# 脑卒中疾病电子病历实体及实体关系标注语料库构建*

常洪阳[1,2]，眷红英[1,2]，马玉团[1,2]，张坤丽[1,2]
1.郑州大学　信息工程学院，河南　郑州
2.鹏城实验室,广东　深圳
1175975869@qq.com；iehyzan@zzu.edu.cn；
458882652@qq.com；ieklzhang@zzu.edu.cn

## 摘要

本文探讨了在脑卒中疾病中文电子病历文本中实体及实体间关系的标注问题，提出了适用于脑卒中疾病电子病历文本的实体及实体关系标注体系和规范。在标注体系和规范的指导下，进行了多轮的人工标注及校正工作，完成了158万余字的脑卒中电子病历文本实体及实体关系的标注工作。构建了脑卒中电子病历实体及实体关系标注语料库（Stroke Electronic Medical Record entity and entity related Corpus, SEMRC）。所构建的语料库共包含命名实体10,594个，实体关系14,457个。实体名标注一致率达到85.16%，实体关系标注一致率达到94.16%。

**关键词：** 语料库构建；脑卒中疾病；命名实体；实体关系

# Corpus Construction for Named-Entity and Entity Relations for Electronic Medical Records of Stroke Disease

**Hongyang Chang [1,2]，Hongying Zan[1,2]，Yutuan Ma[1,2]，Kunli Zhang[1,2]**
1.School of Information Engineering, Zhengzhou University, Zhengzhou, Henan, China
2.The Peng Cheng Laboratory,Shenzhen, Guangdong, China
1175975869@qq.com；iehyzan@zzu.edu.cn；
458882652@qq.com；ieklzhang@zzu.edu.cn

## Abstract

This paper discussed the labeling of Named-Entity and Entity Relations in Chinese electronic medical records of stroke disease, and proposes a system and norms for labeling entity and entity relations that are suitable for content and characteristics of electronic medical records of stroke disease. Based on the guidance of the labeling system and norms, this carried out several rounds of manual tagging and proofreading and completed the labeling of entities and relationships more than 1.5 million words. The entity and entity relationship tagging corpus of stroke electronic medical record（Stroke Electronic Medical Record entity and entity related Corpus, SEMRC）is fromed. The constructed corpus contains 10,594 named entities and 14,457 entity relationships. The consistency of named entity reached 85.16%, and that of entity relationship reached 94.16%.

**Keywords:** Corpus construction , Stroke disease , Named entity , Entity relations

河南省医学科技攻关计划省部共建项目（SB201901021），郑州市协同创新重大专项科技攻关项目（20XTZX11020）

## 1 引言

脑卒中(cerebral stroke) (胡钟竞, 2018)俗称脑中风, 是由于脑部血管突然破裂（即脑出血）或血管阻塞导致血液不能流入大脑（即脑梗塞）而引起脑组织损伤的一组疾病。据2020年世界卫生组织[0]公布，脑卒中是全球范围的第二大杀手，占世界死亡总人数11%，而在中国，根据科普中国网[1]显示，脑卒中已然成为中国死亡原因第一位，同时也是中国成年人致残的首要元凶。因此，对脑卒中疾病进行研究是非常有必要的，而构建脑卒中电子病历实体及实体关系标注语料库是智能分析的基础。

电子病历是指医务人员在医疗活动过程中，使用医疗机构信息系统生成的文字、符号、数据、图表、图形、影像等数字化信息，并能实现存储、管理、传输和重现的医疗记录，是病历的一种记录形式，包括门（急）诊病历和住院病历，记录了病人从入院到出院期间诊断治疗全部过程的诊疗信息，包含了大量真实可靠的病情信息 (中华人民共和国国家卫生和计划生育委员会, 2017)，如“于外伤后出现头懵不适感”、“头CT（2018-07-21我院）:1.硬膜下出血2.蛛网膜下腔出血。”等。对这些电子病历文本进行实体及实体关系标注并构建语料库对后续的相关研究具有重大意义。（本文出现电子病历若无特殊说明都是指中文电子病历）

对海量的电子病历进行人工标注的代价是昂贵的，因此对电子病历的处理常常需要借助自然语言处理、机器学习和深度学习等技术进行自动抽取。由于医学文本信息的领域特点，使用通用语料库训练过的网络模型不能很好的应用到医学文本的信息抽取中，而且电子病历属于半结构化文本，不方便机器自动处理。因此，构建脑卒中疾病电子病历的实体及实体关系标注语料库将为脑卒中疾病的健康咨询、智能辅诊等相关研究提供可靠的数据基础。

