



# 中华人民共和国医药行业标准

XX/T XXXXX—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

Artificial intelligence medical device—Quality requirements and  
evaluation—Part 2: General requirements for datasets

点击此处添加与国际标准一致性程度的标识

(征求意见稿)

XXXX—XX—XX 发布

XXXX—XX—XX 实施

发布

# 目 次

前言 .....	II
1 范围 .....	1
2 规范性引用文件 .....	1
3 术语 .....	1
4 数据集文档要求 .....	5
5 数据集质量要求 .....	8
6 数据集质量符合性评价 .....	11
附录 A（资料性附录） 数据集控制等级 .....	15
附录 B（资料性附录） 数据入排样例 .....	16
附录 C（资料性附录） 数据集质量测度举例 .....	17
参考文献 .....	18
图 1 数据集质量评价流程图 .....	12
表 1 质量特性分类与评价方式 .....	12
表 A.1 数据集控制等级与数据集分类对照表样例 .....	15
表 B.1 骨龄影像评价样表 .....	16
表 C.1 数据集质量测度举例 .....	17

## 前 言

本文件按照GB/T 1.1—2020给出的规则起草。

本文件的附录A、附录B、附录C为资料性附录。

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。

本标准由国家药品监督管理局提出。

本标准由人工智能医疗器械标准化技术归口单位归口。

本标准起草单位：

本标准主要起草人：

# 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 范围

本文件描述了人工智能医疗器械全生命周期使用的数据集的通用质量要求和评价方法，适用于研发、生产、测试、日常质控等环节、对人工智能医疗器械质量有直接影响的数据集开发与评价。

## 2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是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 36344-2018	信息技术 数据质量评价指标
GB/T 34960.5: 2018	信息技术服务 治理 第5部分：数据治理规范
GB/T 35295: 2017	信息技术 大数据 术语
GB/T 25000.12: 2017	系统与软件工程 系统与软件质量要求和评价(SQuaRE) 第12部分：数据质量模型
GB/T 2828.1-2012	计数抽样检验程序 第1部分：按接收质量限(AQL)检索的逐批检验抽样计划
GB T 18391.1-2009	信息技术 元数据注册系统(MDR) 第1部分：框架
GB/T 8054-2008	计量标准型一次抽样检验程序及表
GB/T 10111-2008	随机数的产生及其在产品质量抽样检验中的应用程序
GB/T 11457-2006	信息技术 软件工程术语
YY/T 0287: 2017	医疗器械 质量管理体系 用于法规的要求
YY/T 0316-2016	医疗器械风险管理对医疗器械的应用
WS/T 305-2009	卫生信息数据集元数据规范

## 3 术语

下列术语和定义适用于本文件。

### 3.1

数据 data

信息的可再解释的形式化表示，以适用于通信、解释或处理。

[GB/T 25000.12-2017，定义4.2]

### 3.2

数据质量 data quality

在指定条件下使用时，数据的特性满足明确的和隐含的要求的程度。

[GB/T 25000.12-2017，定义4.3]

## 3.3

数据集 data set

具有一定主题，可以标识并可以被计算机化处理的数据集合。

[WS/T 305-2009, 定义3.1.2]

## 3.4

数据元 data element

用一组属性规定其定义、标识、表示和允许值的数据单元。

[WS/T 305-2009, 定义3.1.6]

## 3.5

数据治理 data governance

数据资源及其应用过程中相关管控活动、绩效和风险管理的集合。

[GB 34960.5-2018, 定义3.1]

## 3.6

元数据 metadata

定义和描述其他数据的数据。

[GB 18391.1-2009, 定义3.2.16]

## 3.7

数据生存周期 data life cycle

数据获取、存储、整合、分析、应用、呈现、归档和销毁等各种生存形态演变的过程。

[GB 34960.5-2018, 定义3.7]

## 3.8

数据质量特性

对数据质量有影响的数据质量属性的类别。

[GB/T 25000.12-2017, 定义4.4]

## 3.9

数据特征层次

从不同粗细粒度表征数据特征的数据层次结构。

[GB/T 35295-2017, 定义2.1.51]

## 3.10

计数检验 inspection by attributes

关于规定的一个或一组要求，或者仅将单位产品划分为合格或不合格，或者仅计算单位产品中不合格数的检验。

[GB/T 2828.1-2012, 定义3.1.3]

## 3.11

计量质量特性 variables quality characteristic  
被检的单位产品特性能用连续尺度进行度量的质量特性。  
[GB/T 8054-2012, 定义3.1.3]

