ICS 11.040.99
C 40

YY

中 华 人 民 共 和 国 医 药 行 业 标 准

YY/T XXXX—××××

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

征求意见稿

（本稿完成日期：2021.7.21）

××××－ ××－ ××发布　　××××－ ××－ ××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。

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：

本文件主要起草人：

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

本文件提出了人工智能医疗器械领域的数据标注通用要求和评价方法。

2 规范性引用文件

下列文件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 XXXX.1-XXXX 人工智能医疗器械质量要求和评价 第1部分：术语

YY/T XXXX.2-XXXX 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 XXXX.1《人工智能医疗器械质量要求和评价 第1部分：术语》、YY/T XXXX.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》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标注的活动。

3.2

标注对象 annotation object

标注任务的目标，如数据的类型、特征、属性等。

3.3

图像标注 image annotation

对图像数据进行分析，建立外部知识的过程

注：如病灶区域、中心点位置、边界、分类等。

3.4

视频标注 video annotation

对视频数据进行分析，建立外部知识的过程。

注：如对动态超声、胚胎视频进行标注等。

3.5

生理信号标注 physiological signal annotation

对生物体中包含生命现象、状态、性质、变量和成份等信息的物理量进行分析，建立外部知识的过程。

注：如心电图的心拍标记、分类等。

3.6

结构化标注 structured annotation

一种标注任务，使用固定格式、固定规则记录结果。

3.7

非结构化标注 non-structured annotation

一种标注任务，记录结果的格式、规则不固定。

3.8

半结构化标注 semi-structured annotation

一种标注任务，记录结果的格式固定，但规则不固定。

3.9

手工标注 manual annotation

完全由人工执行的标注任务。

3.10

自动标注 automatic annotation

完全由机器执行的标注任务。

注：标注结果需要人工确认。

3.11

半自动标注 semi-automatic annotation

由人工和机器混合完成的标注任务。

3.12

多标签 multi-label

同一个数据具有多种标注对象。

3.13

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。

3.14

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、对标注结论有直接贡献的人员。

注：包括初级标注人员、审核人员、仲裁人员等。

3.15

初级标注人员 initial annotator

执行标注任务、给出初级标注结论的人员。

3.16

审核人员 annotation reviewer

对初级标注结论进行审核和质控的人员。

3.17

仲裁人员 arbitrator

当多名标注人员对同一原始数据的标注结果不一致时，负责给出最终判定的人员。

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。

3.18

标注人员表现 annotator performance

标注人员执行标注任务的绩效。

3.19

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 标注任务定义

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。

标注对象的数据模态分为图像、信号、视频等类型。标注任务的执行主体分为：人工、自动、半自动标注等类型。标注结果的格式分为：结构化、非结构化、半结构化等类型。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 标注任务描述文档

4.2.1 标注规则

标注责任方应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；

——标注对象的名称符合医学规范；

——不同标注对象之间是可区分的；

——标注对象的定性特征应是可验证的；

——标注对象的定量特征应是可测量的；

——列举标注规则依从的法规文件、技术标准、医学规范、专家共识、专家评议、文献分析等；如根据专家评议、文献分析确定标注规则，应描述分析过程；

——如标注规则来自试验测量、临床统计等渠道，应提供客观数据；

——对标注规则可能导致的偏倚风险进行分析。

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价指标；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、标注软件的名称、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、亮度、机械振动、电磁干扰等。

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；

——标注信息、测量信息的存储格式、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、人员。

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。

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，明确对标注准确性的度量。评价方式包括专家论证、专家比对、定量计算等。

对具体标注场景，对准确性的度量可使用下列指标进行计算：

——检出：召回率、精确度；

——分类：灵敏度、特异性、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

标注责任方应评估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、标注信息、原始数据。

标注责任方应评估标注人员表现的一致性。

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信息的精度。

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信息安全的措施，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、标注平台信息，如名称、型号、软件完整版本、序列号等；

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，保存相应的标注信息、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

标注人员根据标注规则执行标注任务。

标注责任方应对实际的标注进度进行监控，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和效率进行监督，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重新进行培训和考核。

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对同一个数据的标注结果，计算重复标注完全一致的样本在重复标注样本中的比例。

注：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：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 结果验收

标注责任方应根据数据标注质量特性和具体的标注任务，明确标注结果的验收准则。

验收准则应考虑以下维度：

a)准确性：标注结果是否与金标准或仲裁人员（交叉仲裁、第三方仲裁）给出的标注结果一致；

b)一致性：信息、数据、记录是否保持内部一致性；标注人员的表现是否具备充分的外部一致性；

c)精度：标注信息的精度是否满足需要；

d)可理解性：用户能否预览和理解标注信息的内容；

e)可访问性：用户能否对标注信息进行访问；

f)可移植性：在标注责任方声称的不同操作环境下，标注信息的性质是否保持不变；

g)保密性：标注信息的存储、使用是否安全；

h)可追溯性：标注过程是否具有完整的记录，是否可溯源；

对标注结果的验收可采用抽样检验、专家评议、第三方检查等方式，应形成验收报告。

7 标注工具及平台要求

7.1 功能要求

7.1.1 处理对象

标注工具及平台应明确定义处理对象的范围，包括数据采集方式、存储格式。

1）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)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；

2）根据数据存储格式，处理对象可分为：

——图像格式：Dicom格式、Dicom-RT格式、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

标注工具及平台应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；

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

标注工具及平台应提供标注任务需要的标注功能，如：

——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；平台界面应提供标注工具，如紧密包裹矩形框、直线、圆环、手工轮廓、区域填充、单点标记、关键词标记、三维立体标记、时间线标记等；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，支持自动标注功能及其人工审核功能，对自动标注结果进行特殊标记或提示。

7.1.4 结果导入导出

标注工具及平台应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，标注工具及平台应是可配置的，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

