重症监护医学信息数据库隐私保护技术探讨

刘宁远 成福春 冯 佳 周蜜果 邵 茵 朱 亮

（上海中医药大学附属岳阳中西医结合医院 上海 200437）

〔摘要〕 介绍《健康保险流通与责任法案》（HIPAA）对隐私的定义以及去标识化过程和方法，从属性删除、日期平移、自由文本处理几方面阐述并分析遵循 HIPAA 原则的重症监护医学信息数据库（MIMIC－III）去标识化及脱敏技术规则制定、应用及改进方面。

〔关键词〕 隐私保护；重症监护医学信息数据库；重症监护数据库；去标识化；匿名技术

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2020.02.005

Discussion on Privacy Protection Technology of the MIMIC－Ⅲ Database *LIU Ningyuan, CHENG Fuchun, FENG Jia, ZHOU Miguo, SHAO Yin, ZHU Liang, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai 200437, China*

〔**Abstract**〕 The paper introduces the definition of privacy as well as de－identification process and method in *Health Insurance Portability and Accountability Act* (HIPAA), elaborates and analyzes the de－identification and formulation, application and improvement of desensitization approach rules of the Medical Information Mart for Intensive Care (MIMIC－III) database conforming to HIPAA principle in terms of attribute removal, date shifting, and free text processing.

〔**Keywords**〕 privacy protection; Medical Information Mart for Intensive Care (MIMIC－III); intensive care database; de－identification; anonymous technology

1 引言

重症监护医学信息数据库（Medical Information Mart for Intensive Care，MIMIC）是由麻省理工学院计算生理实验室建立的大样本、单中心危急重症监护数据库。目前 MIMIC 数据集包括 MIMIC－II 和 MIMIC－III，本文针对 MIMIC－III 进行分析。MIMIC－III 涉及患者生命体征、用药情况、护理记录、手术操作记录、检验结果在内共 26 张表[1]。随着信息化进程的加快，医疗系统中的数据呈爆炸式增长，但是目前这类系统在最初设计时没有考虑到医疗数据再利用问题，更多的只是满足医院收费和运营等日常工作需要，因此大部分医疗机构在现有临床数据库的基础上对医疗数据开展共享研究还缺乏系统、有效的手段。然而随着 MIMIC－III 等科研数据集的出现，更多的国内学者和临床医生对数据利用产生浓厚兴趣。先进的数据挖掘技术在提高信息使用率的同时也必然导致隐私泄露问题。MIMIC－III 对患者数据的隐私保护完全遵循《健康保险流通与责任法案》（Health Insurance Portability and Accountability Act，HIPAA）原则，采用多种技术手段

〔收稿日期〕 2019－07－25

〔作者简介〕 刘宁远，工程师，发表论文 1 篇；通讯作者：朱亮，博士，副教授。

对敏感信息进行匿名化处理，为满足临床科研需求的多样性，涉及的数据包括患者症状、诊断、用药、检查、检验、手术治疗等，既有结构化数据，也有自由文本、医学影像等非结构化数据。这些数据来自于不同信息系统，涉及不同来源的数据融合，患者数据的隐私保护成为科研数据利用分析和临床数据共享的关键[3]。

2 HIPAA 与隐私

2.1 隐私条款

美国关于隐私安全的立法较早，1974 年通过《隐私权法》保护公民个人信息的隐私权。1996 年美国通过著名的 HIPPA 法案，2003 年 HIPAA 中的隐私规则和安全规则生效。随后几年对其补充法案进一步发布，美国形成针对个人健康信息的隐私安全法律保护体系。HIPAA 分为不同部分，每个部分解决医疗保险改革中的一个独特问题。其中两个主要的部分是便携性和简化管理。便携性是指允许个人在调换工作时医疗保险不会因为工作变动而承保中断。简化管理这一部分是建立用于接收、传送和维护医疗信息的规则，确保隐私和个人身份信息的安全标准，这部分的焦点即是 HIPAA 中的隐私条款。

2.2 HIPAA 中对隐私的定义

HIPAA 中提出受保护的健康信息（Protected Health Information，PHI）概念，其定义为：主要由医疗服务提供商等适用主体或其商业伙伴持有或传输、以任何形式或媒体存在的可识别的个人健康信息。而可识别的个人健康信息是健康信息的一个子集，是指个人过去、目前和未来的生理和心理健康状况、医疗护理状况及与医疗护理相关的支付信息，这些信息至少包含法律规定的能够识别出个人的 18 项身份识别信息中的一项。法案规定向外提供 PHI 时必须遵循最小必要原则，即能不披露尽量不披露，以治疗为目的、向患者本人和依据患者意愿的披露除外。适用主体可以将去标识化后的数据提供给第 3 方，去标识化必须符合专家决定原则或者避风港原则。专家决定原则是指由行业内的相关专家决定哪些信息必须去除并且提供书面分析结果；避风港原则是指 18 项必须要去除的 PHI，见表 1。HIPAA 所指定去除的 PHI 是绝对的，但匿名手段并不唯一，因此需要最大限度地考虑到在不同科研需求及医疗环境下其数据所独有的研究价值，进而实行不同的匿名化方式。从另一个角度来看在国内开展临床科研数据集工作是否一定要遵循 HIPAA 原则也值得思考，如特殊的宗教信仰，敏感的检验、检查结果等，这些标识符属性或敏感信息是否有必要进行处理，都需结合实际情况再做决定。

