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

团体标准

淋巴水肿辅助诊断数据处理标准

2023 - XX - XX发布

2023 - XX - XX实施

XXX学会  发布

目次

[前言 II](#_Toc20798)

[引言 III](#_Toc17708)

[1 范围 4](#_Toc3698)

[2 规范性引用文件 4](#_Toc26759)

[3 术语和定义 4](#_Toc17257)

[4 数据处理过程 5](#_Toc15188)

[4.1 数据准备工作 5](#_Toc11796)

[4.2 数据脱敏标准 5](#_Toc10684)

[4.3 数据处理标准 7](#_Toc20664)

[4.4 数据加密标准 7](#_Toc3290)

[4.5 数据处理质量控制 8](#_Toc2143)

[参考文献 9](#_Toc31871)

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处理标准

* 1. 范围

本文件主要规定了淋巴水肿数据集建设中数据处理内容、数据处理过程、数据处理方法、数据处理安全与质控的关键内容。

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

* 1. 术语和定义

下列术语和定义适用于本文件。

数据采集 data acquisition

从数据源中得到原始数据，通过标准化处理并转化为满足数据共享和利用需求的过程。

数据处理 data processing

对数据进行分析和加工的技术过程，也是对数据的采集、存储、检索、加工、变换和传输，使数据转换为有效信息的过程。

数据脱敏 data masking

对数据中某些铭感信息通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护。

数据清洗 data cleaning

发现并纠正数据文件中可识别错误的过程，将数据中重复、多余部分的数据进行筛选并清除，把缺失部分的数据补充完整，将不正确的数据纠正或删除。

数据加密 data encryption

将数据信息经过加密钥匙及加密函数转换，变成无意义的密文，数据接收方可通过解密函数、解密钥匙进行数据还原的过程。

* 1. 数据处理过程
     1. 数据准备工作

数据处理开始前，先要进行数据准备，明确待处理的数据内容、数据格式等，并进行分类汇总，以便正式数据处理所需。

本标准涉及处理的淋巴水肿电子病历，参考国家电子病历及信息化行业标准以及领域诊疗指南，数据内容包含11个模块分类，263个数据元字段。具体的数据元字段如下：

1. 基本信息：包括患者标识、性别、出生日期、年龄、民族、身高、体重等。
2. 就诊记录：包括入院日期、出院日期、入院科室、出院科室、就诊日期、就诊科室等。
3. 现病史：包括患者主诉信息等。
4. 既往史：包括既往相关手术史，放化疗史以及治疗史。
5. 个人史：是否有吸烟史、饮酒史等。
6. 家族史：包括家族疾病名称、家族疾病亲属关系以及患病年龄等。
7. 体格检查：包括检查部位、检查结果、异常描述等。
8. 专科检查：受累部位，皮肤颜色、肿胀是否可凹、纤维化、皮温、关节、活动、合并畸形。包括肢体测量的数据。
9. 诊断：包括淋巴水肿诊断ICD-10编码、共病诊断等。
10. 实验室检查：包括检验日期、检验项目名称、检验定性结果、检验定量结果、检验定量结果单位、检验结论；检验细类包括血常规（红细胞计数、血红蛋白、红细胞压积、平均红细胞体积、平均红细胞HGB含量、平均红细胞HGB浓度、RBC分布宽度SD、RBC分布宽度CV、血小板计数、白细胞计数、中性分叶核粒细胞百分率、淋巴细胞百分率、单核细胞百分率、嗜酸性粒细胞百分率、嗜碱性粒细胞百分率、原始细胞百分率、中性分叶核粒细胞绝对值、淋巴细胞绝对值、单核细胞绝对值、嗜酸细胞绝对值、嗜碱细胞绝对值）；血生化（总胆红素、直接胆红素、间接胆红素、丙氨酸氨基转移酶、门冬氨酸氨基转移酶、AST/ALT、总蛋白、白蛋白、球蛋白、白球比例、葡萄糖、尿素、肌酐、估算肾小球滤过率、血清胱抑素C测定、尿酸、甘油三酯、胆固醇、高密度脂蛋白、低密度脂蛋白、碱性磷酸酶、谷氨酰转肽酶、肌酸激酶、乳酸脱氢酶、羟丁酸脱氢酶、钠、钾、氯、二氧化碳结合力、阴离子间隙、血清β羟基丁酸测定、血钙、血镁、血清无机磷、总胆汁酸）；小便常规（外观检查：颜色、浊度；尿液化学分析：比重、酸碱度、隐血、白细胞、尿蛋白定性、尿葡萄糖、尿胆原定性、尿胆红素定性、酮体定性、亚硝酸盐；尿沉渣定量分析：红细胞、白细胞、上皮细胞、管型、病理管型、细菌电导率；尿沉渣镜检：白细胞、红细胞、脓细胞、粘液丝、一般上皮细胞、小圆上皮细胞、颗粒管型）；大便常规（外观检查：颜色、性状；镜检：白细胞、红细胞、脓细胞、吞噬细胞、淀粉颗粒、脂肪滴、似酵母菌、隐血）；肿瘤标记物（CEA、AFP、CA19-9、CA-50、CA-125、CA-153、CA-724、SCC）等。
11. 影像学检查：包括淋巴核素显像、MRI检查、超声检查（检查日期、检查名称、检查部位、检查结果、检查结论）等。
    * 1. 数据脱敏标准
         1. 数据脱敏范围

