团 体 标 准

2型糖尿病专病数据集标注与质控规范

（框架草案）

Standards for data annotation and quality control of type 2 Diabetes Mellitus

|  |
| --- |
|  |
|  |

发布

 实施

四川省生物信息学学会  发布

版权保护文件

版权所有归属于该标准的发布机构，除非有其他规定，否则未经许可，此发行物及其章节不得以其他形式或任何手段进行复制、再版或使用，包括电子版，影印件，或发布在互联网及内部网络等。使用许可可于发布机构获取。

目次

[前 言 3](#_Toc137798327)

[引 言 4](#_Toc137798328)

[1 范围 5](#_Toc137798329)

[2 规范性引用文件 5](#_Toc137798330)

[3 术语和定义 5](#_Toc137798331)

[4 数据标注规范 7](#_Toc137798332)

[4.1人员管理 7](#_Toc137798333)

[4.1.1标准人员 7](#_Toc137798334)

[4.1.2培训 7](#_Toc137798335)

[4.1.3考核 7](#_Toc137798336)

[4.2基础设施管理 8](#_Toc137798337)

[4.2.1标注场所 8](#_Toc137798338)

[4.2.2标注软件 8](#_Toc137798339)

[4.3 标注方法 8](#_Toc137798340)

[4.4 数据标注流程 9](#_Toc137798341)

[5 数据质控规范 9](#_Toc137798342)

[5.1 人员 10](#_Toc137798343)

[5.2 方法 10](#_Toc137798344)

[5.3 指标 10](#_Toc137798345)

[5.4 通过准则 10](#_Toc137798346)

[5.5 数据验证方法 10](#_Toc137798347)

[2型糖尿病专病数据集标注与质控规范编制说明 12](#_Toc137798348)

[一、任务来源及计划要求 12](#_Toc137798349)

[二、编制过程 12](#_Toc137798350)

[三、调研和分析工作情况 13](#_Toc137798351)

[四、主要技术内容的说明 14](#_Toc137798352)

前 言

本标准按照GB/T 1.1-2020《标准化工作导则 第一部分：标准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。

本文件由四川省生物信息学学会提出。

本文件由四川省生物信息学学会知识产权与标准化工作委员会归口。

本标准起草单位：四川大学华西医院

本标准主要起草人：任艳

引 言

糖尿病是一种以慢性高血糖为特征的代谢性、进展性疾病，严重威胁人类健康。近30多年来，我国糖尿病患病率显著增加，我国一项大型流行病学研究提示，2015至2017年中国糖尿病患病率高达11.2%，患病人口数量为世界第一，其中绝大多数为2型糖尿病患者，1型糖尿病和其他类型糖尿病较为少见。2型糖尿病及相关的心、脑、肾脏等慢性并发症给家庭与社会带来严重的经济负担，因此早期发现、规范治疗对防治2型糖尿病相关慢性并发症、降低致死致残率意义重大。

本2型糖尿病数据标准规范，以中华医学会糖尿病学分会的《中国2型糖尿病防治指南（2020版）》作为主要参考依据，旨在形成共识性的2型糖尿病领域数据集规范体系，为医疗数据的采集提供一致性的行业标准，实现同质化的数据存储和使用，并以高质量的数据库支撑包括临床医学研究及人工智能医疗器械等研发及应用的需求。

2型糖尿病专病数据集标注与质控规范（框架草案）

1 范围

本文件主要规定了2型糖尿病专病数据集标注与质量控制涉及的相关要求,数据类型包括基本信息表、就诊记录表、体格检查表、实验室检查表、家族病史表、现病史表、既往史表、个人烟酒史表、血糖检测表、诊断表、处方治疗表等。

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 1833.1-2022 《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

下列术语和定义适用于本文件。

3.1 糖尿病诊断标准

根据中华医学会糖尿病学分会的《中国2型糖尿病防治指南（2020版）》，糖尿病诊断依据静脉血浆葡萄糖测定结果的诊断标准如下：

|  |  |
| --- | --- |
| **诊断标准** | **静脉血浆葡萄糖或HbA1c水平** |
| 典型糖尿病症状 |  |
| 加上随机血糖 | ≥11.1mmol/L |
| 或加上空腹血糖 | ≥7.0 mmol/L |
| 或加上OGTT 2h血糖 | ≥11.1mmol/L |
| 或加上HbA1c | ≥6.5% |
| 无典型糖尿病症状者\*，需改日复查确认 |

