

T/CQAE

团 体 标 准

T/CQAE XXXX—XXXX

人工智能 辅助识别系统 儿童先天性心脏 病的数据质量控制要求

Artificial intelligence - Assisted recognition system - Data quality control
requirements for congenital heart disease in children

(报批稿)

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX - XX - XX 发布

XXXX - XX - XX 实施

中国电子质量管理协会 发布

目 次

前 言	II
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 缩略语	1
5 数据采集要求	1
5.1 采集设备	1
5.2 采集操作	1
5.3 采集质量控制要求	3
5.4 录入者管理要求	3
6 数据分析要求	4
6.1 数据可视化要求	4
6.2 音频规格要求	4
6.3 数据分析流程	4
7 数据标注要求	5
7.1 标注信息要求	5
7.2 标注质量控制要求	6
7.3 标注团队要求	6
7.4 审核委员会团队要求	6
7.5 不良数据处理	6
7.6 心音标注软件要求	6
8 数据集质量控制要求	7
8.1 心音数据集有效性与质量要求	7
8.2 心音数据集的隐私性	7
附 录 A （规范性） 心音数据的值域标注信息	9

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专利的责任。

本文件由中国电子质量管理协会计算机硬件和信息系统质量测评分会提出，由中国电子质量管理协会归口。

本文件起草单位：上海计算机软件技术开发中心、上海交通大学医学院附属新华医院、上海医药行业协会、道普信息技术有限公司、上海市软件行业协会。

本文件主要起草人：孙锬、陈文捷、沈颖、孙新、陈敏刚、吴耀卫、刘梦云、姚宝敬、陈笋、马泽宇、赵鹏军、蔡立志、赵毅、朱蓓芬、魏丽丽。

人工智能 辅助识别系统 儿童先天性心脏病的数据质量控制要求

1 范围

本文件规定了儿童先天性心脏病的数据采集、数据分析、数据标注和数据集质量控制要求。本文件适用于儿童先天性心脏病的心音辅助识别数据采集者、分析者、标注者和管理者。

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。

GB/T 35273—2020 信息安全技术 个人信息安全规范

YY/T 1035—2021 听诊器

3 术语和定义

YY/T 1035—2021界定的以及下列术语和定义适用于本文件。

3.1

心音 heart sound

由心肌收缩、心脏瓣膜关闭和血液撞击心室壁、大动脉壁等引起的机械波现象所产生的声音。

3.2

心音样本 heart sound sample

同一受检者一次采集的心音音频文件、受检者基本信息、心脏超声报告合并称为一份心音采集样本，作为收录的最小数据单元。

3.3

心音数据集 heart sound dataset

多个已标注且经过审核的心音样本组成的数据集合。

3.4

上位机 master computer

与听诊器连接，能够控制听诊器进行心音数据采集，并对听诊器采集的心音数据进行接收、管理的计算机。

4 缩略语

下列缩略语适用于本文件。

JPEG: 联合图像专家组 (Joint Photographic Experts Group)

MFCC: 梅尔倒谱系数 (Mel-Frequency Cepstral Coefficients)

PCM: 脉冲编码调制 (Pulse Code Modulation)

PDF: 便携式文档格式 (Portable Document Format)

PNG: 便携式网络图形 (Portable Network Graphics)

5 数据采集要求

5.1 采集设备

应采用电子听诊器进行心音数据采集，心音数据采集设备在数据采集方面应满足YY/T 1035—2021的要求。

5.2 采集操作

5.2.1 心音采集对象

心音采集对象为0-18岁新生儿、婴幼儿、学龄期、青春期儿童人群。

5.2.2 心音采集环境要求

心音采集环境有如下要求。

- a) 采集时应确保周围环境安静，例如：减少旁人说话、开关门声、背景音乐等背景音干扰；
- b) 采集时应保证受检者情绪平稳，应与受检者做好沟通，例如：对于婴幼儿应尽可能安抚，避免哭闹，对于能配合体位的受检者，应采取坐位或仰卧位，并露出胸口，同时做好受检者隐私保护工作。

