|  |  |
| --- | --- |
| ICS |  |
| CCS | |  | | --- | |  | |

团体标准

淋巴水肿辅助诊断数据标注与质控标准

2023 - XX - XX发布

2023 - XX - XX实施

XXX学会  发布

目次

[[前 言 II](#_Toc20447)](#_Toc25588)

[[引 言 III](#_Toc20447)](#_Toc9641)

[[1 范围 4](#_Toc20447)](#_Toc28054)

[[2 规范性引用文件 4](#_Toc20447)](#_Toc29654)

[[3 术语和定义 4](#_Toc20447)](#_Toc26010)

[[4 数据标注规程框架 5](#_Toc20447)](#_Toc15968)

[[5 数据标注规则 5](#_Toc20447)](#_Toc4703)

[[5.1 定义所需标注数据与数据量统计 5](#_Toc20447)](#_Toc397)

[[5.2 实体标注规则 6](#_Toc20447)](#_Toc21550)

[[5.3 命名实体标注细则 7](#_Toc20447)](#_Toc22854)

[[5.4 分类混淆处理细则 11](#_Toc20447)](#_Toc15620)

[[5.5 数据标注执行及注意事项 12](#_Toc20447)](#_Toc5749)

[[5.6 标注说明中术语体系规范化 13](#_Toc20447)](#_Toc28340)

[[5.7 标注工具和标注平台 13](#_Toc20447)](#_Toc29582)

[[5.8 标注任务的创建、分发、开展和回收 13](#_Toc20447)](#_Toc817)

[[5.9 标注结果质量检查和控制 14](#_Toc20447)](#_Toc27368)

[[5.10 标注结果交付验收 15](#_Toc20447)](#_Toc17197)

[[参考文献 16](#_Toc20447)](#_Toc31944)

1. 前言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由中国生物医学工程学会提出。

本文件由中国生物医学工程学会知识产权与标准化工作委员会归口。

本文件起草单位：首都医科大学附属北京世纪坛医院、中国科学院自动化研究所

本文件主要起草人：沈文彬、孙宇光、夏松、苏万春、徐毅、汤昊宬、冀秀元、丁振甲

1. 引言

淋巴水肿是由于淋巴系统的先天性或获得性障碍，导致局部体液滞留和组织肿胀，是严重影响人们健康和生活质量的疾病之一。据国际淋巴学会统计，目前全世界患者数约1.7亿，我国的患病人数超过一千万，尤其是继发性淋巴水肿患者数量正伴随着癌症或肿瘤患者人数快速增长，每年接受乳腺肿瘤、妇科肿瘤、男性泌尿生殖系统手术治疗的患者出现淋巴水肿的比例高达20%-60%，通常在术后几个月甚至数年后出现，开始时没有明显症状，但随着病情逐步加重，多数患者将面临肢体功能障碍，严重者可致残。及时准确的筛查和诊断淋巴水肿是早治疗的基础，对防止疾病的恶化和提升治疗效果的至关重要。从研究通用数据集角度来看，全球范尚无通用公开的淋巴系统疾病数据，淋巴水肿文本类辅助诊断通用数据集领域完全空白。

规范化的淋巴水肿数据集是支撑淋巴学科研究、推进学科向智能化发展的核心基石，也是促进临床医学+工程信息学+人工智能三元融合，形成同质性、可及性医疗服务开展的关键要素。本文件制定的淋巴水肿数据集数据采集要求，填补了目前世界范围内对淋巴水肿文本数据库及数据采集标准化的空白，对支撑产业界研发淋巴外科辅助诊断类人工智能医疗器械产品，推动人工智能技术在淋巴外科临床实际应用和持续健康发展具有重要意义。

淋巴水肿辅助诊断数据标注与质控标准

* 1. 范围

本文件主要规定了淋巴水肿数据集建设中数据标注内容、数据标注规则、数据标注过程、数据标注技术、数据标注质量控制及数据标注安全控制的关键内容。

本文件适用于淋巴水肿AI（人工智能）医疗器械的研发、生产、检测过程中的训练、测试、评估数据集建立及其评估。

* 1. 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 5271.1 信息技术 词汇 第1部分：基本术语

