ICS 11.040.99
CCS C 40

YY

中华人民共和国医药行业标准

YY/T 1833.3—2022

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

Artificial intelligence medical device—Quality requirements and evaluation—Part 3:General requirement for data annotation

2022-08-17 发布　　2023-09-01 实施

国家药品监督管理局　发布

目　　次

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。

本文件是YY/T 1833《人工智能医疗器械　质量要求和评价》的第3部分。YY/T 1833已经发布了以下部分：

——第1部分：术语；

——第2部分：数据集通用要求；

——第3部分：数据标注通用要求。

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。

本文件由人工智能医疗器械标准化技术归口单位归口。

本文件起草单位：中国食品药品检定研究院、国家药品监督管理局医疗器械技术审评中心、上海长征医院、中国人民解放军总医院、中山大学中山眼科中心、四川大学华西医院、广东省人民医院、中国医学科学院皮肤病医院（中国医学科学院皮肤病研究所）、中国科学院深圳先进技术研究院、浙江大学、广州大学、深圳大学、北京大学、中国科学院自动化研究所、中国生物医学工程学会、河南省医疗器械检验所、腾讯医疗健康（深圳）有限公司、上海联影智能医疗科技有限公司、飞利浦（中国）投资有限公司、上海西门子医疗器械有限公司、通用电气医疗系统贸易发展（上海）有限公司、推想医疗科技股份有限公司、北京安德医智科技有限公司。

本文件主要起草人：李静莉、彭亮、刘士远、何昆仑、郑海荣、田捷、吴健、周晓华、林浩添、步宏、林彤、万遂人、梁会营、刘凯、孟祥峰、倪东、殷丽华、萧毅、李佳戈、李澍、王珊珊、王晨希、王晶、葛鑫、颜子夜、钱天翼、崔征、秦川、詹翊强、王少康、郝烨、范丽、张楠、张培芳、刘畅、王浩。

引　言

近年来，人工智能医疗器械不断发展，成为医疗器械标准化领域的一个新兴方向。我国已初步建立人工智能医疗器械标准体系。在该标准体系中，YY/T 1833《人工智能医疗器械　质量要求和评价》是基础通用标准，为开展细分领域的标准化活动提供指导，拟由八个部分构成。

——第1部分：术语。目的在于为人工智能医疗器械的质量评价活动提供术语。

——第2部分：数据集通用要求。目的在于提出数据集的通用质量要求与评价方法。

——第3部分：数据标注通用要求。目的在于提出数据标注环节的质量要求与评价方法。

——第4部分：可追溯性。目的在于对利益相关方明确人工智能医疗器械可追溯性的含义、要求与评价方法。

——第5部分：算法安全要求。目的在于规范人工智能医疗器械采用的人工智能算法的安全要求与评价方法。

——第6部分：环境要求。目的在于规范人工智能医疗器械的运行环境条件要求与评价方法。

——第7部分：隐私保护要求。目的在于加强人工智能医疗器械保护受试者隐私的能力。

——第8部分：伦理要求。目的在于从技术层面实现人工智能伦理的要求，保护人的权益。

数据标注是基于监督学习的人工智能医疗器械在研发、测试阶段常用的一种技术服务，决定了参考标准的准确性和可靠性，从而对数据集的质量和产品质量产生重要影响。本文件作为YY/T 1833的第3部分，对数据标注说明文档、质量特性、标注与质控流程、标注工具和质量评价方法进行规范。

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　范围

本文件规定了人工智能医疗器械数据标注通用要求和评价方法。

本文件适用于人工智能医疗器械数据标注活动。

2　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。

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

3　术语和定义

YY/T 1833.1、YY/T 1833.2 界定的以及下列术语和定义适用于本文件。

3.1

标注任务　annotation task

有目的地对一批数据进行分析、添加外部知识的活动。

3.2

标注对象　annotation object

标注任务分析的具体信息,如数据的类型、特征、属性等。

3.3

结构化标注　structured annotation

使用固定格式、固定规则记录结果的标注任务。

3.4

非结构化标注　non-structured annotation

使用不固定的格式、规则记录结果的标注任务。

3.5

半结构化标注　semi-structured annotation

使用固定的格式、不固定的规则记录结果的标注任务。

3.6

手工标注　manual annotation

完全由人工执行的标注任务。

3.7

自动标注　automatic annotation

完全由机器执行的标注任务,标注完成后由人工审核。

3.8

半自动标注　semi-automatic annotation

由人工和机器混合完成的标注任务。

3.9

语义标注　semantic annotation

以数据代表的含义和关系为标注对象的标注任务。

3.10

标注人员　annotator

具备完成特定标注任务目标并满足质量要求的能力、执行标注任务、对标注结果有直接贡献的人员。

注：包括初级标注人员、审核人员、仲裁人员等。

3.11

初级标注人员　initial annotator

执行标注任务、给出初步标注结果的人员。

3.12

审核人员　annotation reviewer

对初步标注结果进行审核和质控的人员。

3.13

仲裁人员　arbitrator

当多名标注人员对同一数据的标注结果不一致时，负责给出最终结果的人员。

注：一般情况下，仲裁人员的资质要求＞审核人员≥初级标注人员。

3.14

标注人员表现　annotator performance

标注人员执行标注任务的能力表征。

3.15

标注责任方　annotation responsible organization

组织开展标注任务、对标注质量有直接责任的实体。

4　标注任务说明文档

4.1　标注任务分类

在标注任务开始前，标注责任方应明确标注任务的分类，包括数据模态、执行主体、标注结果格式、标注结果性质、标注结果形式等维度。

标注任务的数据模态分为图像、信号、视频、文本等类型。标注任务依据其执行主体，可分为人工标注、自动标注、半自动标注等类型。依据标注结果的格式，标注任务可分为结构化标注、非结构化标注、半结构化标注等类型。标注结果性质可分为GT值、参考标准、金标准等类型。标注结果的形式分为检出、分类、分割、语义等类型。

注：语义标注常用于描述目标之间的关系或联系，如超声图像上的肌肉、脂肪相对位置。

4.2　标注任务描述

4.2.1　标注规则

标注责任方宜陈述标注任务依据的规则，符合以下要求：

——各标注对象的定义唯一、无歧义；
——标注对象的名称具有依从性文件；
——不同标注对象之间是可区分的；
——标注对象的定性特征宜可验证；
——标注对象的定量特征宜可测量；
——提供标注规则的依从性文件，如根据专家评议、文献分析确定标注规则，宜描述分析过程；
——如标注规则来自试验测量、临床统计等渠道，宜提供客观数据；
——对标注规则可能导致的偏倚风险进行分析。

注：依从性文件包括法规文件、技术标准、医学规范、专家共识、专家评议、文献分析等。

4.2.2 标注人员

标注责任方应描述对标注人员的要求，包括人员资质、选拔依据、培训内容、对标注人员表现的评估指标；如适用，应按照初级标注人员、审核人员、仲裁人员等角色分别展开描述。

标注责任方应描述标注与质控流程中的人员分工、决策机制（审核、仲裁、分歧处理）、人员比对。

4.2.3 标注工具

标注责任方应对标注过程使用的硬件、软件、平台等进行描述，如设备的型号，软件的名称、型号、版本号、功能、参数设置、平台名称、访问地址等；如采用算法提供辅助标注，应描述算法性能指标与验证方法。

4.2.4 标注环境

标注责任方应分析标注环境对标注人员、标注过程、数据质量、标注工具的影响，描述对标注环境的要求，如温湿度、照明条件、噪声干扰等。

4.2.5 数据

标注责任方应对标注过程输入、输出的数据进行描述，包括：
——待标注数据的适用范围、质量要求和选择依据；
——标注对象的定义和示例，如阳性样本、阴性样本、目标区域、非目标区域、主要征象、次要征象、干扰项、疑难情形示例等；
——标注结果、测量结果的存储格式、预览方法、颗粒度、精度等。

标注责任方应描述数据整理方案，如数据清洗、数据查重等。

对来自实验室测量的数据，标注责任方应描述测量方法、测量装置、测量条件及人员等。

对于来自仿真合成的数据，标注责任方应描述计算过程及确认方式。

注：附录 A 给出了标注任务说明文档的示例。

5 数据标注质量特性

5.1 准确性

标注责任方应根据标注结果的形式，声称标注结果的准确性。

如适用，在具体标注场景下，可使用下列指标：
——检出：召回率、精确度；
——分类：灵敏度、特异度、准确率；
——分割：Dice 系数、交并比、Hausdorff 距离；

