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

|  |
| --- |
|  |

 |

团体标准

淋巴水肿辅助诊断数据采集标准

2023 - XX - XX发布

2023 - XX - XX实施

XXX学会  发布

目次

[前言 II](#_Toc21290)

[引言 III](#_Toc27341)

[1 范围 4](#_Toc1503)

[2 规范性引用文件 4](#_Toc10346)

[3 术语和定义 4](#_Toc22648)

[4 缩略语 4](#_Toc26733)

[5 数据采集内容 5](#_Toc25248)

[6 数据采集过程 6](#_Toc5766)

[7 数据采集质量控制 8](#_Toc28998)

[8 数据采集安全控制 8](#_Toc9179)

[参考文献 10](#_Toc4368)

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采集标准

* 1. 范围

本文件主要规定了淋巴水肿数据集建设中数据采集过程、数据采集内容、数据采集技术、数据采集质量控制及数据采集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

GB/T 22239 信息安全技术 网络安全等级保护基本要求

GB/T 31916.1—2015 信息技术 云数据存储和管理 第1部分：总则

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。

数据融合 data fusion

集成多个数据源以产生比单独数据源更有价值信息的过程。

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。

* 1. 缩略语

以下缩略语适用于本文件：

AI：人工智能 (Artificial Intelligence)

FTP：文件传输协议（File Transfer Protocol）

HTTP：超文本传输协议（Hypertext Transfer Protocol）

HTTPs：超文本传输安全协议（Hypertext Transfer Safe Protocol）

MRI：磁共振成像（Magnetic Resonance Imaging）

HGB：血红蛋白（Hemoglobin）

ICD：国际疾病分类（International Classification of Diseases）

RBC：红细胞（Red Blood Cell）

AST/ALT：天门冬氨酸转氨酶/谷丙转氨酶(Aspartate Transaminase/Alaninetransaminase)

CEA：癌胚抗原（carcinoembryonic antigen）

AFP：甲胎蛋白（Alpha-FetoProtein）

CA19-9：糖类抗原 19-9,(Carbohydrate Antigen 19-9)

CA-50：糖类抗原-50(Carbohydrate Antigen 50)

CA-125：糖类抗原-125 (Carbohydrate antigen 125)

CA-153：糖类抗原-153(Carbohydrate antigen 153)

CA-724：糖类抗原-724(Carbohydrate antigen 724)

SCC：鳞状上皮细胞癌抗原（squamous carcinoma of the cervix）

* 1. 数据采集内容
		1. 数据采集源

本标准规范从病种与临床过程、应用场景和数据模态相互映射的维度来标定数据采集源，数据采集主要涵盖临床过程中筛查与诊断环节，数据源包含淋巴水肿电子病历、MRI、核医学和超声。

**表1 病种和临床流程映射表**

|  |  |
| --- | --- |
| 病种 | 临床过程 |
| 病源 | 预防 | 筛查 | 诊断 | 治疗 | 预后 |
| 淋巴水肿 |  |  | √ | √ |  |  |

**表2 应用场景和数据模态映射表**

|  |  |
| --- | --- |
| 场景 | 临床过程 |
| 电子病历 | CT | MRI | 核医学 | 超声 | 病理 |
| 淋巴水肿 | √ |  | √ | √ | √ |  |

* + 1. 数据采集字段

淋巴水肿电子病历标准数据集的模块参考国家电子病历及信息化行业标准以及领域诊疗指南。数据集共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、RBC分布宽度、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. 采集设备要求
			1. 电子病历采集

采集系统需要兼容医院内各类数据接入源，包括原业务数据库的备份库、数据中心、集成平台和文件接口。数据库方面支持Oracle、MySQL、SQL Server等主流数据库平台。

* + - 1. MRI数据采集

（1）采用1.5T或3.0T MRI扫描仪；

（2）核磁报告阅片人员和审核人员均要求在三甲医院阅片工作3年以上；

（3）MRI检查区域：上肢或下肢。

* + - 1. 核素显像数据采集

（1）采用Siemens Symbia T16双探头SPECT／CT或SPECT仪进行显像；

（2）核素显像方法：于患者双足或双手第1及第4趾/指间皮下缓慢注射99Tcm-DX，3-5 mCi/肢体，0.1-0.2 ml/注射点；分别于注射后10 min，l h、3 h和6 h采集图像；扫描速度为14-16 cm/分钟。