YY/T 1833.1 人工智能医疗器械 质量要求和评价 第 1 部分：术语

YY/T 1833.2 人工智能医疗器械 质量要求和评价 第 2 部分：数据集通用要求

YY/T 1833.3 人工智能医疗器械 质量要求和评价 第 3 部分：数据标注通用要求

* 1. 术语和定义

下列术语和定义适用于本文件。

数据标注 data annotation

对文本、图像、语音、视频等待标注数据进行归类、整理、编辑、纠错、标记和批注等操作，为待标注数据增加标签，生成满足机器学习训练要求的机器可读数据编码。

数据标签 data label

标识数据的特征、类别和属性等，用于建立数据及机器学习训练要求所定义的机器可读数据编码间的联系。

标注任务 annotation task

按照数据标注规范对指定数据集进行标注的过程。

数据标注员 data labeler

对文本、图像、语音、视频等待标注数据进行归类、整理、编辑、纠错、标记和批注等操作的工作人员。

标注工具 annotation tool

数据标注员完成标注任务产生标注结果时所需的工具和软件。

注1：标注工具可生成标签并提供参考模板。

注2：不同数据类型和标注任务需要不同的标注工具。

标注平台 annotation platform

开展标注任务的系统化工作台。

注：标注平台在包含标注工具全部功能的基础上将所有标注环节工具化，可有效地对标注任务进行全局管理和跟踪。

标注说明规则 annotation instruction

数据需求方用于明确标注任务和标注数据的书面陈述，包含执行标注任务所需的标注工具、任务描述、标注方法、正确示例、常见错误等内容。

标注方法 annotation method

定义数据标注员进行数据标注时的环境和规程，包含标注对象定义、所用标注工具和标注平台、标注格式、标注前的准备工作、标注后的处理工作等。

* 1. 数据标注规程框架

数据标注规程框架应包含：

1. 标注任务前期准备：
2. 定义淋巴水肿数据集待标注规模，准备待标注数据；
3. 制定标注说明规则，包括标注规则定义、标注规则内容、标注工具和平台等；
4. 确定标注人员的选拔要求，开展标注前培训；
5. 标注任务中：
6. 创建淋巴水肿数据集标注任务；
7. 淋巴水肿数据集标注任务的分发；
8. 淋巴水肿数据集标注任务的开展；
9. 淋巴水肿数据集标注任务的回收；
10. 标注结果输出：
11. 淋巴水肿数据集标注结果的质量检查和控制；
12. 淋巴水肿数据集标注数据的交付验收。
    1. 数据标注规则
       1. 定义所需标注数据与数据量统计

数据标注前应完成的准备工作：

1. 分析数据。确定机器学习和模型训练过程中所需的标注数据类型、量级、用途及应用场景等。本标准制定的淋巴水肿数据集构建以淋巴外科淋巴水肿专病电子病历文本为主体，数据规模拟定在3,000例，初步用于淋巴水肿辅助诊断CDSS产品的研发，未来可支撑面向淋巴水肿疾病的所有人工智能辅助产品的研发、生产、测试、评估；
2. 整理数据。明确数据与标签文件存放的目录结构，在任务分配与回收时，按照指定的目录进行数据组织；
3. 明确命名规则。应明确数据与标签文件的命名方式，命名规则避免数据更新迭代时的重名，便于数据追踪、标注追踪，且数据文件名与标签文件名保持一致。本标准指定的淋巴水肿数据为关键敏感数据，因此在命名规则上，以患者姓名+患者唯一识别码+患者入院日期为表头，以SHA-256算法做加密处理，形成唯一的特征码做文件名。
4. 评估数据量。根据标注任务的人员分配模式、工具选择、标注任务类型、算法选择对数据量对应的工作任务进行评估。
5. 标注数据定义与需求量。明确标注数据定义并确定最终的标注需求量。
   * 1. 实体标注规则
        1. 标注实体分类

本标准将医学实体划分为十五大类，包含“疾病”、“疾病诱因”、“疾病诊断分型”、“临床症状”、“医疗程序”、“医疗设备”、“药物”、“医学检验项目”、“身体”、“科室”、“病程”、“频次”、“微生物类”、“当前修饰”和“否认修饰”。本指南借鉴UMLS（一体化医学语言系统 Unified Medical Language System）语义类型界定实体涵盖的范围，但不局限于UMLS的定义。

* + - 1. 标注分类定义
         1. 疾病

第一类标注命名实体是疾病（dis），疾病是指导致病人处于非健康状态的原因或者医生对病人做出的诊断，并且是能够被治疗的。包括疾病或综合征、中毒或受伤、器官或细胞受损，其对应的UMLS语义类型有疾病或者综合征（disease or syndrome）、中毒或受伤（injury or poisoning）等；常见的疾病标注实体如“淋巴水肿”、“高血压”、“糖尿病”等。

* + - * 1. 疾病诱因

第二类标注命名实体是疾病诱因（di），疾病诱因是指某个有害因素作用于机体、达到一定的强度和之间会产生某个特定疾病，本标准中特定疾病统一指的是淋巴水肿。常见的疾病诱因标注实体如“冶游史”、“放化疗史”、“饮酒史”、“血吸虫病疫接触史”等。

* + - * 1. 疾病诊断分类

第三类标注命名实体是疾病诊断分型（dt），疾病的具体分类，表示疾病的进展程度，疾病诊断分类一般出现在诊断中；常见的疾病诊断分类标注实体如“继发性”、“原发性”、“轻度”、“慢性”等。