标注工具及平台应具有过程记录功能，符合条款5.8可追溯性的要求。

7.2 安全

7.2.1 数据安全

标注过程相关数据的输入输出宜符合下列要求：

a）数据来源：待标注的数据预先进行脱敏，标注前应进行确认。

b）数据传输安全：数据传输宜使用加密技术、身份验证技术和数据完整性校验技术保证数据以安全的方式传输给指定的对象；

c）数据存储安全：标注平台应采取安全措施保障数据安全，如加密存储。原始数据和标注信息宜分开存储为原始数据文件和标注数据文件。标注责任方向标注平台上传原始数据前，应对原始数据文件建立独立备份，确保该备份不被修改、删除。

d）数据销毁：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上所有数据彻底删除，并无法恢复。

7.2.2 网络安全

标注工具及平台应采取必要的措施，确保网络安全，如：

a）身份鉴别：对用户进行标识并对标识信息进行管理和维护；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；定期更换用户登录密码

b）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。

c）安全措施：宜采用防火墙、边界防护、入侵防护等安全措施。

7.3 验证方法

7.3.1 功能验证

对照标注软件的说明文档，编写测试用例；如标注工具、标注平台使用AI算法进行辅助标注，标注工具、标注平台的制造商可参考附录C，提交算法性能研究资料。对标注软件开展操作检查；如适用，对算法性能研究资料开展文档检查，指标定义、测试流程、测试集描述应准确、清晰、无歧义，符合7.1的要求。

7.3.2 安全验证

对照标注平台的网络安全文档，编写测试用例，开展操作检查，应符合7.2的要求。

7.3.3 测试用例的要求

测试用例宜包含测试目标、测试用例唯一标识、测试对象、测试步骤、测试环境、测试边界、预期的响应或产出、测试结果解释、用于判断测试用例是否通过的准则。

附 录 A
（资料性）
标注任务描述示例

本附录对基于不同模态的标注任务描述进行举例，仅作为参考信息。

A.1 可穿戴心电

A.1.1 标注任务分类

标注任务分类	数据模态：可穿戴心电
	执行主体：人工标注
	标注结果格式：HL7
	标注结果性质：参考标准
	标注结果形式：分类

A.1.2 标注对象

本标注任务的标注对象是心电信号的质量（每 10s 一段心电信号的整体质量），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”。

标注对象的定义由心电图临床专家和工程技术专家组成的专家组给出，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。

主要参考文献包括：

[1] Joachim Behar, Julien Oster, Qiao Li, et., al.. ECG Signal Quality During Arrhythmia and Its Application to False Alarm Reduction. IEEE Transactions on biomedical engineering, 2013, 60(6).

[2] Yuanbo Shi, Ning Han, PeiyaoLi, et., al.. Robust Assessment of ECG Signal Quality for Wearable Devices. 2019 IEEE International Conference on Healthcare Informatics, 2019.

[3] Chengyu Liu, Xiangyu Zhang, Lina Zhao, et., al.. Signal Quality Assessment and Lightweight QRS Detection for Wearable ECG SmartVest System, IEEE Internet of Things Journal-Special Issue on Wearable Sensor Based Big Data Analysis for Smart Health, 2018.

[4] GD Clifford, J Behar, Q Li, et., al.. Signal Quality Indices and Data Fusion for Determining Clinical Acceptability of Electrocardiograms. Physiological Measurement, 2012, 33(9).

A.1.3 标注规则

组织 3 名心电图医生，根据制定好的标注标准，经培训后，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。上述标注流程可用图 A.1 表示。

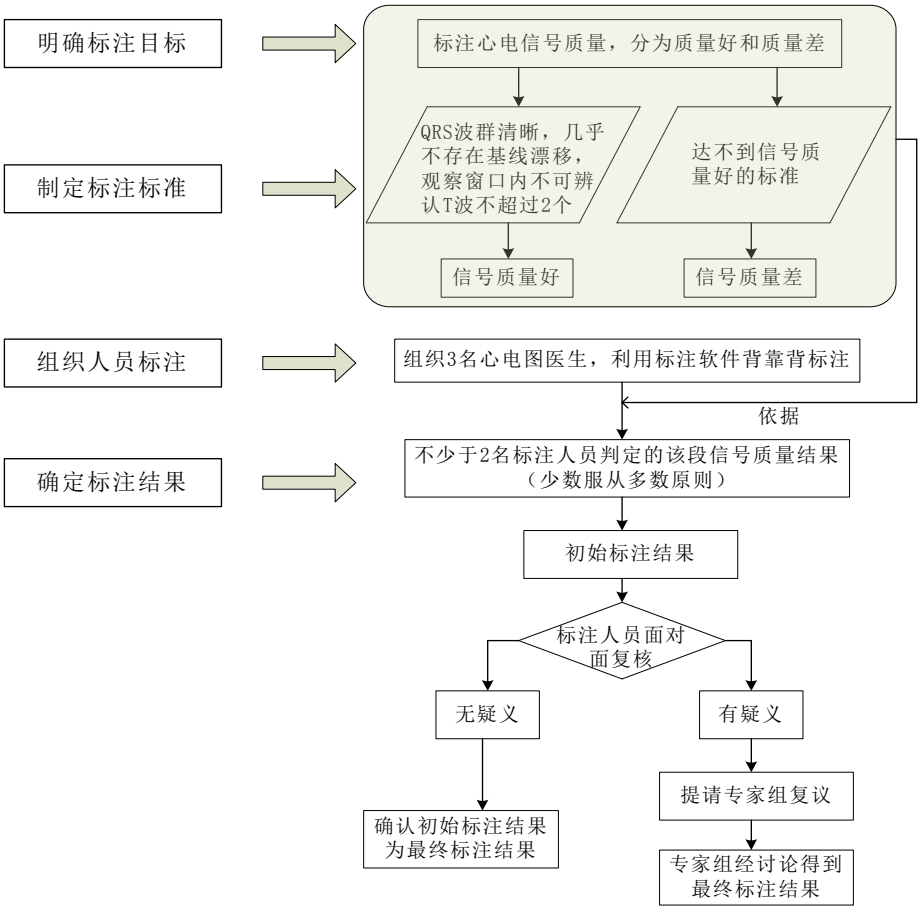

图 A.1 可穿戴心电标注流程示意图

A.1.4 标注人员

心电图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 90%。

A.1.5 标注软件

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.1.6 标注环境

标注任务在解放军总医院南病房楼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.1.7 数据

数据采取日期为 2020 年 1 月~7 月，采集设备某品牌的穿戴式单导联心电监护设备，数据格式为二进制文件，采样率为 200Hz。数据采集的地点为解放军总医院高压氧科。标注前需将每一个病人采集的数据按照每 10 秒一段、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