5.2.3 心音采集操作要求

心音采集操作按如下要求进行。

- a) 采集时应采集如图 1 所示的听诊区。

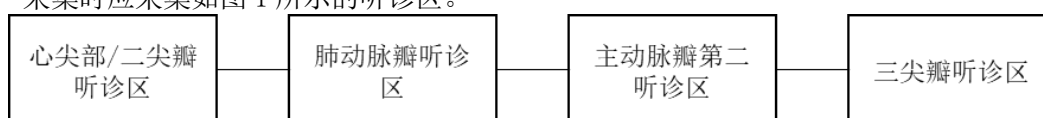


图1 心音采集听诊区

- b) 采集时应遵循各听诊区的临床听诊定位：
 - 1) 心尖部/二尖瓣听诊区定位方式：乳头位与第4肋间，乳头下1肋间听诊即第5肋间隙，可以触摸肋骨定位，锁骨中线即经过锁骨中点的垂直线，心尖部位于左侧第5肋间+左侧锁骨中线内侧0.5-1.0cm。
 - 2) 肺动脉瓣听诊区定位方式：胸骨自上而下可触及一骨性突出部位，为胸骨角，此处平对第2肋，胸骨左侧边缘即为肺动脉瓣听诊区。
 - 3) 主动脉瓣第二听诊区定位方式：胸骨左缘第3肋间，即肺动脉瓣听诊区下一肋间。
 - 4) 三尖瓣听诊区定位方式：胸骨左缘第4~5肋间，即平对乳头的胸骨左缘或下一肋间。
- c) 采集时应在每个听诊区采集不少于 15s 的心音，每个听诊区的采集时间间隔不低于 2s。
- d) 切换听诊区期间，应使听诊器远离受检者皮肤表面、衣物等干扰源，处于悬空状态，确保各听诊区之间有一段空白录音间隙。
- e) 受检者基本信息直接录入系统或用电子表格记录。基本信息包含如下：
 - 受检者编号；
 - 出生日期；
 - 性别；
 - 身高（cm）；
 - 体重（kg）；
 - 临床诊断信息，如先心病初筛阳性、先心病初筛阴性。

5.2.4 心音信号采集器处理要求

应在采集器进行一系列的信号增强及降噪处理。

- a) 模拟部分：
 - 1) 通过传感器将机械波转换为正负电压变化的模拟信号，通过模拟低通滤波电路和模拟高通滤波电路的组合进行初始滤波，衰减 20Hz 以下及 2000Hz 以上的模拟信号；
 - 2) 通过运算放大器将模拟信号整体放大，放大倍率宜设计为 5-10 倍。
- b) 数字部分：
 - 1) 通过音频芯片，对模拟电压信号进行采样，转换为数字信号。音频芯片的采样率应不低于 8000Hz，采样位宽应不低于 16 位；
 - 2) 通过音频芯片，对数字信号进行 20Hz-800Hz 的带通滤波处理、50Hz 的陷波处理、100Hz-200Hz 的带通增益处理，达到降低高频段噪音、降低工频干扰以及增强心音核心频段的效果；

- 3) 通过音频芯片将数字信号整体放大，放大倍率宜设计为10-20倍。

5.2.5 心音录入软件处理要求

心音录入软件在音源管理、档案管理、数据处理、数据格式方面有如下要求。

- a) 音源管理要求：
 - 1) 宜支持通过无线方式与电子听诊器进行通讯；
 - 2) 宜支持本地导入符合采集要求的心音音频文件；
 - 3) 应支持心音数据质量的自动检测，输出心音数据质量报告，提示并过滤不合格数据。
- b) 档案管理要求：
 - 1) 宜支持单个或批量受检者档案建档；
 - 2) 宜支持心脏超声报告文件的导入及匹配；
 - 3) 应支持受检者档案的增、删、查、改。
- c) 数据处理要求：
 - 1) 应支持实时心音音频码流的接入、存储，宜支持可视化；
 - 2) 应支持多种格式心音音频文件的存储，宜支持可视化；
 - 3) 宜支持心音图、频谱图等多维度的心音数据可视化；
 - 4) 宜支持适合心音信号的多种滤波组件及信号增强组件。
- d) 数据格式要求：
 - 1) 应支持至少两种音频格式的处理，包括PCM原始数据格式、WAV无损格式；
 - 2) 应支持PDF格式文件的处理；
 - 3) 应支持JPEG、PNG格式图片的处理；
 - 4) 应支持Excel格式文件的处理。

5.3 采集质量控制要求

5.3.1 心音样本录入方式

心音样本录入方式分为同步和异步录入。同步录入方式，即听诊器与心音录入软件实时连接，在心音录入软件界面进行受检者信息录入及心脏超声报告导入；异步录入方式，即所有信息与文件均存储于上位机，通过一定的规则一次性导入心音录入软件。