* + - * 1. 临床症状

第四类标注命名实体是临床症状（sym），临床症状是疾病的表现，泛指患者不适感觉以及通过检查得知的异常表现。主要包括症状、体征，其对应的UMLS语义类型有症状或体征（sign or symptom）、异常检查结果（abnormal test results）等；常见的临床症状分类标注实体如“肿胀”、“触痛”、“皮肤破溃”、“肿大淋巴结”等。

* + - * 1. 医疗程序

第五类标注命名实体是医疗程序（pro），医疗程序泛指为诊断或治疗所采取的措施、方法及过程。主要包括检查程序、治疗或预防程序，其对应的UMLS语义类型有化验过程（laboratory procedure）、治疗或预防过程（therapeutic or preventive procedure）等；常见的医疗程序分类标注实体如“手术治疗”、“预防接种”、“放化疗”、“吸脂减容术”等。

* + - * 1. 医疗设备

第六类标注命名实体是医疗设备（equ），医疗设备泛指为诊断或治疗所使用的工具、器具、仪器等。主要包括检查设备、治疗设备，其对应的UMLS语义类型有医疗设备（medical device）、药物传输设备（drug delivery device）等；常见的医疗设备分类标注实体如“弹力袜”、“弹力袖套”、“绷带”、“气压泵”等。

* + - * 1. 药物

第七类标注命名实体是药物（dru），药物是指用来预防、治疗及诊断疾病的物质，其对应的UMLS语义类型有临床药物（clinical drug）、抗生素（antibiotic）等；常见的药品分类标注实体如“抗生素”、“二甲双胍”、“青霉素”、“阿司匹林”等。

* + - * 1. 医学检验项目

第八类标注命名实体是医学检验项目（ite），医学检验项目是指检查涉及到的体液检查项目、重要生理指标以及其他检查项目，本标准规定医疗检验项目主要针对人体而言，是能够通过设备或实验检测出的项目，并且是能够被量化，有其对应的测量值或指标值。其对应的UMLS语义类型有实验室检查（laboratory test）等；常见的医学检验项目分类标注实体如“淋巴现象”、“核磁”、“心脏彩超”、“术后病理”等。

* + - * 1. 身体

第九类标注命名实体是身体（bod），身体泛指细胞、组织、及位于人体特定区域的由细小物质成分组合而成的结构、器官、系统、肢体，另外包括身体产生或解剖身体产生的物质等。主要包括身体部位、身体物质，其对应的UMLS语义类型有身体部位（body part）、组织（organ）、组织成分（organ componment）等；常见的身体分类标注实体如“左下肢”、“双下肢”、“大腿”、“右上肢”等。

* + - * 1. 科室

第九类标注命名实体是科室（dep），科室主要是指医院或医疗机构所设有的科室其对应的医疗保健相关组织（healthcare related organization）等；常见的科室分类标注实体如“当地医院”、“我院门诊”、“中国医学科学院肿瘤医院”、“北京协和医院”等。

* + - * 1. 病程

第十一类标注命名实体是病程（dp），病程指的是患某种疾病的整个过程，本标准中特指淋巴水肿疾病；常见的病程分类标注实体如“1年余”、“5年前”、“3年余”、“2年余”等。

* + - * 1. 频次

第十二类标注命名实体是频次（qd），频次指的是频率和次数，本标准中频次包含患者疾病症状发生频率、用药给药频率等，既包含定性描述又包含定量数据；常见的频次分类标注实体如“偶有”、“不断”、“反复”、“1片/天”等。

* + - * 1. 微生物类

第十三类标注命名实体是微生物类（mic），微生物类包括细菌、病毒、真菌以及一些小型的原生生物、显微藻类等在内的一大类生物群体，另外包括微生物类产生的毒素、激素、酶等，其对应的UMLS语义类型有细菌（virus）、真菌（fungus）等；微生物类分类标注实体如“抗酸杆菌”。

* + - * 1. 当前修饰

第十四类标注命名实体是当前修饰（pr），当前修饰属于实体修饰，反应了实体与患者（病情）的时间关联关系；常见的当前修饰分类标注实体如“目前”、“至今”、“当前”等。

* + - * 1. 否认修饰

第十五类标注命名实体是否认修饰（neg），否认修饰也属于实体修饰，反应了实体与患者的否定逻辑关联关系；常见的否认修饰分类标注实体如“否认”、“无”、“未见”等。

* + 1. 命名实体标注细则
       1. 疾病
          1. 疾病或综合征

疾病或综合征是指疾病或综合征名称。比如高血压、肺炎、心脏病、败血症、畸形等。

测定结果的分析：[肺炎]dis、[败血症]dis、严重[先天性心脏病]dis或[畸形]dis影响新生儿代谢和循环功能，特别是[严重感染]sym时，可[微循环障碍]dis和[DIC]dis

* + - * 1. 器官或细胞损伤

器官、细胞等发生异常或损伤后，如果能够危及人的机体，此时虽然它们属于身体的一部分，但是已成为一种致病因素，危害人体健康。还有[颅内出血]dis（[产伤]dis、[外伤]dis），[颅脑损伤]dis，[脑血管畸形]dis。