表 1 HIPAA 隐私条例规定的 18 种 PHI

PHI 类型	分析
名字	可直接确认患者实体
位置	定义为所有州以下区域，包括街道、城市、辖区、邮编或其他等价的地理编码，国内应用时应考虑省市区县的泛化范围
日期	所有直接与患者相关的、除年的日期元素
年龄 >89 岁患者	因是极少数元素所以可间接猜测出患者实体
电话号码	可通过其他途径确认患者实体，从多个渠道获取不同匿名数据表，通过不同数据集的背景信息结合其余数据集推测出个体敏感信息
传真号码	同上
电子邮件地址	同上
社会安全号码	同上
病历号	同上
健康计划受益号	同上
账户号	同上
证书及证书许可号	同上
车标识符	同上
设备标识符和序列号	不限于医疗设备
统一资源定位符	可直接确认人员实体以及物理地址
IP 地址	同上
生物标识符	包括手指指纹和声音
其他唯一的标识性数字、编码和符号	如全脸照片、疤痕、刺青等图像

3 去标识化过程

3.1 概述

所谓去标识化，简单来说就是断开数据和个人

信息主体的关联。去标识化过程包括确定目标、识别标识、处理标识和导出数据，见图 1。经过处理后的数据必须保证可逆性，即通过严格授权机制可逆，以满足不同业务需求。首先确定去标识化对象，目标数据集中存在标识符属性时，根据事先决定的策略、法规标准、业务背景、数据用途等要素确定哪些数据属于去标识化对象[4]。确定目标后通过查表标识法、规则分析法和专家判断法对目标数据进行处理。

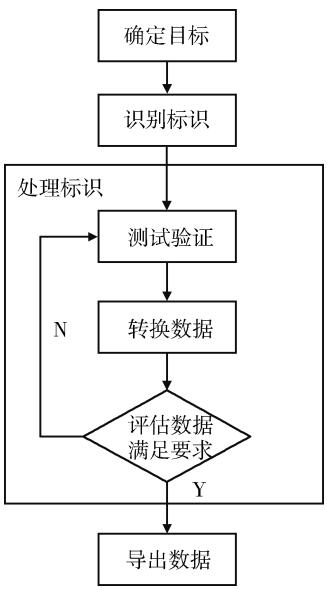

图 1 去标识化过程

3.2 查表标识法

预先建立对照数据表，存储需要去标识化的标识符元素，在识别标识数据时将需识别数据的各个属性名或字段逐个与对照表中的数据进行对比，再对其中的标识符属性做相应处理。查表标识法适用于数据集与目标标识符属性明确的关系型数据库，如已明确姓名与身份证号等属性。

3.3 规则分析法

通过建立规则，编写特定程序，对不同数据集使用不同算法从目标中自动发现需去标识化的属性。适用于非单一记录的自由文本元素，包括以自由文本形式记录的病史及检查报告等非结构化数据。

3.4 专家制定法

如字面所示即通过专家审查，人工发现和确定去标识化的数据。适用于有特殊含义或特殊值的表。

4 MIMIC – III 中去标识化技术分析

4.1 属性删除

在 MIMIC – III 数据集中标识符属性包括姓名、电话、住址、社会安全卡号等。这些数据主要集中在人员字典表之中，在发布前 MIMIC 均对表中的这类属性进行删除处理。这种方式看似简单粗暴，但能够有效阻止攻击者从发布的表中直接获取到患者隐私。然而进一步的研究表明仅仅使用属性删除这一单一技术手段并不能有效保证患者隐私不遭到泄露，这种仅对特殊属性进行处理的方式不能有效防范外界攻击者通过连接攻击等间接攻击形式对个体隐私信息进行获取[6]。随着信息技术的日益进步以及互联网的普及，攻击者对数据的获取途径也在不断增多。在对单个表进行攻击时攻击者很难获取到患者隐私，但是若攻击者将多个从不同渠道、不同系统中获取到的匿名表通过某些特殊字段相互关联，再经算法以及链接自身数据库后就极有可能推测出患者的隐私。因此数据集的发布不能只是用单一脱敏手段，必须综合使用多种技术手段对数据进行复合处理。

