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

团体标准

T/CSBME 050—2022

基于医院电子病历的抑郁症数据处理标准

Standard for depression data processing based on hospital electronic medical records

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](#_Toc26081)

[引言 III](#_Toc14526)

[1 范围 4](#_Toc17003)

[2 规范性引用文件 4](#_Toc21958)

[3 术语和定义 4](#_Toc9466)

[4 数据处理标准 4](#_Toc3137)

1. 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。

本文由四川省生物信息学会提出。

本文由四川省生物信息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：袁敏兰

引言

1990年至2017年，精神疾病成为了伤残损失健康生命年的三大主要原因之一，我国精神疾病的患病率较30年前更高。全球有近3亿人受抑郁症的困扰，5岁至25岁年龄段抑郁症的综合患病率更是高达14.3%，严重影响人民的社会功能和生活质量。而双相情感障碍抑郁发作、脑器质性疾病共病抑郁（癫痫共病抑郁、卒中后抑郁）等疾病同样给家庭与社会带来严重的的经济负担，需要与原发性抑郁症进行鉴别并进行针对性、规范化的诊疗。

本数据标准规范以《精神障碍诊断与统计手册（第五版）》（DSM-5）、《2016 ACP临床实践指南：药物治疗与非药物治疗治疗成人重度抑郁症》作为主要参考依据，旨在形成共识性的抑郁症诊疗领域数据集标准体系，为医疗数据的处理提供一致性的行业标准。

基于医院电子病历的抑郁症数据处理标准

* 1. 范围

本文件主要规定了抑郁症数据处理涉及的相关要求。抑郁症数据包括基本信息，心理测评量表，患者主述，脑部磁共振，脑电信号。

* 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 人工智能医疗器械 质量要求和评价 第 1 部分：术语

YY/T 1833.2 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. 术语和定义

下列术语和定义适用于本文件。

（1）汉密尔顿抑郁量表

|  |  |
| --- | --- |
| 分值 | 结果分析 |
| 总分< 7分 | 正常 |
| 7分<总分≤17分 | 轻度抑郁 |
| 17分<总分≤24 | 重度抑郁 |
| 总分>24分 | 极重度抑郁 |

（2）汉密尔顿焦虑量表

|  |  |
| --- | --- |
| 分值 | 结果分析 |
| 总分< 7分 | 正常 |
| 7分<总分≤20分 | 可能有焦虑 |
| 21分<总分≤28 | 明显有焦虑 |
| 总分>29分 | 严重焦虑 |

1. 患者主述:就诊时主动提出的问题、症状或困扰，反映了患者对自身精神健康状 况的感受和描述。

心境问题：例如情绪低落、焦虑、易怒、情绪波动等。

睡眠障碍：如入睡困难、睡眠中断、睡眠过多或过少等。

思维问题：包括注意力集中困难、思维迟缓、思维混乱或妄想等。

记忆和认知问题：例如记忆力减退、注意力不集中、判断力下降等。

自杀意念或自伤行为：表现为对生命失去兴趣、自杀念头或自伤行为。

社交和人际关系问题：涉及孤独感、人际关系困扰、社交回避等。

身体症状：有时精神疾病患者可能报告身体上的症状，如头痛、胃痛、胸闷等，尽管这些症状在医学上无法找到明确的生理原因。

1. 脑部磁共振（Magnetic Resonance Imaging，MRI）

一种医学影像技术，用于获取详细的脑部结构和组织的高分辨率图像。它利用强大的磁场和无害的无线电波来生成具有较高空间分辨率的详细脑部图像。

1. 脑电信号（Electroencephalogram，EEG）

脑电信号是由大脑中的神经元之间的电活动产生的微弱电流所形成的。这些电流在大脑皮层中的神经元之间传递和交互，形成了复杂的电活动模式。脑电信号可以反映大脑中的神经元活动，并提供有关大脑功能和活动的信息。

（6）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]

* 1. 数据处理标准
     1. 脱敏

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。

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。

（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：

1. 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”，
2. 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。
3. 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
4. 用“\*\*\*”代替。
5. 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”。
6. 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。
7. 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，常用的去标识话方法举例如下：

1. 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。
2. 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。
3. 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”。
4. 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：

1. 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”。
2. 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。
3. 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。
4. 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。
5. 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。
6. 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。
7. 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：

1. 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。
2. 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。
3. 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。
4. 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。
5. 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。
6. 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”。
   * 1. 数据清洗

（1）清洗规则

1. 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。
2. 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。
3. 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。
4. 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。
5. 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。
6. 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。

（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。

* + 1. MRI数据处理

检查数据扫描质量，数据是否存在跳层，少层问题、存在伪影遮挡问题：检查数据扫描完整性，所要求的序列有无缺失、无超范围扫描。

MRI检查区域：头部

MRI设备情况：设备场强、名称及厂家：

线圈计数：通道线圈

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **MRI检查扫描参数** | | | | | | | | |
| 序列名称 | 检查方位 | TR（ms） | TE （ms） | 层厚/层间距 | 矩阵 | FOV | 翻转角 | 采集时间（s） |
|  |  |  |  |  |  |  |  |  |
| MRI扫描基础参数设定由各中心自行调试，如需协助可联系XXXX。 | | | | | | | | |

* + 1. 脑电数据处理

利用合法稳定的工具读取DICOM3.0协议内所携带的病人信息，参照法规要求，判定哪些信息为敏感字段，确定敏感信息后，利用程序将该敏感信息清洗。隐私保护程序采用隐私删除方法，通过编辑头文件隐私信息实现。

电生理采集区域：头部

电子图像存储格式： 囗DICOM 囗PNG 囗JPEG 囗其他

仪器设备：

仪器类型：

仪器厂家：

探头： 频率： MHz

成像模式： 机械指数：

采集时长