在标注“疾病”实体时，需要注意：

（1）一般有些疾病名称很长，前面会有“XX性”、“XX状”、“XX型”等，以及身体部位（一个或多个）的修饰，在保证疾病完整性和具体性的情况下，在标注时应该与这些前缀一起标注。

（2）大部分统称均标注，比如营养性疾病、代谢性疾病、化脓性和非化脓性综合征等，此类虽然是统称，但是有对应的疾病范畴，所以应该标注。特殊情况：像常见病、多发病、疾病等单独出现时，此类统称范围太广，不应标注。

（3）当疾病有若干种分型时，“疾病+分型”或“分型+疾病”整体标注，分型单独出现不标注。可将[MPS]dis分为I~VII型，除[MPSII型]dis为X连锁隐性遗传外

* + - 1. 疾病诱因

患者因其他疾病或经历医疗操作，例如淋巴清扫术后、放射治疗史等，造成或诱导引发的疾病，用以判断淋巴疾病的原发性或继发性。

有些[急性病毒感染]dis引起明确的相关性疾病，如[HAV感染]dis与[急性黄疸型肝炎]dis；[轮状病毒感染]dis与[季节性婴幼儿腹泻]dis；小儿患有[肝脏、肾脏、甲状腺疾病]dis。

标注“疾病诱因”时需要注意与“疾病”区分，疾病诱因通常以“XX疾病史”、“XX术后”出现，在标注时应与后缀一起标注。

* + - 1. 疾病诊断分类

在诊断里，通常出现对某个诊断疾病的分类信息，如Ⅱ型、极高危组。这类信息不是疾病名，而是对疾病的一个具体分类，表示疾病的进展程度,这类实体通常出现在诊断里，并且一般紧跟一个具体的疾病。例如[肝硬化]dis[失代偿期]dt、[多发性骨髓瘤]dis[轻链型]dt[Ⅲ期A]dt、[糖尿病]dis[Ⅱ型]dt

* + - 1. 临床症状
         1. 症状

症状是指病人自己向医生陈述（或是别人代述）的不适或痛苦表现。通常是病人主观感觉的不适，如腹痛、头晕等，或是自己发现的病理改变，如血尿便血、活动障碍等。

* + - * 1. 体征

体征是指医生观察到的或者通过检查程序或设备检查到的发生于病人的异常变化以及异常检查结果。通常是指医师利用自己的感官（视触叩听）或者医疗器具（血压计叩诊锤等医疗设备）发现的病人的病理生理变化。

在标注“临床症状”实体时，需要注意：

1. 在“临床症状”实体内部，若包含除“临床症状”之外的其他实体（“疾病”、“身体”、“医学检验项目”等），内部实体应作为方括号嵌套成分（如“[[肺动脉干]bod突出]sym”）。
2. 如出现[[肺动脉]bod高压]sym，[[肺动脉干]bod突出]sym
3. 如果前后文中有“表现为”、“表现有”、“有”、“不良反应有”、“症状有”、“等症状”、“反应”等描述症状出现的词，则标注对应实体为“临床症状”。
4. 不良反应有[[气管]bod痉挛]sym、[[心功能]ite不全]sym、[恶心]sym、[呕吐]sym
5. 对于“临床症状”的修饰成分，通常表示其严重程度、频率等，为保证标注完整性，在标注时应该将修饰和症状一起标注。
6. 遇紧急情况， [气管]bod异物导致[严重呼吸困难]sym
7. 对于“体征”，一般都是通过医疗设备观察到的病理或生理改变等客观表现，因此多出现在表示检查的词后面，比如“见”，“可见”、“及”、“闻及”、“显示”等，这里有两种情况：一种是“医疗程序/医疗设备+（检查词）+体征”，则在标注时仅将对应的体征标注为“临床表现”，不用标注此类检查词；另一种是“身体部位/主体+（检查词）+体征”，为保证标注意义完整，则在标注时应该将“身体部位/主体+（检查词）+体征”作为整体标注成“临床表现”。
8. 行[头颅CT]pro显示[[双侧额部]bod或[额顶部]bod有[蛛网膜下腔]bod增宽]sym
9. 描述“临床症状”时，通常是对病人（一个或多个）身体部位进行描述，在标注时应该与这些身体部位一起标注。若出现“部位/主体+有/无+临床表现”，应该整体标注为“临床症状”。若出现“无+症状/体征”，“无”作为描述临床表现的一种修饰成分，应该整体标注为“临床症状”。
10. 常伴有[[膀胱逼尿肌]bod无抑制性收缩]sym，其中25%患儿有[尿失禁]sym
11. 如果临床症状实体后面紧跟“症状”、“体征”、“反应”，此时应该将实体和此类词语整体标注；若是非临床症状实体后面紧跟此类词语，则不标注；若单独出现“症状”、“体征”、“反应”表示的是一种统称含义，则不标注。
12. [[中毒]dis症状]sym与[[颅高压]dis征象]sym明显、[[神经系统]bod局灶定位体征]sym出现，[神经影像学检查]pro帮助诊断。
13. 临床上的“指征”，一般是指手术指征。在标注工作中，指征标注为“临床症状”。
14. 临床指征为： [血便]sym； 有[里急后重]sym；
    * + 1. 医疗程序
           1. 检查程序