本文的主要工作是针对脑卒中疾病电子病历文本，探讨实体及实体间关系，创立脑卒中疾病电子病历标注规范体系，构建脑卒中疾病电子病历文本的实体及实体关系标注语料库。

## 2 相关研究

### 2.1 医学文本信息抽取及语料标注

对于医学信息的抽取，i2b2(Informatics for Integrating Biology & the Bedside)举行的公开评测引起了大家浓厚的兴趣。在2006年举办的患者抽烟状态识别任务中 (Özlem et al., 2008)，i2b2把患者抽烟的状态定义成了五个类别，该评测在2008年又加入了对电子病历中肥胖及其并发症进行抽取的任务，同时在标注中引进了推断机制，检查实体的属性值如血糖值、血脂值等能够对患者状态进行定量表述的描述，对于这些数值型的描述也进行了标注 (Özlem, 2009)。在2009年i2b2组织的评测任务加入了对电子病历中药物相关信息的抽取 (Zlem et al., 2010)。2010年i2b2的评测任务发起了倡议，希望参与评测的队伍可以在电子病历中抽取出医疗概念、医疗问题及对问题的修饰，并且能够识别出医疗问题与治疗、检查之间存在的关系 (Zlem et al., 2011)。在2012年i2b2举行的评测任务中加入了抽取电子病历中的时间信息及医疗事件与时间之间的关系 (Sun et al., 2013)。在2014年i2b2组织的评测任务中进行了糖尿病类患者的电子病历中心脏病风险因素的抽取 (Stubbs and Uzuner, 2015)。除了i2b2，还有一些其他研究者做了相关的工作，Meystre等 (2015)构建了对医疗问题标注相关修饰词信息的医疗术语标注语料库、梅奥诊所 (Savova et al., 2010)首次对实体及关系的修饰信息进行细致分类，Leonardo Campillos等 (2018a)构建了法语语种的命名实体及实体关系语料库，以及一些其他的相关工作，如对医疗事件之间的关系 (Rink et al., 2011)、电子病历中的时间信息 (Styler et al., 2014)、医疗术语和实体 (Meystre et al., 2017)、对实体和实体关系进行修饰的信息 (Campillos et al., 2018b)等信息做了处理。

中文医疗信息抽取领域近些年来也取得了许多的成果。杨锦峰等 (2014)在构建中文电子病历命名实体和关系语料库过程中采用了以预标注的方法训练标注人员更新标注规范的模式在标注结果上取得了较好的一致性。Lei等人 (2014a; 2014b)借鉴i2b2组织2010年的实体分类，在2013年把病历中出现的治疗进一步划分为了过程及药物，并于2014年抽取研究了电子病历中

[0]World Health Organization. The top 10 causes of death, available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[1]我国脑卒中发生居高不下，请收下这份“协和医生秘籍”：https://cloud.kepuchina.cn/newSearch/imgText?id=6750802040109207552

出现的检查、药物、治疗过程及医疗问题等。Wu等人 (2015)在Lei等人 (2014a)标注的语料库上使用深度学习算法识别电子病历中的命名实体。昝红英等人 (2020a; 2020)在所构建的面向儿科疾病的实体及实体关系语料库中抽取多元组，构建了儿科医学知识图谱；基于目前国内医学领域信息抽取发展现状对深度学习模型在这一领域的应用及未来发展趋势做了总结。张坤丽等人 (2020)于2019年以构建中文医学知识图谱任务为基础,构建出能够实现半自动化的实体及关系标注平台，即本文标注过程中所采用的平台。

### 2.2 中文医学语料库

Lei等人 (2014a)于2013年收集了协和医院的800份电子病历并由两名专家医生进行标注构建了命名实体标注语料库。2016年杨锦峰等人 (2016)在922份病历文本基础上构建了中文电子病历命名实体和实体关系语料库。2017年苏嘉等人 (2019)在中文健康信息处理领域构建了第一份关于心血管疾病风险因素的语料库。昝红英等人 (2020a; 2020b)利用自行开发的标注工具构建了包含常见疾病504种的面向儿科疾病的实体及关系标注语料库；并于2019年在原有的医学命名实体及关系标注体系的基础上结合了症状的特征、概念等及症状在医学影像中所发挥的作用，构建了一个共包含了8,772种症状和146,631条关系的症状知识库。Tongfeng Guan等人 (2020)基于教科书、电子病历等多种数据来源构建了中国医学信息提取数据集(Chinese Medical Information Extraction，CMeIE)

## 3 脑卒中疾病电子病历实体及实体关系标注体系的制定

参考昝红英等人 (2020b; 2020a)提出的中文电子病历命名实体和实体关系标注规范和面向儿科疾病的实体及关系标注语料库中使用的标注规范及Tongfeng Guan等人 (2020)使用的标注规范，在临床医生的专业指导下，本文制定了适用于脑卒中疾病电子病历内容特点的标注规范。图 1为脑卒中电子病历标注体系示意图。按照疾病及其与症状之间的关系以及疾病和症状分别与检查、手术治疗、药物治疗、其他治疗、修饰和时间等实体之间关系来说明脑卒中疾病电子病历标注规范。