### 3.12

计量抽样检验 sampling inspection by variables  
按规定的抽样方案从批中随机抽取一定数量的单位产品。用测量、试验或其他方法取得它们的质量特性值，与质量要求进行对比，并判断该批产品能否接收的过程。  
[GB/T 8054-2012, 定义3.1.4]

### 3.13

批 lot  
按抽样目的，在基本相同条件下组成的总体的一个确定部分。  
[GB/T 10111-2008, 定义3.1.4]

### 3.14

完整性  
保护数据准确性和完备性的性质  
注：改写GB/T 25000.12-2017, 定义4.12

### 3.15

准确度 accuracy  
对数据内容正确、形式有效的一种度量。  
注：改写GB/T 11457-2006, 定义2.22c

### 3.16

精度 precision  
对于说明的量的精确或差异的程度。例如2位十进制数字对5位十进制数字。  
[GB/T 11457-2006, 定义2.1160]

### 3.17

一致性 consistency  
在数据集的各部分之间，一致、标准化、无矛盾的程度。  
注：改写GB/T11457-2006, 定义2.320

### 3.18

可达性 accessibility  
组成数据集的各部分便于选择使用或维护的程度。  
注：改写GB/T 11457-2006, 定义2.20

### 3.19

可用性 availability  
数据集在投入使用时可操作或可访问的程度。

注：改写GB/T 11457-2006, 定义2.115

### 3.20

安全[性] security

保密[性]

对数据进行的保护，以防止其受到意外的或蓄意的存取、使用、修改、毁坏或泄密。

安全性也涉及对受试者隐私、元数据、通信以及计算机安装的物理保护。

注：改写GB/T 11457-2006, 定义2.1420

### 3.21

可移植性

数据集能从一种硬件或软件环境转换至另外一种环境的特性。

注：改写GB/T 11457-2006, 定义2.1149

### 3.22

数据集制造者

以其名义制造预期用于人工智能医疗器械的数据集并负有设计和/或制造责任的自然人或法人，无论此数据集的设计和/或制造是由该自然人或法人进行或有另外的一个或多个自然人或法人代表其进行。

注：改写YY/T 0287-2017, 定义3.10

### 3.23

数据质量测度

变量，为其赋值作为数据质量特性的测量结果

[GB/T 25000.12-2017, 定义4.5]

### 3.24

缺失数据 missing data

按照研究方案要求收集但未观测到的数据

[药物临床试验数据管理与统计分析的计划和报告指导原则2016版]

### 3.25

离群值 outliers

样本中的一个或几个观测值，它们离其他观测值较远，暗示它们可能来自不同的总体。

[《药学术语》第二版]

### 3.26

参考标准 reference standard

诊断和治疗过程或基于标注过程建立的基准。

注：参考标准可包含疾病、生理状态或生理异常以及位置和程度等信息标签。

### 3.27

无歧义

数据文档描述的语义、语境仅表达一种含义。

## 4 数据集文档要求

### 4.1 数据集描述

#### 4.1.1 数据集整体描述

##### 4.1.1.1 数据集类型

数据集文档应明确数据集的类型，按照预期用途可分为模型训练、模型验证、性能独立测试、临床评价、产品质控等类型；按照数据来源可分为公有数据集、私有数据集；按照用户类型可分为自用数据集、他用数据集；按照访问管理方式可分为开放数据集、封闭数据集；按照更新形式可分为静态数据集和动态数据集。

注1：自用数据集是指数据集制造者为本机构建立的数据集，用于模型训练、模型验证、内部测试、产品质控等。自用数据集在数据自身不提供给其他机构的前提下，可由其他机构间接使用（如联邦学习场景）。

注2：数据集制造者直接提供给其他机构使用的数据集为他用数据集。

注3：预期用于性能独立测试的数据集应确保测试数据与被测系统所用的训练、内部验证数据（包含被测系统通过预训练模型等方式间接使用的数据）不存在交叉；此类数据集仅用于封闭测试，在测试完成后应确保从被测系统和第三方系统中彻底删除，避免针对性调优。