（3）核素显像报告阅片人员要求在三甲医院阅片工作3年以上，审核人员要求在三甲医院阅片工作5年以上；

（4）核素淋巴显像类型/目标：上肢或下肢。

* + - 1. 血管超声数据采集

（1）采用彩色多普勒超声诊断仪进行受累肢体血管数据信息采集。

（2）超声诊断报告要求在三甲医院阅片工作3年以上。

（3）超声诊断类型/目标：上肢或下肢。

* 1. 数据采集过程
		1. 采集方式要求
			1. 采集方式

在院内部署数据采集工具，对接院方信息科提供的原业务数据库的备份库、数据中心、集成平台或者文件接口。

* + - 1. 采集协议

采用HTTP协议或者数据库连接协议来采集标准的电子病历数据。

* + - 1. 采集参数

按照淋巴水肿病例数据标准要求的数据参数指标进行采集。

* + - 1. 采集精度

按照淋巴水肿病例数据标准要求的数据精度进行采集。

* + 1. 采集人员要求
			1. 人员选拔

电子病历数据采集人员需经过医院信息科和相关部门同意的人选。

淋巴核素显像、MRI检查、超声检查类数据采集技师的资质建议要求从事图像采集工作3年以上。

* + - 1. 人员培训

根据数据采集的完备性、准确性、一致性等要求，对参与数据采集的人员进行事前培训。

在文本、影像类数据采集过程中，定期按照淋巴外科专业医师的意见，对数据采集质量进行抽检，并对采集人员进行专项培训。

* + - 1. 人员考核

在正式开始数据采集前制定采集人员考核标准，要求熟悉文本、影像（包括淋巴核素显像、MRI检查、超声检查）数据采集的相关技术要点，能根据实际操作场景的不同，调整参数以获得质量最佳的数据。考核标准通常以操作规范性、操作娴熟度、操作稳定度三方面进行综合评价。

在文本、影像类数据采集过程中，定期按照淋巴外科专业医师的意见，对数据采集质量进行抽检，对经重复培训后仍出现数据采集不过关的人员进行人员变更。

* + 1. 采集过程要求
			1. 采集前准备

数据源选择：根据需要采集数据的数据源类型（如：文件、数据库、传感器等），却确定数据源链接通讯的方式，与医院信息中心、临床相关科室明确采集标准范围及属性，明确需要支持的结构化数据与非结构化数据类型；

采集服务器部署：采集之前，在医院部署相应的物理服务器，在服务器上部署相应的采集工具，与医院信息科确认采集接口规范，医院传输通讯的要求以及临床数据传输加密的形式。

* + - 1. 采集过程

数据采集方式：数据采集分为人工录入和系统采集两种，通过分析数据源类型，选择数据采集方式。基于医院内网环境下的数据源，依照医院数据安全规范，统一按照信息中心的要求进行系统桥接。医院外网环境数据源，可依据实际需要选择采集方式。

数据汇聚：对不同数据源的原始数据进行清洗、转换、分析等处理，确保数据的完备性、准确性。

数据存储：基于国家规定的“三法一条例”，即《数据安全法》、《网络安全法》、《个人信息保护法》及《关键信息基础设施保护条例》的有关规定，严格依照医院院内数据管理规范进行存储，处理后的数据存储应满足安全、有效、高可靠、易管理的要求。

数据质量：数据采集周期内，要充分保障医疗数据的安全性、完备性、准确性、一致性、可追溯性要求，对采集数据进行严格的质控管理。

数据安全：数据采集周期内，应按照“三法一条例”的有关规定，严格确保医疗数据安全及安全控制要求，实现授权访问、可定位溯源、数据脱敏加密、定期不定期安全审计及监测等。

* 1. 数据采集质量控制
		1. 数据质量控制原则

对数据的质量控制应严格贯穿于整个数据采集过程，遵循但不限于以下原则：

1. 完备性：应包含数据规则要求的数据的必要元素；
2. 准确性：应真实反映数据所描述的实体；
3. 一致性：应保证数据与其他特定上下文中使用的数据无矛盾；
4. 时效性：应保证数据发生变化后及时被更新；
5. 可访问性：应保证数据在需要时能被安全访问；
6. 可追溯性：应保证数据能够被追踪和监管。
	* 1. 数据质量控制方式
			1. 数据清洗

