|  |  |
| --- | --- |
| ICS | 11.040.99 |
| CCS | |  | | --- | |  |   C30 |

团体标准

T/CSBME 050—2022

继发性高血压人工智能辅助风险识别数据采集标准

Data collection standard for artificial intelligence-assisted risk identification of secondary hypertension

2022 - 01 - 15发布

2022 - 03 - 01实施

四川省生物信息学学会  发布

目次

[前言 II](#_Toc93077075)

[引言 III](#_Toc93077076)

[1 范围 1](#_Toc93077077)

[2 规范性引用文件 1](#_Toc93077078)

[3 术语和定义 1](#_Toc93077079)

[4 缩略语 2](#_Toc93077080)

[5 数据集总要求 4](#_Toc93077084)

[6 数据采集 4](#_Toc93077084)

[参考文献 8](#_Toc93077088)

1. 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。

本文件起草单位：四川大学华西医院。

本文件主要起草人：曾锐、陈蕾、殷晋、李东泽、贾禹、李芳卉、李文韬、童瑶、公奥博。

1. 引言

高血压是心血管病最为常见，也是临床常见的慢性非传染性疾病之一。高血压带来严重的疾病负担。其中，继发性高血压占高血压患者的5%-15%。继发性高血压除高血压本身对机体的影响外，与之伴随的内分泌紊乱、低钾血症、肾功能不全、低氧血症等还可导致独立于血压之外的心脑血管损害，因此早识别、早诊断、早治疗尤为重要，然而目前对继发性高血压的诊断存在一定的困难。本数据标准规范以《中国高血压健康管理规范（2019）》作为主要参考依据。继发性高血压主要包括肾实质性高血压、肾血管性高血压、库欣综合征、原发性醛固酮增多症、阻塞性睡眠呼吸暂停综合征、嗜铬细胞瘤等。

本继发性高血压数据标准规范旨在形成共识性的继发性高血压领域数据集标准体系，为医疗数据的采集、标注和质控等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。

继发性高血压人工智能辅助风险识别数据采集标准

* 1. 范围

本文件主要规定了继发性高血压人工智能辅助风险识别所使用的高血压数据集（包裹基本信息、体格检查、病史、实验室检查等数据）在数据采集、数据整理、数据标注、数据集构建过程中所涉及的相关要求。

本文件适用于继发性高血压人工智能辅助风险识别的研发、生产、检测过程中的数据集建立及其评估，以及继发性高血压的初步筛查与风险评估。

* 1. 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 医学与医疗器械 质量要求和评价 第 1 部分：术语

YY/T 1833.2 医学与医疗器械 质量要求和评价 第 2 部分：数据集通用要求

* 1. 术语和定义

下列术语和定义适用于本文件。

术语

（1）人工智能医疗器械 artificial intelligency medical device; AIMD

采用AI技术实现其预期用途的医疗器械

注1：如采用机器学习、模式识别、规则推理等技术实现其用途的独立软件。

注2：如采用内嵌AI算法、AI芯片实现其医疗用途的医疗器械。

[YY/T 133.1-2022]

（2）继发性高血压风险识别人工智能辅助诊断技术

继发性高血压风险识别人工智能辅助诊断技术是指医务人员采用人工智能(AI)技术用于辅助继发性高血压风险识别的一项医疗技术。该技术基于患者基本特征、生命指征、血压波动特征、常规血液学检查结果（包括血常规、生化1、生化4等）此类易于采集的数据，协助临床医生尤其是基层医院医生，在患者首次确诊高血压后进行继发性高血压风险识别，筛选出继发性高血压高风险患者，建议进一步进行继发性高血压病因诊断。

（3）继发性高血压

继发性高血压是病因明确的高血压，当查出病因，并有效去除或控制病因后，作为继发症状的高血压可被治愈或明显缓解。对继发性高血压的筛查、鉴别成为高血压诊断评估和治疗的重要内容。临床上，当患者有相应临床表现、体格检查线索，或是难治性高血压时，建议筛查、排除继发性高血压。

本AI产品采集患者基本特征、生命指征、血压波动特征、常规血液学检查结果，并综合考虑以下因素： （1）血压中、重度升高的年轻高血压患者；（2）舒张压高于100 mmHg 的老年高血压患者（年龄≥65岁）；（3）症状、体征或实验室检查有怀疑线索，例如肢体脉搏搏动不对称性减弱或缺失，腹部听到粗糙的血管杂音，既往患有肾脏疾病史等；（4）不明原因的高血压伴低血钾。基于先前构建的继发性高血压风险识别算法，筛选出继发性高血压高、中、低危患者。