注4：产品质控指的是人工智能医疗器械产品在使用阶段的期间核查、日常质控等用途。

注5：访问管理方式指的是数据集制造者是否允许外部用户访问。数据集制造者内部的访问管理可参考本文件进行设计。

注6：仅在研发早期阶段用于概念验证、对产品和算法模型无直接影响的数据集属于例外情形，制造商可参照本文件的要求，声称所适用的条款。

##### 4.1.1.2 控制等级

数据集文档应描述数据集的控制等级。

注：控制等级与数据集类型、密切相关，数据集制造者可参照附录A中的表A.1进行划分。

##### 4.1.1.3 数据形态

数据集文档应描述数据集的数据模态、数据格式、数据量、存储方式。

##### 4.1.1.4 特征层次

数据集文档应描述数据集的数据特征层次、可能的子集分解和组合变化。

##### 4.1.1.5 临床适用场景

数据集文档应描述数据集的临床适用场景，包括但不限于病种组成与比例、流行病学统计分布、受试者人群分布特征、应用场景等。

##### 4.1.1.6 版本控制

数据集文档应描述数据集的版本号，记录每个版本的创建时间、修改时间、创建人、状态、存储位置；适当时，描述版本号随数据集变化的规则。



#### 4.1.1.7 新增数据

数据集文档应描述数据集是否接收新增数据，以及新数据的入组规则。

#### 4.1.1.8 文档管理

数据集文档宜通过计算机技术（如硬件、普通网络、云服务），对文档内容进行发布、交换、管理和查询。这些数据可以由数值、日期、字符、图像等不同的物理存储格式来实现。

#### 4.1.1.9 元数据属性

数据集文档应描述数据集元数据属性，包括但不限于数据集名称、标识符、数据集发布方、数据集语种、数据集分类类目名称、数据集摘要等。

#### 4.1.1.10 数据元属性

数据集文档宜描述数据元的共用属性、专用属性，包括但不限于数据集版本、注册机构、分类模式、主管机构等。

#### 4.1.1.11 标注信息

数据集文档应描述数据集是否具有标注信息。

### 4.1.2 数据采集信息

#### 4.1.2.1 合规性陈述

数据集文档应提供数据来源的合规性陈述，其中涉及人类遗传资源数据的部分应提供伦理审批、伦理豁免等信息。

注：人类遗传资源材料是指含有人体基因组、基因等遗传物质的器官、组织、细胞等遗传材料。人类遗传资源信息是指利用人类遗传资源材料产生的数据等信息资料。

#### 4.1.2.2 隐私保护

数据集文档应描述用于保护受试者隐私的技术手段，包括但不限于数据脱敏、数据匿名化等。适当时，数据集文档应描述数据脱敏或者数据匿名化的规则。

#### 4.1.2.3 多样性

数据集文档应提供数据来源多样性的描述，包括但不限于受试者人群、采集场所、采集设备、参数设置、临床数据采集人员资质、实验数据采集人员资质、采集流程、采集时间。

#### 4.1.2.4 依从性

数据集文档宜提供数据采集依据的法规、技术标准、临床规范、专家共识或其他参考文献。

#### 4.1.2.5 数据入排

适当时，数据集文档应描述临床数据的入组和排除标准，并对数据的入排情况进行记录。

注：数据入排情况样例见附录B

### 4.1.3 数据预处理

数据集文档应明确描述数据集进行预处理的操作步骤和内容。数据集文档应能区分和追溯原始数据和预处理后的数据。

#### 4.1.4 数据标注信息

##### 4.1.4.1 依从性

数据集如具有标注信息，适当时，文档应描述数据标注依据的法规、技术标准、临床规范、专家共识或其他参考文献。

##### 4.1.4.2 参考标准

数据集如具有标注信息，文档应明确数据集参考标准制定规则、形式、范围、数据格式与数据规范。适当时，数据集制造者应描述参考标准的验证方式。

##### 4.1.4.3 标注流程

数据集如具有标注信息，文档应明确数据标注与质控流程，明确决策机制。适当时，应明确标注人员分歧的仲裁机制。

##### 4.1.4.4 标注信息

数据集如具有标注信息，文档应描述数据集参考标准、标注信息的数据格式、数据规范。

#### 4.1.5 数据存储信息

数据集文档应提供数据集存储方式与存储路径、备份、恢复的说明，如单机存储、普通网络存储、云服务存储。如果数据集制造者使用云服务存储，应提供云服务提供商的名称、资质、访问路径、使用权限说明等。

#### 4.1.6 数据集用户访问

##### 4.1.6.1 访问控制

数据集文档应描述用户访问控制机制，包括但不限于用户类型、权限分配、授权机制。

##### 4.1.6.2 访问条件

数据集文档应描述访问数据集需要的资源、环境、方法、数据接口、协议、工具。

##### 4.1.6.3 可视化

数据集文档宜描述数据集信息的可视化呈现方式。

#### 4.1.7 数据集变更

##### 4.1.7.1 变更规则

适当时，数据集文档应明确数据集更新规则。数据集间以及内部子集间如果发生数据更新与交换，应从控制等级高的集合向等级低的集合单向流动，例如数据可从封闭数据集进入开放数据集，反之应禁止。