检查程序包括通用检查方法、专项检查、医学影像检查等，检查方法是医生为达到化验目的而采取的某种手段；专项检查是病人通常情况下做的某种检查；医学影像检查是放射科或核医学部门的医疗程序。

* + - * 1. 治疗程序

治疗或预防程序是医生为达到治疗目的而采取的某种手段，如化疗、放疗、手术、透析、紧急救治等。

在标注“医疗程序”实体时，需要注意：

1. 当医疗程序前面有身体部位（对某部位进行检查或治疗），应该整体标注。
2. 辅助治疗和非药物治疗也标注为“医疗程序”。
   * + 1. 医疗设备

“医疗设备”泛指为诊断或治疗所使用的器具、或仪器等，包括检查设备、治疗设备。

* + - * 1. 检查设备

检查设备通常指的是医院中用来检查或检验的仪器，比如血细胞分析仪、生化分析仪等。

* + - * 1. 治疗设备

治疗设备是医生为达到治疗目的而单独或者组合使用于人体的仪器、设备、器具，如注射器、供养面罩、呼吸器等。

1. [血管内支架]equ在[先天性心脏病]dis中的应用：常用[4通道测压导管]equ。

在标注“医疗设备”实体时，需要注意：

1. 医疗设备的属性不可标注为“医疗设备”。

例如[注射器]equ范围大， 包括10~50ml[注射器]equ均可适用。

* + - 1. 药物

药物广义上是指用来预防、治疗及诊断疾病的物质，另外也包括临床诊断试剂。

在标注“药物”实体时，需要注意：

1. 药物的属性不可标注为“药物”。
2. 可用[9α-氟氢可的松]dru 0.05～0.1 mg/d
3. 大部分药物的统称均标注，比如营养素、抗菌药物、急救药物等，此类虽然是统称，但是有对应的治疗范畴。像常用药、药物等单独出现时，此类统称范围太广，不应标注。
4. 配备[急救药物]dru可增加安全性。
   * + 1. 医学检验项目

检查涉及到的体液检查项目、重要生理指标以及其他检查项目，本文规定“医疗检验项目”主要针对人体而言，是能够通过设备或实验检测出的项目，并且是能够被量化，有其对应的测量值或指标值。

在标注“医学检验项目”实体时，需要注意：

1. 测量值不应该标注为“医学检验项目”。
2. [输血]pro指征： [[心率]ite＞ 110次/分]sym（“＞110次/分”不标注）
3. 上下文有关键的提示文字的检查：“检查”、“检”、“查”、“测定”、“检验”等，其对应的检查项目标注为“医学检验项目”。
4. 对[腹泻]sym较重的患儿，应及时检查[血pH]ite、[二氧化碳结合力]ite、[碳酸氢根]ite、

[血钠]ite、[血钾]ite、[血氯]ite及[血渗透压]ite。

1. 在“临床症状”中，有时会涉及到对医学检验项目的描述，比如一些生理指标等，标注时应该将这些生理指标嵌套标注为“医学检验项目”。
2. 患者的临床症状有所好转， [[血清CK]ite下降]sym
3. 常规的检查项目：尿常规、血常规等标本采集类项目，应该标注为“医疗检验项目”。
   * + 1. 身体

“身体”身体泛指细胞、组织、及位于人体特定区域的由细小物质成分组合而成的结构、器官、系统、肢体，另外包括身体产生或解剖身体产生的物质等，包括身体部位和身体物质。

* + - * 1. 身体部位

身体部位包括器官或器官组成、身体系统、身体位置或区域。常见于尿布包裹处及光滑[皮肤]bod相互直接摩擦部位， 如[肛周]bod、[臀部]bod、[外阴]bod、[腹股沟]bod以及[腋窝]bod、[下颏]bod等处[皮肤]bod

* + - * 1. 身体物质

身体物质可由组织、细胞（包括细胞成分、细胞结构）、生物大分子、以及身体或解剖身体产生的物质。

在标注“身体”实体时，需要注意：

1. 表示身体部位具体位置的方位词或者数量词，如“上”、“下”、“左”、“右”、“部”、“侧”、“双”、“多”等应当一同标注。
2. 尤以[双下肺]bod明显， 严重病例可合并[[胸腔]bod积液]sym或[脓胸]sym。
3. 当出现多个身体部位的组合，分开标注会失去原本含义，应作为整体标注为“身体”。
4. [指、趾甲]bod
5. 病变细胞亦标注为“身体”
6. [肿瘤细胞]bod、[白血病细胞]bod、[狼疮细胞]bod
   * + 1. 科室