定义

(1)血压：血压（blood pressure，BP）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，通常所说的血压是指体循环的动脉血压。

(2)偶测血压：被测者在没有任何准备的情况下测得的血压。

(3)动态血压：使用动态血压记录仪测定一个人昼夜24小时内，每间隔一定时间内的血压值。动态血压包括收缩压、舒张压、平均动脉压、心率以及它们的最高值和最低值等项目。

(4)高血压：动脉血压超过正常值的异常升高。本指南定义为收缩压≥140mmHg和(或)舒张压~>90mmHg。

(5)收缩压：心室收缩时，主动脉压急剧升高，在收缩期的中期达到最高值，这时的动脉血压值称为收缩压，也称为“高压”。

(6)舒张压：心室舒张时，主动脉压下降，在心舒末期动脉血压的最低值称为舒张压，也称为“低压”。

(7)脉压：收缩压减舒张压的差值。

(8)平均动脉压：一个心动周期中每一瞬间动脉血压的平均值，大约等于舒张压加1／3脉压。

(9)KPa：千帕，通常用于表示血压数值。

(10)mmHg：毫米汞柱，人们用水银血压计来测量血压时用水银柱的高度“毫米汞柱”来表示血压的水平。

1mmHg(毫米汞柱)=0．133kPa(千帕斯卡)

7．5mmHg(毫米汞柱)=1kPa(千帕斯卡)

(11)理想血压：收缩压<120ramHg和舒张压<80mmHg。

(12)正常血压：收缩压应<130mmHg，舒张压<85mmHg。

(13)血压正常高限或高血压前期：收缩压在130～139mmHg和／或舒张压在85～89mmHg。

(14)临界高血压：收缩压在140～160mmHg(18．6～21．3kPa)，舒张压在90～95mmHg(12．0～12．6kPa)。

(15)肌酐：人体内肌酸的代谢产物。肌酸量与肌肉容量呈正比，以稳定的速度产生，并以稳定的速度向血液中释放肌酐，由血循环带到肾脏，从尿中排出体外。长期高血压损害肾功能，使肾脏对血中肌酐的排出减少，使血肌酐升高。

(16)心脏的超声波检查：又叫超声心动图，可以反映出各瓣膜的形态、活动是否正常，各心房心室的大小及形态，心脏周围出入心脏的大血管的情况，有无先天性心脏病或风湿性心脏病造成的心脏损害、心包积液、各种心肌病，检出高血压所致的左心室肥厚和评价心功能。

(17)原发性高血压：原因不明的高血压，占90％以上，目前尚难根治但能被控制。也叫高血压病。

(18)继发性高血压：也叫症状性高血压。由于某些疾病(肾脏疾病、内分泌疾病如肾上腺肿瘤或增生和其他原因所致)在其发展过程中产生的，原发疾病治愈后，血压也会随之下降，占高血压患者的5％～10％。

(19)肾血管性高血压：指各种原因引起的肾动脉或其主要分支的狭窄或闭塞性疾病，引起肾血流量减少或缺血所致的高血压。

(20)肾实质性高血压：肾脏实质疾病，如急、慢性肾小球肾炎，肾盂肾炎，多囊肾等引起的高血压。

(21)库欣综合征：是由于长时间接触皮质醇或其分泌量过高而造成的一系列临床表现与体征。其临床表现与体征可能包括：高血压、向心型肥胖（肚腩有赘肉但双臂与双腿较细）、紫纹（红色或紫色的短纹）、满月脸（脸又圆又红）、脂失养（水牛肩，两肩中间有一较大肿块）、肌肉无力、骨质疏松、粉刺、皮肤脆弱难以愈合等。

(22)原发性醛固酮增多症：是由于肾上腺的皮质肿瘤或增生，醛固酮分泌增多所致。原醛症是一种继发性高血压症，约占高血压症中0.4～2%。发病年龄高峰为30～50岁，女性较男性多见。引起本病最常见的原因为醛固酮瘤，约占原醛症的60%～80%，大多数为单个腺瘤，左侧多见。其次有双侧肾上腺皮质增生，又称为特发性醛固酮增多症，约占20%～30%。少见原因有地塞米松可抑制性醛固酮 增多症，醛固酮癌、异位分泌醛固酮的肿瘤。