##### 4.1.7.2 新增数据

数据集若接受新增数据，数据集文档应记录新增数据的统计描述。

#### 4.1.7.3 数据退役

数据集若允许数据退役，数据集文档应保留记录并单独存储，不可删除。

### 4.2 标识和标示

#### 4.2.1 标识

数据集应**应随附专用文件**，记录数据集的标识，体现数据集名称、版本号、数据集制造者信息，可在数据集文档中进行详细描述。数据集如发生变更，标识应同步变更。数据集制造者应支持数据集标识的创建、修改、删除和查询。

#### 4.2.2 数据集制造者信息

数据集文档宜包含数据集制造者的单位名称和联系方式。适当时，陈述数据集制造者是否对数据集的使用提供技术支持和维护，以及技术支持和维护的具体内容。

#### 4.2.3 文件命名与编码

数据集文档应陈述数据集中的文件命名和编码规则。

### 4.3 数据集制造者资质

数据集文档应描述数据集制造者的数据集质量管理能力。适当时，应声称数据集制造者满足的技术标准、管理标准。

## 5 数据集质量要求

### 5.1 概述

数据集质量要求一方面包括数据本身的质量特性，另一方面包括对数据集整体满足用户需求的程度，与数据来源、数据集开发和管理过程密切相关。当数据集作为产品提供给其他用户时，还需要关注数据集制造者提供服务的能力。当数据集作为产品提供给用户，或在监管活动中提供证据时，第三方或质量审核人员宜根据数据集的预期用途、临床适用场景确定质量特性具体指标，对数据集开展评价，形成技术报告，用于数据集的验证。适当时，数据集制造者可在数据集文档中列出描述数据集质量特性的客观指标。

### 5.2 质量测度

#### 5.2.1 完整性

##### 5.2.1.1 准确性

数据与真值相符合的程度应符合数据集制造者声称的准确性，例如：

- a) 记录临床观测结果的准确性；
- b) 记录实验室检验结果的准确性；
- c) 记录临床结论或参考标准的准确性；
- d) 记录其他相关信息的准确性；
- e) 文字描述应准确、清晰、无歧义；

- f) 影像及其检查报告的准确性；
- g) 病理检查报告的准确性；
- h) 采集设备的有效性；
- i) 人员操作的有效性；
- j) 数据形式的合理性。

注：附录C给出了各个质量测度的示例。

### 5.2.1.2 完备性

数据集应包含实现数据集预期用途、满足临床适用场景需要的信息，例如：

- a) 数据结构与格式应支持数据元、标注信息、元数据的保存、传输与访问，避免信息空白、遗漏或丢失；
- b) 对不可避免的缺失数据、离群值，应给出明确的处理方式，证明临床场景的合理性，对相关数据元加以标识；
- c) 元数据应描述数据集标识信息、内容信息、数据质量特性等相关信息；
- d) 数据包含的信息应支持数据标注与参考标准溯源的需要。

### 5.2.2 唯一性

同一数据集内的数据元应是唯一的。同一数据集内的数据唯一性宜以受试者个体为识别特征，对同一个体的不同时间、空间的数据应明确描述处理方法，并证明临床场景的合理性，记录之间的关联性。

### 5.2.3 一致性

数据集的内部一致的程度和外部一致的程度均应符合数据集制造者声称的一致性指标。

内部一致性指的是相同来源数据之间的相关性，如：

- a) 同一数据元在数据集生命周期不同阶段应保持一致；
- b) 数据的特征应保持一致；
- c) 数据的采集、预处理、标注方法应保持一致；
- d) 原始记录、中间记录与最终记录应保持一致。

外部一致性指的是不同来源数据之间的相关性，如：

- a) 不同来源的数据应在数据特征方面保持一致；
- b) 离群值应是可解释的；
- c) 不同来源的数据在采集、标注环节应依从相同的法规、技术标准、医学规范或其他文献要求。

### 5.2.4 确实性

数据、元数据应是真实和可信的，如：

- a) 遗传资源应来自真实的临床数据采集流程；适当时，数据采集涉及的设备、人员、方法应符合技术标准、临床规范或专家共识；
- b) 非遗传资源应来自可溯源的实验、模型、数据扩增等活动；
- c) 元数据应如实对数据进行描述。

### 5.2.5 时效性

数据的采集、标注、流转、归档、变更等活动应遵守数据集制造者声称的时限。动态数据集应明确数据的更新周期、更新方法和更新比例。若数据涉及临床诊断、治疗中的动态时序过程，应证明数据在临床时效上的合理性。