“科室”主要是指医院或医疗机构所设有的部门以及科室。

1. [外科]dep[血液/肿瘤科]dep和[放疗科]dep为基本组成单位
   * + 1. 病程

疾病和症状的持续性状态修饰设置，在发生于患者本人的确定程度这个方面的修饰

1. 病程：肯定发生或已经发生于患者本人的疾病和症状。
2. 患有[淋巴水肿]sym[1年余]dp
3. 自诉[5年前]dp通过[CT检查]ite检测出[乳腺癌]dis
   * + 1. 频次

疾病和症状的发生频率状态修饰设置，在发生于患者本人的确定程度这个方面的修饰

（1）频次：指症状或者疾病出现的频率。

1. 患者[走路不稳，[偶]occ有头晕]sysm
2. [大便[偶]occ有一过性发白]sym
3. [[时]occ有胸闷气短]sym
   * + 1. 微生物

微生物类包括细菌、病毒、真菌以及一些小型的原生生物、显微藻类等在内的一大类生物群体。微生物类亦可分为致病微生物和非致病性微生物。其他[细菌]mic也可产生[肠毒素]mic， 如[耶尔森菌]mic、[鼠伤寒沙门菌]mic

在标注“微生物类”实体时，需要注意：

1. 出现“病毒+病毒的一部分（基因、DNA）”应该整体标注为“微生物类”。
2. 此法能发现不完整[病毒]mic如潜伏[病毒DNA]mic
   * + 1. 当前修饰

疾病和症状的时间状态修饰设置，在发生于患者本人的确定程度这个方面的修饰

1. 当前的：肯定发生或正在发生于患者本人的疾病和症状。
2. [头晕、呕吐[伴]pre[右下肢]bod无力]sym
3. 自诉[有]pre[冠心病]dis史
   * + 1. 否认修饰

疾病和症状的逻辑状态修饰设置，在是否发生方面的修饰

1. 否认：患者主动否认、或肯定不发生于患者身上。
2. [[各瓣膜区]bod[未闻及]abs病理性杂音]sym
3. [[全腹]bod[无]abs压痛、反跳痛及肌紧张]sym
4. [腹壁静脉曲张]dis：[无]abs]dis
   * 1. 分类混淆处理细则
        1. 疾病(dis)和临床症状(sym)

疾病和临床症状的最大区别是：疾病是通过鉴别诊断的，疾病实质上就是身体受损；而临床症状实质上是身体受损后所表现出来的现象，比如说病人的不适感觉、身体出现的异常变化，但是这些往往是病人或者医生看到的表面现象。而作为医生则需要通过进一步的鉴别诊断来确认病人所患疾病，这也就说明疾病和临床表现存在着本质性的差异。

在遇到“感染”相关的实体时，有以下几点需要注意一下：

1. 若出现明确致病原因（病毒、细菌或身体部位等名称）与“感染”组合成词，则整体标注为“疾病”。如[HAV感染]dis、[球菌感染]dis、[上呼吸道感染]dis。
2. 当单独出现“感染”一词时，若上下文明显表示是对某种疾病的指代，则标注为疾病，否则标注为“临床症状”
3. 若“感染”一词前面的修饰词表明程度或频率时，则整体标注为“临床症状”
   * + 1. 医疗程序(pro)和医疗设备(equ)

医疗程序是指检查过程以及预防或治疗过程，描述的是医生为诊断或者治疗而采取的一系列操作或过程。而医疗设备是指诊断或者治疗而使用的设备，描述的是具体的设备实体（工具、器具、仪器或机器），是医生进行诊断或者治疗的工具。标注者在标注时应该谨慎区分。

① 临床发现有些[头颅]bod较大的婴儿， 行[头颅CT]pro和[MRI检查]pro

* + - 1. 药物(dru)和身体(bod)

标注实体是某种身体物质时，但是有“口服”、“注射”等字眼显式地表明该实体是一种药物（是外来的），此时应该将该类实体标注为“药物”。否则，如果只是表明该实体在人身体中的一种状态（是内在的），应该标注为“身体”。

1. [糖尿病]dis患儿由于[[胰岛素]bod分泌不足或缺如]sym
2. 纠正[高血钾]dis： [葡萄糖]dru 0.5g/kg 加[胰岛素]dru 0.3U/kg [静滴]pro
   * + 1. 医学检验项目(ite)和医疗程序(pro)

医学检验项目是指体液检查项目、生理测量、重要生理指标以及其他检查项目，通常是名词。但如果医学检验项目名称后面紧跟着“检查”、“测定”、“诊断”、“分析”等，表明是医生为诊断或者治疗而采取的一系列操作或过程，故应将医学检验项目名称和这些词语作为整体一起作为标注为“医疗程序”。

1. 监护包括[脉搏]ite、[血压]ite、[尿量]ite、[血乳酸含量测定]pro和[血气分析]pro
   * + 1. 医学检验项目(ite)和身体(bod)