(23)阻塞性睡眠呼吸暂停综合征：是一种病因不明的睡眠呼吸疾病，临床表现有夜间睡眠打鼾伴呼吸暂停和白天嗜睡。由于呼吸暂停引起反复发作的夜间低氧和高碳酸血症，可导致高血压，冠心病，糖尿病和脑血管疾病等并发症及交通事故，甚至出现夜间猝死。

(24)嗜铬细胞瘤：为起源于神经外胚层嗜铬组织的肿瘤，主要分泌儿茶酚胺，根据肿瘤是来自交感神经或副交感神经将副神经节瘤分为副交感神经副神经节瘤（包括化学感受器瘤、颈动脉体瘤等）及交感神经副神经节瘤（包括腹膜后、盆腔及纵隔后的副神经节瘤）。某些患者可因长期高血压致严重的心、脑、肾损害或因突发严重高血压而导致危象，危及生命，但如能及时、早期获得诊断和治疗，是一种可治愈的继发性高血压病。

* 1. 缩略语

以下缩略语适用于本文件：

AI：人工智能（Artificial Intelligence）

AIMD：人工智能医疗器械（Artificial intelligency medical device）

SHT：继发性高血压（Secondary hypertension）

BP：血压（Blood pressure）

* 1. 数据集总要求
     1. 数据来源

数据来源于四川大学华西医院与2011-2021年住院的诊断包含“高血压”的患者

* + 1. 数据类型

数据类型应包含宫颈TBS报告系统中涉及的所有病变类型，包括继发性高血压相关信息，如患者年龄、性别、病史、生化指标、疾病ICD编码等。每一类型宜尽可能包括更多的细分类。具体数据集分类及医学和风险级别定义见附录A。

* + 1. 数据分布

为保证数据集具有充分的临床代表性，控制偏倚，数据分布宜充分考虑数据多样性，包括如下情况：

1. 考虑受试者人群差异：数据应覆盖全国各部分地域包括沿海与内陆、南部与北部地区。数据集宜覆盖5个不同省市的10家以上不同医疗机构，各医疗机构样本比例不宜高于10%；
2. 考虑采集场所差异：数据应主要以门诊病人为主，可包含部分体检及宫颈癌筛查人群；
3. 考虑设备与配置差异：数据应采用两种或以上的液基制片技术，选取其中一种作为数据集的主要组成部分。数据宜尽可能包含不同层级的代表性临床机构，来自多家不同厂商、设备型号、参数的制片设备和数字化采集设备；
4. 考虑预期使用者的代表性：数据应符合继发性高血压的流行病学特征，包括疾病构成、人群分布等。患者年龄、疾病分型、分级、分期的具体数据分布和数据量应符合流行病学和统计学要求。疾病分型、分级、分期数据应满足5.1相关要求；
5. 考虑疑难少见病例：数据集中应包括一定数量的疑难少见病例，？
   * 1. 数据完备性

训练数据集数据和测试数据集应具备受试者基本信息，包括年龄、民族、医疗机构名称、医疗机构性质、省份、细胞学判读结果。若为阳性病例还应具备组织学结果、门诊/体检/两癌筛查。

预期用途为性能测试、临床评价、产品质控等的数据集宜具备HPV检测结果和免疫组化结果。

* 1. 数据采集
     1. 伦理与患者隐私保护

从志愿者、体检中心、医疗机构、科研项目等途径收集的数据应符合伦理与隐私保护要求。数据应脱敏以保护患者隐私。

* + 1. 数据采集标准

继发性高血压数据必须来自有资质的医疗机构或实验室，采集过程应符合相关规范和标准。元信息包括收集数据的时间、数据来源、数据精度、数据完整性、数据一致性和数据保密性等。在收集患者的生病史和体征数据时，必须包括个人的基本信息、病史、生活方式、家族史、收缩压和舒张压等，并且整合在一起，以获取更完整准确的患者信息。

* + 1. 数据格式标准

在收集高血压的基本信息时，必须注意高血压的定义、分类和诊断策略，并且将这些信息记录到标准表格中。数据文件应命名规范，包括患者ID、采样时间、样本编号等信息，确保数据的完整性和可追溯性。

* + 1. 数据质量管理标准

针对数据采集过程中可能出现的错误和失误，采取相应的技术和方法，保证数据质量，。所有数据必须进行质量控制，包括数据来源的质量、数据清洗的质量、数据标注的质量等。数据质量必须符合相关规范和标准，确保数据的可靠性和可信度。