### 5.2.6 可访问性

数据集制造者应明确数据集访问控制等级，适当时，制定明确的访问控制策略。

数据集可被访问的程度应满足数据集的预期用途和临床适用范围，可支持多种灵活的访问方式，包括本地访问和远程访问方式，非必要情况下不应扩大访问权限。

### 5.2.7 依从性

数据格式、数据接口、数据加密、隐私保护机制应符合数据集制造者声称的技术标准或规范；

适当时，数据标注过程和参考标准的定义应依从明确的标准规范、专家共识、操作规程或其他参考文献。

### 5.2.8 保密性

数据集制造者应制定明确的信息安全策略，对数据集的保密性负责。

封闭使用的数据集应具有数据集授权访问机制、隔离保护机制，开发与使用过程应防止数据泄露、数据篡改、数据丢失，以文件形式记录脱敏、网络安全、物理隔离、数据审计等关键措施。

### 5.2.9 效率

在数据集制造者规定的运行环境下，处理、使用数据集的效率宜满足数据集制造者声称的水平。

### 5.2.10 精度

适当时，数据定量特征、数据集定量特征、数据标注结果的精度应满足数据集制造者声称的水平。

### 5.2.11 可追溯性

在数据集开发、管理、使用过程中，数据集制造者宜保证数据访问踪迹和数据变更踪迹的可审计性，例如：

- 1) 原始数据来源、元数据来源、合规性证明；
- 2) 数据采集活动记录；
- 3) 人员管理记录；
- 4) 数据标注流程记录；
- 5) 盲态管理记录；
- 6) 数据流通记录；
- 7) 数据质疑、审计、停用、纠正记录；
- 8) 标注工具、平台使用记录；
- 9) 数据集标注结果的统计信息查询，包括标注进展、标签统计、标注者进展统计、难例集等；
- 10) 数据服务异常与故障记录；
- 11) 数据维护与备份记录；
- 12) 数据更新记录。
- 13) 云服务提供商名称、联系方式、云服务类型（公有云、私有云）。

### 5.2.12 可理解性

适当时，数据集能被授权用户预览和解释的程度应符合数据集制造者声称的水平。

### 5.2.13 可用性

适当时，数据集能被授权用户访问和检索的程度应符合数据集制造者声称的水平。

#### 5.2.14 可移植性

数据能被存储、替换或从一个系统转移到另一个系统并保持已有质量的程度应符合数据集制造者声称的水平。

#### 5.2.15 可恢复性

适当时，数据集在使用过程中保持质量并抵御失效事件的程度应符合数据集制造者声称的水平。

#### 5.2.16 代表性

数据集的数据特征层次、流行病学统计、样本来源多样性、数据多样性等应符合数据集制造者声称的临床适用场景。

### 5.3 数据集风险分析

#### 5.3.1 选择偏倚

数据集制造者应对数据集的代表性（受试者人群、使用者、设备选择、参数设置等能否代表临床实际）进行分析。

#### 5.3.2 覆盖偏倚

数据集制造者应对数据集病种组成能否覆盖临床适用场景下的目标人群全部情形（单发、多发、并发、复杂病理等）进行分析。

#### 5.3.3 参考标准偏倚

数据集制造者应验证数据标注结果或参考标准的确立过程，分析参考标准偏倚。

#### 5.3.4 验证偏倚

数据集制造者应描述参考标准的验证方法；分析数据标注过程与临床诊断过程的关联、差异和影响。

#### 5.3.5 标注顺序偏倚

数据集制造者应描述数据标注流程设计，分析标注者分工、标注者工作顺序、记忆效应对标注结果的影响。标注过程如使用计算机辅助识别（CAD）工具，应评估由CAD引起的偏倚情况，给出预防措施，避免CAD对人的干扰。

#### 5.3.6 情景偏倚

数据集制造者应描述数据集患病率与临床目标人群患病率的差异，提供补偿、纠正方案。

## 6 数据集质量符合性评价

### 6.1 总体原则

数据集质量符合性评价包括对数据集文档、质量特性和数据集整体的评价，评价流程如图1所示。适当时，数据集制造者应提供数据集、原始数据、元数据、标注工具、存储介质和其他工具的访问权限。各质量特性的评价方式与抽样方式可参照表1的内容。