标注任务分类	数据模态：眼底彩照
	执行主体：人工标注
	标注结果格式：分类：图像类型不限
	标注结果性质：参考标准
	标注结果形式：分类

A.2.2 标注对象

本标注任务的标注对象是糖尿病视网膜病变（DR）的分类（DR 病变的疾病分期）。无明显 DR 的定义为：散瞳眼底检查所见无异常；非增生性 DR 的轻度增生型定义为：散瞳眼底检查所见仅有微动脉瘤；非增生性 DR 的中度增生型定义为：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：（1）4 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 定义为：出现以下 1 种或多种体征：新生血管形成、玻璃体积血或视网膜前出血。

标注对象的定义根据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。

主要参考文献包括：《眼科学》、《眼底病学》、《糖尿病视网膜病变的国际临床分级标准》。

A.2.3 标注规则

组织 2 名以上眼底医生，根据制定好的标注标准，经培训后，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。

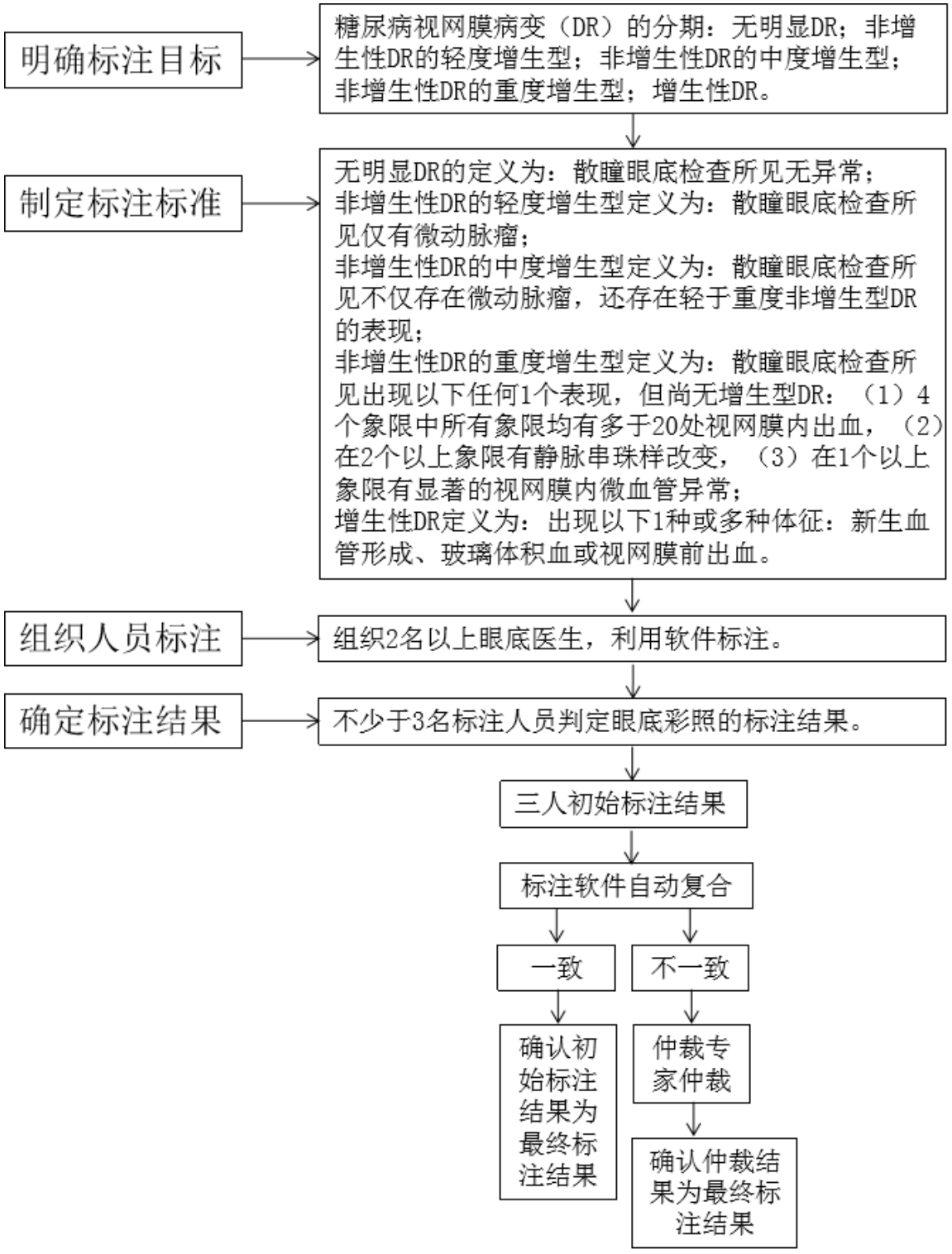

图 A.2 眼底彩照标注与质控流程示意图

A.2.4 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 100%。

A.2.5 标注软件

标注软件不限制造商，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.2.6 标注环境

标注任务在中山大学中山眼科中心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.2.7 数据

数据采取日期为 2020 年 1 月-12 月，采集设备某品牌的 45 度眼底相机。数据格式为 jpg、tiff、dcm、png。数据采集的地点为中山大学中山眼科中心人工智能研发部。标注前需将每一个病人采集的数据整理，只选取一张用于标注。

A.3 宫颈细胞病理图像

A.3.1 标注任务分类

标注任务分类	数据模态：宫颈液基细胞涂片的数字病理图像
	执行主体：人工标注
	标注结果格式：jpeg, tiff, jpeg2000
	标注结果性质：参考标准
	标注结果形式：分类

A.3.2 标注对象

本标注任务的标注对象是宫颈液基细胞涂片的数字病理图像（Whole Slide Image, WSI）中的细胞。采用目前国际广泛使用的宫颈细胞学 TBS（The Bethesda system） 2014 版为标准进行细胞学诊断，诊断结果具体描述为：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 (ASCUS)，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)。诊断结果为①则为阴性涂片，②-⑩则为阳性涂片，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由临床细胞病理专家和工程技术专家组成的专家组给出,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。

主要参考文献包括：

（1）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。中华病理学杂志.2021.50(4):319-322.

（2）The Bethesda System for Reporting Cervical Cytology: A Historical Perspective. Acta Cytol. 2017;61(4-5):359-372.