——测量、计数：绝对误差、相对误差；

——动态曲线评估：Pearson 相关系数、2-范数误差。

5.2 一致性

标注责任方应声称标注过程各个环节输入输出数据、信息的内部一致性，包括人员信息、标注结果、原始数据。

标注责任方应声称标注人员之间的一致性，如：

分类任务：使用 Kappa 系数描述人员之间的一致性。

5.3 精度

对于可定量描述的标注结果，标注责任方应声称标注结果的精度。

5.4 可理解性

标注责任方应说明标注结果能被授权用户理解的程度，并以书面形式展示可验证的证据。

5.5 可访问性

标注责任方应陈述标注结果可被授权用户访问的程度，并以书面形式展示可验证的证据。

5.6 可移植性

标注责任方应陈述标注结果能被安装、替换或从一个系统移动到另一个系统中，并保持已有质量的属性的能力。

5.7 保密性

标注责任方应陈述确保标注结果安全的措施，并以书面形式展示可验证的证据。

5.8 可追溯性

标注责任方应陈述标注任务可被追溯和记录的程度，如：

a) 标注任务、质控流程涉及的人员信息，如标注任务创建者、管理者、标注人员、审核人员、仲裁人员等；

b) 标注任务包含的操作信息，如初始标注、比对、合并、补充、修改、删除、审核、仲裁等；操作信息也包括标注数据的流转动作，如传输、复制等；

c) 标注工具信息，如名称、型号规格、完整版本、制造商、运行环境、软件确认等；

d) 标注任务的时间信息，如每个样本完成标注、审核、仲裁的时间节点。

6 标注与质控流程

6.1 业务架构

标注责任方应根据数据流向和人员分工，描述标注与质控的业务架构。标注责任方应根据业务架构所描述的输入输出节点，保存相应的标注结果、人员操作记录。标注责任方应明确在哪些条件下对标注结论进行审核、仲裁。当初级标注人员的结论一致时，宜对标注结论进行抽样审核。当初级标注人员的结论不一致时，宜提交仲裁。

注：附录 B 给出了具体示例。

6.2 过程组织

6.2.1 任务生成

标注责任方应根据标注任务的定义，收集和整理待标注的数据，准备标注工具和环境，选拔标注人员，明确标注流程、决策机制与工作量，围绕标注规则开展培训，形成记录。

如适用，标注责任方应记录标注任务的创建者、管理者信息。

6.2.2 任务分配

标注责任方应为标注人员分配标注工具和操作场地，设置操作权限，下发待标注的数据。

6.2.3 任务实施

标注人员应根据标注规则执行标注任务。

标注责任方宜对标注进度进行监控，对标注人员的任务进行调度，确保初级标注人员、审核人员、仲裁人员的协调性。

6.2.4 质量控制

在标注过程中，标注责任方应对标注人员的标注质量进行监督，评估标注人员的表现，考虑重复性指标和准确性指标。当标注人员表现出现显著下降时，标注责任方应对标注人员进行休整、培训和再评估。

对重复性指标的评价可采用埋题验证的方式，统计同一个标注人员在每次连续标注过程中对同一个数据的标注结果，计算重复标注一致或误差在允许范围内的样本在重复标注样本中的比例。

注 1：例如每完成 20 张糖网图像的分类标注后，随机抽选其中一张重新标注。

对准确性指标的评价可对比标注人员与仲裁结论，计算仲裁人员认为正确的初级标注样本比例。

注 2：例如每完成 20 张糖网图像的标注，随机抽选一张由仲裁人员仲裁和对比，以统计准确性。

6.2.5 安全管理

标注责任方应执行如下安全管理措施：

a) 在标注前，标注责任方应确保待标注数据已完成数据脱敏；应建立待标注数据的独立备份，确保该备份不被修改、删除；

b) 执行数据标注、计算和存储的设备在停用、退役或退出标注任务前应将其中所有数据彻底删除，并无法恢复；

c) 标注责任方应保证标注过程的网络安全，如采用防火墙、边界防护、入侵防护等安全措施。

7 标注工具

7.1 功能

7.1.1 处理对象

标注工具宜明确定义处理对象的范围，包括数据采集方式、存储格式。

a) 根据数据的采集方式，处理对象可分为：

——影像数据：CT、MR、PET、X 线、乳腺钼靶、超声、内窥镜、病理等；

——信号数据：心电图(ECG)、脑电图(EEG)、肌电图(EMG)、心肺音等；

——文本数据(如适用)：门急诊记录、住院记录、实验室记录、用药记录、手术记录、随访记录。

b) 根据数据存储格式，处理对象可分为：

——图像格式：Dicom、Dicom-RT、png、jpg、tif 等；

——信号格式：xml、HL7 等；

——视频格式：avi、mp4 等；

——文本格式（如适用）：txt、doc、pdf 等；

——其他格式：制造商自定义的数据格式。

7.1.2 数据显示

标注工具应具有数据显示界面，符合以下要求：

a) 标注工具宜支持数据读取范围内的数据显示功能，如：

——Dicom 格式数据：序列翻页、窗宽窗位调整、多窗格显示、平移、整体缩放、反色、局部放大、直线测量、角度测量、图像旋转/翻转、序列播放、恢复原图、影像渲染、图像增益、动态范围等；

——视频格式数据：视频播放暂停、帧率调整、整体缩放、局部放大、对比度调整、饱和度调整等；

——图片格式数据：平移、旋转、整体缩放、局部放大、对比度调整、饱和度调整等；

——文本格式数据：字体大小调整、字体类型调整、局部放大、单栏显示、多栏显示、整页显示、滚动显示等。

b) 数据显示界面应防止数据的未授权获取，如复制、下载、另存、打印等。

7.1.3 数据标注

标注工具宜提供标注任务需要的标注功能，如：

——提供标注工具，支持基本标注任务类型，包括分类标注、分割标注和检出标注等；

——分类标签可根据标注任务的颗粒度进行设置，如病例维度、检查维度、图像维度、病灶维度等；

——支持标签模板配置及版本管理，包括标签模板创建、查看、删除、修改、组合等；

——支持标注质控量化方法配置，包括全检、抽检等；

——如适用，支持自动标注功能、半自动标注功能及其人工审核功能，对自动标注结果进行特殊标记或提示；允许审核人员对自动标注结果进行编辑、修改、保存等操作。对不具有审核权限的人员限制其对自动标注结果进行操作。

注：附录 C 给出了 AI 辅助标注性能评价的一般思路。

7.1.4 结果导入导出

标注工具及平台宜提供标注结果的导入导出功能，如：

——支持标注结果的查看、筛选、统计、下载和导出等操作；

——支持标注结果条件筛选功能，如数据类型、标注结果类型、标注人员、标注进度等；

——支持标注结果统计功能，如标注数量、标注时间范围等；

——支持标注结果下载和导出内容自定义配置，包括项目、病人、数据、标签等；

——支持标注结果下载和导出文件数据格式可选的功能；

——支持标注结果导入功能，应建立数据与标注结果的关联，对格式不符、未匹配或者重复匹配的标注结果进行提示；

——支持结果导入导出权限设置，包括人员权限、数据权限、项目权限等配置。

7.1.5 进度显示

标注工具及平台宜提供具有显示标注任务进度的能力，如：

——支持数据标注状态显示，包括未标注和已标注等；

——支持项目或者数据集标注进度统计与显示功能，包括百分比显示、柱形图显示、饼图显示等；
——支持条件检索的标注进度统计与显示功能，检索条件包括项目、数据集、数据类型、标注人员等。

7.1.6 任务调度

标注工具及平台宜具备标注任务调度功能，如：
——支持标注任务的创建、查看、暂停、恢复、重启、删除、修改及相应权限配置；
——支持标注任务的权限配置，包括人员权限、数据权限、项目权限、操作流程权限等配置；
——支持标注任务的逻辑配置，包括交叉标注方法、仲裁标注条件与方法、审核标注条件与方法等。

7.1.7 审核与仲裁

对于需要审核、仲裁的标注任务，标注工具及平台宜支持自定义配置功能，如：
——支持仲裁条件与方法的自定义配置，包括仲裁触发条件、仲裁人员设置、仲裁数据设置；
——支持审核条件与方法的自定义配置，包括审核触发条件、审核人员设置、审核数据设置。