4.2 日期平移

4.2.1 时间处理 在 MIMIC – III 数据集中几乎所有业务表都包含时间这一属性，这些记录中的时间都是受保护的 PHI，然而这些时间信息对于临床研究及数据分析又是非常重要的元素，所以 MIMIC – III 数据集将所有日期按照每位患者的标识属性 SUBJECT_ ID 按规则进行平移，每个 SUBJECT_ ID 对应一个随机偏移量 N 来使日期元素迁移到未来的某个时间点，保留业务时间的连续性以及这一属性所独有的周期特性，从而在保证该元素的利用及分析价值同时也保护患者隐私。为保证日期在医疗数据中的分析挖掘价值，该随机数 N 有以下特征：N

是7 的倍数，使得转换后的日期与真实日期具有相同的工作日周期，允许以星期为单位对数据进行分析；当 N 转换为时间单位后应大于 1 个世纪，这样可避免转换日期和真实日期混淆，简化从记录中识别和去除遗留真实日期数据的任务；N 对于单个患者的所有就医数据都是相同的，但在患者之间是互不相同的。

4.2.2 年龄处理 MIMIC – III 数据集中删除患者年龄这一属性，但可以通过入院时间或者记录时间与出生时间的关联推导出业务发生时患者的年龄。当患者年龄 > 89 岁 MIMIC 会将其出生日期由入院日期向前调整 300 年，模糊处理以遵守 HIPAA 原则，这部分患者年龄中位数为 91.4 岁。研究者只需将 Patients 表中的出生时间和 Admissions 中的各类业务发生时间两两相减后便可确定患者的入院、出院及死亡年龄。总而言之，日期平移技术使得第3 方或攻击者无法直接界定患者当前真实年龄及行为发生时间这类特殊属性。从物理角度来看，时间连续平移对称性保证客观定律不会随发生时间改变而改变，即考察的时间不同，物理系统服从的规律相同。因此日期平移技术不会改变数据的有效性、准确性和时效性，而 MIMIC – III 对时间数据处理的特殊手段进一步细化其变量范围，使经处理的数据与其原始数据在属性上更为接近，从而允许研究者在季度特征、特殊时间节点上对数据进行有效分析。时间平移技术开销较少，对设备也无特殊要求，行之有效地顾及到适用主体的需求，值得在建设科研数据库的过程中加以借鉴。

4.3 自由文本去标识化

4.3.1 概述 在病史记录表 NOTEEVENTS 中保留着患者的详细病史、护理记录、检验检查报告及出院报告，这些文本信息记录在其 Text 属性之中，这些自由文本信息包含着大量的标识化内容，MIMIC – III 利用模式识别算法对这些数据进行遍历，本质上该算法适用于任何医疗文本。

4.3.2 屏蔽 模式识别算法遍历文本时根据空格进行分词，然后与已知受保护的健康信息查找表进行关键词比对，直接识别住院患者和医护人员的姓名。由于姓名误拼、昵称使用和探视人员姓名不在已知查找表内，还需与常用姓名、医院名称等潜在查找表做关键词匹配，识别潜在的命名实体[5]。得到的标识化内容被屏蔽替换后用“[]”与其他文本进行区分。

4.3.3 泛化 因 HIPAA 中明确规定超过 89 岁的年龄属于标识化信息，所以在文本中涉及 89 岁以上的年龄关键字也需处理，统一用 [* * Age over 89 * *] 代替，另外 MIMIC – III 对新生儿和 < 14 周的儿童也使用相同方法。这种方法称为泛化技术，简单来说泛化是将原始值划分进与其属性相似的一组值中，这组值存在一个范围，通过不同范围的划分可以有效地与表中其他数据区分，以满足去标识化要求。总体来看这种技术保证被泛化后的属性不会发生改变，不会对研究产生影响，更可通过泛化范围的约束和控制满足不同精度的研究需求，但在泛化范围上要小心取值避免造成过度泛化。记录案例如下：

nit No: [* * Numeric Identifier 69098 * *]

Admission Date: [* * 2172 – 9 – 22 * *]

Discharge Date: [* * 2172 – 10 – 19 * *]

HISTORY OF PRESENT ILLNESS: Baby boy [* * Known lastname 44129 * *] is a 31 and [* * 1 – 14 * *] – week boy born to a 27 – year – old G1/P0 (to 1) mother with [* * Name2 (NI) * *] type O +, antibody negative…