首先明确数据集使用的数据必须是获得医院伦理委员会批准或者豁免的临床脱敏数据，患者的隐私保护应当满足法规的要求，保证数据脱敏、患者隐私安全和患者利益，故数据应严格遵循法规和医院数据管理规范要求进行脱敏处理。

依据待处理数据的内容，确定数据脱敏范围包含：病人姓名、年龄、性别、ID号、检查日期、出入院日期等涉及敏感信息的数据。

* + - 1. 脱敏方法
         1. 姓名去标识

姓名是一种常用的标识符，适用的去标识化方法举例如下：

1. 泛化编码。使用概括、抽象的符号来表示，如使用“张先生”来代替“张三”，
2. 或使用“张某某”来代替“张三”。这种方法是用在需要保留“姓”这一基本特征的应用场景。
3. 抑制屏蔽。直接删除姓名或使用统一的“\*”来表示。如所有的姓名都使
4. 用“\*\*\*”代替。
5. 随机替代。使用随机生成的汉字来表示，如使用随机生成的“辰筹猎”、来取代“张三丰”。
6. 假名化。构建常用人名字典表，并从中选择一个来表示，如先构建常用的人名字典表，包括龚小虹、黄益洪、龙家锐、······等，假名化时根据按照顺序或随机选择一个人名代替原名。如使用“龚小虹”取代“张三丰”。这种方法有可能用在需要保持姓名数据可逆变换的场景。
7. 可逆编码。采用密码或其他变换技术，将姓名转变成另外的字符，并保持可逆特性。如使用密码和字符编码技术，使用“SGIHLIKHJ”代替“张三丰”，或使用”Fzf”代替”Bob”。
   * + - 1. 地址去标识

对于地址，常用的去标识话方法举例如下：

1. 泛化编码。使用概括、抽象的符号来表示，如”江西省吉安市安福县”使用“南方某地”或“J省”来代替。
2. 抑制屏蔽。直接删除地址或使用统一的“\*”来表示。如所有的地址都使用“\*\*\*\*\*\*”代替。
3. 部分屏蔽。屏蔽地址中的一部分，以保护地址信息。如使用“江西省xx市xx 县”来代替“江西省吉安市安福县”。
4. 数据合成。采用重新产生的数据替代原地址数据，数据产生方法可以采用确定性方法或随机性方法。如使用“黑龙江省鸡西市特铁县北京路23号”代替 “江西省吉安市安福县安平路1号”。
   * + - 1. 数值型标识符去标识

常用的数值型标识符的去标识化包括：

1. 泛化编码。使用概括、抽象的符号来表示，如“有四个人，他们分别是蓝色、绿色和浅褐色的眼睛”来代替“有1个人是蓝色眼睛，2个人是绿色的眼睛，1个人是浅褐色的眼睛”。
2. 抑制屏蔽。直接删除数值或使用统一的“\*来表示”。如所有的数值都使用“\*\*\*\*\*\*”代替。
3. 顶层和底层编码。大于或小于一个特定值的处理成某个固定值。例如，年龄超过70岁的一律用“大于70岁”描述，以保障满足此条件的人数多于20000人。
4. 部分屏蔽。使用数值的高位部分代替原有数值，如百分制考试成绩全部使用去掉个位数、保留十位数的数值代替。
5. 记录交换。使用数据集中其他记录的相应数值代替本记录的数值。如设定规则，将记录集中的所有的身高数据取出并全部打乱位置后（其他属性数据位置不变）放回原数据集中。这种方法可以保持数据集的统计特性不变。
6. 噪声添加。相对原始数据，产生微小的随机数，将其加到原始数值上并代替原始数值。如对于身高1.72m， 产生随机数值-0.11m，加到原始数值后将其变为1.61m。
7. 数据合成。采用重新产生的数据替代原始数据，数据产生方法可以采用确定性方法或随机性方法。如使用“19”岁年龄代替“45”岁年龄。
   * + - 1. 日期去标识

在数据集中，日期有多种存在形式，包括出生日期、开始日期、纪念日等。常见的对日期的去标识化，方法包括：

1. 泛化编码。使用概括、抽象的日期来表示，如使用1880年代替1880年1月1日。
2. 抑制屏蔽。直接删除日期数据或使用统一的“\*”来表示。如所有的数值都使用“某年某日”代替。
3. 部分屏蔽。对日期中的一部分做屏蔽，如1880年某月1日代替1880年1月1日。
4. 记录交换。使用数据集中其他记录的相应数值代替本记录的数值。如设定规则，将记录集中的所有的日期数据取出并全部打乱位置后（其他属性数据位置不变）放回到原数据集中。这种方法有利于保持数据集的统计特性。
5. 噪声添加。相对原始数据，产生微小的随机数，将其加到原始数值上并代替原始数值。如对于出生日期1880年1月1日，产生随数值32天，加到原始数值后将其变为1880年2月2日。
6. 数据合成。采用重新产生的数据替代原日期数据，如使用“1972年8月12日”代替“1880年1月1日”。
   * 1. 数据处理标准
        1. 数据清洗规则