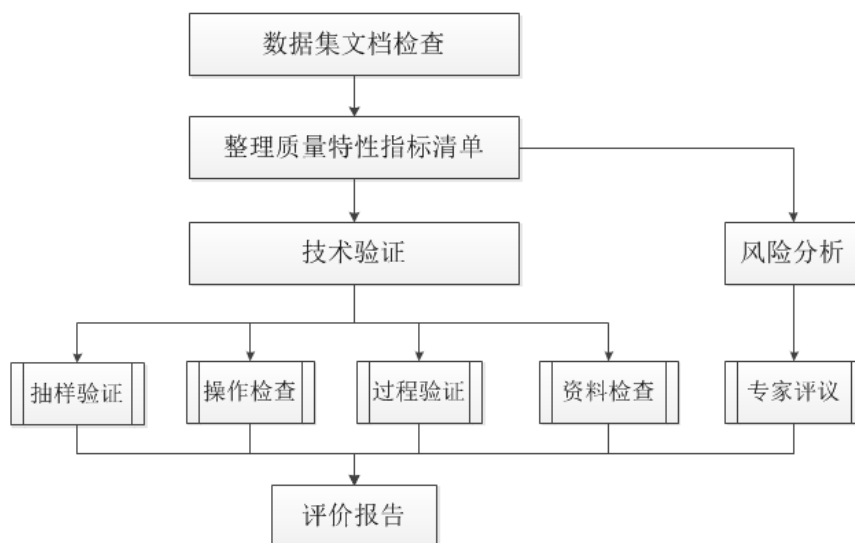


图1 数据集质量评价流程图

表1 质量特性分类与评价方式

质量特性	评价方式	抽样方式
准确性	试验验证、过程验证	计量型抽样
完备性	试验验证+资料检查	计数型抽样
精度	试验验证	计数型抽样
唯一性	查重	计数型抽样
一致性	试验验证	计量型抽样
确实性	检查	计数型抽样
时效性	试验验证	计数型抽样
可访问性	试验验证	计数型抽样
依从性	资料检查	/
保密性	资料检查	/
效率	试验验证	/
可追溯性	资料检查	计数型抽样
可理解性	操作检查	/
可用性	操作检查	/
可移植性	操作检查	/
可恢复性	检查	/
代表性	统计分析	/

数据集质量验证可按照特征采用全部验证或抽样验证方案。

抽样验证方案应根据具体质量特性指标制定。属性数据宜采用计数型抽样检验，数据集制造者应声称接收质量限、使用方风险、生产方风险等指标，具体方法见GB/T 2828.1-2012；变量数据宜采用计量型抽样检验，数据集制造者应声称极限质量、接收常数、使用方风险、生产方风险等指标，具体方法见GB/T 8054-2008。

## 6.2 质量特性评价

使用全部数据或者6.1的方法得到抽样样本集，作为技术验证的对象。

### 6.2.1 准确性

根据5.2.1.1中适用的情形，计算抽样样本中被认为是准确的样本比率，应符合5.2.1.1的要求。

注：如果数据集参考标准由人工标注确定，标注结果的准确率评估可与临床金标准进行比对。当临床金标准难以获取时，可与第三方仲裁专家组的标注结果进行对照。在此情形下，应明确第三方仲裁专家组的资质、选拔要求与仲裁机制，记录仲裁过程与仲裁的频次，形成仲裁报告。

### 6.2.2 完备性

对照数据集文档中的描述，编写测试用例，对抽样样本集的数据、元数据和标注内容进行检查，应符合5.2.1.2的要求；

### 6.2.3 唯一性

对抽样样本集进行查重验证，结果应符合5.2.2的要求。

### 6.2.4 一致性

定量比对抽样样本集在数据集生命周期各个环节之间的变化情况，计算内部一致性；定量比较不同来源的样本数据特征的一致性，计算外部一致性。指标可选用Pearson相关系数、Kappa系数等，由数据集制造者确定，应符合5.2.3的要求。