\*注：典型糖尿病症状包括烦渴多饮、多尿、多食、不明原因体重下降；随机血糖指不考虑上次用餐时间，一天中任意时间的血糖，不能用来诊断空腹血糖受损或糖耐量减低；空腹状态指至少8 h没有进食热量。OGTT为口服葡萄糖耐量试验；HbA1c为糖化血红蛋白。

3.2 2型糖尿病

2型糖尿病是糖尿病的主要类型，病因和发病机制不明确，其显著的病理生理学特征为胰岛素调控葡萄糖代谢能力的下降（胰岛素抵抗）伴胰岛 β 细胞功能缺陷所导致的胰岛素分泌减少（相对减少）。

3.3 糖尿病性酮症酸中毒

是由于胰岛素不足和升糖激素不适当升高引起的糖、脂肪和蛋白质代谢严重紊乱综合征，临床以高血糖、高血酮和代谢性酸中毒为主要特征。1型糖尿病（T1DM）有发生DKA 的倾向；2 型糖尿病（T2DM）亦可发生 DKA。

3.4 高渗性高血糖状态

是糖尿病的严重急性并发症之一，临床以严重高血糖而无明显酮症酸中毒表现、血浆渗透压显著升高、脱水和意识障碍为特征。

3.5 糖尿病患者合并心血管疾病

主要包括动脉粥样硬化性心血管疾病和心力衰竭，其中动脉粥样硬化性心血管疾病包括冠心病、脑血管疾病和周围血管疾病，糖尿病患者的心血管疾病也是糖尿病患者的主要死亡原因。

3.5 糖尿病肾病

慢性肾脏病包括各种原因引起的慢性肾脏结构和功能障碍。糖尿病肾病是指由糖尿病所致的慢性肾脏疾病，病变可累及全肾（包括肾小球、肾小 管、肾间质等），我国约 20%-40% 的糖尿病患者合并糖尿病肾病，现已成为慢性肾脏病和终末期肾病的主要原因。

3.6 糖尿病视网膜病变

是常见的糖尿病慢性并发症，也是成人失明的主要原因，糖尿病视网膜病变尤其是增殖期视网膜病变是糖尿病特有的并发症，罕见于其他疾病。

3.7 糖尿病神经病变

是糖尿病最常见的慢性并发症。2型糖尿病患者神经病变的发生发展与糖尿病病程、血糖控制状况、肥胖、胰岛素抵抗和慢性低度炎症等因素相关，病程10年以上者易出现明显的神经病变临床表现。糖尿病神经病变以远端对称性多发性神经病变最具代表性。

3.8 糖尿病下肢动脉病变

通常指糖尿病下肢动脉粥样硬化。下肢动脉病变是外周动脉疾病的一个组成成分，表现为下肢动脉的狭窄或闭塞，其主要病因是动脉粥样硬化。与非糖尿病患者相比，糖尿病患者更常累及股深动脉及胫前动脉等中小动脉。

3.9 糖尿病足病

指初诊糖尿病或已有糖尿病病史的患者，足部出现感染、溃疡或组织的破坏，通常伴有下肢神经病变和（或）周围动脉病变，是糖尿病严重和治疗费用高的慢性并发症之一，重者可以导致截肢和死亡。

3.10 低血糖

非糖尿病患者诊断标准为血糖＜2.8mmol/L，接受药物糖尿病患者血糖只要低于3.9mmol/L即属于低血糖。

3.11 2型糖尿病的综合控制目标（见下表）

|  |  |
| --- | --- |
| 测量指标 |  |
| 毛细血管血糖（mmol/L） |  |
|  空腹 | 4.4-7.0 |
|  非空腹 | <10.0 |
|  糖化血红蛋白（%） | <7.0 |
|  血压（mmHg） | <130/80 |
|  总胆固醇（mmol/L） | <4.5 |
|  高密度脂蛋白（mmol/L） |  |
|  男性 | >1.0 |
|  女性 | >1.3 |
|  甘油三酯（mmol/L） | <1.7 |
|  低密度脂蛋白（mmol/L） |  |
|  未合并动脉粥样硬化性心脏病 | <2.6 |
|  合并动脉粥样硬化性心脏病 | <1.8 |
| 体质指数（kg/m2） | <24 |

注：毛细血管血糖：指毛细血管末梢全血的葡萄糖值；空腹状态指至少8 h没有进食热量。HbA1c为糖化血红蛋白。血压：是指血液在血管内流动时作用于单位面积血管壁的侧压力，它是推动血液在血管内流动的动力，在不同血管内被分别称为动脉血压、毛细血管压和静脉血压，此处所说的血压是指体循环的动脉血压。体质指数：指身体质量指数，衡量人体胖瘦程度以及是否健康的一个标准，计算方式为身高除以体重的平方（体重单位：千克；身高单位：米），正常值为18-24 kg/m2。

