ICS 11.020
CCS C 07

团　　体　　标　　准

T/LXLY 31—2024

多模态老年共病数据集构建方法

Construction method of multimodal elderly comorbidity dataset

2024-12-30 发布　　2024-12-30 实施

中国老年学和老年医学学会　发布
中　国　标　准　出　版　社　出版

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专利的责任。

本文件由首都医科大学附属北京友谊医院提出。

本文件由中国老年学和老年医学学会归口。

本文件起草单位：首都医科大学附属北京友谊医院、中国科学院自动化研究所、中国医学科学院阜外医院、郑州大学第一附属医院、北京市朝阳区疾病预防控制中心、华东师范大学、广州互云医院管理有限公司。

本文件主要起草人：李虹伟、汤雯、杨雪冰、陈炜、孙颖、马丽红、田蕊、李书明、张志忠、李陇豫、万端畅、王学栋、李明达、潘博、刘冰。

多模态老年共病数据集构建方法

1 范围

本文件规定了多模态老年共病数据集的构建流程、构建方法、质量评估、安全和隐私。

本文件适用于医疗机构、医疗数据科研机构和应用机构等相关机构和个人对多模态老年共病数据集的构建、研究、应用和质量控制。

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 5271.31 信息技术 词汇 第31部分：人工智能 机器学习