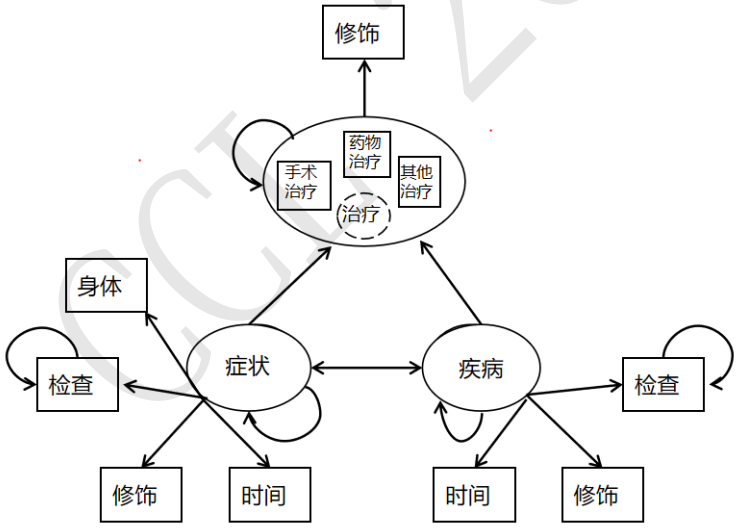


图 1: 脑卒中疾病电子病历实体及实体关系标注体系示意图

### 3.1 疾病、症状及其之间的关系

在脑卒中疾病电子病历标注过程中，疾病实体是指患者在一定条件下受到病因的损害作用后，机体因自稳调节紊乱而引发的异常生命活动的过程或是医生针对患者情况做出的诊断。疾病的概念范围用ICD-10和MeSH此表中编码为C的疾病概念来界定，但同时不局限于词表中的概念，借助百度百科和医学百科等辅助确认疾病概念。

在此次症状的标注过程中，主要参考了《中文症状库》 (昝红英et al., 2020b)和《诊断学》中的症状实体。本文没有专门区分症状与异常检查结果（体征），而是统一当作症状标注，即

患者自述或家属转述或者医生通过观察、仪器等方法检查到患者出现的异常结果都标为症状实体。

疾病与症状之间存在的关系：疾病导致症状。实例："4余年胃镜检查提示胃溃疡，后复查胃镜恢复，平时易出现胃部不适"中出现的疾病与症状之间的三元组为<胃溃疡，疾病导致症状，胃部不适>。

### 3.2 治疗及其与疾病、症状之间的关系

治疗是指因疾病或症状而施加给患者的治疗程序、药物给予、干预实施等。本文认为治疗可以通过治疗的定义、手段和方法等可以再做更加精细的划分，因此在本次研究中不单独出现"治疗"实体，而是分别拆分成为了"手术治疗"、"药物治疗"和"其他治疗"。

手术治疗指通过针、刀、剪等医疗器械在患者身体局部进行割、切、缝合等操作来完成维持患者健康目的的过程，通常用于外科治疗。本次标注中主要通过ICD-9-CM和MeSH词表中E编码的手术概念以及病历中明确指出患者通过某种手术进行治疗来界定手术实体的范围。药物是指能够对机体的生理功能或代谢活动产生影响的化学物质，此次标注对药物实体范围的界定主要为ATC、MeSH词表中D编码的药物以及病历中明确指出患者使用过或出现在用药指导部分的药物。其他治疗主要包括放射治疗、辅助治疗、化疗以及其他要完成一定治疗目的，如：营养神经、清除自由基、改善循环等。