A.3.3 标注流程

每张宫颈液基细胞涂片的 WSI 组织 1 名病理专业研究生（总共 12 名），根据制定好的标注标准，经培训后，使用软件标注正常细胞（>1000 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，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。上述标注流程如图 A.3 所示：

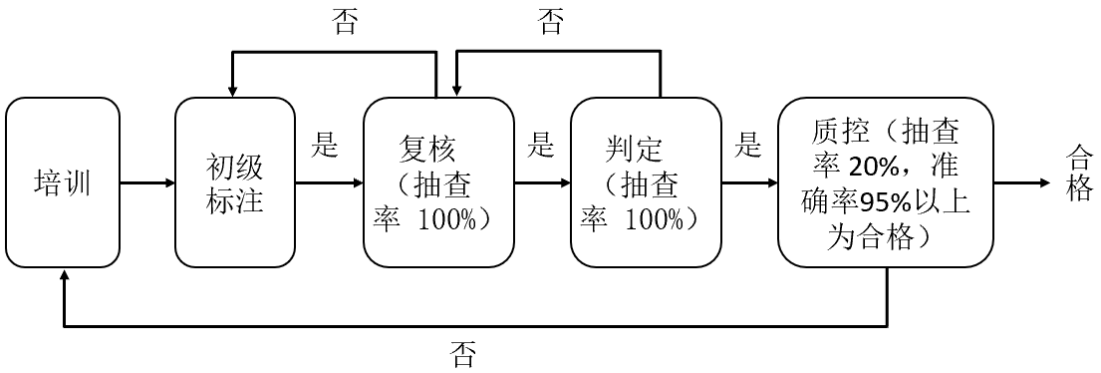

图 A.3 标注流程示意图

A.3.4 标注人员及分工

本标注任务的人员和分工见表 A.1。

表 A.1 标注人员职级及分工明细

职责	人数	职称	具体分工
标注者	12	病理专业硕士或博士研究生	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

其中，人员的考核指标为标注分类的准确率不低于 95%。

A.3.5 标注软件

标注软件来自商用现货软件 COTS/开源软件/自制软件，名称为 Automated Slide Analysis Platform (ASAP)，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.3.6 标注环境

标注任务在四川大学华西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.3.7 数据采集

数据采集时间段为 2019 年 1 月~2020 年 12 月；采集地点为四川大学华西医院病理科；数据采集设备为某品牌的全数字切片扫描仪，采用 0.25μm/pixel 的扫描分辨率获得宫颈液基细胞涂片 WSI。标注前，按照《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》对每一张宫颈液基细胞涂片 WSI 的整体质量进行评估，对于满意的图像样本，可用于标注。

A.4 皮肤彩照

A.4.1 标注任务分类

标注任务分类	数据模态：皮肤照片
	执行主体：手工标注
	标注结果格式：JSON
	标注结果性质：临床参考
	标注结果形式：分类

A.4.2 标注对象

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，标注点有皮损和分类两类：皮损：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）；分类：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），4 个标注目标。

标注对象的定义由皮肤科临床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均大于 20 年，从事数据标注的年限不低为 2 年。

主要参考文献包括：

《中国银屑病诊疗指南(2018 版)》中华医学会皮肤性病学分会银屑病专业委员会

A.4.3 标注规则

组织 3 名皮肤科医生，根据制定好的标注标准，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。

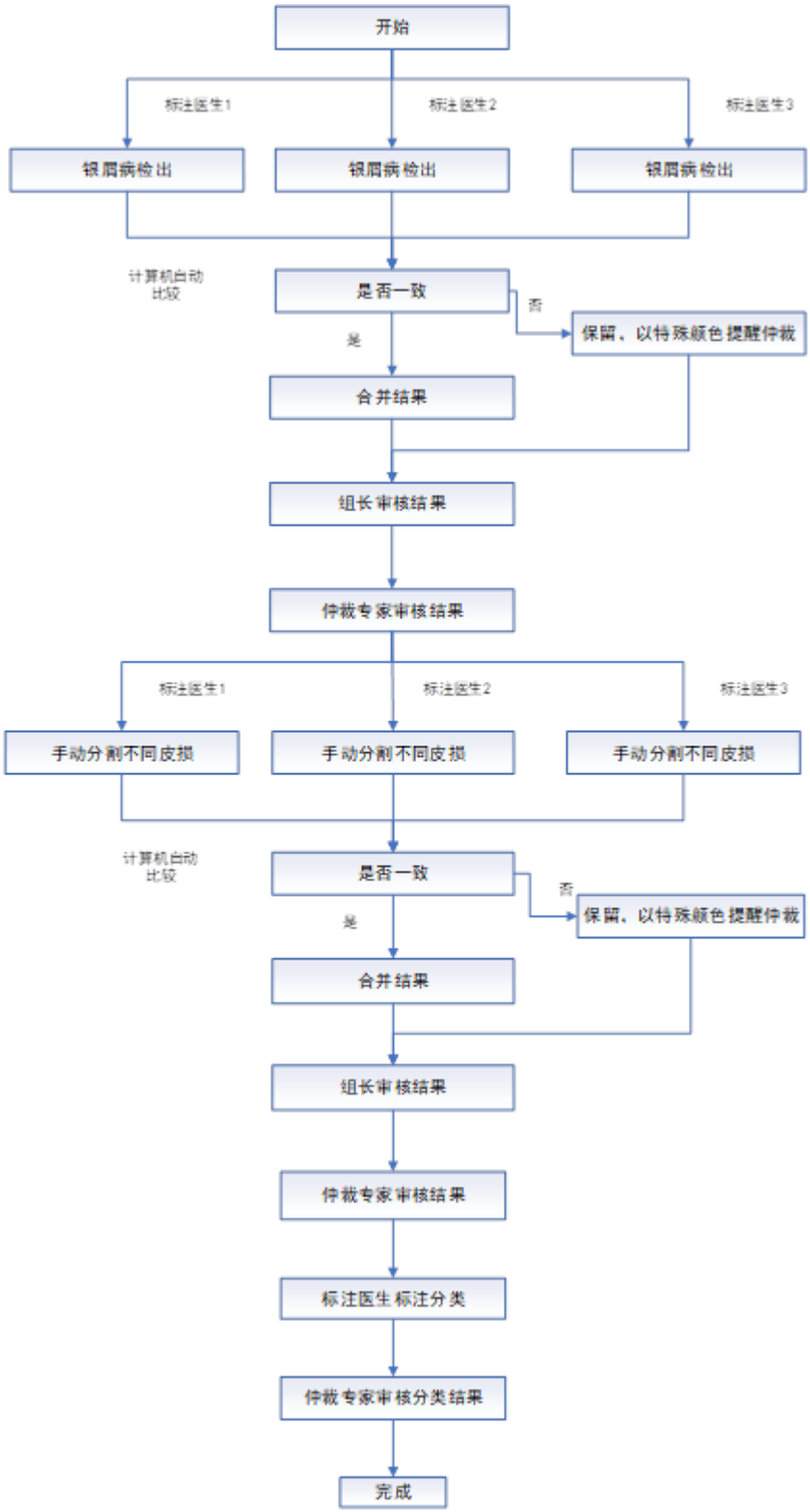

图 A.4 皮肤影像标注流程示意图

A.4.4 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.4.5 标注软件

标注软件来自商用现货软件，名称为元医绘，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.4.6 标注环境