7.1.8 过程记录

标注工具及平台宜具有过程记录功能，符合 5.8 可追溯性的要求。

7.1.9 安全功能

标注工具应具备以下安全功能：
a) 数据传输安全：数据传输应保证数据以安全的方式传输给指定的对象，如使用加密技术、身份验证技术、数据完整性校验技术等；
b) 数据存储安全：标注工具应具备安全措施保障数据安全，如加密存储；原始数据和标注结果应分开存储为原始数据文件和标注数据文件；
c) 身份鉴别：应对用户进行标识并对标识信息进行管理和维护；应确保用户在信息系统生存周期内的唯一性，应在用户提出动作要求前成功地进行身份鉴别；应定期更换用户登录密码；
d) 访问控制：应具备访问控制策略并实现策略控制下主体与客体间操作的控制。

8 评价方法

8.1 标注任务说明文档

查阅标注责任方提供的文件，应满足第 4 章的要求。

8.2 标注任务质量特性

8.2.1 准确性

在具体标注场景下，可按 YY/T 1833.2—2022 第 6 章对标注结果进行抽样；通过专家论证、专家比对、定量计算等方式对抽样样本或全体样本进行评价，计算标注责任方规定的指标，应满足 5.1 的要求。

8.2.2 一致性

通过抽样检验的方式，检查标注结果与过程文件的一致性，应满足 5.2 的要求。

8.2.3 精度

根据标注责任方的声称，检查标注结果包含的数据定量特征，应满足 5.3 的要求。

8.2.4 可理解性

从语言、符号和(计量)单位等方面对标注结果进行预览、操作,检查用户能否预览和理解标注信息的内容,应满足5.4的要求。

8.2.5 可访问性

编写测试用例,进行实际操作,验证用户能否对标注结果进行访问,应满足5.5的要求。

8.2.6 可移植性

对照标注责任方的陈述,对标注结果的安装、替换、转移进行实际操作,验证不同操作环境下标注信息的性质是否保持不变,应满足5.6的要求。

8.2.7 保密性

检查原始数据及标注结果的授权访问机制、隔离保护机制等,应满足5.7的要求。

8.2.8 可追溯性

对标注过程产生的记录进行检查,应满足5.8的要求。

8.3 标注与质控流程

对标注流程文件进行检查,应满足第6章的要求。

8.4 标注工具

编写测试用例,进行实际操作验证,应满足第7章的要求。

附 录 A
（资料性）
标注任务描述示例

A.1 可穿戴心电

A.1.1 标注任务分类

本标注任务根据数据模态属于生理信号标注，数据模态为单导联可穿戴心电波形信号；执行主体为人工标注。本标注任务属于结构化标注。标注结果的存储格式为 HL7。标注结果给出信号质量的分类，作为参考标准。

A.1.2 标注规则

本标注任务的标注对象是心电信号的质量（每 10 s 一段心电信号的整体质量）。心电信号质量的定义和标注规则由心电图临床专家和工程技术专家组成的专家组依据临床文献和讨论给出，专家职称均为副高级以上，其中医疗系列专家从事临床工作的年限为 10 年以上，从事数据标注相关工作的年限为 1 年以上。标注结果包含两种分类，即“信号质量好”和“信号质量差”。“信号质量好”的定义为心电信号观察窗口中 QRS 波群清晰；几乎不存在基线漂移，即基线漂移幅度不超过信号幅值 1/3，且不影响 QRS 波判断；观察窗口内 T 波清晰，不可辨认的 T 波不超过 2 个；高频噪声干扰极小。病理性改变不影响对信号质量水平的判断，如早搏、心动过速等病理过程，只要波形清晰，判断为“信号质量好”。不符合上述情形的心电信号被判断为“信号质量差”。

标注规则如下：组织 3 名心电图医生，提前培训心电信号质量的定义和软件操作。标注时，各医生使用软件背靠背标注信号质量。记录每名标注人员的标注结果。先采用少数服从多数法，即以不少于 2 名标注人员判定的该段信号质量结果，作为该段信号初始标注结果。标注人员面对面复核信号质量的初始标注结果，如对初始标注结果没有疑义，则初始标注结果即作为最终标注结果；如果初始标注结果存在分歧或疑义，则提请专家组仲裁（3 位专家组成），专家组结合初步标注结果，经讨论给出最终标注结果。在标注过程中，可周期性地重复出现某段信号，观测标注人员的结果是否保持一致；如出现自相矛盾，则使用休整、培训等手段进行干预。

A.1.3 标注人员

心电图医生从事临床工作的年限不低于 1 年，取得医师以上职称，接受过本次标注规则培训。

仲裁专家组的职称不低于中级职称，从事临床工作的年限不低于 8 年，从事标注的年限不低于 1 年。

标注人员的考核指标包括分类的准确率，要求不低于 90%。

A.1.4 标注工具

标注软件为自编软件，软件主要功能包括心电数据的读取、显示、添加标注、标注审核与修改、保存标注结论等。

A.1.5 标注环境

标注任务在某医院医学人工智能实验室进行，使用医用显示器及办公电脑进行，无特殊环境要求。

A.1.6 数据

数据采集日期为2020年1月—7月，采集设备为某品牌的穿戴式单导联心电监护设备，已取得境内医疗器械二类注册证，满足YY 0885—2013等医疗器械标准。数据存储格式为二进制文件，采样率为200 Hz。本标注任务的标注对象定义见A.1.2。数据采集的地点为某医院高压氧科，数据来源为该临床科室的患者，对患者的年龄、职业、籍贯等特征无特殊要求。数据的清洗围绕数据采集的有效性展开，如心电监护设备的佩戴情况、信号强度，人工剔除无效数据；数据的查重主要是对文件名、患者编号(ID)和文件内容的重复性进行检查。具体细节见医院的数据采集、清洗、查重操作规程。标注前需将每一个病人采集的数据按照每10 s一段、非重叠的方式分段，然后标注每一段信号的整体质量。

A.2 眼底彩照

A.2.1 标注任务分类

本标注任务根据数据模态属于图像标注，数据模态为眼底彩照；执行主体为人工标注。本标注任务属于结构化标注。标注结果以字符的形式存储，可使用csv、xml或json格式进行存储。标注结果给出眼底彩照的分类，作为参考标准。

A.2.2 标注规则

本标注任务的标注对象是糖尿病视网膜病变眼底彩照(diabetic retinopathy，DR)的分类(DR病变的疾病分期)。

各分类的含义如下：

——无明显DR：散瞳眼底检查所见无异常；

——非增生性DR的轻度增生型：散瞳眼底检查所见仅有微动脉瘤；

——非增生性DR的中度增生型：散瞳眼底检查所见不仅存在微动脉瘤，还存在轻于重度非增生型DR的表现；

——非增生性DR的重度增生型定义为散瞳眼底检查所见出现以下任何1个表现，但尚无增生型DR的情形，包括：①4个象限中所有象限均有多于20处视网膜内出血，②在2个以上象限有静脉串珠样改变，③在1个以上象限有显著的视网膜内微血管异常；

——增生性DR：出现以下1种或多种体征，包括新生血管形成、玻璃体积血或视网膜前出血。

标注对象的定义和标注规则参考了糖尿病视网膜病变的国际临床分级标准、眼科学诊断规范，由眼科临床专家和眼底病专家组成的专家组给出，专家职称为主任医师，从事临床工作的年限为10年。

标注时需组织2名以上眼底医生，提前培训图像分类和软件操作，使用软件标注DR分类。记录每名标注人员的标注结果。先采用交叉标注，每张彩照需要2名以上标注医师进行独立标注。如果对于各个疾病分类的标注结果一致，则结束标注流程，并将该彩照及其标注结果纳入数据库，可对标注结果进行抽样审核，作为质控；若在交叉标注阶段，标注医师对于单个或多个疾病的标注不一致，则将该彩照送入仲裁标注环节，仲裁标注医生对需要仲裁的标注结果进行复核并出具最终标注结果。在标注过程中，对标注医师与自身的一致性进行周期性的监控，采用埋题验证的方式进行，例如每完成20张糖网图像的分类标注后，随机抽选其中一张重新标注。