* + 1. 数据共享

为了支持产业界研发更先进的继发性高血压人工智能辅助风险识别产品，对于数据共享者需要签署数据使用合同，确保数据仅用于研究和学术目的。共享前需要进行数据脱敏、匿名或去标识化处理，保护个人隐私。

本文件制定的继发性高血压人工智能辅助风险识别数据采集要求，旨在通过规范数据采集、格式、质量、共享和伦理，支持产业界研发更先进的继发性高血压人工智能辅助风险识别产品，进一步推动人工智能技术在继发性高血压的风险识别中的实际应用和持续健康发展。

6.6 釆集设备标准

6.6.1 兼容性要求

兼容医院内各类数据接入源，包括原业务数据库的备份库、数据中心、集成平台和文件接口。数据库方面支持Oracle、MySQL、SQL Server等主流数据库平台。

6.6.2 釆集要求

①采集方式

在院内部署数据采集工具，对接院方信息科提供的原业务数据库的备份库、 数据中心、集成平台或者文件接口。

②采集协议

采用HTTP协议或者数据库连接协议来采集标准的电子病历数据。

③采集参数

按照高血压病例数据标准要求的数据参数指标进行采集。

④采集精度

按照高血压病例数据标准要求的数据精度进行采集。

6.6.3 釆集过程标准

①釆集人员要求

选拔：经过医院信息科和相关部门同意的人选。

培训：对采集人员按照医院的安全和标准作业流程进行培训。

考核：在采集过程和结果中，按照相关标准进行考核。

②釆集过程要求

采集准备：采集之前，在医院部署相应的由清华提供的物理服务器，在服务器上部署相应的采集工具，跟医院信息科商定采集接口，这部分工作做完即已做好采集前的准备工作。

采集过程：采集过程严格按照相关安全标准进行。

6.6.4 数据脱敏标准

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。

②内容

被脱敏信息：病人姓名、年龄、性别、ID号、检查日期、出入院日期等所有涉及敏感信息的数据。

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

1）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”。

2）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”。

3）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，以保障满足此条件的人数多于20000人。

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-0.11 m,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。

4）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，产生随数值32天，加到原始数值后将其变为1880年2月2日。

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”。

参考文献

[1]国家心血管病中心. 中国心血管病报告2018 [M]. 北京: 中国大百科全书出版社,2019.

[2]国家药品监督管理局医疗器械技术审评中心. 深度学习辅助决策医疗器械软件审评要点[Z], 2019.

[3]Wang Z, Chen Z, Zhang L, et al. Status of hypertension in China: results from the China hypertension survey, 2012‐2015[J]. Circulation, 2018, 137(22): 2344‐2356.

[4]Su M, Zhang Q, Bai X, et al. Availability, cost, and prescription patterns of antihypertensive medications in primary health care in China: a nationwide cross‐sectional survey [J]. Lancet, 2017,39

[5]刘明波，李镒冲，刘世炜，等．2010年中国人群高血压疾病负担[J].中华流行病学杂志,2014,35(6):680-683.

[6]中国高血压防治指南修订委员会．中国高血压防治指南（2018年修订版）[J]中国心血管杂志，2019,24(1):24-56.

[7]Sandström YK, Ljunggren G, Wändell P, et al. Psychiatric comorbidities in patients with hypertension‐‐a study of registered diagnoses 2009‐2013 in the total population in Stockholm County,Sweden[J]. J Hypertens, 2016, 34(3): 414‐420; discussion 420.

[8]Bautista LE, Vera‐Cala LM, Colombo C, et al. Symptoms of depression and anxiety and adherence to antihypertensive medication[J]. Am J Hypertens, 2012, 25(4): 505‐511.

[9]Piercy KL, Troiano RP, Ballard RM, et al. The physical activity guidelines for Americans[J]. JAMA, 2018, 320(19): 2020‐2028.

[10]国家药品监督管理局医疗器械技术审评中心. 人工智能医疗器械注册审查指导原则（征求意见稿）[Z], 2021.

[11]中国食品药品检定研究院. 人工智能医疗器械质量要求和评价 第1部分：术语(报批稿)，2020.

[12]中国食品药品检定研究院. 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(报批稿)，2020.

[13]人工智能医疗器械质量要求和评价 第3部分：数据标注通用要求(征求意见稿)