标注任务在中国医学科学院皮肤病医院进行，使用普通显示器及办公电脑进行，无特殊环境要求。

A.4.7 数据

数据采取日期为 2020 年 1 月~12 月，采集设备单反相机（Canon EOS 800D），数据格式为 jpg。

数据采集的地点为中国医学科学院皮肤病医院激光科。

A.5 电子病历文本

A.5.1 标注任务分类

标注任务分类	数据模态：电子病历文本
	执行主体：人工标注
	标注结果格式：BIO
	标注结果性质：参考标准
	标注结果形式：实体提取

A.5.2 标注对象

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 UMLS 语义类型有疾病或者综合征（disease or syndrome）、中毒（injury or 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 语义类型有症状或体征（sign or 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）、治疗或预防过程（therapeutic or 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，指检查涉及到的体液检查项目、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，主要是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。

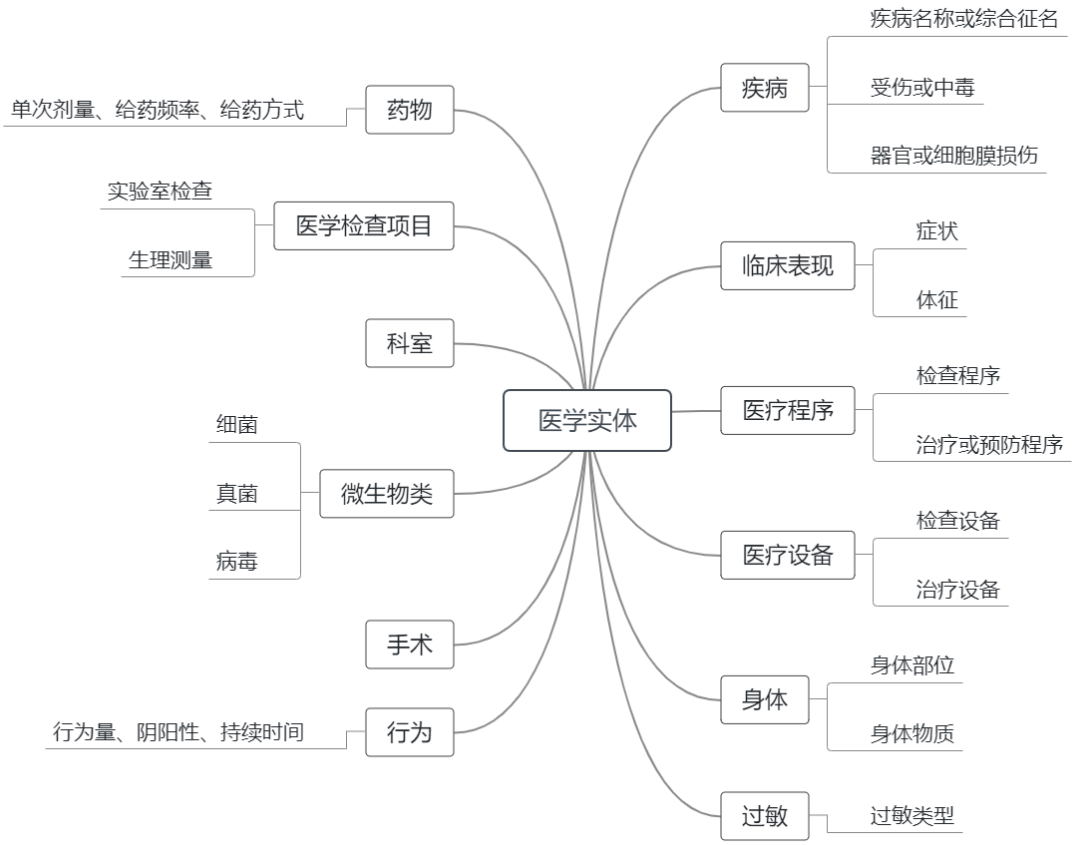

图 A.5 标注实体

标注对象的定义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。

主要参考文献包括：

[1] Nadkarni P. Natural language processing: an introduction [J]. Journal of the American Medical Informatics Association, 2011, 18(5): 544-551.

[2] Aronson AR. Elective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program [J]. Proceedings of the AMIA Symposium. American Medical Informatics Association, 2001, 2001(1): 17-21.

[3] Zou Q, Chu W W, Morioka C, et al. Index Finder: a method of extracting key concepts from clinical texts for indexing [J]. Proceedings of the AMIA Symposium. American Medical Informatics Association, 2003, 2003: 763-767.

[4] Rindflesch T C, Aronson A R. Ambiguity resolution while mapping free text to the UMLS Metathesaurus [J], 1994: 240.

[5] Weeber M, Mork JG, Aronson AR. Developing a Test Collection for Biomedical Word Sense Disambiguation [J]. Proceedings of the AMIA Symposium. American Medical Informatics Association, 2001, 8(1): 746.

[6] Chapman WW, Bridewell W, Hanbury P, et al. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries [J]. Journal of Biomedical Informatics, 2001, 34(5): 301-310.

[7] Huang Y, Lowe HJ. A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports [J]. J Am Med Inform Assoc, 2007, 14(3): 304-311.

[8] Tao C, Solbrig HR, Sharma DK, et al. Time-Oriented Question Answering from Clinical Narratives Using Semantic-Web Techniques[M]. Heidelberg: Springer Berlin Heidelberg, 2010: 241-256.