A.2.3 标注人员

眼底医生的职称不低于主治医师，从事临床工作的年限不低于3年，从事标注的年限不低于1年，接受过眼底标注分类培训。

仲裁专家组的职称不低于主任医师，从事临床工作的年限不低于10年，从事标注的年限不低于

3 年。

人员的考核指标包括分类的准确率，要求不低于 90%。

A.2.4 标注工具

标注时使用的软件不限制造商，软件主要功能包括眼底图像数据的读取、显示、添加标注、标注审核与修改、保存标注结论。软件界面展示详见具体标注软件的说明书。

A.2.5 标注环境

标注任务在某医院人工智能研发部进行，使用医用显示器及办公电脑进行，无特殊环境要求。

A.2.6 数据

数据采取日期为 2020 年 1 月—12 月，采集设备为某品牌的视场角为 45°的眼底彩色相机（具有医疗器械注册证）。眼底彩照的数据格式为 jpg、tiff、dcm、png。本标注任务的标注对象定义见 A.2.2。数据采集的地点为某医院人工智能研发部。数据来源为该医院就诊、声称视力下降的糖尿病患者，平均年龄大于 40 岁，性别不限。数据的清洗围绕图像质量展开，如眼底彩照的分辨率、亮度、视野范围、图像中心位置等应满足阅片的需要，人工剔除质量不达标的图像；数据的查重主要是对文件名、拍照部位、数据来源和数据本身的重复性进行检查。医院具有数据采集、清洗、查重的详细方案和操作规程。标注前需将每一个病人采集的眼底彩照进行整理，只选取一张用于标注。具体技术细节见该医院的数据采集、清洗与查重操作规范。

A.3 宫颈细胞病理图像

A.3.1 标注任务分类

本标注任务按照数据模态属于图像标注，数据模态为宫颈液基细胞涂片的数字病理图像。本标注任务的执行主体为人工标注。本标注任务属于结构化标注，标注结果以 jpeg、tiff、jpeg2000 的格式进行存储。标注结果给出病理图像的分类，作为参考标准。

A.3.2 标注规则

本标注任务的标注对象是宫颈液基细胞涂片的全切片图像（whole slide image，WSI）的检测与分类。依据目前国际广泛使用的 2014 版子宫颈细胞学报告系统（the bethesda system，TBS）对细胞进行检出和分类，具体描述为：①未见上皮内病变或恶性细胞（NILM），②意义不明的非典型鳞状细胞（ASC-US），③非典型鳞状细胞，不除外高度鳞状上皮内病变（ASC-H），④低度鳞状上皮内病变（LSIL），⑤高度鳞状上皮内病变（HSIL），⑥鳞状细胞癌（SCC），⑦非典型腺细胞-非特异（AGC-NOS），⑧非典型腺细胞-倾向于肿瘤（AGC-FN），⑨原位腺癌（AIS），⑩腺癌（ADC）。细胞分类结果为①则涂片的分类结果为阴性涂片，细胞分类结果为②～⑩则涂片的分类结果为阳性涂片，同时标注阴性涂片中的正常细胞和阳性涂片中的异常细胞。标注对象中各标注分类的定义和标注规则由临床细胞病理专家和工程技术专家组成的专家组根据文献调研和讨论给出，临床细胞病理专家职称为主任医师，从事临床工作的年限为 10 年以上，从事数据标注的年限为 2 年以上。

具体标注时，每张宫颈液基细胞涂片的 WSI 由 1 名病理专业研究生（经过宫颈液基细胞学 TBS 图像培训 2 个月，并通过识图考核）（总共 12 名）执行初级标注。根据制定好的标注标准，初级标注人员经培训后，使用软件标注（圈出）正常细胞（>1 000 个）和异常细胞（<100 个则全部标注；>100 个则至少标注 100 个）。每 3 名初级标注人员的标注结果交由 1 名具有初级职称的细胞病理学医师进行复核，复核后的结果作为初级标注结果；之后交由 2 名具有中级及以上职称的细胞病理学医师进行最终判定，判

定结果即作为最终标注结果。

每完成100张WSI的标注后，交由1名具有高级职称的细胞病理学专家，按照20%的抽查率进行抽查复核，要求准确率95 %以上(标注100个细胞，标注错误细胞不超过5个)，则该批切片记为标注合格，否则需重新对标注进行校准并再次抽查直至合格。如遇到疑难病例，由具有高级职称的细胞病理学医师进行诊断复核；如对病例诊断或标注结果有疑义，则提请专家组(3位具有高级职称的细胞病理学专家组成)仲裁，专家组结合初步标注结果，经讨论给出最终标注结果。

A.3.3 标注人员及分工

本标注任务的人员和分工见表A.1。其中，人员的主要考核指标为：对细胞分类的准确率不低于95%。

表A.1 标注人员职级及分工明细

职责	人数	职称	具体分工
初级标注人员	12	病理专业硕士或博士研究生，经过宫颈液基细胞学TBS图像培训2个月，并通过识图考核	将WSI中的正常细胞和异常细胞按要求进行标注
审核者	4	初级职称、拥有1年以上诊断经验的细胞病理医师	对标注者的结果进行详细审核，及时地将结果反馈给标注者，并将确认后的结果作为该WSI的初始标注结果
判定者	2	中级职称及以上、拥有5年以上诊断经验的细胞病理学专家	对审核者的结果进行判定，并将确认后的结果作为该WSI的最终标注结果
质控专家	1	高级职称，全国知名细胞病理学专家	针对疑难的病例进行诊断复核，对判定者审核后的切片按照20%的比例进行抽查，对有疑义病例诊断或标注结果进行仲裁

A.3.4 标注工具

标注软件来自某商用现货软件COTS/开源软件/自制软件，发布版本号为1.8，软件主要功能包括WSI的读取、显示、添加标注、标注审核与修改、保存标注结果等，详见标注软件说明书。

A.3.5 标注环境

标注任务在某医院临床病理研究所进行，使用医用显示器及办公电脑进行，无特殊环境要求。

A.3.6 数据

数据采集时间段为2019年1月—2020年12月；采集地点为某医院病理科；数据采集设备为某品牌的全数字切片扫描仪(已获得医疗器械注册证)，采用0.25 μm/像素的扫描分辨率获得宫颈液基细胞涂片数字图像，存储格式为厂家自有格式，可转化为jpg、dcm等其他图像格式。本标注任务的标注对象见A.3.2。数据来源为该医院宫颈癌筛查的女性患者群体。标注前，按照参考文献[13]对每一张宫颈液基细胞涂片WSI的整体质量进行数据清洗，筛选整体质量达标的图像样本用于标注。数据的查重主要是对涂片编号、标识、成像视野、数据来源和图像本身的重复性进行检查。具体参数见该医院的数据采集、清洗、查重的操作规程。

A.4 皮肤彩照

A.4.1 标注任务分类

本标注任务按照数据模态属于图像标注，数据模态为皮肤照片。本标注任务按照执行主体属于手工标注。本标注任务属于结构化标注。标注结果以JSON格式进行存储。标注结果给出图像的分类，作为参考标准。

A.4.2 标注规则

本标注任务的标注对象是寻常型及脓疱型银屑病临床图像中的皮损检测与分类、图像整体分类。皮损的分类标签包括：丘疹、斑丘疹(有鳞屑覆盖或无)；红色斑块、浸润性红斑(有鳞屑覆盖或无)；脓疱(包括脓疱部位的红斑)。图像分类标签包括：寻常型银屑病——点滴状、寻常型银屑病——斑块状、脓疱型银屑病(可标注1类或多类)。

标注对象的定义和标注规则由皮肤科临床专家组成的专家组根据文献调研及讨论给出，专家职称为主任医师，从事临床工作的年限均大于20年，从事数据标注的年限不低为2年。

标注时，组织3名皮肤科医生，提前培训标注对象的含义和标注软件界面。经培训后，使用软件背靠背标注。根据银屑病临床图像特征，采用多人盲标＋分阶段审核方法进行。即检出环节：3名标注医师背靠背独立标注，然后用计算机自动判断检出的一致性，以所有人标注结果的并集作为结果；皮损分割环节：3名标注医师背靠背独立标注，然后用计算机自动判断检出的一致性，以所有人标注结果的并集作为结果；分类环节：3名标注医师背靠背进行分类，分类结果同样由计算机自动判断一致性和进行合并，同时保留不同意见；审核环节：由其他标注组长和仲裁专家各自独立对检出和分类结果进行审核与修改，纠正漏诊、误诊和误判。

在标注过程中，可周期性地重复出现某张照片，观测标注人员的结果是否保持一致；如出现自相矛盾，则使用休整、培训等手段进行干预。

A.4.3 标注人员

皮肤科医生的职称不低于主治医师，从事临床工作的年限不低于3年，从事标注的年限不低于1年，接受过标注培训。仲裁专家组的职称不低于副主任医师，从事临床工作的年限不低于6年，从事标注的年限不低于1年。

人员的考核指标包括分类的准确率，要求不低于90%。

A.4.4 标注工具

标注软件来自某商用现货软件，发布版本号为3.0，软件主要功能包括皮肤病影像数据统一校验、导入、格式转换和数据关联拼接、显示、标注、质控、数据存储与管理、数据安全与保密、数据溯源、后台管理等。软件界面详见标注软件说明书。