### 6.2.5 确实性

对照数据集文档中的描述，对抽样样本集的数据、元数据、头文件和标注内容进行检查，应符合5.2.4的要求；

### 6.2.6 时效性

提取和检查抽样样本集的时间信息，应符合5.2.5的要求。

### 6.2.7 可访问性

根据数据集文档，编写测试用例，对抽样样本集进行实际操作，应符合5.2.6的要求。

### 6.2.8 依从性

编写测试用例，检查抽样样本集与相关法规、技术标准、技术规范的符合性，应符合5.2.7的要求。

### 6.2.9 保密性

编写测试用例，检查抽样样本集的授权访问机制、数据的隔离保护机制、网络安全工具，应符合5.2.8的要求。

### 6.2.10 效率

编写测试用例，在数据集制造者规定的运行环境下，读取、传输抽样样本集各三次，计算单个样本的平均处理时间，应符合5.2.9的要求。

### 6.2.11 精度



对数据集标注工具进行过程验证，检查抽样样本集中的数据、元数据、标注结果包含的数据定量特征、数据集定量描述的精度，应当满足5.2.10的要求。

#### 6.2.12 可追溯性

对数据集的开发、管理、使用过程记录进行检查，应符合5.2.11的要求。

#### 6.2.13 可理解性

对照数据集文档，对抽样样本集进行预览和解释，应符合5.2.12的要求。

#### 6.2.14 可用性

对照数据集文档，对抽样样本集进行访问和检索，应符合5.2.13的要求。

#### 6.2.15 可移植性

对照数据集文档，把抽样样本集从一个系统移植到另一个系统，计算准确率，应符合5.2.14的要求。

#### 6.2.16 可恢复性

对照数据集文档，编写测试用例，模拟抽样样本集在使用过程中的失效事件，应符合5.2.15的要求。

#### 6.2.17 代表性

对照数据集文档，对整个数据集开展统计分析，检查统计分层、样本组成、数据多样性等，应符合5.2.16的要求。

#### 6.2.18 数据集风险评价

根据数据集统计数据和适用的技术标准、医学规范，参照YY/T 0316-2016开展数据集的风险分析，至少包括选择偏倚、覆盖偏倚、参考标准偏倚、验证偏倚、标注顺序偏倚、情景偏倚，形成研究资料，应符合5.3的要求。

注：可采用专家评议法，对数据集的整体质量、偏倚问题进行问卷评估。

附 录 A  
(资料性附录)  
数据集控制等级

表A.1 数据集控制等级与数据集分类对照表样例

等级	预期用途	用户类型	访问管理类型	管理要求	访问控制要求	存储要求	追溯要求
1	模型训练、模型调优	自用	封闭	对采集、预处理、标注、存储、更新进行管理	由数据集制造者定义	网络存储、本地存储	符合 5.2.11
2	模型训练、模型调优	他用	开放、封闭			网络存储、本地存储	
3	产品日常质控	自用	封闭		访问者需要授权	专用介质存储	
4	产品日常质控	他用	封闭				
5	性能独立测试	自用、他用	封闭	除上述要求外，数据来源应可溯源	访问者需要获得授权；访问活动应形成记录；必要时进行数据审计	独立存储，物理隔离；必要时进行加密	符合 5.2.11；数据来源应溯源至采集机构
	临床评价	自用	封闭	除上述要求外，应符合《医疗器械临床试验质量管理规范》		独立存储，物理隔离；必要时进行加密	符合 5.2.11；数据来源应溯源至采集机构，必要时可通过采集机构进一步追溯至受试者

注1：1级为最低级，5级为最高级；高等级数据集的数据退役后可流入低等级数据集，不允许从低等级流入高等级。

**附 录 B**  
(资料性附录)  
数据入排样例

临床样本入排条件样例（儿童骨龄辅助诊断数据集）

1、入选标准

- ① 年龄 0.5-18 周岁（含 0.5 岁到 17 岁 364 天）的骨龄片，性别不限
- ② 如为前瞻性试验，需要受试者及其法定代理人或法定代理人自愿参加，并签署知情同意书

2、排除标准

- ①X 光片显示受试者手腕部有严重的骨折病史，例如指骨、桡骨、尺骨破裂；
- ③ X 光片显示受试者手腕部有金属植入物；
- ④ X 光片不清晰，骨骼边界无法辨别，有伪影；
- ⑤ 研究者认为不适合入组的 X 光片：如由于锥形骨骺，软骨瘤等；
- ⑥ 不符合影像质量评价要求。

3、剔除标准

- ① 金标准专家组无法识别判断；
- ② 病例信息记录不完整，结果无法评价；
- ③ 不符合入选标准/符合排除标准的病例。

骨龄影像质量评价要求如表 B.1 所示

表B.1 骨龄影像评价样表

评价项目	是	否
1. 图像中无其他干扰物，例如其他手在按压；	<input type="checkbox"/>	<input type="checkbox"/>
2. 五指张开，不存在部分手指重叠现象；	<input type="checkbox"/>	<input type="checkbox"/>
3. 手部无遮挡；	<input type="checkbox"/>	<input type="checkbox"/>
4. 手占图片面积的 30%；	<input type="checkbox"/>	<input type="checkbox"/>
5. 左手正位片的拍摄标准：		
（1）拍摄对象：拍摄左手手腕部正位 X 光片，除拍摄手部诸骨外，还应包括尺、桡骨远端骨干 2-3cm；	<input type="checkbox"/>	<input type="checkbox"/>
（2）拍摄体位：手掌面向下，紧贴暗盒，中轴与前臂轴成直线，五指自然分开，拇指与手掌呈约 30 度角；	<input type="checkbox"/>	<input type="checkbox"/>
（3）投照中心：X 线投照中心垂直正对第三掌骨头。	<input type="checkbox"/>	<input type="checkbox"/>
<b>注：以上评价项目若有一项选择“否”，则该受试者不入组。</b>		

