、病情变化趋势及整体治疗进展;③发作相关信息(clm),详细区分了发作频率(clm-frq、clm-frq-amb、clm-frq-num)、持续时间(clm-dur)、“发作前兆”与“发作结束”状态标识(clm-sta-bef、clm-sta-aft)、发作症状(clm-sym)及其复述表达(clm-sym-cpy),便于精准刻画患者的发作症状及其演变过程细节;④药物信息(dru)涵盖药物名称(dru-nam)、剂量(dru-dos)、用法(dru-mtd)、不良反应(dru-adv)及药物依从性(dru-com),并特别设置了“剂量引用”标签(dru-dos-cpy)用于处理文中引用先前药物信息的情形;⑤治疗信息(trt)用于标注药物治疗之外的癫痫治疗方式,主要包括脑部手术,以及针灸、埋线等中医干预手段;⑥检查相关信息分为检查项目(exa-itm)与检查发现(exa-fnd),便于识别脑电图、影像学等关键检查及其结果;⑦诱发因素及状态信息(fac)细化了诱因标签(fac-inc),并增设了明确诱因(fac-inc-def)、非明确诱因(fac-inc-amb)、无诱因(fac-inc-spo)三个子类,同时针对癫痫高发背景和特殊时期考虑设置了睡眠状态(fac-con-slp)、孕产期状态(fac-con-preg)两类环境状态标签。详细的标注示例库,特殊情况处理规则以及歧义鉴别说明见网络资源附件1。

表 1 标签分级方案
Table 1 Hierarchical label schema

Level-1 category	Level-1 tag	Level-2 category	Level-2 tag	Level-3 category	Level-3 tag
Disease	dis ^a	Epilepsy subtype	dis-ept	-	-
		Comorbidities and medical history	dis-his	-	-
Temporal information	tim	-	-	-	-
Clinical manifestation	clm ^a	Symptom	clm-sym	Symptom reference	clm-sym-cpy
		Seizure-stage indicator	clm-sta ^a	Preictal (Aura)	clm-sta-bef
				Postictal	clm-sta-aft
				-	-
		Symptom duration	clm-dur	-	-
		Seizure frequency	clm-frq	Number of seizures	clm-frq-num
				Ambiguous frequency expression	clm-frq-amb
Examination	exa ^a	Examination item	exa-itm	-	-
		Examination finding	exa-fnd	-	-
Treatment	trt	-	-	-	-
Drug	dru ^a	Drug name	dru-nam	-	-
		Drug dosage	dru-dos	Drug reference	dru-dos-cpy
		Drug administration	dru-mtd	-	-
		Adverse drug reaction	dru-adv	-	-
		Medication adherence	dru-com	-	-
Factor	fac ^a	Condition	fac-con	Sleep status	fac-con-slp
				Pregnancy/ peripartum status	fac-con-preg
				Spontaneous	fac-inc-spo
		Precipitating factor	fac-inc	Definite trigger	fac-inc-def
				Possible trigger	fac-inc-amb
				-	-

^a Hierarchy demonstration label only, not used in annotation.

标注体系具有任务适配性强、标签粒度细、结构清晰的特点,能够有效支持多源信息整合和下游模型训练。体系在设计中坚持实体非嵌套原则,强调以短语或句子为基本标注单位,贴合中文医学文本中实体的表达形式。此外,制定体系时也充分考虑标注的可执行性与一致性,对存在省略词或模糊用法的情形制定了明确处理策略,有效提升了标注质量与一致性。整体上该体系为构建结构化癫痫知识库和自动抽取系统提供了坚实的语料基础与技术保障。

2.2 人工标注结果统计

在人工标注结果方面,3位标注人员IAA最终为0.86~0.88,标签标注可靠。本研究共完成对癫痫相关临床文本的命名实体识别标注任务,涉及25类实体标签,标注实例总量804条,实体数达28 400个。

如表2所示,从整体统计来看标注数据的标签分布呈明显的不均衡特征,反映出临床记录中各类信息的重要性与出现频率。其中,发作症状类标签(clm-sym)数量最多,达到8 931条,占总标注数的31.45%,显示癫痫发作表现是记录的核心内容之一。药物相关信息(包括dru-nam、dru-dos、dru-mtd等)总量也较高,说明药物治疗在癫痫管理中的关键地位。时间信息(tim)标签共2 739条,