5.3.2 采集质量控制要求

心音样本采集质量控制有如下要求。

- a) 应保障每例心音样本都进行数据质量控制，严格遵守质量控制要求。其中，数据类型和完整性要求应包括：不应出现断层、缺层等情况；不应进行任何修改、编辑；不应进行有损压缩等；
- b) 单次心音采集时长不低于60s，应按照顺序完整覆盖二尖瓣听诊区、肺动脉瓣听诊区、主动脉瓣第二听诊区、三尖瓣听诊区共四个听诊区的音频信号；
- c) 心音存储文件格式应为无损音频格式；
- d) 心音存储文件码流不低于8kbps，数据位宽不低于16bit；
- e) 预处理阶段应进行心音质量评定，心音音频信号应能明确区分出第一心音和第二心音，才能进行后续转换。

5.3.3 录入质量控制要求

心音样本录入质量控制有如下要求。

- a) 应确保受检者基本信息的真实性、有效性；
- b) 心脏超声报告应为PDF格式文件，可进行查阅及自动化识别处理；
- c) 同一样本中的所有信息，应一一对应，并具有唯一标识号。

5.4 录入者管理要求

心音样本数据的录入者有如下管理要求。

- a) 参与数据上传的操作者应进行严格的身份认证，包括但不限于手机号码、身份证号码、就职医院/科室、资格证书；
- b) 所有样本的录入动作、过程可溯源。

6 数据分析要求

6.1 数据可视化要求

应构建心音时频能量图，如图2所示，心音时频能量图的生成分为以下几个步骤：

- a) 从 WAV 文件中获取 PCM 码流数据；
- b) 对 PCM 码流进行降采样、滤波、归一化等预处理；
- c) 时频转换，通过 MFCC、傅立叶变换、小波分解等方法将心音时域信号转换为频域信号，此时图谱数据为二维矩阵，一个维度代表时间，一个维度代表频率；
- d) 能量计算，计算每一个时间点在每一个频点的能量值，形成一个新的维度；
- e) 能量着色，选取任意渐变色带，将能量值最小值对应色带的一端，能量值最大值对应色带的另一端，将能量值等比例用相应的色号替代；
- f) 完成构图，形成一个第一维度代表时间、第二维度代表频率、第三维度代表能量的图谱。

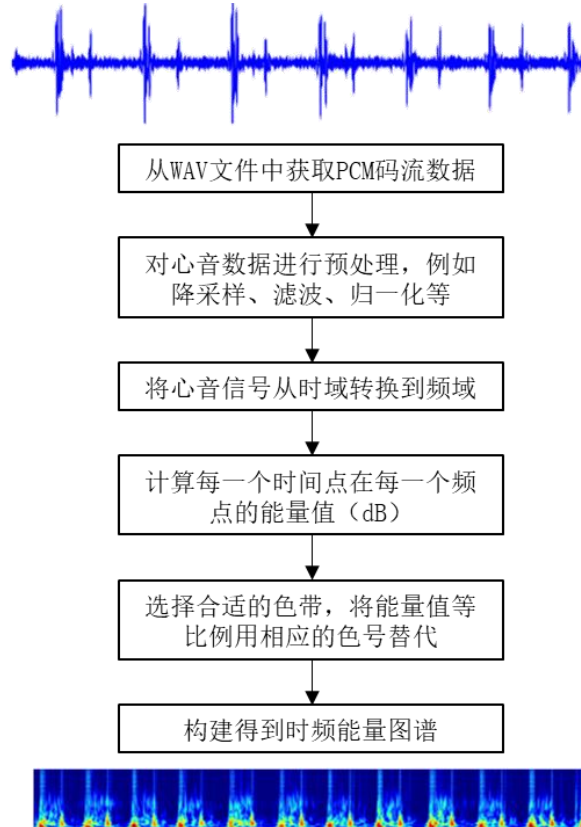


图2 心音时频能量图生成步骤

6.2 音频规格要求

电子听诊器采集的原始心音数据应由音频芯片AD采样得到的原始值构成，针对心音数据的无损传输及存储的需求，音频码流应使用PCM进行编码，音频存储文件格式应使用无损WAV文件格式。