A.4.5 标注环境

标注任务在某皮肤病医院进行，使用普通显示器及办公电脑进行，无特殊环境要求。

A.4.6 数据

数据采取日期为2020年1月—12月，采集设备为某型号的单反相机，皮肤彩照的数据格式为jpg。

数据采集的地点为某皮肤病医院激光科。本标注任务的标注对象定义见A.4.2。数据来源为疑似患有银屑病的患者。标注前需开展数据清洗，对皮肤彩照的分辨率、亮度、拍摄位置、视野进行检查，剔

除图像模糊或缺乏有效信息的照片。同时，按照数据来源、文件名、文件内容对照片进行查重。具体技术参数见该科室的数据采集与质控的操作规范。

A.5 电子病历文本

A.5.1 标注任务分类

本标注任务按照数据模态属于文本标注，数据模态为电子病历文本。本标注任务按照执行主体属于人工标注。本标注任务属于半结构化标注。标注结果以 BIO 格式存储，标注结果为医学实体，作为参考标准。

A.5.2 标注规则

本标注任务的标注对象是电子病历文本的医学实体，包含约 12 类实体。

第一类实体是疾病，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，其对应的医学一体化语言系统(unified medical language system，UMLS)语义类型有疾病(disease)或者综合征(syndrome)、中毒(injury，poisoning)等。

第二类实体是临床表现，临床表现是疾病的表现，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(sign)或体征(symptom)、异常检查结果(abnormal test results)等。

第三类实体是医疗程序，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)、治疗过程(therapeutic procedure)或预防过程(preventive procedure)等。

第四类实体是医疗设备，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。

第五类实体是身体，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 component)等。

第六类实体是过敏，指外来物质进入体内或者内生物质引起机体免疫系统发生异常反应。常见的变应原有食物、吸入物、微生物以及昆虫毒素、药物、异种血清和物理因素等。

第七类实体是药物，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。

第八类实体是医学检验项目，指检查涉及的体液检查项目、重要生理指标以及其他检查项目，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。

第九类实体是科室，主要指医院或医疗机构所设有的科室，其对应的 UMLS 语义类型有医疗保健相关组织(healthcare related organization)等。

第十类实体是微生物，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，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(bacterial)、真菌(fungus)、病毒(virus)等。

第十一类实体是手术，指医生用医疗器械对病人身体进行的切除、缝合等治疗。

第十二类实体是行为，是指或者有意识的活动，如吸烟、饮酒等。

标注对象的定义和标注规则由医学领域专家和自然语言处理技术专家共同组成的专家组根据文献调研和讨论给出，专家职称副高级以上，从事医疗工作或医学文本数据处理经验的年限为 5 年及以上，

有从事医学文本数据标注的工作经历。

标注按照三个阶段进行。

第一阶段提出标注规范的初稿。深入分析中文医学文本的特点，制定中文医学文本的句子边界检测、字符串切分、分词、词性标注、浅层句法分析等简单任务语料，以及拼写/语法错误识别和纠正、命名实体识别、词义消歧、实体修饰信息识别、关系分类、时序信息抽取等复杂任务语料的标注规范初稿。要求在标注规范初稿中，列出样例的正反例和经过充分讨论后的标注歧义项，同时开发标注工具引入标注提示。

第二阶段采用迭代式的标注方法来训练标注人员和更新标注规范。每一轮迭代都从未标注数据集中随机选出一定数量的医学文本数据作为训练样本。标注不一致的情况均由所有标注人员一同讨论来达到标注的统一，并将这些讨论结果更新到现有标注规范中。在每一轮标注培训中，通过计算两组标注人员的标注一致性来评估培训质量。在标注一致性连续三次处于较高水平时，表明标注规范已经趋于稳定且标注人员对标注规范的认识趋于一致，可以开始将最终标注结果用于语料库的构建。

第三阶段正式构建标注结果语料库。语料库构建过程中，将采取多种措施来保证标注质量，例如：

a) 两组标注人员被分配的数据中加入了一定数量的重复数据，该数据会被两组标注人员标注并可用来计算该阶段的一致性评估结果；

b) 标注工具中有不确定标注的选项，标注人员可以对自己不确定的标注进行标记，这些不确定的标注可以在标注结束后统一讨论后决定；

c) 标注人员按阶段性提交已标注的数据，审核人员将对这些数据进行随机抽样检查，并将与现有规范冲突的情况取出来进行讨论。

A.5.3 标注人员

标注人员的职称不低于医师，从事临床工作的年限不低于1年，从事标注的年限不低于1年，接受过医学文本数据自然语言处理语料规范化生成培训。

审核人员的职称不低于主治医师，从事临床工作的年限不低于3年，有从事1年以上医学文本数据标注工作经验。

标注结果质量的考核指标包括标注任务的准确性、完整性，要求不低于培训要求的98%。

A.5.4 标注工具

标注软件来自开源工具，名称为BRAT(v1.3及兼容版本)，基于Web网页使用，其生成的标注结果可以将非结构化的原始文本结构化，实现对文本的结构化标注并供计算机处理。BRAT既支持用户对文本进行手工工标注，也可以利用其配置的工具对文本进行自动标注，或者对其他标注工具的标注结果进行可视化展示。通过对配置文件进行修改可定义标注的实体名称以及实体间的关系类型。详见标注工具说明书。

A.5.5 标注环境

标注任务在某医院进行，使用普通办公电脑进行，标注辅助工具BRAT应在Linux系统或Windows内的虚拟机Linux系统中使用，其余无特殊环境要求。

A.5.6 数据

数据采取日期为2016年2月—2019年6月，收集妇科门诊的半结构化电子病历文本，将其转换为csv文件格式后包含16个字段，各大实体从所属字段中标注出来。标注对象的定义见A.5.2。数据来源为该医院妇科就诊的患者，对年龄等个人特征无特殊要求。标注前，对文本格式、语义的合理性与完整性进行数据清洗，剔除自相矛盾、内容缺失的电子病历文本；基于电子病历的来源、内容进行查重。其

他细节参见该医院大数据中心的数据采集与质控规范。

A.6 乳腺超声

A.6.1 标注任务分类

本标注任务依据数据模态属于视频标注，数据模态为乳腺超声图像序列。本标注任务依据执行主体属于半自动标注。本标注任务为结构化标注。标注结果为乳腺结节出现的图像及其在图像中的位置，使用 csv 格式保存，作为参考标准。

A.6.2 标注规则

本标注任务的标注对象是超声动态图像序列中乳腺结节的定位。标注对象的定义与标注规则的设计主要依据参考文献[14]，具体考虑的结节类型包括增生结节、乳腺囊肿、纤维腺瘤、乳腺癌。

标注时，组织 3 名以上超声医生，培训对标注对象的认识和标注工具的使用。经培训后，使用软件对超声动态图像序列包含的乳腺结节位置进行标注。记录每名标注人员的标注结果。标注过程共两轮，第一轮采用盲标注，即每位医生独立地对所有数据进行标注。第二轮采用可见标注，即第一轮的标注结果对每位医生可见，每位医生在参考其他医生上一轮标注结果的基础上对自己之前的标注结果进行修正。完成两轮标注后，将第二轮标注结果中同时被两名以上医生标注出的乳腺结节作为最终的标注结果。

第一轮的标注过程使用了目标跟踪算法，辅助医生标注。对于图像序列中的某一个待标注结节，标注人员首先记录下该结节出现的起始帧与结束帧，然后在起始帧与结束帧之间选择任意一张或多张图像作为标注帧，并在该图像中标出结节所在的位置，然后从标注帧开始分别应用目标跟踪算法逐帧跟踪计算该结节在标注帧与起始/结束帧之间出现的位置并进行标注记录。当自动填充过程结束后，标注人员需要对自动标注的结节位置进行确认和必要的修改来完成整个标注过程。通过这种半自动标注模式，标注人员只需要进行少量的标注操作就可以完成结节在大量图像上的位置标注工作。

A.6.3 标注人员

标注医生的职称不低于副主任医师，过去 3 年内每年累计不少于 500 例乳腺超声检查，接受过乳腺超声标注培训。

A.6.4 标注工具

标注软件为自制软件，主要功能包括乳腺超声动态图像序列数据的读取、显示、半自动辅助标注、标注审核与修改、保存标注结论。标注软件界面详见标注软件说明书。其中，半自动辅助标注算法在厂家自有的测试集上开展过性能确认，以算法输出的结节中心点到人工标注的中心点之间的距离作为主要指标。