占比9.64%,在所有实体中排名第三,表明临床文本中对治疗、发作等事件的时间定位具有较高的记录频率。检查相关标签(exa-fnd和exa-itm)合计超过4 000条,显示检查项目及结果是诊疗过程中不可或缺的信息来源。以上实体在临床文本中频次高,临床关注度高,因此在后续的NER任务中优先关注识别表现,以辅助评估语料库构建的有效性。

表 2 各标签数量统计
Table 2 Entity label frequency statistics

Label	Count	Percentage	Label	Count	Percentage
dis-ept	182	0.64%	trt	113	0.40%
dis-his	715	2.52%	dru-nam	2 454	8.64%
tim	2 739	9.64%	dru-dos	1 648	5.80%
clm-sym	8 931	31.45%	dru-dos-cpy	83	0.29%
clm-sym-cpy	671	2.36%	dru-mtd	476	1.68%
clm-sta-bef	247	0.87%	dru-adv	38	0.13%
clm-sta-aft	854	3.01%	dru-com	196	0.69%
clm-dur	1 483	5.22%	fac-con-slp	360	1.27%
clm-frq	840	2.96%	fac-con-preg	12	< 0.10%
clm-frq-num	652	2.30%	fac-inc-def	361	1.27%
clm-frq-amb	369	1.30%	fac-inc-amb	114	0.40%
exa-itm	1 815	6.39%	fac-inc-spo	711	2.69%
exa-fnd	2 336	8.23%			

相比之下,诱因和状态信息(fac-inc与fac-con)标签分布较低,如fac-con-preg(孕产期)仅出现12次,尽管数量较少,但这类标签往往涉及癫痫发作的关键诱因或背景状态,具有较高的临床关注度。例如,fac-con-slp(睡眠状态)与夜间或清晨癫痫发作密切相关,有助于识别睡眠相关癫痫亚型;fac-con-preg则可能提示妊娠期癫痫管理的风险变化,影响治疗策略的制定。因此,尽管这类信息在文本中出现频率较低,但具有重要临床意义,对此类实体的保留与细化具有重要实用价值。考虑到其低频特性,在NER验证中不强求较高的识别性能,但可作为观察模型覆盖能力的辅助指标。

总体而言,该标注数据集中标签体系覆盖全面、细粒度设计合理,能够充分支持癫痫临床文本的结构化建模和自动抽取研究,为后续模型训练与任务评估提供数据基础。

2.3 NER测试结果

在模型验证方面,本研究选用BERT类的五种预训练

模型分别结合BiLSTM-CRF与GlobalPointer两种常用序列标注框架开展对比实验,旨在初步验证语料库标签体系的可识别性与实际应用潜力。

整体表现方面,结合BiLSTM-CRF的5种验证模型在验证集上取得了Micro-F1为0.865、Macro-F1为0.635、Weighted-F1为0.865的均值结果;而结合GlobalPointer的验证模型在相同BERT模型基础上取得更优的表现, Micro-F1为0.906、Macro-F1为0.760、Weighted-F1为0.904,表现出更好的实体边界建模能力和类别间平衡性。

从标签级别分析,如图2所示,采用表现较优的GlobalPointer框架结合不同BERT模型组合进行评估,见表3,结果显示在高频标签(如clm-sym、dru-nam、tim等)上取得了较高的F1值(均超过0.90),表明语料库中结构清晰、表达规范的实体具备良好的可识别性。部分低频标签(如fac-inc-amb、trt、dru-adv)的F1值相对较低(分别为0.31、0.57、0.47),主要受到样本量有限和语言表达差异较大的影响。