标注实体是某种身体物质时，但是在检查中涉及到对该实体的指标限定，显式地表明该类实体应该是某种检查项目，并且后面通常紧跟着测量值或者指标值，此时应该将该实体标注为“医学检查项目”。否则，如果只是表明该实体在人身体中的一种状态，则标注为“身体”。

1. [输血]pro指征： [[心率]ite＞ 110次/分]sym； [[红细胞]ite＜ 3×1012/L]sym
   * 1. 数据标注执行及注意事项
        1. 人员选拔

数据标注人员需具备该领域的业务和专业知识，并能根据标注规则，检查所标注数据是否满足数据集建设需要。

淋巴水肿数据集数据标注人员资质建议要求在三甲医院从事淋巴外科工作3年以上。

* + - 1. 人员培训

加强数据标注员相关数据标注规则培训，保证每个标注人员理解标注说明规则，满足标注技能要求。

在实际标注开始前，应对数据标注员进行检验，检验标注培训结果，及时发现问题，并将问题及应对措施整理归档。宜先对小样本数据集进行预标注，并对标注结果进行审核，在满足审核标准后，再正式分发标注任务。

* + - 1. 标注执行

数据标注员按照标注规则标注时，遇存疑数据应及时记录。由专业人员（副主任医师以上）对存疑数据进行确认，并提出后续处理办法。宜采用RCT双盲对照组进行标注，并定期进行集中反馈，对标注差异较大的数据，由专业人员出具最终处理意见。

标注规则及细则具有可调整性，对调整后的规则细则，应保证所有参与标注的人员充分理解。发现规则未涵盖的情况或实例时，标注员应及时向上反馈、沟通和处理。

* + 1. 标注说明中术语体系规范化

术语体系的规范化至少应满足：

1. 遵从国家法规和行业规范；
2. 建立统一的标注术语字典，确保数据标注员对术语和定义理解的一致性；
3. 在标注说明规范和细则进行相应的培训后，数据标注员应具备规范使用标注术语完成任务的技能；
   * 1. 标注工具和标注平台

标注工具选择应满足以下条件：

1. 易操作性：标注工具应降低标注人员的操作难度，提供自交方式的自有标签；
2. 规范性:标注工具的数据导出格式，应满足或可转换到格式要求；
3. 高效性：标注工具应能提升标注任务的完成效率。

以上述三项要点的满足为导向，本标准构建的标注任务采用中国科学院自动化研究所自研的标注系统，围绕首都医科大学附属北京世纪坛医院淋巴外科的实际需求，以通用doccano标注平台为蓝本，以Go-zero国产框架为基础，开发了一套文本标注工具平台，可以通过该平台对NLP（自然语言理解）任务的语料库进行标注。平台包含了文本分类、序列标注和序列到序列的标注功能，可以为情绪分析、命名实体识别、文本摘要等内容创建标记数据。平台具有易操作性，只需创建项目，导入淋巴水肿病历数据即可开始标注，同时支持文本、图像、语言多模态的标注任务，可按标注单位需求提供定制化的技术服务。由于涉及淋巴水肿临床数据的标注使用，平台已通过医院机构的审查论证规程。

* + 1. 标注任务的创建、分发、开展和回收
       1. 标注任务的创建

创建标注任务前，将待标注的淋巴水肿数据上传到架设在医院内网的数据标注平台。

标注数据上传成功后，启动标注任务的创建，由医院科室负责人向已遴选的数据标注员进行标注人任务目的及标注规范的宣导，根据上传数据的类型、量级，划分不同任务模块，进行相关任务创建。

创建任务包括：

1. 明确任务基本信息：标注任务目的、任务需求、任务描述等；
2. 任务配置：根据不同的任务需求，添加与标注任务相关的标注标签；
3. 将数据路径上传至平台；
4. 通过标注平台对标注数据进行版本控制。
   * + 1. 标注任务的分发

根据任务参数（包括类型、量级、难度等）及需求，将标注任务下发给参与数据标注的人员。

每项标注任务的分发对象包含数据标注员和数据标注审核员。在数据标注员进行任务操作时，由审核员进行标注完数据的审核与纠错。

由于淋巴水肿标注任务涉及专业领域知识，标注标签的理解或多或少存在非一致性的问题。因此在分发标注任务时，采用分批多人参与的标注形式，将同一项标注子任务分发给两个病区的标注员（中、高年资医生）同步进行标注，每批次子任务回收后进行结果比对，并由主任医师与副主任医师进行标注结果确认，尽可能保证标注结果的一致性。

* + - 1. 标注任务的开展

标注任务中数据标注方法分为两种：全人工标注和半自动标注。

全人工标注方式依赖人力进行标注，对专业性要求越高的数据，全人工标注结果越精准，但当数据量较大时，人力成本耗费较大。

半自动标注方式采用训练好的数据标注模型对目标数据进行检测，并由标注工具自动进行标注，再由标注审核员对自动标注数据进行审核和纠错。半自动标注通常适用于标注数据量大但标注难度较低的任务。