4 数据标注规范

本标准中数据集数据来源于四川大学华西医院电子病历系统。

4.1人员管理

4.1.1标准人员

标注医师应满足以下要求：

1. 执业资质：具有执业医师资格的内分泌代谢科专科医师；
2. 工作经验：从事内分泌代谢领域专科诊疗8年以上，至少为主治医师，对2型糖尿病的诊疗有较为丰富的经验；
3. 参与标注前，应统一接受必要的培训并通过考核。

复核医师、仲裁专家一般应由数据集建设实施单位聘请行业内知名专家担任。

4.1.2培训

信息部门或者大数据其他主管部门的培训，主要包含用户授权、软件安装、数据加密和反加密工具使用、数据质量监控、数据接入监控、数据统计工具使用、数据导出管理、标注流程、术语定义、样本总体描述、相关案例等。

培训由信息部门或者大数据平台主管部门组织，由数据采集方进行，培训前应准备培训环境和培训资料。

建议入选的标注医师接受相关培训，统一对标注规则的认识，熟悉标注软件操作。

4.1.3考核

通过制定相关的考核制度进行考核，内容包括数据安全管理制度、数据使用管理规范、使用违规惩罚制度、标注规则等四部分：

1. 数据安全管理制度： 数据安全管理从系统安全和审批管理两方面进行考虑。系统安全包括对信息部门数据备份策略、用户权限、账户弱密码、账户有效期和数据导出管理等方面进行管理；审批管理主要指账户权限审批流程和数据申请审批流程。
2. 数据使用管理规范：数据使用人需严格遵守国家有关法律法规，采集数据时需注明申请数据范围、用途、使用事件等重要信息，对使用数据有保密责任，数据采集结束要立即向管理部门报备，不可随意使用数据或进行其他违规操作。
3. 使用违规惩罚制度：对于擅自传播、转让、更改数据用途、未经许可使用数据或其他违规行为，医院管理部门有权利追究法律责任。
4. 标注规范：宜设置标注医师培训合格的测试及接受标准，宜选择与标准判读结果吻合率在90%以上的医师作为标注医师

4.2基础设施管理

4.2.1标注场所

在医院内部提供的安静、私密的工作环境，以便数据标注人员集中进行工作，在医院内网环境下对数据进行管理，防止数据外泄。某些数据标注工作可能需要与信息中心等其他部门进行密切合作，以收集和标注与特定临床领域相关的数据，在这种情况下，数据标注人员可能需要前往相应的科室或部门进行标注工作。

4.2.2标注软件

使用专门用于标注数据的软件工具，可以根据不同的数据类型和标注任务进行选择。具体的软件选择应根据医院的需求和实际情况进行评估和决策。医院可以根据数据类型、标注任务的复杂性、用户友好性、数据安全性等因素来选择适合的软件工具。同时，确保选择的软件工具符合医院的信息安全和合规要求，以保护医疗数据的安全和隐私。

4.3 标注方法

1. 对于一个病历资料，标注员应根据病史和检查结果，标注所有与2型糖尿病相关的ICD编码。

2. 对于同一份病历资料，标注员应根据临床共识和病史分析，排除非2型糖尿病的诊断，只选择与继2型糖尿病相关的ICD编码进行标注。

3. 所有标注结果应经过初步的质量控制，排除标注错误和遗漏。

4. 所有标注结果应经过二次审核，确保数据的准确性和一致性。

5. 对标注数据进行质量评估，评估项包括但不限于：数据完整性、准确性和一致性等。

6. 对标注数据进行统计分析，确定数据的分布情况、特征等，为后续分析提供支持。

7. 对数据进行可视化，发现数据的异常情况和规律，为后续分析提供支持。

4.4 数据标注流程

标注医师应对所接收的数据进行审查和检查，确保数据的完整性和准确性。数据集的标注信息包含三大类，包含“病史”、“诊断信息”、“处方信息”，标签包括但不限于：现病史、既往史、并发症、合并症、药物种类及治疗方案疗效判断等。标注任务说明文档包含数据标注内容、数据标注规则、数据标注人员、数据标注工具、数据标注质量控制方法。

每例2型糖尿病患者相关数据应由至少2位标注医师标注，标注结果一致数据方可入库；若结果不一致，数据应再由复核医师进行标注，复核结果若与任何一位医师标注结果相同，数据方可入组。若与原始标注任何一方都不相同，应再由仲裁专家进行仲裁，以仲裁结果作为最终结果入组。其流程如下图所示：