治疗分别与疾病和症状之间的关系见下表 1。实例："4年前因声带息肉行"声带手术"；"一句中存在手术治疗与疾病之间三元组<声带手术，治疗施加于疾病，声带息肉>。

| 实体1 | 关系类型 | 实体2 | 实体1 | 关系类型 | 实体2 |
|---|---|---|---|---|---|
| 手术治疗 | 治疗改善了症状 | 症状 | 手术治疗 | 治疗改善了症状 | 疾病 |
| 手术治疗 | 治疗恶化了症状 | 症状 | 手术治疗 | 治疗恶化了症状 | 疾病 |
| 手术治疗 | 治疗导致了症状 | 症状 | 手术治疗 | 治疗导致了症状 | 疾病 |
| 手术治疗 | 治疗施加于症状 | 症状 | 手术治疗 | 治疗施加于症状 | 疾病 |
| 手术治疗 | 因为症状而没有采取治疗 | 症状 | 手术治疗 | 因为症状而没有采取治疗 | 疾病 |
| 药物治疗 | 治疗改善了症状 | 症状 | 药物治疗 | 治疗改善了症状 | 疾病 |
| 药物治疗 | 治疗恶化了症状 | 症状 | 药物治疗 | 治疗恶化了症状 | 疾病 |
| 药物治疗 | 治疗导致了症状 | 症状 | 药物治疗 | 治疗导致了症状 | 疾病 |
| 药物治疗 | 治疗施加于症状 | 症状 | 药物治疗 | 治疗施加于症状 | 疾病 |
| 药物治疗 | 因为症状而没有采取治疗 | 症状 | 药物治疗 | 因为症状而没有采取治疗 | 疾病 |
| 其他治疗 | 治疗改善了症状 | 症状 | 其他治疗 | 治疗改善了症状 | 疾病 |
| 其他治疗 | 治疗恶化了症状 | 症状 | 其他治疗 | 治疗恶化了症状 | 疾病 |
| 其他治疗 | 治疗导致了症状 | 症状 | 其他治疗 | 治疗导致了症状 | 疾病 |
| 其他治疗 | 治疗施加于症状 | 症状 | 其他治疗 | 治疗施加于症状 | 疾病 |
| 其他治疗 | 因为症状而没有采取治疗 | 症状 | 其他治疗 | 因为症状而没有采取治疗 | 疾病 |

表 1: 治疗与疾病、症状间关系

### 3.3 检查及其与疾病、症状之间的关系

检查指为了查清证实患者是否患有某种疾病或具有某些症状而通过特定的技术、医疗仪器设备而进行的检查项目、手段、过程等，为医生的临床诊断和治疗提供依据。为界定检查覆盖范围，避免标注歧义检查限于以下三种：1）诊疗计划、辅助检查及治疗过程中提到的检查手段，如："头CT"、"头颈联合CT"、"头颅磁共振"等；2）体液检查项目、生理指标、生理测量及其他检查项目，后面通常根由表示指标值或测量值的数值。如："体温36.7℃"、"血压134／87mmHg"、"甘油三酯2.32mmol/L"等；3）病历中直接指出的检查，如："查"、"检查"、"示"、"查体"、"试验"等。由于我们在后续工作中采用深度学习算法，这些算法对于数值数字并不敏感，因此在检查项目出现的指标数值结果我们没有进行标注，只标注了其中的检查项目。

检查与疾病间存在的关系：检查证实了疾病、为了证实疾病而采取的检查；检查与症状间存在的关系：检查证实了症状、为了证实症状而采取的检查。实例："查体：伸舌右偏，"一句

中存在检查与症状之间三元组<查体，检查证实了症状，伸舌右偏>。

### 3.4 身体及其与症状之间的关系

身体包括部位、器官或身体位置、区域及身体系统、器官

参考中文电子病历命名实体和实体关系标注规范中没有身体或者部位的实体，但经过我们对脑卒中疾病电子病历的分析认为身体实体是有必要的尤其是当症状与身体部位之间并不直接相连，如“双侧额顶叶、双侧侧脑室周围脑白质脱髓鞘”、“双侧小脑半球、左侧桥小脑结合臂含铁血黄素沉淀”等，部位与部位之间有间隔，如果不添加身体实体会造成大量的信息确实并影响了电子病历本身的严谨真实性。

身体与症状之间存在的关系：位置。当身体部位与症状不能够直接相连时，则将其标注为：<身体，位置，症状>。实例：“双侧额叶、左侧顶叶点状白质脱髓鞘”一句中出现身体与症状之间三元组<双侧额叶，位置，点状白质脱髓鞘>、<左侧顶叶，位置，点状白质脱髓鞘>。

### 3.5 修饰及其与疾病、症状和治疗之间的关系

电子病历中的一些对疾病、症状及治疗等实体进行定性或定量非数值的描述，如：“无饮水呛咳”中无字、“头晕稍好转”中的稍好转、“脑梗死可能性大”中的可能性大等，在脑卒中电子病历标注过程中我们将其标注为修饰实体。