心音数据编码码流不低于8kbps、数据位宽不低于16bit。

6.3 数据分析流程

6.3.1 分析步骤

心音信号分析应按照如图3所示的步骤进行：

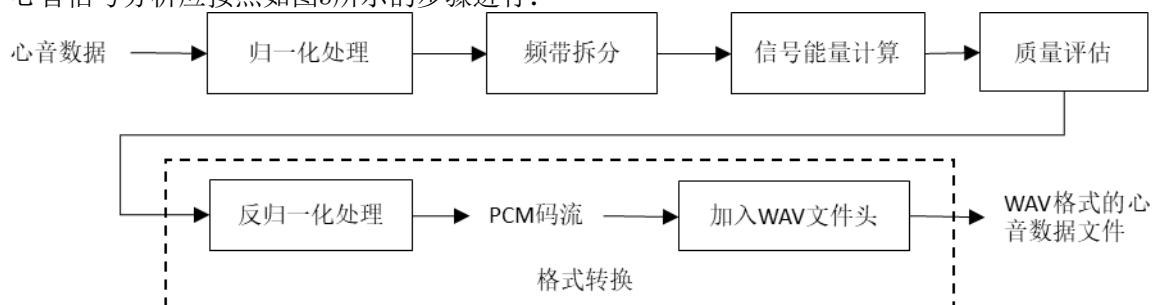


图3 心音信号分析步骤

6.3.2 归一化处理

选用常见的归一化函数对心音数据进行归一化处理，要求基本度量单位应统一。

6.3.3 频带拆分

归一化处理后的的心音信号分别经过20Hz低通，20~500Hz带通，500Hz高通IIR滤波器进行频带拆分（截止频点衰减值为3dB），得到不同的信号幅度值。

6.3.4 信号能量计算

计算不同频率通道下的心音信号的能量值。

6.3.5 质量评估

质量评估的步骤如下：

- 根据不同频率通道下的心音信号的能量值计算出基础分；
- 对基础分进行若干个约束，得到不同频率通道下心音信号质量分数；
- 综合不同频率通道下心音信号质量分数求出平均值，作为心音信号质量等级划分的依据；
- 根据计算结果进行心音信号质量等级的划分，作为心音质量的评定结果。

6.3.6 格式转换

将质量评估完成后的心音数据进行反归一化处理，还原PCM码流，加入WAV文件头，生成WAV格式的心音音频文件。

7 数据标注要求

7.1 标注信息要求

7.1.1 综述

心音标注信息要求标注者对两部分信息进行标注：时域信息部分、值域信息部分。

7.1.2 时域信息标注

在时域上，将心音信号分为四个瓣膜听诊区，每个瓣膜听诊区分为若干心动周期。

- 瓣膜区包括：二尖瓣听诊区、肺动脉瓣听诊区、主动脉瓣第二听诊区和三尖瓣听诊区；
- 每个心动周期包括第一心音期、收缩期、第二心音期、舒张期。

心音数据的时域信息标注，主要是对各瓣膜听诊区的起始时间点和心动周期各分割期对应的起始时间点进行标注。

7.1.3 值域信息标注

心音数据的值域标注，主要是对心音质量、心音是否有异常及异常的具体类型进行标注。具体标注内容按附录A的规定执行。

7.2 标注质量控制要求

对于心音数据的人工标注，应建立规范的标注流程，建立多人标注与复查机制，对分歧样本有明确处理，保证标注质量。

同一个心音样本，由系统分派至少2位标注医师分别进行标注，标注的结果由标注初审医师审核并确认，并对标注结果进行校正。当2位标注医师的标注结论不一致时，由标注初审医师提供校准意见，应提请标注复审医师复核，并提交最终标注结论。

当心音标注结果与心脏超声确诊结果无法匹配或临床上无法解释时，由心音标注数据审核委员会进行裁定，取得共识后的标注结果方可用于后续的数据建模和分析。

7.3 标注团队要求

标注团队由标注医师、标注初审医师及标注复审医师组成，其中：

- a) 标注医师：儿科心血管专科医生工作年限 5 年及以上，接受心音标注标准操作流程、标注软件及云平台系统的相关培训，负责心音音频的首轮标注。
- b) 标注初审医师：儿科心血管专科医生工作年限 10 年以上，接受心音标注标准操作流程、标注软件及云平台系统的相关培训，针对标注医师的标注结果进行审核并确认，对存在标注误差的结果予以校准，如遇有争议的标注结果，提交复审医师重审。
- c) 标注复审医师：儿科心血管专科医生工作年限 15 年以上，接受心音标注标准操作流程、标注软件及云平台系统的相关培训，对于提请复审的心音样本进行标注审核，对存在争议的心音标注结果提交审核委员会仲裁。