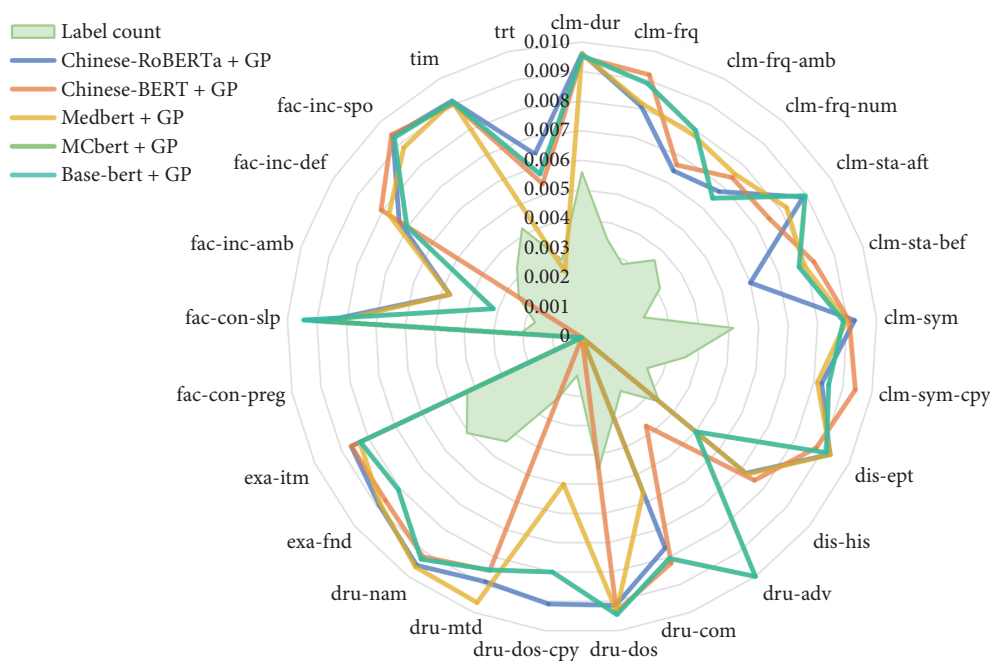


图2 最优五种模型组合的命名实体识别效果对比(F1分数)

Fig 2 Comparison of named entity recognition performance among the top 5 model combinations (F1 scores)

GP: GlobalPointer. Label counts are $\log_{10}$ -transformed for visualization.

在文献病例的跨域验证中,使用表现最优的chinese-Roberta+GlobalPointer框架进行测试,模型在高频核心标签(如clm-sym、dru-nam、tim)上仍保持较高准确率和F1值(均在0.80以上),趋势与内部验证类似。部分低频标签(如fac-inc-amb、trt、dru-adv)表现略有下降,主要受外部文本样本量有限以及语言表述差异较大的影响。总体

来看,外部验证进一步支持语料库设计的跨域适配性,同时提示低频类别仍有优化空间。

在进一步分析不同标签的误差结构时,低频实体(如fac-con-preg、fac-inc-amb、dru-adv、trt等)与高频实体在错误类型上呈现出明显差异。高频实体(如clm-sym、dru-nam、tim)通常具备规范、稳定的医学表达形式。模

表 3 各实体标签命名实体识别结果
Table 3 Named entity recognition results by entity label

Label	Test Set F1 ^a	External Validation F1 ^b	Label	Test Set F1	External Validation F1
dis-ept	0.913	0.471	trt	0.569	0.250
dis-his	0.641	0.675	dru-nam	0.937	0.902
tim	0.910	0.848	dru-dos	0.935	0.781
clm-sym	0.901	0.866	dru-dos-cpy	0.602	0.533
clm-sym-cpy	0.857	0.837	dru-mtd	0.879	0.854
clm-sta-bef	0.753	0.880	dru-adv	0.474	0.000
clm-sta-aft	0.850	0.680	dru-com	0.752	0.581
clm-dur	0.956	0.940	fac-con-slp	0.905	0.500
clm-frq	0.864	0.881	fac-con-preg	0.000	0.000
clm-frq-num	0.695	0.777	fac-inc-def	0.744	0.714
clm-frq-amb	0.739	0.414	fac-inc-amb	0.315	0.211
exa-itm	0.844	0.773	fac-inc-spo	0.924	0.990
exa-fnd	0.851	0.780			

GP: GlobalPointer. ^a The test set F1-score refers to the average F1 of the top five GlobalPointer-based models with optimal performance on the test set (Chinese-RoBERTa + GP, Chinese-BERT + GP, MedBERT + GP, MC-BERT + GP, Base-BERT + GP). ^b The external validation F1-score corresponds to the results of the Chinese-RoBERTa + GlobalPointer model tested separately on literature case samples.