GB/T 22239 信息安全技术 网络安全等级保护基本要求

GB/T 25069 信息安全技术 术语

GB/T 35273 信息安全技术 个人信息安全规范

GB/T 39725 信息安全技术 健康医疗数据安全指南

GB/T 42755 人工智能 面向机器学习的数据标注规程

3 术语和定义

GB/T 5271.31、GB/T 25069、GB/T 39725界定的以及下列术语和定义适用于本文件。

3.1

多模态 multimodal

多种类型(模态)的数据或信息源。

注：多种类型如图像、文本、音频、视频等。

3.2

老年共病 elderly comorbidity

2种或2种以上慢性健康问题同时发生在一个老年人个体，影响老年人个体健康状况持续1年及以上的情况。

注：慢性健康问题可能是脏器疾病、精神心理问题、老年综合征，也能是其他影响老年人健康的问题。

3.3

多模态老年共病数据集 multimodal elderly comorbidity dataset

从不同的老年共病医疗数据源采集到的多种类型(模态)的数据集。

示例：图像、文本、生理参数。

3.4

时间序列数据 time series data

在不同时间收集到的数据，反映某一事物、现象等随时间的变化状态或程度。

示例：患者的生命体征数据。

4 缩略语

下列缩略语适用于本文件。

ATC：解剖学治疗学及化学分类系统（Anatomical Therapeutic Chemical）

CT：计算机断层扫描（Computed Tomography）

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

NYHA：美国纽约心脏病学会（New York Heart Association）

PCA：主成分分析（Principal Components Analysis）

5 构建流程

多模态老年共病数据集构建流程包括数据收集、数据预处理、数据标注、数据整合、特征提取、特征构建、特征融合、导出与存档，构建流程见图 1。

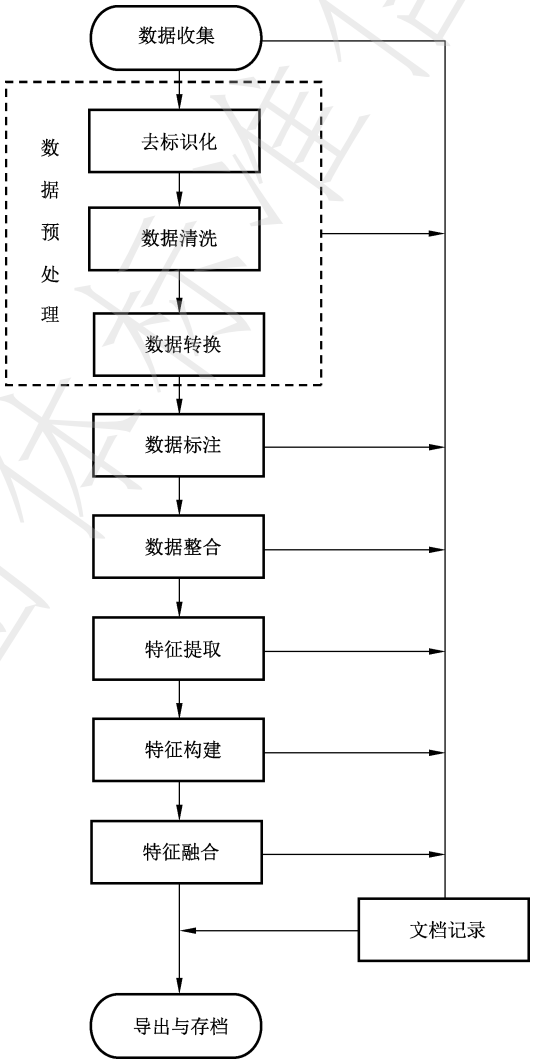

图 1 多模态老年共病数据集构建流程图

6 构建方法

6.1 数据收集

6.1.1 用于构建多模态老年共病数据集的数据来源可为各级医院电子信息平台、自建数据采集平台、中长期照护机构数据平台、智能养老设备等。数据收集应按相关规定，遵循公开、透明、合法、正当的原则进行收集。

6.1.2 数据为结构化数据和非结构化数据类型。数据包括但不限于表 1 规定的类型。

表 1 数据类型

数据类型	范围
基线资料	性别、年龄、身高、体重、诊断、手术史、用药史等
生命体征	常规生理检查数据，包括体温、心率、血糖、血压、血氧等
目前症状和体征	意识状态、临床体征、临床症状等
老年综合评估	体力状态评估、日常生活能力评估、衰弱评估、营养评估、认知评估、疼痛评估、跌倒风险评估、视听力障碍评估等
实验室检查	血尿便常规、肝肾功能、电解质、血脂、血糖、糖化血红蛋白等
特殊操作或治疗	冠脉造影、动脉内球囊反搏、起搏器植入、呼吸机辅助通气、穿刺置管等
康复方案	饮食模式、运动频率、运动时间、运动强度等
辅助检查	肺功能、骨密度检查等
新发不良事件	肺炎、呼吸衰竭、急性冠脉综合征、骨折、跌倒、消化道出血、再入院等
药物治疗	处方药及保健药品等
疾病诊断	—
其他模态	心电图、胸部 CT、腹部超声、超声心动图等

6.2 数据预处理

6.2.1 去标识化

使用哈希算法为老年患者生成唯一标识，对老年患者个人信息进行保护。

6.2.2 数据清洗

数据清洗遵循完整性、合法性、一致性、唯一性、权威性的原则，处理方式包括以下内容。

a） 缺失值处理：

1） 使用插补（均值、中位数等）或删除缺失值较多的样本；

注：如生命体征、实验室检查数据等关键变量。

2） 人工核实后补充。

b） 异常值处理：

1） 逻辑错误：变量之间不符合逻辑关系，核实后修订，且保留修订记录；

示例：出院时间早于入院时间。

2） 自然离群值：采用稳健统计方法进行分析；

3） 人为离群值：人工核实后纠正。

c） 重复值处理：

1） 重复数据结合唯一标识核对并清理，删除完全重复的数据；

2） 由于内部标识编码重复导致数据重复时，人工核实后纠正。

6.2.3 数据转换

将非结构化数据（如文本或图像）转换为结构化格式，并统一格式。

6.3 数据标注

6.3.1 对数据进行标注，包括：

a） 确定老年共病患病情况；

b） 老年共病相关的临床指标、症状及药物使用情况；

c） 老年共病患者不良事件发生情况。

6.3.2 数据标注按 GB/T 42755 的规定执行。

6.4 数据整合

6.4.1 通过哈希算法为老年患者生成的唯一标识字符串，将不同来源的数据进行融合，整合多模态数据为统一的数据表。

6.4.2 整合后的数据应符合以下要求：

a） 数据字段命名规则一致；

b） 数值型数据单位一致；

c） 编码规则一致；

d） 日期时间格式一致，按时间轴对齐。

6.5 特征提取

6.5.1 根据不同类型数据，选择表 2 规定的方法进行特征提取，将特征值转化为由数字和编码组成的混合矩阵。

表 2 数据特征提取方法

数据类型		处理方法	说明
数值型	单一型	直接提取	年龄、血压、实验室检查数据等
	范围型	取均值	
文本型	单一实体描述型	药物→ATC 编码 诊断、症状→ICD 编码	诊断、手术史、用药史等
	固定类别型	固定文本选择题→Onehot 编码	
	多实体描述型	NER 模型→诊断文本和手术文本→诊断 ICD 和手术 ICD 编码	
分级/分类型	分类型	Onehot 编码	性别等
	分级型	根据级别总数分配级别，级别对应指标程度	NYHA 心功能分级等