5 数据质控规范

* + 1. 人员

仲裁专家由具有内分泌代谢领域专科工作经验10年以上的副主任医师或主任医师担任。

* + 1. 方法

分阶段审核：

1. 初步审核阶段：主要考察数据项内容的完整性和规范性，包括字段数据是否存在、数据格式、数据长度、时间范围、数值范围是否正确等。此阶段的目的是筛选出明显错误或不合规的数据，以便后续审核。
2. 逻辑审核阶段：主要考察数据项内容的正确性和一致性，包括数据项内容是否正确、数据项间内容是否一致等。此阶段的目的是确保数据的逻辑正确性和一致性。

3、标准审核阶段：主要考察数据项的标准性，即数据是否与字典一致。此阶段的目的是确保数据的标准化和一致性。

* + 1. 指标

数据是否按照5.4的要求进行完整性、规范性、标准性、正确性等验证

* + 1. 通过准则

1、 完整性：数据项内容的完整情况，主要考察字段数据是否存在，NULL等同于空。

2、 规范性：数据项的数据格式，数据长度，时间范围，数值范围是否正确。

3、 标准性：数据项的数据是否与字典一致，如性别的取值应该是男，女。

4、 正确性：数据项内容是否正确，如入院记录中月经史患者性别不应该为男。

5、 一致性：数据项间内容是否一致，如考察病案首页上的患者、手术、诊断、费用等信息是否同HIS或其他系统中的信息一致。

6、 时序性：时间逻辑合理性（流程时间分布），如出院时间需要晚于入院时间。

7、 整合性：相关系统对应数据项目可对照或关联，如医疗费用明细表里的医嘱 标识应该在中西医处方表、药品类医嘱表、草药处方表或者非药品医嘱表里。

8、 唯一性：记录的唯一性，如PatientID在Patient表不能重复，相同的医嘱不应该重复。

* + 1. 数据验证方法

1、采集方提供数据报告，包括接入数据范围（系统数量）、数据量（需要分类， 比如患者、医嘱、处方、检验检查数量等）、数据纵深（起始到终止日期）。医院信息部门工程师可进行宏观数据核验。

1. 数据质量抽样检查：比照HIS、EMR等业务系统的单个患者进行数据核验。
2. 病案数据抽样检查：随机抽取一定数量的病例数据，跟院内数据进行核对。
3. 隐私数据检查：查看是否按照要求进行去隐私化处理，比如患者姓名、地址、身份证号等，可有意识地搜索隐私数据看可否能检索到。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **编号** | **数据域** | **数据项名称** | **值** | **说明** |
| 1 | 标注人员 | 姓名 |  |  |
| 2 | 医院 |  |  |
| 3 | 年资 | 8年 |  |
| 4 | 培训状态 | 是/否 |  |
| 5 | 考核状态 | 是/否 |  |
| 6 | 基础设施 | 标注场所 |  |  |
| 7 | 标注软件名称 |  |  |
| 8 | 标注软件版本 |  |  |
| 9 | 标注指标 | 空腹血糖、餐后2h血糖 |  |  |
| 10 | 糖化血红蛋白 |  |  |
| 11 | 血压 |  |  |
| 12 | 血脂 |  |  |
| 13 | 体重 |  |  |
| 14 | 质控人员 | 质控方法 |  |  |
| 15 | 医院 |  |  |
| 16 | 年资 | 10年 |  |
| 17 | 培训状态 | 是/否 |  |
| 18 | 质控方法 | 质控流程 | 疗效判断 |  |
| 19 | 疗效判断质控 |  |  |

## 2型糖尿病专病数据集标注与质控规范编制说明

# 一、任务来源及计划要求

科技创新2030-“新一代人工智能”重大项目“医学人工智能技术临床应用评估标准研究 ”子课题“决策支持临床应用评估标准数据集建设研究”。

# 二、编制过程

包括编制原则、工作分工、征求意见、各阶段工作过程等。

l、编制原则

在中国知网、维普网、万方数据库等数据库中，以“2型糖尿病标准数据集”、 “2型糖尿病数据集标准”等为检索词进行文献检索，遴选符合研究需求的相关的文献和报道。通过谷歌、百度等搜索引擎，检索2型糖尿病数据集标准的研发单位、文献书籍、标准指南等相关资料。在文献研究的基础上，采取集中研讨，数据集建设过程、术语、纳入字段等主题进行头脑风暴，明确2型糖尿病专病数据集标注、质控规范定位。每次讨论参加人数 7-10 人不等，对讨论内容进行及时记录，会后进行整理、提炼和分析。

