团 体 标 准

2型糖尿病专病数据集处理规范

（框架草案）

Standards for data processing 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](#_Toc136729905)

[1 范围 5](#_Toc136729906)

[2 规范性引用文件 5](#_Toc136729907)

[3 术语和定义 5](#_Toc136729908)

[4 数据处理 6](#_Toc136729909)

[4.1 数据脱敏 6](#_Toc136729910)

[4.1.1类型 6](#_Toc136729911)

[4.1.2内容 6](#_Toc136729912)

[4.1.3方法 6](#_Toc136729913)

[4.2 数据清洗 8](#_Toc136729914)

[4.2.1 清洗规则 8](#_Toc136729915)

[4.2.2 文本脏数据处理 8](#_Toc136729916)

[4.3数据核验 8](#_Toc136729917)

[2型糖尿病专病数据集标注与质控规范编制说明 9](#_Toc136729914)

[一、任务来源及计划要求 9](#_Toc136729915)

[二、编制过程 9](#_Toc136729916)

[三、调研和分析工作情况 10](#_Toc136729917)

[四、主要技术内容的说明 10](#_Toc136729917)

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处理规范（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数据处理涉及的相关要求。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。

YY/T 1833.1-2022 《人工智能医疗器械 质量要求和评价 第1部分：术语》

DB51/T 2989—2023 《四川省健康医疗大数据应用指南》

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 |
| HbA1c（%） | <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 数据处理

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 数据脱敏

4.1.1类型

数据集使用的数据必须是获得医院伦理委员会批准或者豁免的临床脱敏数据，患者的隐私保护应当满足法规的要求，保证数据脱敏、患者隐私安全和患者利益，故数据应严格遵循法规要求进行脱敏处理。

4.1.2内容

被脱敏信息：病人姓名、ID号、出入院日期等所有涉及敏感信息的数据。

4.1.3方法

确认具体敏感字段之后，针对相关的信息以下面的方式进行脱敏操作：

**a、 姓名的去标识化**

姓名是一种常用的标识符，适用的去标识化方法举例如下：

a） 泛化编码。使用概括、抽象的符号来表示，如使用“张先生”来代替“张三”，或使用“张1某某”来代替“张三”。这种方法是用在需要保留“姓”这一基本特征的应用场景。

b） 抑制屏蔽。直接删除姓名或使用统一的“\*”来表示。如所有的姓名都使用“\*\*\*”代替。

c） 随机替代。使用随机生成的汉字来表示，如使用随机生成的“辰筹猎” 来取代“张三丰”。

d） 假名化。构建常用人名字典表，并从中选择一个来表示，如先构建常用的人名字典表，包括龚小虹、黄益洪、龙家锐等，假名化时根据按照顺序或随机选择一个人名代替原名。如使用“龚小虹”取代“张三丰”。这种方法有可能用在需要保持姓名数据可逆变换的场景。

e） 可逆编码。采用密码或其他变换技术，将姓名转变成另外的字符，并保持可逆特性。如使用密码和字符编码技术，使用“SGIHLIKHJ”代替“张三丰”，或使用“Fzf”代替“Bob”。

**b、 地址的去标识化**

对于地址，常用的去标识化方法举例如下：

a） 泛化编码。使用概括、抽象的符号来表示，如“江西省吉安市安福县”使用“南方某地”或“J省”来代替。

b） 抑制屏蔽。直接删除地址或使用统一的“\*”来表示。如所有的地址都使用“\*\*\*\*\*\*”代替。

c） 部分屏蔽。屏蔽地址中的一部分，以保护地址信息。如使用“江西省XX 市XX县”来代替“江西省吉安市安福县”。

d） 数据合成。采用重新产生的数据替代原地址数据，数据产生方法可以采用确定性方法或随机性方法。如使用“黑龙江省鸡西市特铁县北京路23 号”代替“江西省吉安市安福县安平路1号”。