6.5.2 时间序列数据对应患者编号，单独整理为表格。

6.6 特征构建

6.6.1 为老年共病患者构建特征，包括但不限于：

a) 基础疾病信息；
b) 基础生命体征；
c) 基础用药信息；
d) 基础老年综合征情况；
e) 不良事件发生情况。

6.6.2 检查特征之间的相关性，去除冗余或高相关的特征，减少多重共线性。使用信息增益或特征重要性等指标评估特征对目标变量的贡献，剔除不重要或噪声特征。

6.6.3 对所有特征进行处理，确保不同模态的特征在同一尺度上。

6.7 特征融合

6.7.1 从不同模态中提取的特征进行整合，形成一个统一的高维特征矩阵。通过对各模态特征的规范化和对齐处理，从不同模态中提取的特征连接形成一个完整的特征表格，在该表格中，按行排列样本，按列组织多模态特征。

6.7.2 为进一步优化特征的表达和利用，特征融合方法宜与特征提取方法相结合，以剔除冗余信息，提升特征表达的有效性。特征融合示意见图 2。

注：包括但不限于 PCA、最大相关最小冗余算法、自动解码器、Feature Map 拼接、加权融合等方法。

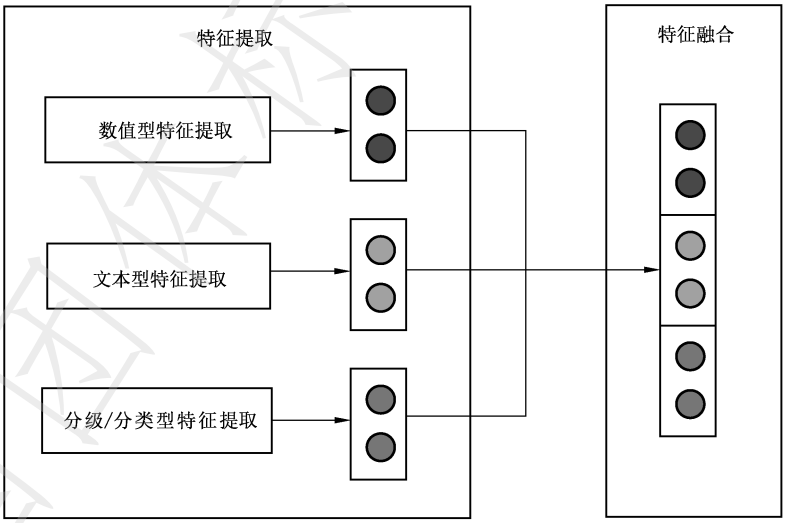

图 2 特征融合示意图

6.7.3 特征融合后应检查特征数量的一致性。

注：特征选择技术如 L1 正则化、决策树、基于模型的选择方法等。

6.8 导出与存档

选择多模态老年共病数据集存储格式，与构建过程中形成的文档记录一并导出并存档，文档记录应包括以下内容。

a) 多模态老年共病数据集概述：
   1) 多模态老年共病数据集名称；

2） 版本信息：记录多模态老年共病数据集的版本号和发布日期；
3） 描述：概述多模态老年共病数据集的目的、应用场景和目标用户，如机器学习；
4） 数据来源：公开数据、调研、收集方式等；
5） 模态种类：多模态老年共病数据集中包含的模态类型，如图像、文本、音频。

b） 数据描述：
1） 数据结构：数据的具体组织方式，如文件夹结构、每个数据文件的内容描述、标注格式等；
2） 样本数量：每种模态中的数据样本数量；
3） 数据格式：每种模态的数据格式，如图像为 JPEG/PNG，文本为 TXT/CSV，音频为 WAV/MP3；
4） 模态配对：说明各模态数据之间的关联方式，如图像与文本的对应关系、时间对齐等。