型主要出现的是边界偏差类错误(如漏标并列症状、药物名后缀截断等),属于可通过局部上下文增强或标注一致性优化解决的问题。相较之下,低频实体的错误更具结构性特征:样本量有限使模型难以充分学习类内分布,典型表现为召回率显著偏低、模型无法激活对应标签(如初版dru-adv的F1为0的情况);部分低频标签本身存在较强表达歧义,例如fac-con-preg(孕产期相关状态)常与一般性的生理描述、时间节点(如“分娩后一天”)或症状性文本共现,模型容易误判为非状态类信息;fac-inc-amb由于意图表达主观性、上下文依赖强,模型在区分“明确诱因”与“不确定诱因”时容易产生类别混淆;trt标签的语言呈现高度多样化,如中医治疗、手术治疗、脑刺激等表述差异较大,导致编码特征难以聚合。

对错误样例的聚类分析也进一步显示,低频标签的F1偏低往往并非单一因素导致,而是类内样本不足,表达形式不一致,上下文依赖度高共同作用的结果。例如,在fac-con-preg标签中,部分描述涉及“妊娠期并发症”、“月经相关发作”等跨类别表达,使模型在不同语义层次之间难以确定边界。上述问题在高频实体中明显较轻,因为其语言模板更为固定,预测可以弱化对动态上下文的依赖^[15]。综合来看,低频实体的识别瓶颈主要源于标注规模与语言结构的双重限制,后续可通过扩充难例样本、引

入上下文敏感的对比学习、或采用标签级别的数据增强策略进一步缓解。例如dru-adv标签初期的F1值为0.00,但在补充13条针对性样本后,模型识别性能显著提升至0.47。

本研究以语料库的结构合理性与任务适配性验证为核心目标,NER实验结果显示,基于BERT+GlobalPointer框架的NER模型在癫痫临床文本的实体抽取任务中展现出较好的性能,尤其在结构化信息丰富的主流标签上识别效果稳定,初步证明语料设计具有较强的应用可行性。

3 讨论

本研究着眼于癫痫领域的实际临床需求,构建了一套面向中文专科病历的细粒度高覆盖的标注体系,该体系在时间阶段划分、模糊频率结构化、影响因素分层结构、上下文引用处理等多个维度实现创新,为癫痫专科领域的信息结构化提供了坚实语料基础。

癫痫临床文本早期研究多采用基于规则与医学本体的方法进行癫痫文本信息抽取。EpiDEA系统^[16]通过整合EpSO本体与cTAKES工具,从104份出院记录中抽取癫痫相关信息,达到88.53%的F1值。Decker等^[17]利用模式匹配和正则表达式扫描给定笔记中的癫痫发作频率短语,提取发作事件类型和定量频率。然而, Moonstone系统的评估^[18]指出,尽管规则提取方法在结构化文本中表现优异,但迁移到急诊记录等非训练环境,尤其是非结构化环境时会显著失效,表明此类方法对文本风格和语境高度敏感。随着机器学习的引入,信息抽取的适应性得以提升。Xia等^[19]构建了癫痫半定型本体(ESO)与实体提取工具,可自动抽取发作持续时间、症状表现等关键信息。Biswal等^[20]进一步通过朴素贝叶斯模型识别EEG报告中的癫痫发作与癫痫样放电。近年来, BERT、GPT等预训练语言模型在医学文本中逐步展现出优越性能。Xie等^[21]探索了基于BERT模型的预训练双向编码表征的微调,以分类癫痫无发作、提取癫痫发作频率,并提取最后一次癫痫发作的日期。Abeyasinghe等^[22]借助GPT-4实现了对发作频率相关短语的高质量抽取, F1值分别达到85.79%和91.84%,进一步印证了预训练模型在癫痫信息抽取中的有效性。目前基于仅编码器模型(如BERT及其衍生)的序列标记方法是此类自然语言理解(NLU)任务的首选模型^[23]。然而,在中文场景中,癫痫文本处理仍面临更多结构性与表达复杂性挑战,如嵌套实体、歧义表达和词序变动等特点^[24],传统方法难以准确解析。Man等^[25]提出层次化标注方法(Hierarchical NER),通过span-level分类策略,有效提升了嵌套实体的识别性能。BERN2等^[26]多语言系