修饰与疾病、症状和治疗之间的关系见表 2。实例：“主诉：视物不清2天加重1天”一句中存在修饰与症状之间三元组<加重1天，严重程度，视物不清>。

| 实体1 | 关系类型 | 实体2 | 实体1 | 关系类型 | 实体2 | 实体1 | 关系类型 | 实体2 |
|---|---|---|---|---|---|---|---|---|
| 修饰 | 性质 | 症状 | 修饰 | 性质 | 疾病 | 修饰 | 既往的 | 手术治疗 |
| 修饰 | 严重程度 | 症状 | 修饰 | 严重程度 | 疾病 | 修饰 | 否认的 | 手术治疗 |
| 修饰 | 否认 | 症状 | 修饰 | 否认 | 疾病 | 修饰 | 当前的 | 手术治疗 |
| 修饰 | 非患者本人的 | 症状 | 修饰 | 非患者本人的 | 疾病 | 修饰 | 既往的 | 药物治疗 |
| 修饰 | 有条件的 | 症状 | 修饰 | 有条件的 | 疾病 | 修饰 | 否认的 | 药物治疗 |
| 修饰 | 可能的 | 症状 | 修饰 | 可能的 | 疾病 | 修饰 | 当前的 | 药物治疗 |
| 修饰 | 待证实的 | 症状 | 修饰 | 待证实的 | 疾病 | 修饰 | 既往的 | 其他治疗 |
| 修饰 | 偶有 | 症状 | 修饰 | 偶有 | 疾病 | 修饰 | 否认的 | 其他治疗 |
| | | | | | | 修饰 | 当前的 | 其他治疗 |

表 2: 修饰与疾病、症状和治疗间关系

### 3.6 时间及其与疾病、症状之间的关系

在脑卒中电子病历标注过程中，本文将病历中出现的与疾病或症状有直接关联的时间点、时间段标注为时间实体。时间与疾病之间存在的关系：既往、持续、将来；时间与症状之间存在的关系：既往、持续、将来。实例：“主诉：视物不清2天加重1天”一句中存在时间与症状之间三元组<2天，持续，视物不清>。

### 3.7 同类实体间关系

在标注过程中，我们发现在电子病历中会经常出现在同一区域有多个同类实体对应同一个或多个实体，由于我们使用的可视化图形标注工具 (张坤丽et al., 2020)，为提高标注过程中对标注人员的友好性和标注效率，我们将出现在同一区域且与同一个或多个实体对应的多个同类实体标注为实体组。实体组关系可以根据后期任务需要进行拆分后分别与对应关系实体进行组合。定义的实体组关系有：<疾病，实体组，疾病>、<症状，实体组，症状>、<检查，实体组，检查>、<手术治疗，实体组，手术治疗>、<药物治疗，实体组，药物治疗>及<其他治疗，实体组，其他治疗>。实例：“无头晕头痛，恶心呕吐”一句中“无”字分别修饰“头晕头痛”和“恶心呕吐”，由于文本数量多及平台特征，如果分别单独标注会造成标注人员难以辨认，因此将“头晕头痛”和“恶心呕吐”标注为实体组<头晕头痛，实体组，恶心呕吐>。

### 3.8 标注原则补充

在本次语料库构建过程中遵循的医学实体标注基本原则:

- 非重复标注原则: 即在一段医学文本中所提及到的实体，其只能属于一种确定的实体类型;
- 非嵌套标注原则: 即全部的实体都是相对独立的，不能作为其他实体的子集;
- 规范性原则: 即标注过程中，实体中不应包含普通文本与标点符号的组合且尽量不包含“或、及、和”等连接词。

## 4 SEMRC语料库构建

构建语料库最为主要的工作就是制定合理的标注规范，并严格地依据规范进行语料标注。以上述制定的脑卒中疾病电子病历实体及实体关系标注体系为基础，在领域专家指导下，本文制定了标注规范的初稿，选定基础标注平台并开发出适用于本体系的标注工具。目前主流的预料标注模式有三种：（1）领域专家标注，适用于专业知识储备要求高的专业领域语料的标注，该模式能够极大程度的保证语料标注的质量，但也存在标注成本高，语料构建周期长等弊端。（2）众包标注，这种模式能够较为明显的降低较大规模语料标注的成本，但只能用于简单的预料构建任务，且标注过程中要巧妙设计以保证标注质量。（3）团体标注，该构建语料库的标注模式与信息检索评价集构建较为类似，能够支持标注过程中不依赖领域专家的情况下构建出质量较高的语料，但同时对标注成员的要求较高。为了兼顾标注质量和标注成本和周期，在语料标注模式选择中本文选择了领域专家+团体标注模式。

### 4.1 数据准备

本文从河南某三甲医院筛选了共200份的电子病历，其中每份电子病历选取包括：入院记录、病程记录（拆分为首次病程记录和查房记录）、出院小结及出院医嘱等作为标注数据集。在这200份患者病历中脑出血患者病历有90份，脑梗塞有110份。在标注之前需要先进行数据脱敏处理，即去除电子病历中的敏感信息，如患者姓名、身份证号、联系方式、家庭住址、工作单位和医生姓名等。