**c、 数值型标识符的去标识化**

常用的数值型标识符的去标识化包括：

a） 泛化编码。使用概括、抽象的符号来表示，如“有四个人，他们分别是蓝色、绿色和浅褐色的眼睛”来代替“有1个人是蓝色眼睛,2个人是绿色的眼睛，1个人是浅褐色的眼睛”。

b） 抑制屏蔽。直接删除数值或使用统一的“\*”来表示。如所有的数值都使用“\*\*\*\*”代替。

c） 顶层和底层编码。大于或小于一个特定值的处理成某个固定值。例如，年龄超过70岁的一律用“大于70岁”描述。

d） 部分屏蔽。使用数值的高位部分代替原有数值，如百分制考试成绩全部使用去掉个位数、保留十位数的数值代替。

e） 记录交换。使用数据集中其他记录的相应数值代替本记录的数值。如设定规则，将记录集中的所有的身高数据取出并全部打乱位置后（其他属性数据位置不变）放回原数据集中。这种方法可以保持数据集的统计特性不变。

f） 噪声添加。相对原始数据，产生微小的随机数，将其加到原始数值上并代替原始数值。如对于身高1.72m，产生随机数值-0.11m,加到原始数 值后将其变为1.61m。

g） 数据合成。采用重新产生的数据替代原始数据，数据产生方法可以采用确定性方法或随机性方法。如使用“19”岁年龄代替“45”岁年龄。

**d、 日期的去标识化**

在数据集中，日期有多种存在形式，包括出生日期、开始日期、纪念日等。常见的对日期的去标识化方法包括：

a） 泛化编码。使用概括、抽象的日期来表示，如使用1880年代替1880年1 月1日。

b） 抑制屏蔽。直接删除日期数据或使用统一的“\*”来表示。如所有的数值都使用“某年某日”代替。

c） 部分屏蔽。对日期中的一部分做屏蔽，如1880年某月1日代替1880年1 月1日。

d） 记录交换。使用数据集中其他记录的相应数值代替本记录的数值。如设定规则，将记录集中的所有的日期数据取出并全部打乱位置后（其他属性数据位置不变）放回到原数据集中。这种方法有利于保持数据集的统计特性。

e） 噪声添加。相对原始数据，产生微小的随机数，将其加到原始数值上并代替原始数值。如对于出生日期1880年1月1日，产生随机数值32天，加到原始数值后将其变为1880年2月2日。若脱敏数据有时序性要求，需控制添加噪声的正负，如对出院时间脱敏时应加上正噪声，保证检查、入院等时间在脱敏后的出院时间之前。

f） 数据合成。采用重新产生的数据替代原日期数据，如使用“1972年8月12日”代替“1880年1月1日”。

4.2 数据清洗

4.2.1 清洗规则

包括非空检核、主键重复、非法代码清洗、非法值清洗、数据格式检核、记录数检核、业务约束核验。

a）非空检核：要求字段为非空的情况下，需要对该字段数据进行检核。

b）主键重复：多个业务系统中同类数据经过清洗后，在统一保存时，为保证主键唯一性，需进行检核工作。

c）非法代码、非法值清洗：非法代码问题包括非法代码、代码与数据标准不一致等，非法值问题包括取值错误、格式错误、多余字符、乱码等, 需根据具体情况进行校核及修正。

d）数据格式检核：通过检查表中属性值的格式是否正确来衡量其准确性，如时间格式、币种格式、多余字符、乱码。

e）记录数检核：指各个系统相关数据之间的数据总数检核或者数据表中每日数据量的波动检核。

f）业务约束核验，在实施处理数据过程中，与医院相关人员共同确定。从业务的正确性、一致性、有效性等角度，考虑数据的核验规则，如：建档日期、入院日期、民族信息等的有效性核验。

4.2.2 文本脏数据处理

对于常见的空缺值、离群值和不一致等脏数据，采用人工检测、统计学方法、聚类、分类、基于距离、关联规则等方法来实现数据清洗。根据缺陷类型分类，可以将脏数据分为缺失值数据、错误数据和错误关联数据这三种典型问题数据，分别进行数据清洗。

4.3数据核验

通过人工复核按照下面的方式进行验证：

a）初期的数据要求提供数据报告，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）。医院信息部门可进行宏观数据核验。

b）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。

c） 病案数据抽样检查：随机抽取一定数量的病例数据进行核对。

d）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。

检出问题后及时汇总和分析出问题原因。可能出现的原因主要有三种，即接入对接偏差、工程师理解偏差和后期后结构化归一的次生原因；这些问题将作为后期调整技术方案的依据。

2型糖尿病专病数据集处理规范编制说明

# 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处理规范定位。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处理规范初稿，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处理规范初稿，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处理规范制定。

# 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处理和处理又标准不一，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型糖尿病的数据处理标准。

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处理涉及的相关要求。

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。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哪些需要脱敏? 2、您认为数据清洗过程可以采取哪些方法? 3、您认为脏数据应该如何管理? 4、您认为对数据采集的来源和设备应该有什么要求? 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采集规范。

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》