[9] Hripcsak G, Elhadad N, Chen Y H, et al. Using empiric semantic correlation to interpret temporal assertions in clinical texts [J]. J Am Med Inform Assoc, 2009, 16(2): 220-227.

A.5.3 标注规则

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 low-level 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 high-level 任务语料标注规范初稿。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。

迭代式更新标注规范。该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价培训质量。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，可以开始语料库的正式构建。

语料库正式构建。语料库构建过程中，将采取多种措施来保证标注质量，例如，①两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价结果；②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；③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。

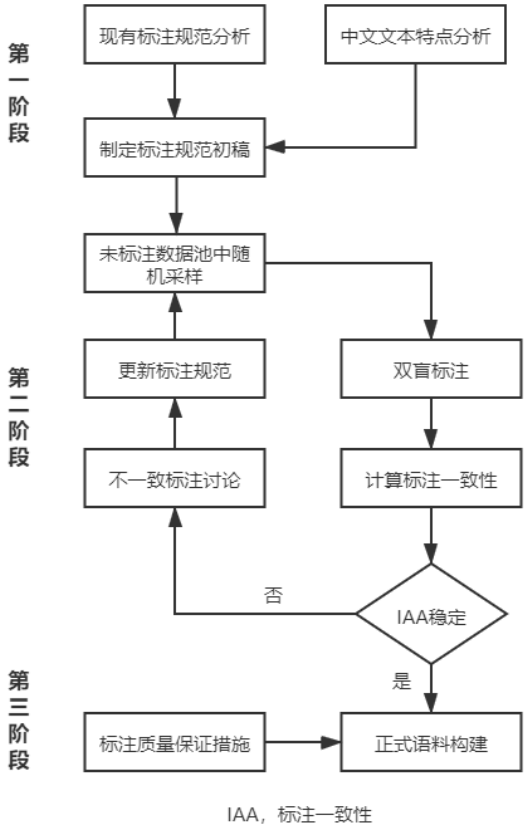

图 A.6 标注流程示意图

A.5.4 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.5.5 标注软件

标注软件来自开源工具，名称为 BRAT，基于 Web 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。详见标注工具官方说明书。

A.5.6 标注环境

标注任务在广东省人民医院医学大数据中心进行，使用普通办公电脑进行，标注辅助工具 BRAT 应于 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.5.7 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。

A.6 乳腺超声

A.6.1 标注任务分类

标注任务分类	数据模态：乳腺超声图像序列
	执行主体：半自动标注
	标注结果格式：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
	标注结果性质：参考标准
	标注结果形式：目标检测

A.6.2 标注对象

本标注任务的标注对象是超声动态图像序列中乳腺结节出现的位置。乳腺结节并没有一个严格的临床定义。通常用来表示发生在乳腺的小肿块。常见的乳腺结节有：增生结节，乳腺囊肿，纤维腺瘤，乳腺癌。乳腺超声 BI-RADS 分级是目前超声、放射、临床上都普遍认可的乳腺结节的分级，具体分级如下：

表 A.2　乳腺超声标注分级

分级	含义
Birads0	需要其他影像学检查(如乳腺X线检查或MRI等)进一步评估
Birads1	阴性。临床上无阳性体征，超声影像未见异常，如无肿块、无结构扭曲、无皮肤增厚及无微小钙化等
Birads2	良性病灶。基本上可以排除恶性病变。根据年龄及临床表现可每6~12个月随诊。如单纯囊肿、乳腺假体、脂肪瘤、乳腺内淋巴结(也可以归类1类)、多次复查图像无变化的良性病灶术后改变及有记录的经过多次检查影像变化不大的结节可能为纤维腺瘤等。
Birads3	可能良性病灶。建议短期复查(3 ~ 6 个月) 及其他进一步检查。根据乳腺X线检查积累的临床经验，超声发现明确的典型良性超声特征如实性椭圆形、边界清、平行于皮肤生长的肿块，很大可能是乳腺纤维腺瘤，它的恶性危险性应该小于2%，如同时得到临床、乳腺X线检查或MRI的印证更佳。新发现的纤维腺瘤、囊性腺病、瘤样增生结节(属不确定类)、未扪及的多发复杂囊肿或簇状囊肿、病理明确的乳腺炎症及恶性病变的术后早期随访都可归于该类
Birads4	可疑的恶性病灶。此级病灶的恶性可能性为2%~95%。评估4类即建议组织病理学检查：细针抽吸细胞学检查、空芯针穿刺活检、手术活检提供细胞学或组织病理学诊断。超声声像图上表现不完全符合良性病变或有恶性特征均归于该类。目前可将其划分为4A、4B及4C。4A类更倾向于良性可能，不能肯定的纤维腺瘤、有乳头溢液或溢血的导管内病灶及不能明确的乳腺炎症都可归于该类，此类恶性符合率为3%~10%；4B类难以根据声像图来明确良恶性，此类恶性符合率为11%~50%；4C类提示恶性可能性较高，此类恶性符合率为51%~94%。
Birads5	高度可能恶性，应积极采取适当的诊断及处理措施。超声声像图恶性特征明显的病灶归于此类，其恶性可能性大于等于95%，应开始进行积极的治疗，经皮穿刺活检(通常是影像引导下的空芯针穿刺活检)或手术治疗。
Birads6	已经活检证实为恶性。此类用于活检已证实为恶性，但还未进行治疗的影像评估。主要是评价先前活检后的影像改变，或监测手术前新辅助化疗的影像改变。

参考文献: 中国抗癌协会乳腺癌诊治指南与规范(2017 年版)

A.6.3 标注规则

组织 3 名以上超声医生，根据制定好的标注标准，经培训后，使用软件对超声动态图像序列中的出现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

医生独立的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。

乳腺超声序列中结节标注需要标注人员在连续的超声影像中标记出结节的所在位置（通常用结节的外接矩形框来表示），一个超声序列通常包含大量的连续采集影像，如果采用纯手动标注，标注人员需要在结节出现的每一张图像中进行标注，标注的时间和人力成本都会快速上升。为了解决这个问题，可以采用人工标注和跟踪算法自动填充相结合来实现一种结节的半自动标注。具体过程如下：