### 4.2 标注规范的制定和标注人员的培训

在脑卒中疾病电子病历实体及实体关系标注体系基础上制定标注规范初稿。在人员培训方面本文采用了预标注+众包标注的模式，在标注人员详细阅读过标注规范初稿后选用另一批不在最终语料库中的相同来源电子病历，将经过格式转换和脱敏等预处理后部署在脑卒中电子病历实体及实体关系标注平台上进行预标注（预标注后的数据不再使用），我们希望通过预标注来达到两个目的：1）完成标注人员对标注过程的熟悉、对标注规范的深入理解；2）集合多位标注人员智慧总结脑卒中电子病历的特点，完成对标注规范初稿的修改完善，形成标注规范v2.0版本。

规范v2.0版本基本完善后在此基础上开始正式标注，对于每一批数据的标注分为四个阶段。第一阶段：在一标文件中完成包括对疾病、症状、手术治疗、药物治疗、其他治疗、修饰、时间、检查及身体等实体的分类标注，同时记录每位标注人员标注过程中的疑惑，对于疑惑问题定期进行讨论解决并在一标文件中进行修改；第二阶段：在一标文件基础上生成二标文件，现有标注人员进行交叉检查，对于有异议的依据规范进行讨论解决，规范没有定义不能很好解决的问题统一讨论解决并修订补充规范；第三阶段：实体标注基本统一后继续由一标人员在二标文件上依据规范v2.0版本中实体关系标注部分进行标注实体间关系，第三、四阶段是对实体关系的标注，过程与第一、二阶段一样。对于选定的198位患者的电子病历我们在标注过程中将其分成了3个批次进行标注，而人员的培训及规范的完善也随着这3个批次的进行而循环进行。整个语料构建过程如图 2所示。

### 4.3 标注中的问题、经验及特点

在标注过程中遇到问题（1）首要问题便是标注人员主要由无医学领域专业基础的计算机专业研究者。为解决这一问题，本文采用了预标注的方法，即在专业医师指导下参考已有标注规

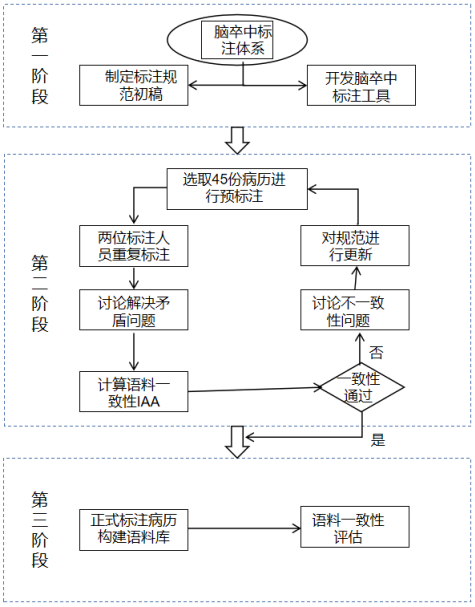


图 2: 脑卒中标注流程图

范和体系制定出最初的标注规范，标注人员学习规范后对病历进行标注，补充专业背景同时发现共性问题，在专业医师共同参与下修订出与特定领域更为契合的新规范，如，在制定最初的标注体系时主要参考借鉴了昝红英等人 (昝红英et al., 2020a)中提出的体系及经验，预标注过程中发现仅以疾病类实体为中心的标注体系与本文收集到的病历数据无法完全融合，因此我们根据预标注中积累的经验提出了以疾病类和症状类实体为双中心词的标注体系。（2）标注中遇到不在规范中的疑问。针对这个问题，我们将疑问进行记录，定期进行集中讨论，更正补充规范后对疑问进行统一化。

本文构建的脑卒中疾病电子病历实体及实体关系标注语料库，数据来源均为脑卒中疾病电子病历，相较于中文电子病历命名实体和实体关系语料库 (杨锦锋et al., 2016)、面向儿科疾病的实体及关系标注语料库 (昝红英et al., 2020a)、中文症状知识库 (昝红英et al., 2020b)等更贴合真实案例且数据特征更加一致，对特定的脑卒中疾病具有充分的针对性，为特定于脑卒中疾病领域的专项深层次研究提供了数据支撑。

## 5 构建结果及分析

文献 (Artstein et al., 2008)中指出当标注一致性评价结果到达0.8时即可判定语料的一致性是可以被接受的。一致性计算公式如式(1)-(3)所表示：

$$P = \frac{A_1 \bigcap A_2}{A_2} \tag{1}$$

$$R = \frac{A_1 \bigcap A_2}{A_1} \tag{2}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{3}$$

A1、A2表示在同一份病历文件上两名不同标注人员的标注结果。

最终标注语料库的一致性评价结果如表 3 所示。本次构建的脑卒中语料库的实体一致性达到了0.85，实体关系一致性达到了0.94.其中脑出血病历标注的实体及实体关系一致性分别为0.84和0.94；脑梗塞病历标注一致性分别为0.86和0.94。表 3结果表明本文最终构建的语料库是可信赖的。