附 录 C  
(资料性附录)  
数据集质量测度举例

表C.1 数据集质量测度举例

质量测度	举例说明
准确性	1、医学影像标注结果与受试者的临床检查报告一致 2、文本信息的编码、格式符合要求
完备性	1、医学影像数据DICOM字段信息包含设备型号、厂家、成像参数等信息，可以支持数据集的细化分层 2、文本数据包含完备的实体、关系信息
精度	1、医学影像标注时，病灶尺寸的测量精度达到1mm以下 2、眼底相机CCD采用的A/D转换器为16位
唯一性	1、数据集中的病例是唯一的 2、不同医学影像序列之间不存在重复的图像
一致性	1、内部一致性：同一受试者的心电数据在采集、标注、存储、使用等各个阶段保持不变 2、外部一致性：来自不同医院的心电数据具有一致的采样率
确实性	1、糖尿病视网膜数据集使用真实的受试者眼底图像，排除动物眼底图像、眼前节图像等情形。 2、心电数据的心率、波形符合临床认知，排除异常干扰和漂移。
时效性	单个影像序列从采集到完成标注的时间间隔不大于7天。
可访问性	数据集可被授权用户预览、复制。
依从性	医学影像数据格式满足Dicom3.0标准。
保密性	1、封闭管理的数据集应具有物理隔离机制。 2、数据集的远程传输采用加密算法。
效率	在推荐的软硬件配置下，读取一个影像序列的时间不超过30秒。
可追溯性	1、数据的收集、清洗、预处理、入库形成原始记录并归档。 2、标注工具的使用时间、人员、事件形成原始记录并归档。
可理解性	1、心电异常事件的标注结果可以直观地显示在心电图上 2、心电异常事件的各个分类具有明确的含义
可用性	数据集可以被算法模型读取、分析，可以建立索引并被分层抽取
可移植性	数据集可以在Windows、Linux等不同操作系统下使用
可恢复性	数据集的使用与管理过程采用备份、容灾机制，可以从故障中恢复
代表性	1、糖尿病视网膜数据集中的阳性样本比例能够符合我国糖网流行病学统计 2、采集糖网图像的眼底相机能够覆盖临床常用机型

## 参 考 文 献

- [1] ISO 19157-2013 Geographic Information-Data quality
- [2] ISO 3951-1:2013 Sampling procedures for inspection by variables — Part 1: Specification for single sampling plans indexed by acceptance quality limit (AQL) for lot-by-lot inspection for a single quality characteristic and a single AQL
- [3] Mats Bergdahl, Manfred Ehling, Eva Elvers, et al. Handbook on Data Quality Assessment Methods and Tools[M]. EUROPEAN COMMISSION, 2007.
- [4] DAMA UK. THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT[R]. DAMA UK, 2013
- [5] Michael G. Kahn, Marsha A. Raebel, Jason M. Glanz, et al. A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research[J]. Med Care. 2012 July ; 50(0): . doi:10.1097/MLR.0b013e318257dd67.
- [6] Nicole Gray Weiskopf, Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research[J]. J Am Med Inform Assoc 2013;20:144–151. doi:10.1136/amiajnl-2011-000681
- [7] Lee K, Weiskopf NG, Pathak J, et al. A Framework for Data Quality Assessment in Clinical Research Datasets[C]. american medical informatics association annual symposium, 2017: 1080–1089.
- [8] David Boone, Suzanne Cloutier, Sergio Lins. Measuring the Quality of HIV/AIDS Client-level Data Using Lot Quality Assurance Sampling(LQAS) [R]. MEASURE Evaluation, University of North Carolina at Chapel Hill, 2019.
- [9] 国家食品药品监督管理总局. 总局关于发布临床试验数据管理工作技术指南的通告(2016年第112号). 2016.
- [10] 国家药品监督管理局医疗器械技术审评中心. 关于发布深度学习辅助决策医疗器械软件审评要点的通告(2019年第7号), 2019.
- [11] Organisation for Economic Co-operation and Development. OECD Recommendation of the Council on Health Data Governance, OECD/LEGAL/0433》, 2017
- [12] World Health Organization. Data quality review: a toolkit for facility data quality assessment. World Health Organization, 2017