A.6.5 标注环境

标注任务使用办公电脑进行，无特殊环境要求。

A.6.6 数据

数据采取日期为 2020 年 7 月—10 月，采集设备为某超声系统及超声探头(已获得医疗器械注册证)。数据采集的地点为某医院超声科。数据来源为接受乳腺癌筛查的人群。成像时，同一个患者会将两侧乳腺分成一共八个区域进行扫描，每个区域存成一个图像序列，每个序列大概在 1 000 帧左右。图像格式为 DICOM。标注对象的定义见 A.6.2。标注前，需要根据图像的分辨率、对比度、视野进行数据

清洗，剔除难以标注的图像序列；根据患者、部位、采集时间和文件内容进行数据查重，避免重复提交标注。具体细节见超声科的数据采集与质控操作规范。

A.7 电子听诊器心肺音

A.7.1 标注任务分类

本标注任务依据数据模态属于生理信号标注，数据模态为听诊心肺音。本标注任务依据执行主体属于人工标注。本标注任务属于结构化标注。标注结果以 wav、pcm 格式存储，给出心肺音的分类，作为参考标准。

A.7.2 标注规则

本标注任务的标注对象是电子听诊器心肺音的质量（每 10 s 一段心肺音信号的整体质量），包含两种分类。其中，“信号质量好”的定义为心肺音信号观察窗口内基本心音成分和基本肺音成分清晰可辨认；几乎不存在基线漂移，或基线漂移不影响对心音与肺音的判断；无背景噪音，或背景噪音功率低、背景噪音功率不超过观察窗口内信号总功率的 5%且背景音成分不与基本或病理心肺音成分在时间上产生交迭；观察窗口内按压削波失真和摩擦音干扰的持续时长不超过总时长的 5%；观察窗口内每个心音周期或肺音周期清晰，不可辨认的心音周期或肺音周期占总数的比率不超过 1/3。病理性改变不影响对信号质量水平的判断，如瓣口狭窄、瓣口返流、肺炎、哮喘等病理过程所导致的病理性心音成分和病理性肺音成分，只要波形清晰，判断为“信号质量好”。不符合上述情形的心肺音信号被判断为“信号质量差”。

标注对象的定义和标注规则由听诊临床专家和工程技术专家组成的专家组结合文献调研与共同讨论给出，专家职称均为副高级以上，其中医疗系列专家从事临床工作的年限为 10 年以上，从事数据标注相关工作的年限为 1 年以上。

组织 3 名听诊医生，培训标注对象的定义和软件操作，使用软件背靠背标注信号质量。记录每名标注人员的标注结果。先采用少数服从多数法，即以不少于 2 名标注人员判定的该段信号质量结果，作为该段信号初始标注结果。标注人员面对面复核信号初始标注结果，如对初始标注结果没有疑义，则初始标注结果即作为最终标注结果；对与自己标注结果不一致的初始标注结果，如有疑义，提请专家组仲裁（3 位专家组成），专家组结合初步标注结果，经讨论给出最终标注结果。在标注过程中，可周期性地重复出现某段心肺音，观测标注人员的结果是否保持一致；如出现分歧和矛盾，则进行干预。

A.7.3 标注人员

听诊医生从事临床工作的年限不低于 1 年，接受过本次标注规则培训。

仲裁专家组的职称不低于中级职称，从事临床工作的年限不低于 8 年，从事标注的年限不低于 1 年。

人员的考核指标包括分类的准确率，要求不低于 95%。

A.7.4 标注工具

标注软件为自编软件，软件主要功能包括听诊心肺音数据的读取、显示、添加标注、标注审核与修改、保存标注结论等。

A.7.5 标注环境

标注任务在某实验室进行，使用医用显示器及办公电脑进行，无特殊环境要求。

A.7.6 数据

数据采取日期为2021年1月—12月，采集设备为某电子听诊器设备（已获得医疗器械注册证），心肺音的数据格式为*.wav或*.pcm，采样率为8 000 Hz。数据采集的地点为国内多家三甲医院病房，数据来源为住院患者。标注对象的定义见A.7.2。标注前，需开展数据清洗，剔除噪声过大、采集时长不足、难以辨识等情形的数据；为避免重复，对数据进行查重验证，包括数据来源、采集时间、文件内容。具体细节见医院的数据采集与质控方案。标注前需将每一个病人采集的数据按照每10 s一段、非重叠的方式分段，然后标注每一段信号的整体质量。

附 录 B
（资料性）
业务架构示例（胸部 CT 肺结节）

胸部 CT 肺结节的标注业务架构图如图 B.1 所示，左侧为标注任务的分解，包括检出、分类、分割（描边界）、测量等四个主要任务，按照初级标注、审核与仲裁两个模块配置标注人员。标注任务的输入为 DICOM 格式的胸部 CT 影像。

以下分别描述各任务的实施及数据输入输出节点。

a) 检出环节：3 名标注医师背靠背独立标注，然后用计算机自动判断检出的一致性，以所有人标注结果的并集作为结果。本环节完成时，系统记录的标注信息为紧密包裹肺结节的标注框（bounding box，简称 bbox）中心坐标、端点坐标，即图 B.1 中的节点 1。

b) 分类环节：3 名标注医师背靠背进行分类，分类结果同样由计算机自动判断一致性和进行合并，同时保留不同意见。本环节完成时，系统记录的标注信息增加了肺结节的分类标签，即图 B.1 中的节点 2。

检出与分类环节均属于初级标注任务。3 名标注医师属于初级标注人员的角色，组成标注小组。其中，对 2 名普通组员的要求是在三甲医院从事阅片工作 5 年以上，职称为住院医师以上；对标注组长的要求是具有副主任医师职称，工作年限在 10 年以上。标注组长在后续环节中对其他标注小组的结果进行交叉审核，行使审核人员的职能。

c) 审核与仲裁环节：由一名其他标注组的组长和一名仲裁专家依次对检出和分类结果进行审核与修改，纠正漏诊、误诊和误判。如果遇到疑难问题，可组织更多的仲裁专家进行集体讨论与确认。本环节过后，每个病例至少由 5 名医师进行过阅片，其中至少由两名具有高级职称的医生进行过审核。本环节完成时，系统记录的标注信息仍为 bbox 中心坐标、端点坐标、肺结节的分类标签。其中，仲裁专家为主任医师职称或具有 15 年以上工作经验的副主任医师，资质高于标注组长。

d) 边界分割与尺寸测量：在检出与分类完成之后，由于边界分割相对简单，建议普通病例的边界分割由 1 名标注医师执行，由 1 名审核专家进行审核，本环节完成时，系统记录的标注信息对应图 B.1 中的节点 3。遇到复杂征象时，可酌情增加审核人数，以保证标注质量。根据医生标注的边界，由计算机辅助测量得出结节的尺寸，标注医师和仲裁专家可以手动修改。本环节完成时，系统记录的标注信息包括 bbox 中心坐标、端点坐标、肺结节的分类标签、肺结节边界端点坐标、肺结节长短径等，作为标注任务的输出。