数据清洗范围包括非空检核、主键重复、非法代码清洗、非法值清洗、数据格式检核、记录数检核和业务约束核验，具体如下：

1. 非空检核：要求字段为非空的情况下， 需要对该字段数据进行检核。
2. 主键重复：多个业务系统中同类数据经过清洗后，在统一保存时，为保证主键唯一性，需进行检核工作。
3. 非法代码、非法值清洗：非法代码问题包括非法代码、代码与数据标准不一致等，非法值问题包括取值错误、格式错误、多余字符、乱码等，需根据具体情况进行校核及修正。
4. 数据格式检核：通过检查表中属性值的格式是否正确来衡量其准确性，如时间格式、币种格式、多余字符、乱码。
5. 记录数检核：指各个系统相关数据之间的数据总数检核或者数据表中每日数据量的波动检核。
6. 业务约束核验：从业务的正确性、一致性、有效性等角度，考虑数据的核验规则，如：建档日期、入院日期、民族信息等的有效性核验。在处理数据过程中，应与医院相关人员共同确定。
   * + 1. 脏数据处理

对于常见的空缺值、离群值和不一致等脏数据，采用人工检测、统计学方法、聚类、分类、基于距离、关联规则等方法来实现数据清洗。

根据缺陷类型分类，可以将脏数据分为缺失值数据、错误数据和错误关联数据这三种典型问题数据，分别进行数据清洗。

* + 1. 数据加密标准
       1. 数据加密规则

数据加密是确保医疗数据安全最经济、最直接、最有效的手段，数据加密在安全认证、加密保护、信任传递等方面的重要作用，可有效满足医疗数据安全流动的需求，实现从离散被动防御向整体主动免疫的本质转换。

1. 物理防护：通过合规、有效加密手段，能系统有效解决网络安全架构鉴别、访问控制、机密控制、抗抵赖等基本安全需求，形成从网络基础资源、信息设施、网络通道、接入终端、设备控制等全体系安全，从底层保障医疗数据的安全可控；
2. 全流程防护：数据加密可以在医疗数据全生命周期发挥作用，从数据产生到传输、存储、处理、分析、共享待再利用，保障数据真实性、完整性、机密性和可追溯性；
3. 隐私计算：加密手段不仅可作用于医疗数据本身，还可有效保证算法计算的可信执行，防止恶意攻击，同时可保证算例可控可管、合法使用，确保计算结果的正确可靠；
4. 信任传递：基于加密的数据标识、数字签名等技术，可确保用户实体身份真实可信、数据来源真实可行，构建医疗数据资源共享及跨机构、跨区域系统应用数据可信共享机制，实现数据资源按需共享、安全交互，保障数据的安全可信及互联互通；
5. 加密监管：基于密码的身份识别、认证鉴权、信任传递、责任认定等手段，在医疗数据采集、存储与处理过程中对数据进行标识，并保存数据处理环节的标识信息，采用递归查询来检索源数据，实现关键数据流向、访问者、访问方式及访问时间的追踪，形成数据流向追踪图，重现数据的历史状态和演变过程，实现追踪溯源和数据使用行为的监管。
   * + 1. 数据加密

本标准处理的医疗数据，都处于医院内网环境下，不存在医院内外网数据交互，因此数据在传输、存储、处理时风险相对较低。

广义而言，本标准所处理的数据，都在医院内网环境；但狭义而言，医院内网环境也分为两个层次，医院信息中心管辖的医疗大数据平台和医生工作站中部署的内网数据处理服务器，两端数据存在交互。为进一步确保数据的安全，在医生工作站上处理的医疗数据，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进行交互，实现数据标注时，所有的患者基本（敏感）信息均被加密隐藏。

* + 1. 数据处理质量控制

数据处理完成后，应按照下面的方式进行人工复核验证：

1. 初期的数据要求提供数据报告，包括接入数据范围（系统数量）、数据量（需要分类，比如患者、医嘱、处方、检验检查数量等）、数据纵深（起始到终止日期）。医院信息部门可进行宏观数据核验。
2. 数据质量抽样检查：比照HIS、EMR等业务系统的单个患者进行数据核验。
3. 病案数据抽样检查：随机抽取一定数量的病例数据进行核对。
4. 隐私数据检查：查看是否按照要求进行去隐私化处理，比如患者姓名、地址、身份证号等，可有意识的搜索隐私数据看可否能检索到。检出问题后及时汇总和分析出问题原因。要么接入对接偏差，要么工程师理解偏差，或是后期后结构化归一的次生原因，作为后期调整技术方案的依据。

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)