统也开始引入癫痫相关实体的归一化处理。尽管如此,目前多数研究仍聚焦于英文语料或通用医学实体类别的识别,对中文癫痫专科文本中频繁出现的实体及模糊频率、时间状态、行为诱因、症状复述等真实表达现象尚缺乏系统的关注。

针对上述不足,本研究从癫痫专科实际出发进行标注体系设计。该体系在“临床表现”层级下引入病程阶段性标签,如“发作前”、“发作后”等,能够精确标注诸如“发作后1小时头痛”这类具有临床时序关系的表达,从而提升模型对发作过程语义依赖的识别能力。为区分精确频率与模糊频率信息,分别通过“精确频率(clm-frq)”与“模糊频率(clm-frq-amb)”标签进行建模,解决了“每周3次”与“偶尔发作”在传统模型中难以区分的问题,保留了真实文本中频率表达的多样性与不确定性,有助于提高数据的语义可解释性与建模实用性。在诱发因素建模方面,本体系系统性地纳入了行为、环境和生理状态等维度的标签,包括“熬夜后”“熟睡中”“妊娠期”等影响发作的因素或常见发作时所处状态的表述,通过结构化表示的方式,将此前难以建模的语义片段转化为具有明确边界的实体信息,为后续发作预测与风险干预研究提供了重要基础。此外,考虑到住院记录与随访文本中频繁存在的复述与省略表达,本体系设计了如“症状-参考(clm-sym-cpy)”与“药物-参考(dru-dos-cpy)”等上下文引用标签,用于标识“症状同前”“继续原剂量”类表述,从而增强了模型对上下文一致性与指代链的解析能力,尤其适用于信息复用率高的住院病程记录处理场景。该标注体系在标签设计、语义分层、上下文建模等多个维度上实现了面向癫痫领域的系统性创新,具备良好的应用拓展能力。

基于该体系构建的语料库,采用当前通用的NER模型进行训练测试,总体上取得较好的识别效果,验证了标注规则的适用性和有效性。然而,部分低频且语义复杂的标签识别依然存在较大的困难,这些问题普遍存在于癫痫临床文本自动结构化任务中,目前尚无成熟方法能够完全解决。未来需要依托更丰富的语料资源、数据增强技术及先进建模方法,结合跨领域研究持续攻克这些共性问题,以提升模型的识别稳定性和泛化能力。

本研究构建的实体体系不仅在结构化临床文本方面表现出良好的适配性,也具有明确的临床落地价值。Yang等^[27]证实基于长期神经生理学数据的发作预测具有临床可行性,关于“发作频率(clm-frq)”及其模糊表达的细粒度标签能够支持癫痫发作风险的动态评估,为基于时间序列的发作预警模型提供关键输入特征^[28-29]。药物相关标签(dru-nam、dru-dos、dru-mtd、dru-com、dru-

adv)可用于构建患者个体化的用药记录链,实现对剂量调整、不良反应识别和依从性下降等临床事件的自动捕捉,有助于临床医师的用药决策与监测^[30-31]。诱发因素(fac-inc)与状态标签(fac-con)能够辅助分析发作的潜在诱因与环境条件,为患者生活方式管理与发作避免策略提供支持^[32-33]。因此本研究的标注体系不仅能够用于构建高质量的训练语料,也为后续开发发作预警系统、用药管理工具及临床辅助诊疗模型奠定了语义基础,具有较强的实际应用潜力。

综上所述,本研究所提出的标注体系在癫痫临床真实文本中表现出良好的适应性与扩展潜力,为构建高质量中文医学语料和推进癫痫领域信息抽取方法研究提供了重要支撑,也为未来开展基于临床文本的精准诊疗工具构建提供了可能路径。

* * *

作者贡献声明 金灵负责论文构思、数据审编、正式分析、调查研究、研究方法、可视化和初稿写作,何孟桥负责数据审编、调查研究、研究项目管理、软件、验证和可视化,巴桑斯甲负责数据审编、调查研究、研究项目管理和验证,冯端宇负责论文构思、数据审编、研究方法、提供资源、监督指导和审读与编辑写作,雷文强负责论文构思、研究方法、提供资源、监督指导和审读与编辑写作,陈蕾负责论文构思、经费获取、提供资源、监督指导和审读与编辑写作。所有作者已经同意将文章提交给本刊,且对将要发表版本进行最终定稿,并同意对工作的所有方面负责。