c） 数据来源：数据的来源，如哪些来自各级医院电子信息平台、自建数据采集平台、中长期照护机构数据平台、智能养老设备。

d） 数据预处理：
1） 预处理步骤：详细记录对原始数据进行的预处理操作；
2） 数据清洗：去除不合格数据样本的具体操作；
3） 数据转换：数据转换的具体操作，格式统一的要求；
4） 对齐和同步：跨模态数据的对齐方式，如不同模态之间的时间同步、空间对齐等；
5） 预处理后的数据。

e） 数据标注：
1） 数据标注的过程：若多模态老年共病数据集包含标注信息，需说明标注过程、标注标准、标注工具、标注员培训情况等；
2） 标注后的数据。

f） 数据整合：
1） 数据整合的方法，一致性检查的操作和要求；
2） 整合后的数据。

g） 特征提取：
1） 特征提取的方法、操作；
2） 特征提取后的数据。

h） 特征构建：
1） 特征的选择、检查、处理方法；
2） 特征构建后的数据。

i） 特征融合：特征融合的方法、冗余信息剔除的方法、特征数量一致性的检查方法。

7 质量评估

7.1 评估原则

多模态老年共病数据集质量评估遵循以下原则：

a） 科学性：反映老年共病状态下的复杂状况及对于机器学习应用性能的影响；
b） 客观性：评估符合实际、客观可信、过程可监控；
c） 系统性：在选择评估指标时考虑指标的系统性和层级关系；
d） 针对性：考虑机器学习应用的需求，在指标的权重和分值上予以区分，体现质量评估对机器学习应用的针对性和导向作用；
e） 引导性：以获取有利于机器学习应用的信息资源为导向。

7.2 评估方法

7.2.1 定性评价法

7.2.1.1 根据评估目的、老年共病和机器学习应用的需求，从主观的角度对多模态老年共病数据集质量进行描述与评估，评估结果可以等级制、百分制或布尔表示。

7.2.1.2 定性评价法的选择：

a) 针对临床数据的合适性和有效性评价，如是否包含了必要的临床特征（如年龄、性别、手术史、用药史等），数据是否符合老年共病的研究需求；

b) 影像数据的分辨率、图像的清晰度、是否存在噪声等。

7.2.2 定量评价法

7.2.2.1 采用确定的量化公式或计算方法作为评估准则，提供客观、直观和具体的结果。可采用数据集质量检测软件检查数据质量，也可通过辅助工具结合人工识别分析方法进行人工检查，一般可分为：

a) 全数检查：针对行业强制要求、特殊要求、其他可能导致严重影响的项目进行；

b) 抽样检查：针对质量比较稳定、数据量较大、检查费用与时间有限的情况进行。

7.2.2.2 定量评价的项目和指标要求见表 3。

表 3 定量评价要求

项 目	指 标
缺失值	≤5%
数据误差	≥97%
一致性冲突率	≤1%
重复数据	无
数据验证规则合格率	≥99%
数据结构化程度	≥90%
合规性	100%

7.2.3 综合方法

将定性和定量两种方法有机地集合起来，从客观和主观两个方面对多模态老年共病数据集质量进行评估。

8 安全和隐私

数据安全和隐私保护应符合以下要求：

a) 符合 GB/T 22239 对数据应用安全的相关要求；

b) 符合 GB/T 35273 对个人信息安全的相关要求；

c) 对传输的数据进行加密；

d) 数据传输过程中使用安全协议；

e) 多模态老年共病数据集构建过程中，使用匿名化或去标识化技术处理个人身份信息；

f) 对数据采集环境、设施和技术采取必要的安全管控措施；

g） 对敏感数据进行加密存储，防止未经授权的访问和数据泄露；
h） 实施访问控制措施，仅允许授权人员访问多模态老年共病数据集；
i） 记录对多模态老年共病数据集的访问和修改行为；
j） 制定数据泄露响应计划；
k） 定期备份。

参 考 文 献

[1] WS 371 基本信息基本数据集 个人信息
[2] WS 372.4 疾病管理基本数据集 第4部分:老年人健康管理
[3] WS 373.1 医疗服务基本数据集 第1部分:门诊摘要
[4] WS 373.2 医疗服务基本数据集 第2部分:住院摘要
[5] WS 445(所有部分) 电子病历基本数据集
[6] WS 538 医学数字影像通信基本数据集
[7] WS 539 远程医疗信息基本数据集