本文主要以脑卒中疾病电子病历作为基础构建标注语料库，介绍了语料标注的过程和体系。整个脑卒中疾病实体及实体关系标注语料库的构建过程历时五个月，共由主任医师1名，副主任医师1名，计算机硕士研究生5名共同参与完成了电子病历标注语料库构建的工作。本次标

| | 脑出血 | | 脑梗塞 | | 脑卒中 | |
|---|---|---|---|---|---|---|
| | 实体 | 关系 | 实体 | 关系 | 实体 | 关系 |
| P | 0.7865 | 0.8942 | 0.8090 | 0.8876 | 0.7983 | 0.8907 |
| R | 0.9035 | 0.9987 | 0.9204 | 0.9986 | 0.9124 | 0.9987 |
| F | 0.8410 | 0.9436 | 0.8611 | 0.9398 | 0.8516 | 0.9416 |

表 3: 脑卒中标注语料库一致性结果

注共完成了标注1,582,962字，实体概念10,594个，实体关系三元组14,457种，具体标注实体及实体关系数量如图3、图4所示。

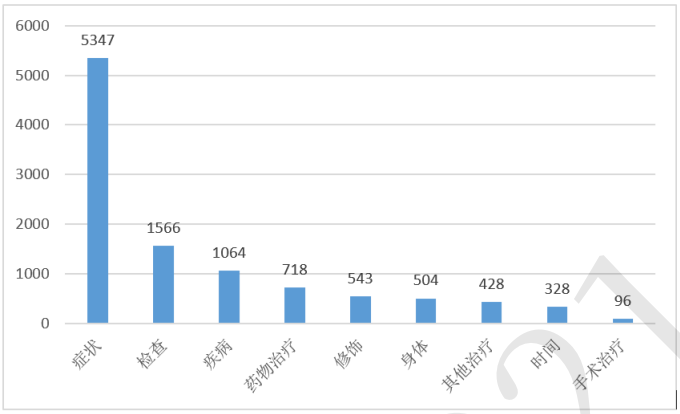


图 3: 标注实体数量

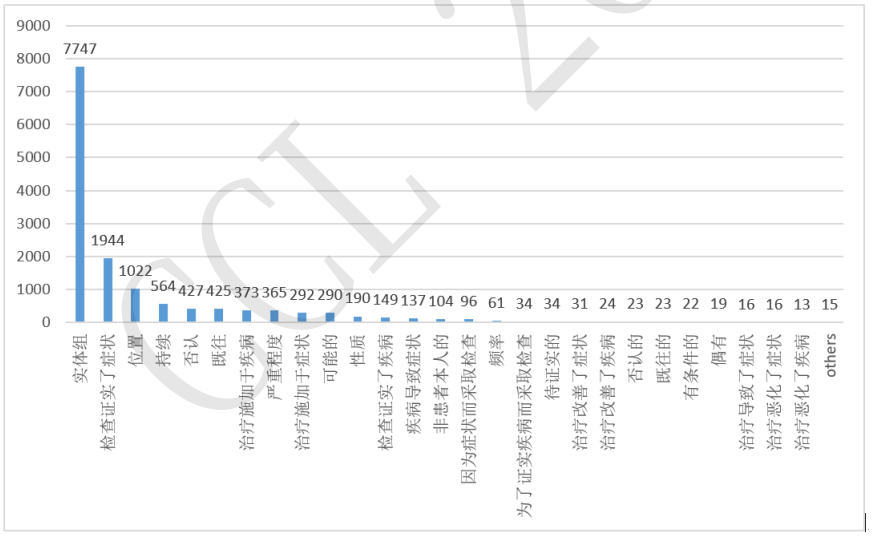


图 4: 标注关系数量

## 6 结语

本文主要对脑卒中疾病电子病历实体及实体关系标注过程进行了探究，具体的从以下的三个方向进行探讨：首先，构建了一套适用于脑卒中疾病电子病历实体及实体关系的标注体系；其次制定了与本文构建体系相对应的语料标注规范；最后根据标注体系和规范构建了中文脑卒中疾病电子病历实体及实体关系标注语料库SEMRC。在体系构建、标注规范的确立及完善、标注过程中所遇到问题的解决都有医学专家的参与和指导，这使本文制定的标注准则具备较强的领域专业性，能够在后续的研究工作中提供一定的科学指导。在语料的标注过程中我们采用

了领域专家+众包的标注模式并结合标注人员预标注培训的标注思想，语料库的较高的一致性结果也肯定了本文的标注方法。在对脑卒中电子病历进行标注的过程中本文根据病历语料的特点提出了以疾病和症状为双头实体，以治疗（包括手术治疗、药物治疗和其他治疗）为副头实体，以时间、修饰、身体、检查等作为从属属性实体的标注体系。本文此次所构建的基于中文电子病历脑卒中实体及实体关系标注语料库可以为使用机器学习算法进行自动抽取、通过计算机技术对脑卒中疾病进行更深层次的探索提供基础。

## 参考文献


Artstein, Ron, Poesio, and Massimo. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596.