Author Contribution Jin Ling is responsible for conceptualization, data curation, formal analysis, investigation, methodology, visualization, and writing--original draft. He Mengqiao is responsible for data curation, investigation, project administration, software, validation, and visualization. Basangsijia is responsible for data curation, investigation, project administration, and validation. Feng Duanyu is responsible for conceptualization, data curation, methodology, resources, supervision, and writing--review and editing. Lei Wenqiang is responsible for conceptualization, methodology, resources, supervision, and writing--review and editing. Chen Lei is responsible for conceptualization, funding acquisition, resources, supervision, and writing--review and editing. All authors consented to the submission of the article to the Journal. All authors approved the final version to be published and agreed to take responsibility for all aspects of the work.

利益冲突 本文作者陈蕾是本刊编委会青年编委。该文在编辑评审过程中所有流程严格按照期刊政策进行,且未经其本人经手处理。除此之外,所有作者声明不存在利益冲突。

Declaration of Conflicting Interests Chen Lei is a member of the Junior Editorial Board of the journal. All processes involved in the editing and reviewing of this article were carried out in strict compliance with the journal's policies and there was no inappropriate personal involvement by the author. Other than this, all authors declare no competing interests.

参 考 文 献

- [1] Thijs R D, Surges R, O'brien T J, et al. Epilepsy in adults[J]. *Lancet (London, England)*, 2019, 393(10172): 689. DOI: 10.1016/S0140-6736(18)