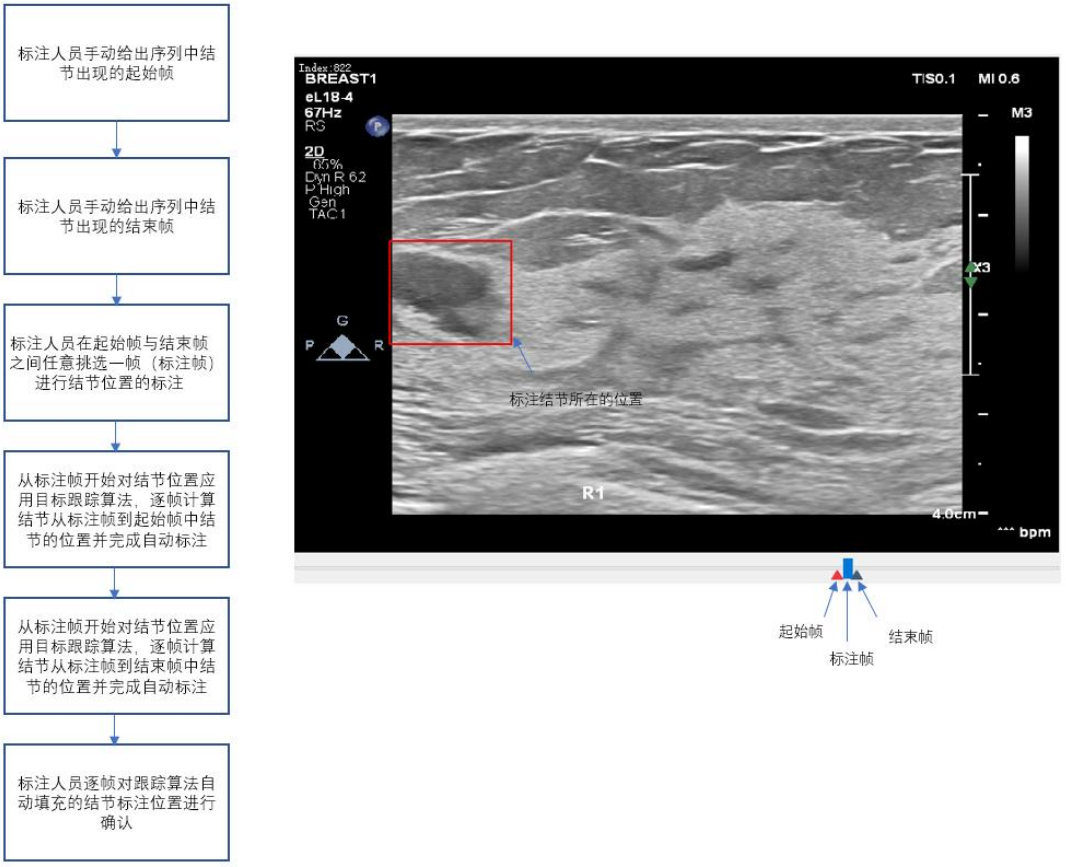

图 A.7 乳腺超声标注流程图

根据上述方法，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，然后再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.6.4 标注人员

标注医生的职称不低于副主任医师，过去三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.6.5 标注软件

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。

A.6.6 标注环境

标注任务使用办公电脑进行，无特殊环境要求。

A.6.7 数据

数据采取日期为 2020 年 7 月-10 月，采集设备为飞利浦 EPIQ-7 超声系统及 EL18-4 和 L12-5 超声探头。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 1000 帧左右。图像格式为 DICOM。数据采集的地点为北京协和医院超声科。