数据清洗过程管理应依照以下步骤进行：

（1）数据清洗包括删除重复和缺失数据、纠正错误、检查数据一致性等。

（2）数据清洗前应明确并统一清洗的规则、方法、结果。

（3）数据清洗时应对在样本采集过程中的数据描述信息进行系统检查校验，并对可能出现的数据重复、部分信息缺失和信息一致性方面进行检查校验，并剔除重复、信息缺失和信息不一致的数据。

（4）数据清洗结果应进行验证，对不满足清洗要求的清洗方法进行调整和改进。

（5）数据清洗过程应进行多次迭代，并进度分析、设计与验证。

* + - 1. 数据转换

应对数据的标准代码、格式、类型等进行转换。必要时，可建立“数据转换规则表”。

* + - 1. 数据分析

通过数据聚合、数据归类、数据关联等方法，分析采集的数据，形成上下文完整有效、前后逻辑关联一致的数据。

* + 1. 数据质量评价方法

数据质量评估方法可分为定性评价法和定量评价法：

（1）定性评价法可根据实现确定的评价指标，对数据的安全性、目的、用途及用户自定义目标进行评价；

（2）定量评价法可采用数据质量检测软件检查查数据质量，也可通过辅助工具结合人工识别分析法进行人工检查，一般可采用全数检查和抽样检查：针对国家强制要求或特殊要求以及其他可能导致严重影响产生的数据质量项目应严格进行全书检查；针对质量比较稳定、数据量较大及时间有限的情况可采用抽样检查。

* 1. 数据采集安全控制
		1. 数据安全要求

医疗数据采集安全应严格遵守《数据安全法》、《网络安全法》、《个人信息保护法》及《关键信息基础设施保护条例》的有关规定，始终贯彻在数据采集全生命周期中，应遵循的控制方法包含但不限于：

1. 严格按照《数据安全法》、《网络安全法》、《个人信息保护法》及《关键信息基础设施保护条例》对医疗数据安全的相关要求；
2. 数据在整个采集、转化、传输过程中应依据授权使用，不被非法冒充、窃取、篡改；
3. 应对数据采集环境、设施和技术采取必要的安全管控措施；
4. 应明确数据采集过程中个人信息和敏感数据的知悉范围和安全管控措施，并采取必要的技术手段和限制措施保证数据不被泄露；
5. 应能够对采集的数据进行定位溯源；
6. 应能够对数据采集过程进行安全审计及检测；
7. 应采用通过检测认证的加密手段，对数据采集过程的安全性进行保障。
	* 1. 安全控制方法

数据采集过程中应全方位防御，避免病毒、攻击、非授权访问和内部泄密，同时应保障访问记录的可查、可追溯和可监督。应包括但不限于：

1. 对不同数据进行分类并标识，采用安全技术进行安全维护；
2. 监控数据的使用情况，防止数据在采集过程中被非法访问、破坏、篡改、丢失；
3. 设立访问和使用权限控制机制；
4. 制定应急响应预案及相应处理措施，并定期进行应急演练，及时发现安全问题并处理；
5. 定期对数据采集的安全性进行风险评价，并据此制定相应的风险处理计划，及时排查安全漏洞，加固安全技术；
6. 采用安全技术维护数据安全，包括但不限于对称与非对称密码技术、VPN技术、身份认证与鉴别技术、完整性验证技术、数字签名技术、密码共享技术等；
7. 制定数据采集操作规范，规范数据采集的数据格式、数据质量、流程和方法等；
8. 制定数据采集原则，明确采集数据的目的和用途，确保数据采集的合法性和正当性；
9. 建立安全管理规范，避免人为因素导致数据泄露、损坏等安全事故的发生。

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)

[4]外周淋巴水肿诊疗的中国专家共识. 中华整形外科杂志, 2020. 36(4): p. 355-360.

[5] The diagnosis and treatment of peripheral lymphedema: 2020 Consensus Document of the International Society of Lymphology. Lymphology, 2020. 53(1): p. 3-19.