- 32596-0.
- [2] Ding D, Zhou D, Sander J W *et al.* Epilepsy in China: major progress in the past two decades[J]. *Lancet Neurol*, 2021, 20(4): 316. DOI: [10.1016/S1474-4422\(21\)00023-5](https://doi.org/10.1016/S1474-4422(21)00023-5).
 - [3] Boleti A P A, Cardoso P H O, Frihling B E F, *et al.* Pathophysiology to risk factor and therapeutics to treatment strategies on epilepsy [J/OL]. *Brain Sciences*, 2024, 14(1): 71. <https://www.mdpi.com/2076-3425/14/1/71>. DOI: [10.3390/brainsci14010071](https://doi.org/10.3390/brainsci14010071).
 - [4] Boßelmann C M. Seizures, semiology, and syndromes: a narrative review[J]. *Seizure*, 2021, 92: 230. DOI: [10.1016/j.seizure.2021.09.019](https://doi.org/10.1016/j.seizure.2021.09.019).
 - [5] Riney K, Bogacz A, Somerville E, *et al.* International League Against Epilepsy classification and definition of epilepsy syndromes with onset at a variable age: position statement by the ILAE Task Force on Nosology and Definitions[J]. *Epilepsia*, 2022, 63(6): 1443. DOI: [10.1111/epi.17240](https://doi.org/10.1111/epi.17240).
 - [6] 中国抗癫痫协会. 临床诊疗指南: 2023修订版. 癫痫病分册[M]. 北京: 人民卫生出版社, 2023.
Chinese Association Against Epilepsy. Clinical practice guidelines: 2023 revised edition. Epilepsy volume[M]. Beijing: People's Medical Publishing House, 2023.
 - [7] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology [J/OL]. *Nucleic acids research*, 2004, 32(Database issue): D267. https://academic.oup.com/nar/article/32/suppl_1/D267/2505235. DOI: [10.1093/nar/gkh061](https://doi.org/10.1093/nar/gkh061).
 - [8] Gaudet-Blavignac C, Foufi V, Bjelogrić M, *et al.* Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for processing free text in health care: systematic scoping review [J/OL]. *J Med Int Res*, 2021, 23(1): e24594. <https://www.jmir.org/2021/1/e24594>. DOI: [10.2196/24594](https://doi.org/10.2196/24594).
 - [9] 杜晋华, 尹浩, 冯嵩. 中文电子病历命名实体识别的研究与进展[J]. *电子学报*, 2022, 50(12): 3030. DOI: [10.12263/DZXB.20220485](https://doi.org/10.12263/DZXB.20220485).
 - [9] Du J, Yin H, Feng S. Research and development of named entity recognition in Chinese electronic medical record[J]. *Acta Electronica Sinica*, 2022, 50(12): 3030. DOI: [10.12263/DZXB.20220485](https://doi.org/10.12263/DZXB.20220485).
 - [10] Tao S, Lhatoo S, Hampson J, *et al.* A bespoke electronic health record for epilepsy care (EpiToMe): development and qualitative evaluation [J/OL]. *J Med Int Res*, 2021, 23(2): e22939. <https://www.jmir.org/2021/2/e22939>. DOI: [10.2196/22939](https://doi.org/10.2196/22939).
 - [11] Tjong Kim Sang E F, De Meulder F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition [C]//Proceedings of CoNLL-2003. Edmonton: Association for Computational Linguistics, 2003: 142-147. DOI: [10.3115/1119176.1119195](https://doi.org/10.3115/1119176.1119195).
 - [12] Artstein R, Poesio M. Inter-coder agreement for computational linguistics[J]. *Computational Linguistics*, 2008, 34(4): 555. DOI: [10.1162/coli.07-034-R2](https://doi.org/10.1162/coli.07-034-R2).
 - [13] Jehangir B, Radhakrishnan S, Agarwal R. A survey on named entity recognition—datasets, tools, and methodologies [J/OL]. *Natural Language Processing Journal*, 2023, 3: 100017. <https://www.sciencedirect.com/science/article/pii/S2949719123000146>. DOI: [10.1016/j.nlp.2023.100017](https://doi.org/10.1016/j.nlp.2023.100017).
 - [14] Su J, Murtadha A, Pan S, *et al.* Global pointer: novel efficient span-based approach for named entity recognition [EB/OL]. arXiv: 2208.03054, 2022. <https://arxiv.org/abs/2208.03054>. DOI: [10.48550/arXiv.2208.03054](https://doi.org/10.48550/arXiv.2208.03054).
 - [15] Geirhos R, Jacobsen J H, Michaelis C, *et al.* Shortcut learning in deep neural networks[J]. *Nat Mach Intell*, 2020, 2(11): 665. DOI: [10.1038/s42256-020-00257-z](https://doi.org/10.1038/s42256-020-00257-z).
 - [16] Cui L, Bozorgi A, Lhatoo S D, *et al.* EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification [C]//AMIA Annual Symposium Proceedings. Chicago: American Medical Informatics Association, 2012: 1191-1200.
 - [17] Decker B M, Turco A, Xu J, *et al.* Development of a natural language processing algorithm to extract seizure types and frequencies from the electronic health record[J]. *Seizure*, 2022, 101: 48. DOI: [10.1016/j.seizure.2022.07.010](https://doi.org/10.1016/j.seizure.2022.07.010).
 - [18] Reeves R M, Christensen L, Brown J R, *et al.* Adaptation of an NLP system to a new healthcare environment to identify social determinants of health[J/OL]. *J Biomed Inform*, 2021, 120: 103851. <https://www.sciencedirect.com/science/article/pii/S1532046421001805>. DOI: [10.1016/j.jbi.2021.103851](https://doi.org/10.1016/j.jbi.2021.103851).