7.4 审核委员会团队要求

心音标注数据审核委员会主要负责针对复审医师提请的复杂心音标注结果进行准确性评价与校准，以及针对后期纳入建模数据集数据可用度进行判定，并针对可能存在的心音标注结果与超声诊断结果的显著分歧进行最终裁定。

心音标注数据审核委员会由3-5名儿科心血管专科工作年限15年及以上的主任医师，与1-2位心脏超声专科工作10年以上的副主任及以上医师组成。

7.5 不良数据处理

对不良数据的标注有如下处理要求。

- a) 无法获得心脏超声诊断报告的音频样本不予以标注；
- b) 心音采集音频质量差的音频样本在标注过程中由标注医生主动剔除，不予以标注；
注：心音采集音频质量差的表现：心音信号被外界的噪声，如肌肉运动、呼吸音或说话的声音所掩盖或部分掩盖，造成原始心音音频信号分析困难。
- c) 心音样本与心脏超声报告无法匹配的数据亦从数据集中剔除，不纳入后续数据建模与分析。

7.6 心音标注软件要求

对心音标注软件有如下要求。

- a) 系统管理要求：
 - 1) 应对参与样本标注、初审、复审、仲裁的医生应进行严格的身份认证，包括但不限于手机号码、身份证号码、就职医院/科室、资格证书、工作年限证明；
 - 2) 应支持多种用户角色的权限管理和编辑；
 - 3) 应支持未标注样本库，已标注样本库、有效样本库、疑难样本库等多类型数据库管理；
- b) 档案管理要求：
 - 1) 应支持受检者档案列表及详情；
 - 2) 应支持标注包、审核包的自动和手动生成和派发；
 - 3) 应支持参与机构、参与医生的档案管理。
- c) 数据处理要求：

- 1) 应支持标注工作台,实现心音播放、时频能量图查看、心音时域自动分割、标注选项框、心脏超声报告查看等功能;
- 2) 应支持审核工作台,实现心音播放、时频能量图查看、心音时域自动分割、已标注历史信息的查看、审核意见框、心脏超声报告查看等功能;
- 3) 应支持采集端、标注工作台、审核工作台之间数据的一致性、完整性。
- d) 数据格式要求:
 - 1) 应支持 WAV 格式心音音频文件播放。
 - 2) 应支持 PDF、JPEG、PNG、Excel 格式文件的处理。

8 数据集质量控制要求

8.1 心音数据集有效性与质量要求

8.1.1 心音数据集的有效性

心音数据集的有效性要求包括:

- a) 清晰度要求:应满足心音音频清晰,且声音频率基本落在人耳听域范围内,满足标注者对音频特征的判断以及标注的执行;
- b) 声音内容要求:对于同一样本,应在二尖瓣听诊区、肺动脉瓣听诊区、主动脉第二听诊区、三尖瓣听诊区分别采集心音,且每个心音音频应至少包含 10 个以上完整心动周期;
- c) 采集技术要求:技术操作正确,设备参数设置符合规范,尽可能降低环境杂音等影响数据有效性的缺陷,并保证因采样损失的信息不对人工智能模型训练造成严重影响。

8.1.2 心音数据集的质量

心音数据集的质量要求包括:

- a) 数据集中应包含所需的多模态数据,包括但不限于:采样后的音频数据,如 WAV 格式的音频文件、心音图、心音频谱能量图;受检者基础信息等;
- b) 数据集中应包含正负样本(其比例接近 1:1),且应包含至少五类不同先天性心脏病细分病种样本,其中每一类先心病样本绝对数量在 50 以上,且不会由于数据量原因严重影响人工智能模型训练。

8.2 心音数据集的隐私性

心音数据集的采集、标注、使用、共享应符合 GB/T 35273—2020 的相关规定,数据处理的各环节应满足:

- a) 心音录入软件应支持对接入数据的脱敏处理,在数据管理的接入、存储、读取等各环节保护受检者的知情权和隐私权,并防止数据篡改和个人敏感信息泄露。
- b) 心音录入软件应在心音样本上传时自动地对受检者隐私信息进行本地数据脱敏处理,脱敏处理可采用数据乱序、数据交换、数据扭曲、数据清洗、数据匿名、数据屏蔽、数据泛化、数据加密等方式。数据脱敏前后应保持数据的特征,包括数据的格式、类型、依存关系、统计特征等。脱敏后的数据应保留原始数据的主外键关系、业务包含关系等。
- c) 心音标注软件应支持对接入样本的脱敏处理,在样本管理、标注等各环节保护受检者的知情权和隐私权,并防止数据篡改和个人敏感信息泄露。
- d) 心音数据采集、处理、应用应保证受检者家长的知情权与授权,责任主体应具有合法、明确的个人信息处理目的。采集过程应符合规范,获得个人信息主体的知情权与授权,责任主体应具有合法、明确的个人信息处理目的,对于心音数据搜集过程,应在受检者家长知情、同意的情况下进行搜集,并对元数据进行脱敏处理。
- e) 心音数据的采集、存储、挖掘、转换、传输、应用等环节应采取合理的隐私安全管理制度、技术规范及操作规程。
- f) 采取管理措施和技术手段,保障数据的保密性、完整性、可用性。采取数据分类、重要数据备份、加密认证等措施保证数据安全。