附 录 B
（资料性）
业务架构示例（胸部 CT 肺结节）

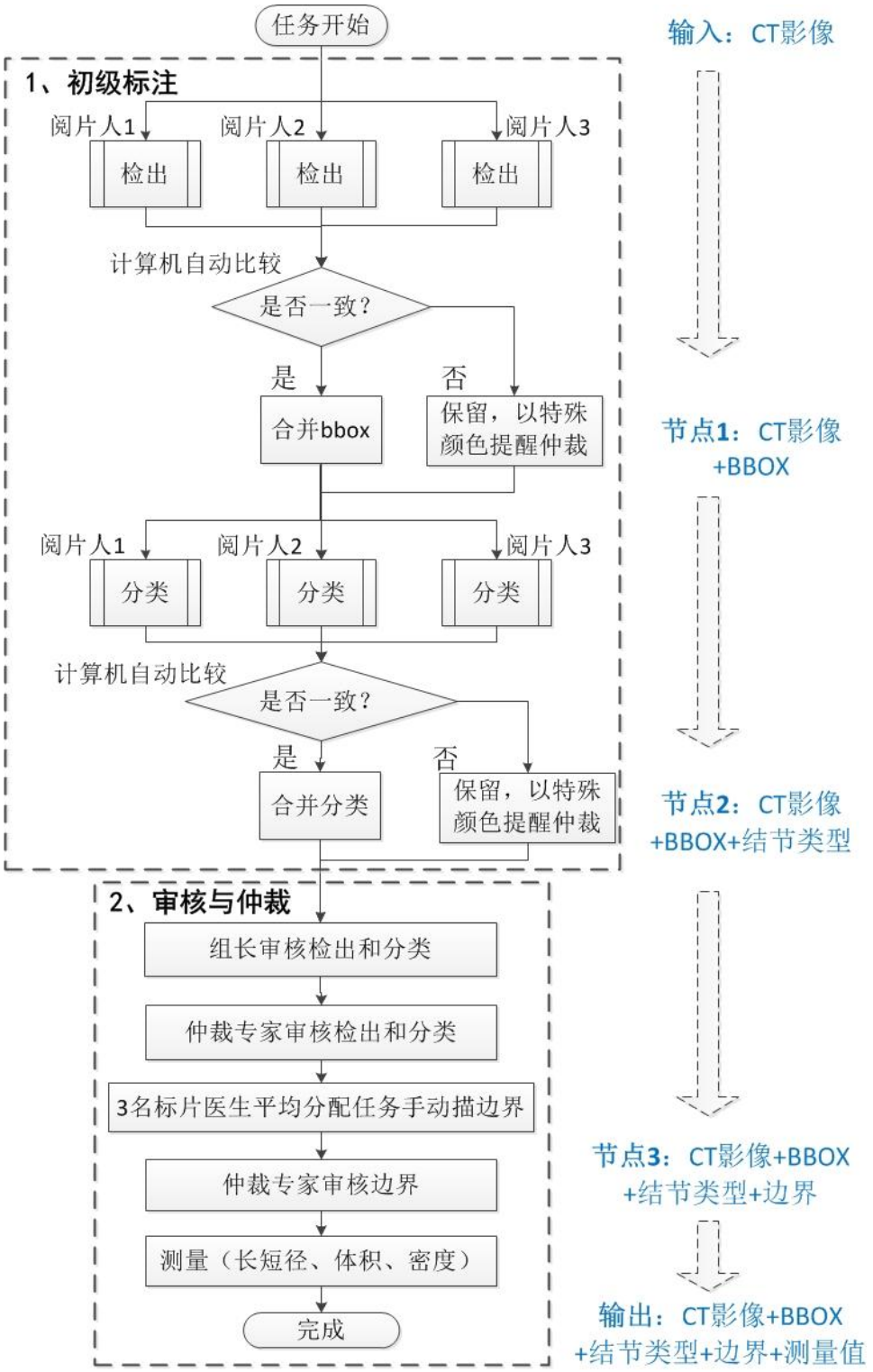

图 B.1 胸部 CT 肺结节标注业务架构示意图

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，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；标注组长要求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：由其他标注组长和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，仲裁专家可以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紧密包裹肺结节的标注框（bounding box，简称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即为节点 3。遇到复杂征象时，可酌情增加审核人数，以保证标注质量。结节的尺寸根据手工边界由计算机自动生成，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紧密包裹肺结节的标注框（bounding box，简称 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。

附 录 C
（资料性）
对 AI 辅助标注性能的评价

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，作为参考。

C.1 总体原则

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。

标注工具或标注平台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估可使用绝对误差、相对误差、相关性 Pearson 值、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×100%；

相关性 Pearson 值：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：

$$\rho X,Y = \frac{\mathrm{cov}(X,Y)}{\sigma X \sigma Y} = \frac{E\left[\left(X-\mu X\right)\left(Y-\mu Y\right)\right]}{\sigma X \sigma Y} \tag{C1}$$

相关 Pearson 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。

MSE 的公式如下：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - d_i)^2 \tag{C2}$$

式中：

y——标签；

d——预测值；

n——样本数量。

MAE 的公式如下：

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left| y_i - d_i \right| \tag{C3}$$

式中：

y——标签；

d——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估。

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、mAP、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、分割相关的评估指标，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估指标。

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，其公式如下：

$$\mathrm{recall} = \frac{TP}{TP + FN} \tag{C4}$$

式中：

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，其公式如下：

$$\mathrm{precision} = \frac{TP}{TP + FP} \tag{C5}$$

式中：

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（Precision）和召回率（Recall）。改变阈值，可画出 Precision - Recall 曲线，该曲线下的面积为平均精度(Average Precison, AP)，其公式为：

$$\mathrm{AP} = \sum_{k=1}^{N} \mathrm{p}(k)\Delta \mathrm{r}(k) \tag{C6}$$

式中：

k——图片数量；p(k)——Precision 值；r(k)——Recall 值。

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

$$\mathrm{mAP} = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{C7}$$

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：

$$Jaccard = \frac{|A \cap B|}{|A \cup B|} \tag{C8}$$

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：

$$Dice = 2 \times \frac{|A \cap B|}{|A| + |B|} \quad \text{（C9）}$$

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：

$$Conformity = 1 - \frac{V_{misclassified}}{TP} \quad \text{(C10)}$$

式中$V_{misclassified}$为错误分割的像素数量，TP 为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 C8。

Hausdorff 距离：用于描述两个分割区域轮廓线的距离，双向 Hausdorff 距离的计算公式为

$$d_H(X,Y) = \max\{d_{XY}, d_{YX}] = \max\left\{\max_{x\in X}\min_{y\in Y} d(x,y), \max_{y\in Y}\min_{x\in X} d(x,y)\right\} \quad \text{.....（C11）}$$

其中d(x, y)为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 C2。

C.4 区域分割

分割的评估使用 Dice 系数、Conformity 系数、交并比、Hausdorff 距离、相关性 Pearson 值、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 C9。

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 C10。

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

$$mIOU = \frac{1}{n}\sum_{i=1}^{n}\frac{|P_i \cap T_i|}{|P_i \cup T_i|} \quad \text{（C12）}$$

式中 P——预测的区域；T——分割标注区域。

d) Hausdorff 距离： 同公式 C12。

e) 相关性 Pearson 值： 同公式 C1。

C.5 辅助分类

分类的评估使用灵敏度、特异性、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：

$$sen = \frac{TP}{TP + FN} \quad \text{(C13)}$$

b) 特异性：

$$spe = \frac{TN}{TN + FP} \quad \text{(C14)}$$

c) 准确率 acc：

$$\mathrm{acc} = \frac{TP + TN}{TP + FP + TN + FN} \tag{C15}$$

C.6 辅助检出

检出的评估使用召回率、精确度、F_1度量、mAP 等作为指标。

a) 召回率：见公式 C4

b) 精确度：见公式 C5

c) F_1度量：

$$F_1 = \frac{2 \times \mathrm{Re}c \times \mathrm{Pr}e}{\mathrm{Re}c + \mathrm{Pr}e} \tag{C16}$$

式中 Rec 为召回率，Pre 为精确度。

d) mAP：见公式 C7。

C.7 关键点定位

关键点定位的评估使用绝对误差、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：

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{C17}$$

b) N 维数据中闵氏距离绝对误差：

$$D(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{\frac{1}{n}} \tag{C18}$$

c) PCK(Percentage of Correct Keypoints)百分比：

$$PCK_{mean}^{k} = \frac{\sum_p \sum_i \delta\left(\frac{d_{pi}}{d_p^{def}} \le T_k\right)}{\sum_p \sum_i 1} \tag{C19}$$

表示正确估计出的关键点比例。其中 i 表示 id 为 i 的关键点，k 表示第 k 个阈值，p 表示第 p 个人，dpi 表示第 p 个人中 id 为 i 的关键点预测值与人工标注的欧式距离，dpdef 表示第 p 个人的尺度因子。

C.8 标注效率

标注工具、标注平台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估指标。

参 考 文 献

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

[2]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

[3]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

[4]T/ISC 0005—2020 针对内容安全的人工智能 数据标注指南

[5]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.

[6]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.

[7]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.

[8]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.

[9]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.

[10]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.

[11] 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..

[12]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.

[13]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.

[14]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.

[15]《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.

[16] 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.