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

|  |
| --- |
|  |

C30 |

团体标准

T/CSBME 050—2022

人工智能辅助脓毒症预警数据处理标准

Artificial intelligence assisted sepsis early warning data collection standard

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：康焰

引言

尽管近年来国际上针对脓毒症(sepsis)做了大量的基础研究工作，并进行了有效临床干预措施的研发与推广，但遗憾的是，无论在发达国家还是发展中国家，脓毒症仍以其高发病率和病死率严重威胁着人类的健康及生命安全，消耗了大量的医疗资源。据不完全统计，全球每年有2-3千万患者罹患脓毒症，每天约有1400人死于脓毒症。因此，提高对脓毒症的认识，正确面对脓毒症所带来的巨大挑战，对改善脓毒症预后意义重大

脓毒症诊治难点在于疾病的异质性，即不同的脓毒症患者可表现为不同的临床表现以及对某种治疗的不同反应，这给临床诊疗和研究带来巨大挑战。脓毒症发生发展机制较为复杂，特别是急危重脓毒症患者，其临床诊治过程中会产生大量的数据，包括生命体征、呼吸力学、体液样本检验、影像学等数据，而人类直觉对于这些数据并不能很好地整合，因此可能无法及时发现潜在的风险。人工智能是用来研究、开发用于模拟、延伸和扩展人类智能的理论、方法、技术的一门新的技术科学。目前，人工智能已经在各行各业发挥了重要作用。在医学方面，人工智能技术在影像诊断、病理诊断中表现出了超过高年资医生的诊断水平，因此具有重要的应用前景。人工智能技术在脓毒症患者诊疗方面也具有较好的应用前景，虽然已经有大量文献报道应用人工智能技术来辅助脓毒症的诊断与治疗，但临床医师对于这类技术缺乏深刻认识。

本数据标准规范以脓毒症3.0标准作为主要参考依据，旨在形成共识性的脓毒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。

人工智能辅助脓毒症预警数据处理标准

* 1. 范围

本文件主要规定了脓毒症数据处理涉及的相关要求。

* 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

2014年1月，ESICM指定19位国际专家组成专家组启动了脓毒症3.0的修订，经过系统文献分析、Delphi调查法，以及脓毒症患者数据库的验证，将脓毒症定义为宿主对感染的反应失调而致的危及生命的器官功能障碍。脓毒性休克是脓毒症的一个亚型，是指脓毒症所致严重的循环障碍和细胞代谢异常，并足以使病死率显著增加。

脓毒症3.0强调致命性的器官功能障碍，专家组分析了SIRS、序贯器官衰竭评分(sequential organ failure assessment，SOFA)、Logistic器官功能障碍评分系统与感染性疾病预后的相关性；结果显示：与SOFA<2分比较，SOFA≥2分患者死亡风险增加2～25倍，同时，SOFA比Logistic器官功能障碍评分简单。因此，工作组推荐将SOFA≥2分作为脓毒症3.0的器官功能障碍标准，即脓毒症3.0的诊断标准为：感染＋SOFA≥2分。另外，符合SIRS标准，未出现器官功能障碍的感染患者将不被诊断为脓毒症。

* 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。