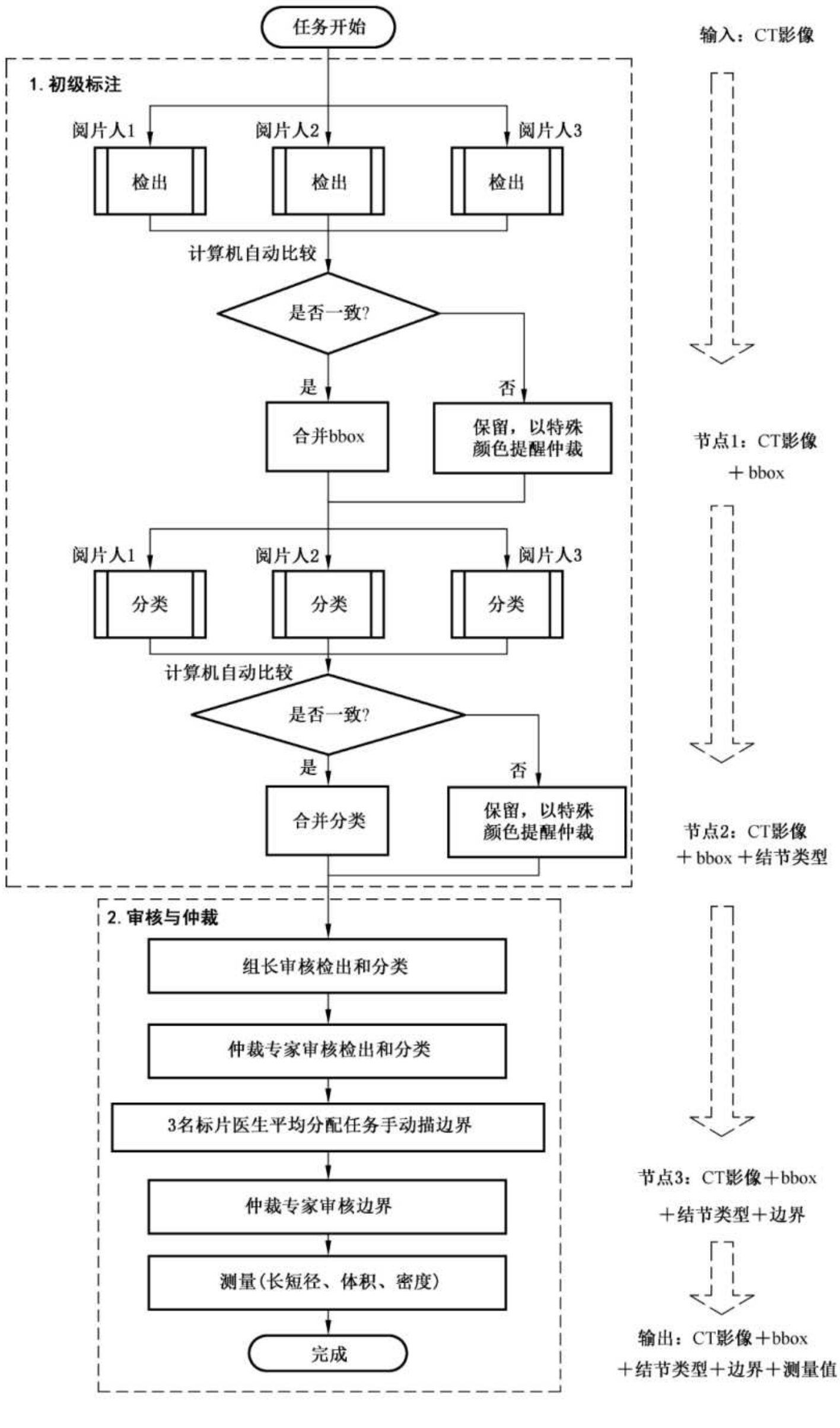

图 B.1　胸部 CT 肺结节标注业务架构示意图

附 录 C
（资料性）
对 AI 辅助标注性能的评价

C.1 总体原则

近年来，基于 AI 算法的辅助标注工具属于研发热点，预期用于提高标注效率。此类工具的形态是多种多样的，如专用的算法模型、已上市的医疗器械软件、公认的第三方开源软件等。在使用前，标注责任方应对辅助标注工具的性能进行确认，但确认方式不唯一，包括算法性能测试、同品种比对、用户反馈等渠道。本附录对性能评价常用的指标进行讨论，仅作为参考。

对辅助标注工具的算法性能评价指标取决于工具的具体功能和应用场景。评价过程一般需要建立测试集。测试人员把测试集输入辅助标注工具，然后对输出的结果进行分析。如辅助标注工具需要对测试集进行预处理，预处理方法应与训练数据的预处理方法一致。

当测试集自带的标注结果具有金标准或参考标准效力时，对辅助标注模型的评价宜采用独立性能评价，直接比较模型的输出与测试数据标注结果。此类测试集也可用于对标注人员进行考核。反之，宜组织专家对标注结果本身进行质控，待建立参考标准后对辅助标注模型进行评价。

标注工具的制造商应描述具体技术指标的定义并给出标称值。

C.2 尺寸辅助测量

对尺寸辅助测量性能的评价可使用绝对误差、相对误差、Pearson 相关系数、均方误差 MSE、平均绝对误差 MAE 等作为指标。

绝对误差：指测量值 X 和真值 Y 之间的差值，其表述为：绝对误差 $=X-Y$；

相对误差：指绝对误差与被测量真值 Y 之比。其表述为：相对误差 $=$ 绝对误差 $/Y\times100\%$；

Pearson 相关系数：指两个变量 X 和 Y 的协方差除以它们标准差的乘积，其计算公式为：

$$\rho(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} \quad \cdots\cdots\cdots\cdots\cdots\cdots (C.1)$$

式中：

$\rho(X,Y)$——Pearson 相关系数；

μ_X——X 的平均值；

μ_Y——Y 的平均值；

σ_X——X 的标准差；

σ_Y——Y 的标准差。

Pearson 相关系数的绝对值越大，相关性越强：相关系数越接近于 1 或 -1，相关度越强，相关系数越接近于 0，相关度越弱。

均方误差 MSE 的公式为：

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-d_i)^2 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (C.2)$$

式中：

MSE——均方误差；

y_i——第 i 个标签；

d_i——第 i 个预测值；

n——样本数量。

平均绝对误差 MAE 的公式为：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-d_i\right| \qquad \cdots\cdots(\mathrm{C.3})$$

式中：

MAE ——平均绝对误差；

y_i ——第 i 个标签；

d_i ——第 i 个预测值；

n ——样本数量。

如果测量对象为明确的实体(如肿瘤直径)，也可使用尺寸单位直接度量误差(如：mm)。

C.3 联想推理

图像领域的联想推理功能，可分两种任务情况做评价。

第一种是基于矩形框的联想推理，可采用类似于视频跟踪的指标，包括召回率、精确度、平均精确度均值、交并比、中心漂移率。

第二种是基于轮廓的联想推理，可采用检出、分割相关的评价指标，包括 Dice 系数、Conformity 系数、交并比、Hausdorff 距离等。同时，针对联想推理，其所需人为提供的初始化帧数、执行联想推理任务的效率或平均耗时也是需要考虑的评价指标。

召回率：表示被正确检测出的目标数量占所有目标数量的比例，其公式为：

$$\mathrm{Rec}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad \cdots\cdots(\mathrm{C.4})$$

式中：

Rec ——召回率；

TP ——正确检测出的目标数量；

FN ——被遗漏的目标数量。

精确度：表示被正确检测出的目标数量占所有被检出对象的比例，其公式为：

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \qquad \cdots\cdots(\mathrm{C.5})$$

式中：

Pre ——精确度；

TP ——正确检测出的目标数量；

FP ——被误认为是目标的对象。

平均精确度：设定正负样本的阈值，可计算出目标检测的精确度和召回率。改变阈值，可画出 Precision-Recall 曲线，该曲线下的面积为平均精确度(average precision，AP)，其公式为：

$$\mathrm{AP}=\sum_{k=1}^{N}p(k)\Delta r(k) \qquad \cdots\cdots(\mathrm{C.6})$$

式中：

N ——曲线节点的数量；

$p(k)$ ——第 k 个节点对应的精确度；

$\Delta r(k)$ ——第 k 个节点对应的召回率步长。

当目标有多种分类时，可计算平均精确度均值，即对所有类别(记为 C 类)的平均精确度求均值，其公式为：

$$\mathrm{mAP}=\frac{\sum_{i=1}^{C}\mathrm{AP}_i}{C} \qquad \cdots\cdots(\mathrm{C.7})$$

式中：

mAP ——平均精确度均值；

AP_i ——第 i 个类别的平均精确度；

C ——类别总数。

矩形框的交并比(Jaccard 系数)：用于评价预测的检测框与真实的检测框的重合程度，其公式为：

$$\text{Jaccard} = \frac{|A \cap B|}{|A \cup B|} \quad \cdots\cdots(C.8)$$

式中：

Jaccard ——Jaccard 系数；

A ——目标区域；

B ——分割区域。

中心漂移率：最后一帧图像的预测目标中心点和真实目标中心点之间的欧式距离。

Dice 系数：是一种集合相似度度量函数，通常用于计算两个分割区域的相似度，其公式为：

$$\text{Dice} = 2 \times \frac{|A \cap B|}{|A| + |B|} \quad \cdots\cdots(C.9)$$

式中：

Dice ——Dice 系数；

A ——目标区域；

B ——分割区域。

Conformity 系数：为错误分割的像素数量占所有被正确分割的目标区域像素之间的比例，其公式为：

$$\text{Conformity} = 1 - \frac{FP}{TP} \quad \cdots\cdots(C.10)$$

式中：

Conformity ——Conformity 系数；

FP ——错误分割的像素数量；

TP ——被正确分割的目标区域像素数量。

分割的交并比：用于评价预测的分割区域与真实的分割区域的重合程度，同公式(C.8)。

Hausdorff 距离：用于描述两个分割区域轮廓线的距离，双向 Hausdorff 距离计算公式见公式(C.11)：

$$d_H(X,Y) = \max\{d_{XY}, d_{YX}\} = \max\left\{\max_{x\in X}\min_{y\in Y} d(x,y), \max_{y\in Y}\min_{x\in X} d(x,y),\right\} \quad \cdots\cdots(C.11)$$