- g) 心音样本数据加密要求如下：对于本地存储模式，应对心音样本中的所有数据进行加密后，存入本地数据库。加密算法包括但不限于国密 SM1、SM2、SM3、SM4、SM7、祖冲之密码、SM9 等，密钥长度不低于 128 位，并且密钥应妥善保存防止泄漏。对于网络传输至服务器存储的模式，应将加密后的结果和当前时间戳及操作者身份信息拼接后再生成数字签名，一并存入服务器数据库。
- h) 严格规范不同等级用户的数据接入和使用权限，并确保数据在授权范围内使用。
- i) 对电子实名认证、数据访问控制、规范数据接入、使用和销毁过程进行痕迹管理，确保对数据访问行为可管、可控。对服务管理全流程留痕，对安全、隐私、风险及事故可查询、可追溯。

附录 A

(规范性)

心音数据的值域标注信息

表A.1~表A.7规定了心音数据的值域标注信息。其中表A.1规定了心音数据的总体标注信息，表A.2~表A.7规定了当相应条件满足时应进一步标注的信息。

表A.1 心音数据的总体标注信息

标注项	标注选项内容	选项内容能否多选	应进一步标注的信息
心音质量	很差（一级）	否	无
	较差（二级）	否	无
	一般（三级）	否	无
	较好（四级）	否	无
	很好（五级）	否	无
	极好（六级）	否	无
心音是否异常	是	否	表A.2 心音异常的标注信息
	否	否	无
先心病具体诊断	正常	否	无
	室间隔缺损	是	无
	房间隔缺损	是	无
	肺动脉瓣狭窄	是	无
	动脉导管未闭	是	无
	复杂先心病	否	无

表A.2 心音异常的标注信息

标注项	标注选项内容	选项内容能否多选	应进一步标注的信息
心音异常类型	第一心音异常	是	表A.3 第一心音异常的标注信息
	第二心音异常	是	表A.4 第二心音异常的标注信息
	存在第三心音	是	无
	存在杂音	是	表A.5 杂音的标注信息

表A.3 第一心音异常的标注信息

标注项	标注选项内容	选项内容能否多选	应进一步标注的信息
第一心音异常类型	第一心音增强	否	无
	第一心音减弱	否	无
	第一心音分裂	否	无

表A.4 第二心音异常的标注信息

标注项	标注选项内容	选项内容能否多选	应进一步标注的信息
第二心音异常类型	第二心音增强	否	无
	第二心音减弱	否	无
	第二心音固定分裂	否	无
	第二心音非固定分裂	否	无

表A.5 杂音的标注信息

标注项	标注选项内容	选项内容能否多选	应进一步标注的信息
杂音类型	收缩期杂音	是	表 A.6 收缩期杂音的标注信息
	舒张期杂音	是	表 A.7 舒张期杂音的标注信息

表A.6 收缩期杂音的标注信息

标注项	标注选项内容	选项内容能否多选	应进一步标注的信息
收缩期杂音类型	收缩早期杂音	否	无
	收缩中期杂音	否	无
	收缩晚期杂音	否	无
	全收缩期杂音	否	无
收缩期杂音响度变化	响度无明显变化	否	无
	响度逐渐增强	否	无
	响度逐渐减弱	否	无
	响度先递增后递减	否	无

表A.7 舒张期杂音的标注信息

标注项	标注选项内容	选项内容能否多选	应进一步标注的信息
舒张期杂音类型	舒张早期杂音	否	无
	舒张中期杂音	否	无
	舒张晚期杂音	否	无
	全舒张期杂音	否	无
舒张期杂音响度变化	响度无明显变化	否	无
	响度逐渐增强	否	无
	响度逐渐减弱	否	无
	响度先递增后递减	否	无