淋巴水肿数据集标注属于专业性强、认知要求高的标注任务，通常需由专业数据标注员进行RCT双盲随机对照组进行标注，再由专业审核员进行审核确认。但考虑到数据量级较大，对人力资源的要求太高，不符合临床实际，因此采用半自动标注方式进行。

首先将3000例待标注数据切分为6批次任务，以300例-300例-400例-500例-500例-1000例的方式进行分割。首批300例采用全人工标注形式，完全由遴选的专业数据标注员（高年资医师）进行手工标注，标注完结果进行分析、归一。在形成300例已标注淋巴水肿数据集后，交由机器进行训练，以6-3-1的比例，60%训练集，30%测试集，10%验证集的方式，对数据进行训练、测试和验证。从第二批数据集开始，逐步推动半自动标注+人工审核的形式进行标注任务，不断打磨机器训练的标注精准性，减少数据标注员的标注工作量。

* + - 1. 标注任务的回收

在协定任务将要完成时，由负责人回收标注作业，对已分配的任务应确保完整回收。

回收环节应注意个别情况和变化的出现，如存在个别标注员未能按时交付，需由候补标注人员继续完成剩余任务，确保标注任务的进度可控。

* + 1. 标注结果质量检查和控制
       1. 标注数据质量检查

质量检查的目的在于确保数据标注结果具有价值，符合需求方的应用目的。根据任务实际情况，质量检查方法可从以下几种中进行选择：

1. 逐条检查：对整个标注项目所包含的所有标注子任务逐一核查并确认，通常适用于项目量级不大、人力充沛、时间节点不紧张、数据标注精度极高的项目。这种方法覆盖的质检范围最全，也适用于任何形式的数据标注场景；
2. 按比例抽查：从全部标注数据种科学的进行样本抽取，对样本中的数据逐条检查，以此评判全部标注数据的质量。样本量选择依据医学统计学原理，应能充分代表全部标注数据的分布。抽查审核时，应由较有经验的审核员完成，确保标注质量；
3. 抽样检查：
   * 1. 简单抽样：以等概率抽取定量待检测样本的方法；
     2. 系统抽样：按批次或每隔一定编号进行检测，每次随机抽取一个或几个样本的方法；
     3. 分层抽样：当不同类型的数据有不同的加工环境（如标注员、不同算法）时，对其质量进行评估抽样方法。
4. 机器验证：通过机器学习，包括使用已训练模型进行检查或使用迁移学习机制对人工标注的数据做质量检查，实现全自动或辅助人工质量检查方式。机器学习方法输出的结果准确率不能完成反应数据集的准确率，但能一定程度上反映数据集的质量。

淋巴水肿数据集的构建采用半自动标注方式进行，已在通用模型上进行定制开发，针对多批次淋巴水肿标注数据集进行多轮训练与实测，具有等同于或约等同于高年资主治医师乃至副主任医师的标注识别能力，因此在最终结果质量检查时，将采用机器验证的方式，对完整3000例淋巴水肿标注数据进行质量检查，反映反馈数据集的标注质量。

* + - 1. 标注数据质量控制

与质量检查面向结果不同，质量控制面向过程，确保标注过程可控，并产生预期结果。在标注过程中，需要对数据质量及标注行为进行规范和监督，及时预警反馈，查明低质量标注数据原因，以此控制标注数据的质量。质量控制的方法可归纳为以下几种：

1. 多人交叉验证：在任务进行过程中，安排一名以上标注员做同一个子任务，通过标注比对和专家辅助选择的方式选出最优的标注结果；
2. 埋题验证：在任务过程中，除了常规标注子任务外，在任务中混杂若干已知结果的测试，以此验证数据标注员的标注水平；
3. 标注员状态验证：通过一定方法对标注员的操作规范性、实时注意力状态、标注准确率等方面进行检查和监测，及时发现操作违规问题，保证数据质量；
4. 机器验证：在任务过程中，同步使用机器学习方法，获取数据准确率信息，一旦发现离群样本或数据准确率明显降低趋势，及时对标注员进行预警。
   * 1. 标注结果交付验收

标注结果交付时，需对最终交付的数据量进行说明，交付验收的内容应包括：

1. 标注结果（必须）
2. 交付和说明文档（可选）
3. 关于标注数据的多元采集信息（可选）
4. 原始数据（可选）

数据标注结构交付后，由需求方对数据标注结果进行验收，验收方法包括抽样验收和逐一验收两种。如验收数据质量未达预期，可要求数据标注员进行修正加工。

参考文献

[1]国家药品监督管理局医疗器械技术审评中心. 深度学习辅助决策医疗器械软件审评要点[Z], 2019.

[2]国家药品监督管理局医疗器械技术审评中心. 人工智能医疗器械注册审查指导原则（征求意见稿）[Z], 2021.

[3]人工智能医疗器械质量要求和评价 第3部分：数据标注通用要求(征求意见稿)