L. Campillos, L. Deleger, C. Grouin, T. Hamon, A. L. Ligozat, and A. Neveol. 2018a. A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52(2):571–601.

L. Campillos, L Deléger, C. Grouin, T. Hamon, and A Névéol. 2018b. A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources & Evaluation*, 52(2):1–31.

Tongfeng Guan, Hongying Zan, Xiabing Zhou, Hongfei Xu, and Kunli Zhang. 2020. Cmeie: Construction and evaluation of chinese medical information extraction dataset. In Xiaodan Zhu, Min Zhang, Yu Hong, and Ruifang He, editors, *Natural Language Processing and Chinese Computing*, pages 270–282, Cham. Springer International Publishing.

J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, and X. Hua. 2014a. A comprehensive study of named entity recognition in chinese clinical text. *J Am Med Inform Assoc*, (5):808–814.

J. Lei, B. Tang, X. Lu, K. Gao, J. Min, and X. Hua. 2014b. Research and applications: A comprehensive study of named entity recognition in chinese clinical text. *Journal of the American Medical Informatics Association Jamia*, 21(5):808.

S. Meystre and P. J. Haug. 2015. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *Journal of Biomedical Informatics*, 39(6):589–599.

S. M. Meystre, K. Youngjun, G. T. Gobbel, M. E. Matheny, R. Andrew, B. E. Bray, and J. H. Garvin. 2017. Congestive heart failure information extraction framework for automated treatment performance measures assessment. *J Am Med Inform Assoc*, (e1):ocw097.

B. Rink, S. Harabagiu, and K. Roberts. 2011. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600.

G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sunghwan, K. C. Kipper-Schuler, and C. G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association Jamia*, 17(5):507.

A. Stubbs and O. Uzuner. 2015. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *Journal of Biomedical Informatics*, 58(Suppl):S78–S91.

William F. Styler, S. Bethard, S. Finan, M. Palmer, S. Pradhan, Piet C De Groen, B. Erickson, T. Miller, C. Lin, and G. Savova. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

W. Sun, R. Anna, and U. Ozlem. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association Jamia*, (5):806–813.

Y. Wu, J. Min, J. Lei, and X. Hua. 2015. Named entity recognition in chinese clinical text using deep neural network. *Studies in health technology and informatics*, 216:624–628.

J. F. Yang, Q. B. Yu, Y. Guan, and Z. P. Jiang. 2014. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. *Acta Automatica Sinica*, 40(8):1537–1562.

Uzuner Zlem, S. Imre, and C. Eithon. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association Jamia*, (5):514.

Uzuner Zlem, B. R. South, S. Shen, and Scott L Duvall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association Jamia*, (5):552.

U Özlem, G. Ira, Y. Luo, and K. Isaac. 2008. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15:14–24.

U Özlem. 2009. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc*, (4):561–570.

中华人民共和国国家卫生和计划生育委员会. 2017. 电子病历应用管理规范(试行). 中国实用乡村医生杂志, (6).

张坤丽, 赵旭, 关同峰, 尚柏羽, 李羽蒙, and 眷红英. 2020. 面向医疗文本的实体及关系标注平台的构建及应用. 中文信息学报, (6):36–44.

眷红英, 刘涛, 牛常勇, 赵悦淑, 张坤丽, and 穗志方. 2020a. 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. 中文信息学报, (5):19–26.

眷红英, 韩杨超, 范亚鑫, 牛承志, 张坤丽, and 穗志方. 2020b. 中文症状知识库的建立与分析. 中文信息学报, v.34(04):33–40.

眷红英、关同峰、张坤丽、奥德玛、穗志方. 2020. 面向医学文本的实体关系抽取研究综述. 郑州大学学报*(*理学版*)*, v.52(04):4–18.

杨锦锋, 关毅, 何彬, 曲春燕, 于秋滨, 刘雅欣, and 赵永杰. 2016. 中文电子病历命名实体和实体关系语料库构建. 软件学报, 27(11):2725–2746.

胡钟竞. 2018. 脑卒中中西医治疗的最新研究进展. 中国医药指南, 016(017):39–41.

苏嘉, 何彬, 吴昊, 杨锦锋, 关毅, 姜京池, 王焕政, and 于秋滨. 2019. 基于中文电子病历的心血管疾病风险因素标注体系及语料库构建. 自动化学报, 45(002):420–426.