式中：

$d_H(X,Y)$ ——双向 Hausdorff 距离；

X ——预测的分割区域；

Y ——人工标注的分割区域；

$d(x,y)$ ——X、Y 两个区域任意两点之间的距离。

对于离散型推理判断，应使用准确率计算推理误差。其中，准确率的表述为：推理准确率＝推理准确的数量/总推理对象。

对于连续型推理判断，应使用 MSE 计算推理误差，表示推理的结果和理想值的差距，同公式(C.2)。

C.4 区域分割

分割的评价推荐用 Dice 系数、Conformity 系数、交并比、Hausdorff 距离、Pearson 相关系数、一致性 ICC 系数作为指标。以下给出对应公式：

a) Dice 系数同公式(C.9)。

b) Conformity 系数:为错误分割的像素数量占所有真实分割像素之间的比例,同公式(C.10)。

c) 平均交并比(mIOU):表示各区域分割结果的交并比(IOU)均值,其计算公式见公式(C.12):

$$mIOU=\frac{1}{n}\sum_{i=1}^{n}\frac{|P_i \cap T_i|}{|P_i \cup T_i|} \quad \cdots\cdots(C.12)$$

式中:

mIOU ——平均交并比;

P_i ——第 i 个预测的区域;

T_i ——第 i 个分割标注区域;

n ——区域总数。

d) Hausdorff 距离:同公式(C.11)。

e) Pearson 相关系数:同公式(C.1)。

C.5 辅助分类

分类的评价使用灵敏度、特异度、准确率作为指标。

记 TP 为真阳,FP 为假阳,FN 为假阴,TN 为真阴准确率。

a) 灵敏度:

$$Sen=\frac{TP}{TP+FN} \quad \cdots\cdots(C.13)$$

式中:

Sen ——灵敏度;

TP ——真阳性样本数量;

FN ——假阴性样本数量。

b) 特异度:

$$Spe=\frac{TN}{TN+FP} \quad \cdots\cdots(C.14)$$

式中:

Spe ——特异度;

TN ——真阴性样本数量;

FP ——假阳性样本数量。

c) 准确率 Acc:

$$Acc=\frac{TP+TN}{TP+FP+TN+FN} \quad \cdots\cdots(C.15)$$

式中:

Acc ——准确率;

TP ——真阳性样本数量;

FN ——假阴性样本数量;

TN ——真阴性样本数量;

FP ——假阳性样本数量。

C.6 辅助检出

检出的评价使用召回率、精确度、F_1 度量、mAP 等作为指标。

a) 召回率:见公式(C.4)。

b) 精确度：见公式(C.5)。

c) F_1 度量：

$$F_1 = \frac{2 \times \mathrm{Rec} \times \mathrm{Pre}}{\mathrm{Rec} + \mathrm{Pre}} \quad \cdots\cdots (\mathrm{C.16})$$

式中：

F_1 ——F_1 度量；

Rec ——召回率；

Pre ——精确度。

d) mAP：见公式(C.7)。

C.7 关键点定位

关键点定位的评价使用绝对误差、相对误差、PCK 百分比作为指标。

a) N 维数据中欧氏距离绝对误差：

$$D(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad \cdots\cdots (\mathrm{C.17})$$

式中：

$D(x,y)$——欧式距离绝对误差；

n ——关键点的总数；

x_i ——第 i 个关键点的预测值；

y_i ——第 i 个关键点的标注值。

b) N 维数据中闵氏距离绝对误差：

$$D(x,y) = \left[\sum_{i=1}^{n} (x_i - y_i)^2\right]^{\frac{1}{n}} \quad \cdots\cdots (\mathrm{C.18})$$

式中：

$D(x,y)$——闵式距离绝对误差；

n ——关键点的总数；

x_i ——第 i 个关键点的预测值；

y_i ——第 i 个关键点的标注值。

c) PCK(percentage of correct keypoints)百分比：用于表示正确估计出的关键点比例，公式见(C.19)

$$\mathrm{PCK}_{\mathrm{mean}}^{k} = \frac{\sum_p \sum_i \delta\left(\frac{d_{pi}}{d_p^{\mathrm{def}}} \leqslant T_k\right)}{\sum_p \sum_i 1} \quad \cdots\cdots (\mathrm{C.19})$$

式中：

$\mathrm{PCK}_{\mathrm{mean}}^{k}$——阈值 T_k 下的算法 PCK 百分比指标；

p ——被标注对象的编号；

i ——关键点的编号；

δ ——括号内的条件成立时，取值为 1，否则取值为 0；

d_{pi} ——第 p 个被标注对象上的第 i 个关键点的预测值与参考标准之间的欧式距离；

d_p^{def} ——第 p 个被标注对象的尺度因子；

T_k ——第 k 个阈值。

C.8 标注效率

标注工具的制造商宜根据典型标注任务的平均耗时作为自动算法标注效率的评价指标。

参 考 文 献

[1] T/CESA 1040—2019 信息技术 人工智能 面向机器学习的数据标注规程

[2] T/CMDA 002—2020 肝胆疾病标准数据规范:肝癌 CT/MRI 影像标注和质控标准

[3] T/ISC 0005—2020 针对内容安全的人工智能 数据标注指南

[4] 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].北京:国家药品监督管理局医疗器械技术审评中心,2019.

[5] 国家药品监督管理局医疗器械技术审评中心.肺炎 CT 影像辅助分诊与评估软件审评要点(试行)[Z].北京:国家药品监督管理局医疗器械技术审评中心,2020.

[6] 国家药品监督管理局.人工智能医疗器械注册技术审查指导原则(征求意见稿)[Z].北京:国家药品监督管理局,2021.

[7] 中国食品药品检定研究院,中华医学会放射学分会心胸学组.胸部 CT 肺结节数据标注与质量控制专家共识(2018)[J].中华放射学杂志,2019,53(1):9-15.

[8] 中华医学会放射学分会,中国食品药品检定研究院,国家卫生健康委能力建设与继续教育中心,等.胸部 CT 肺结节数据集构建及质量控制专家共识[J].中华放射学杂志,2021,55(2):104-110.

[9] 中华医学会放射学分会医学影像大数据与人工智能工作委员会,中华医学会放射学分会腹部学组,中华医学会放射学分会磁共振学组.结直肠癌 CT 和 MRI 标注专家共识(2020)[J].中华放射学杂志,2021,55(2):111-116.

[10] 中华医学会放射学分会医学影像大数据与人工智能工作委员会,中华医学会放射学分会腹部学组,中华医学会放射学分会磁共振学组.肝脏局灶性病变 CT 和 MRI 标注专家共识(2020)[J].中华放射学杂志,2020,54(12):1145-1152.

[11] 《实体瘤病理数据集建设和数据标注质量控制专家共识》筹备组.实体瘤病理数据集建设和数据标注质量控制专家意见(2019)[J].第二军医大学学报,2019,40(5):465-470.

[12] 中华医学会眼科学分会青光眼学组,中国医学装备协会眼科人工智能学组.中国基于眼底照相的人工智能青光眼辅助筛查系统规范化设计及应用指南(2020 年)[J].中华眼科杂志,2020,56(6):423-432.

[13] 中国病理医师协会数字病理与人工智能病理学组,中华医学会病理学分会数字病理与人工智能工作委员会,中华医学会病理学分会细胞病理学组.宫颈液基细胞学的数字病理图像采集与图像质量控制中国专家共识[J].中华病理学杂志,2021,50(4):319-322.

[14] 中国抗癌协会乳腺癌专业委员会.中国抗癌协会乳腺癌诊治指南与规范(2019 年版)[J].中国癌症杂志,2019,29(8):609-680.

[15] ISO/IEC 2382:2015 Information technology—Vocabulary

[16] Saha A,Harowicz M R,Mazurowski M A.Breast cancer MRI radiomics:An overview of algorithmic features and impact of inter-reader variability in annotating tumors[J].Medical physics,2018,45(7):3076-3085.

[17] Dong D,Tang L,Li Z Y,et al.Development and validation of an individualized nomogram to identify occult peritoneal metastasis in patients with advanced gastric cancer[J].Annals of Oncology,2019,30(3):431-438.

[18] Meng L,Dong D,Chen X,et al.2D and 3D CT radiomic features performance comparison in characterization of gastric Cancer:a multi-center study[J].IEEE journal of biomedical and health informatics,2020,25(3):755-763.