2、工作分工

四川大学华西医院负责提供医学意见并编制2型糖尿病专病数据集标注、质控规范初稿，电子科技大学负责提供电子信息相关医院并修改2型糖尿病专病数据集标注、质控规范初稿，国家卫健委卫生发展中心负责提供意见，协调工作。

3、征求意见

无

4、各阶段工作过程

2023年1月-2023年5月 2型糖尿病数据集标准相关文献调研。

2023年4月-2023年5月 2型糖尿病数据集标准“头脑风暴”会。

2023年6月 2型糖尿病专病数据集标注与质控规范制定。

# 三、调研和分析工作情况

1、国内外情况

我国现已成为全球糖尿病患者人数最多的国家，患病率高达11.2%，患病人数超1.4亿人，其中绝大多数为2型糖尿病患者，1型糖尿病和其他类型糖尿病较为少见。2型糖尿病及相关的心、脑、肾脏等慢性并发症给家庭与社会带来严重的经济负担，因此提高2型糖尿病诊治水平对防治2型糖尿病相关慢性并发症、降低致死致残率意义重大。

任何科学研究离不开数据的支撑，要想探索到糖尿病的逆转，完成高质量的科学研究，离不开大量长期的数据。目前我国糖尿病相关临床研究数量众多，但高质量、多中心的研究较为缺乏，而单中心研究的数据采集和标注、质控又标准不一，数据无法“共建共享”。近年来各类与健康相关的平台、系统、软件和设备广泛运用，对改善居民健康生活，提高居民健康水平发挥重要作用。但是，由于各类信息系统独立研发，缺乏统一完备的数据标准，导致各个系统无被法“互联互通”，造成大量数据资源浪费，产生各种“信息孤岛”，不利于健康大数据的整体运用。

国外建立了许多著名的医学数据库及数据集，其中具有代表性的与糖尿病和糖尿病并发症相关的有印度视网膜病变数据集EyePACS，美国营养数据库NHAMES， 英国生物银行数据库UK Biobank等，为标准数据集的发展做出了贡献。

在健康信息数据标准方面，发达国家相关工作起步较早、发展较快，例如，美国疾病控制与预防中心(CDC)为规范公共卫生数据标准，研发了公共卫生概念数据模型(PHCDM)，英国国家医疗服务体系(NHS) 也推出了数据模型、数据字典等数据标准，支持并促进以病人为中心的医疗服务的改善，为一致准确的数据收集与交换构建数据标准框架。但尚未推出针对2型糖尿病的数据标注、质控标准。

国内相关的数据集标准包括：中国食品药品检定研究院发布了《人工智能医疗器械质量要求与评价》的《第2部分：数据集通用要求》（征求意见稿），然而，在2型糖尿病的领域尚未有相关的数据集标准。

2、标准编制的必要性

建立糖尿病相关诊疗数据集标准，对促进人口健康信息互联互通、共建共享具有十分的重要意义，有助于提高糖尿病相关研究的水平，促进行业内学术合作与交流以及成果推广，对完善我国人口健康信息标准体系具有重要作用。

# 四、主要技术内容的说明

包括技术参数和指标的确定依据、修订标准的各修订点及理由、主要技术内容的说明

1. 范围

本文件主要规定了2型糖尿病专病数据集数据标注、质控涉及的相关要求。

1. 主要指标确定依据

本文根据此前检索结果，依据中国食品药品检定研究院发布了《人工智能医疗器械质量要求与评价》的《第2部分：数据集通用要求》，草拟了本数据集规范的基本框架。参考临床经验，并根据2022年美国糖尿病协会（ADA）糖尿病管理指南及中华医学会糖尿病学分会推出的中国 2 型糖尿病防治指南(2020 年版)等，初步确定数据标注范围及标注依据。

本研究邀请了医疗保健、电子信息等方面的专家作为讨论参与者，成功邀请了5位，进行头脑风暴。邀请对象均为副高以上职称，长期从事内分泌代谢病诊治、卫生信息等方面研究和实践工作，熟悉2型糖尿病诊治的发展和卫生相关信息系统研发情况，理论基础深厚，实践经验丰富。在头脑风暴会中，邀请各位专家从一下几点展开讨论：“1、您认为糖尿病数据集的数据有哪些需要标注? 2、您认为数据标注集过程应该是怎样的，应该从哪些方面进行管理? 3、您认为数据标注人员应该如何管理? 4、您认为可以用哪些方法对数据进行治疗控制? 讨论时对访谈内容及时记录，访谈后进行整理和分析，最后结合文献调研结果形成数据集标注与质控规范。

3、规范性引用文件

1. YY/T 1833.1-2022 《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》