 - [19] Xia Y, He M, Basang S, *et al.* Semiology extraction and machine learning-based classification of electronic health records for patients with epilepsy: retrospective analysis [J/OL]. *JMIR Med Inform*, 2024, 12: e57727. <https://medinform.jmir.org/2024/1/e57727>. DOI: [10.2196/57727](https://doi.org/10.2196/57727).
 - [20] Biswal S, Nip Z, Moura Junior V, *et al.* Automated information extraction from free-text EEG reports [C]//2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Milan: IEEE, 2015: 6804-6807. DOI: [10.1109/EMBC.2015.7319956](https://doi.org/10.1109/EMBC.2015.7319956).
 - [21] Xie K, Gallagher R S, Conrad E C, *et al.* Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing[J]. *J Am Med Inform Assoc*, 2022, 29(5): 873. DOI: [10.1093/jamia/ocac018](https://doi.org/10.1093/jamia/ocac018).
 - [22] Abeyasinghe R, Tao S, Lhatoo S D, *et al.* Leveraging pretrained language models for seizure frequency extraction from epilepsy evaluation reports [J/OL]. *NPJ Digital Medicine*, 2025, 8(1): 208. <https://www.nature.com/articles/s41746-025-01592-4>. DOI: [10.1038/s41746-025-01592-4](https://doi.org/10.1038/s41746-025-01592-4).
 - [23] Deuffer T, Hillebrand L, Bauckhage C, *et al.* Informed named entity recognition decoding for generative language models [C]//2024 IEEE International Conference on Big Data (BigData). Washington, DC: IEEE, 2024: 6001-6010. DOI: [10.1109/BigData62323.2024.10825603](https://doi.org/10.1109/BigData62323.2024.10825603).
 - [24] Yew A N J, Schraagen M, Otte W M, *et al.* Transforming epilepsy research: a systematic review on natural language processing applications[J]. *Epilepsia*, 2022, 64(2): 292. DOI: [10.1111/epi.17474](https://doi.org/10.1111/epi.17474).
 - [25] Man J, Hu Z, Liu H, *et al.* A model for epilepsy named entity recognition based on Chinese EEG reports [C]//Proceedings of the 2024 5th International Symposium on Artificial Intelligence for Medicine Science. Chengdu: Association for Computing Machinery, 2024: 467-472. DOI: [10.1145/3706890.3706971](https://doi.org/10.1145/3706890.3706971).
 - [26] Sung M, Jeong M, Choi Y, *et al.* BERN2: an advanced neural biomedical named entity recognition and normalization tool[J]. *Bioinformatics*, 2022, 38(20): 4837. DOI: [10.1093/bioinformatics/btac598](https://doi.org/10.1093/bioinformatics/btac598).
 - [27] Yang H, Müller J, Eberlein M, *et al.* Seizure forecasting with ultra long-term EEG signals[J]. *Clin Neurophysiol*, 2024, 167: 211. DOI: [10.1016/j.clinph.2024.09.017](https://doi.org/10.1016/j.clinph.2024.09.017).
 - [28] Jones F J S, Smith J R, Ayub N, *et al.* Implementing standardized provider documentation in a tertiary epilepsy clinic [J/OL]. *Neurology*, 2020, 95(2): e213. <https://www.neurology.org/doi/10.1212/WNL.0000000000009778>.
 - [29] Linehan C, Benson A, Gunko A, *et al.* Exploring the prevalence and profile of epilepsy across Europe using a standard retrospective chart review: Challenges and opportunities[J]. *Epilepsia*, 2021, 62(11): 2651. DOI: [10.1111/epi.17057](https://doi.org/10.1111/epi.17057).
 - [30] Khemani D B, Malave D S, Shinde S, *et al.* AI-driven pharmacovigilance: enhancing adverse drug reaction detection with deep learning and NLP [J/OL]. *MethodsX*, 2025, 15: 103460. <https://www.sciencedirect.com/science/article/pii/S221501612500305X>. DOI: [10.1016/j.mex.2025.103460](https://doi.org/10.1016/j.mex.2025.103460).
 - [31] Powell J R, Cook J, Wang Y, *et al.* Drug dosing recommendations for all patients: a roadmap for change[J]. *Clin Pharmacol Therap*, 2021, 109(1): 65. DOI: [10.1002/cpt.1923](https://doi.org/10.1002/cpt.1923).
 - [32] Beniczky S, Rylvlin P. Mobile health and digital technology in epilepsy: the dawn of a new era [J]. *Epilepsia*, 2023, 64 (S4): S1. DOI: [10.1111/epi.17813](https://doi.org/10.1111/epi.17813).
 - [33] Nuthalapati P, Thomas L, Donahue M A, *et al.* Improving seizure frequency documentation and classification [J/OL]. *Neurology*, 2023, 13(6): e200212. <https://www.neurology.org/doi/10.1212/CPJ.0000000000200212>.

(2025-07-16收稿, 2026-06-29修回)

编辑 吕 熙



开放获取 本文使用遵循知识共享署名—非商业性使用4.0国际许可协议(CC BY-NC 4.0), 详细信息请访问

<https://creativecommons.org/licenses/by-nc/4.0/>。

OPEN ACCESS This article is licensed for use under Creative Commons Attribution-NonCommercial 4.0 International license (CC BY-NC 4.0). For more information, visit <https://creativecommons.org/licenses/by-nc/4.0/>.

© 2026 《四川大学学报(医学版)》编辑部

Editorial Office of Journal of Sichuan University (Medical Sciences)