4.4 标准化建议

目前国外面向电子病历的命名实体算法已趋近完善，MIMIC 数据集利用模式识别算法识别病史文本中的命名实体以实现患者数据去标识化。其通过查找表内关键词对比、正则表达式和上下文检索的简单启发式算法来移除 PHI[5]。然而国内的病史文本语义识别较之国外还存在诸多挑战，包括电子病历文本的非规范性和专业性，医疗实体的独特性和标注语料的稀缺性，都会对识别算法的可靠度产生影响。在对文本进行关键词检索设计正则表达式时一定要将中文的特殊语法和句法、分词、命名实体的相互嵌套、跳跃、非连续性考虑在内。在对数值类型的 PHI 去标识时可参照 MIMIC 中的算法技术，通过正则表达式识别数字特殊字符，且该正则表达

式必须在识别数字的同时对其文本中所包含的医疗术语进行分析，若识别出包含代表检验、检查结果的关键字时，其中所含数字格式的文本就应被认为是临床数据而加以保留。

5 结语

医疗信息的合理使用与发布需要建立在完善的规则与流程之上，患者隐私保护只是其中一部分。目前国内相关规则制定部门开始注意到医疗大数据平台的构建需求，在中国医院协会发布的《医疗机构医疗大数据平台建设指南》中也建议参考 HIPAA 中的相关规则，但对具体匿名手段并无明确规定。所以完全遵循 HIPAA 原则的 MIMIC－III 数据集在目前阶段更具研究价值。

数据共享要求数据公开，数据公开化是否会导致恶意滥用，从而侵犯个人隐私值得关注。首先，数据共享默认数据的可及性、透明性和可读性；另一方面个人隐私总是要求被一种默认的非透明性所保护。这成了大数据时代存在的悖论。如今通过日益成熟的技术以及对需求环境的完善调研，学习先进案例的成功经验，可以确信隐私并不应该成为共享的对立面。医疗数据的发布需要受到限制，这是因为其所涵盖的内容能够直接辨识到患者主体，但也正是这些个体所独有的属性影响着医学研究成果，因此两者关系不能孤立看待。数据共享的正当与否要综合权衡该数据的使用场合及数据主体的知情权。医疗机构也需对整个发布过程给予适当的约束和安全防护，将信息安全作为数据利用的前提，将临床科研数据集建设成值得信赖的平台，实现隐私与共享、安全与利用之间的“共赢”。

参考文献

1 范勇，赵宇卓，李沛尧，等．危急重症监护数据库 MIMIC－Ⅲ疾病谱分析［J］．中华危重病急救医学，2018，30（6）：531－537.

2 陈静，李保萍．MIMIC－III 电子病历数据集及其挖掘研究［J］．信息资源管理学报，2017（4）：29－37.

3 Johnson AE，Ppllard TJ，Shen L，et al. MIMIC－III，a Freely Accessible Critical Care Database［J］. Sci Data，2016（35）：1－9.

4 谢安明，金涛，周涛．个人信息去标识化框架及标准化［J］．大数据，2017（5）：20－29.

5 郑西川．临床大数据应用系列［EB/OL］．［2018－08－20］．http：//www. hit180. com/32892. html. 2018.

6 姜宝彦．基于多属性泛化的 K－匿名算法的设计与实现［D］．大连：大连理工大学，2015.

7 王剑，张政波，王卫东，等．基于重症监护数据库 MIMIC－II 的临床数据挖掘研究［J］．中国医疗器械杂志，2014，38（6）：402－406.

8 刘英华．基于数据发布的隐私保护模型研究［M］．北京：中国人民出版社，2010：1－25.

9 陈磊．医疗数据隐私保护研究综述［J］．中国数字医学，2013，8（11）：95－98.

10 Peter B Jensen，Lars J Jensen，Soren Brunak. Mining Electronic Health Records：towards better research applications and clinical care［J］. Nature Reviews Genetics，2012，13（6）：395－405.

11 李开源，冯聪. MIMIC 数据库在急诊医学临床研究过程中运用的思考［J］．中华危重病急救医学，2018，30（5）：494－496.

12 王强芬．大数据时代医疗隐私层次化控制的理性思考［J］．医学与哲学，2016，37（5）：5－8.

13 Ross MK，Wei W，Ohno－Machado L. Big Data and the Electronic Health Record［J］. Yearb Med Inform，2014（9）：97－104.

14 Johnson AE，W. Stone，David J，et al. The MIMIC Code Repository：enabling reproducibility in critical care research［J］. Journal of The American Medical Informatics Association，2018，25（1）：32－39.

15 Benitez K，Malin B. Evaluating Re－identification Risks with Respect to the HIPAA Privacy Rule［J］. Am. Med. Inform. Assoc，2010（17）：169－177.

16 Sage，April. Physical Security，HIPAA，and the HHS Wall of Shame［J］. J Healthc Prot Manage，2014，30（1）：85－90.

17 王国臣，高波. HIPAA 法案对医疗机构的作用［J］．国外医学（医院管理分册），2001（4）：158－159.

18 吴兴华．数据共享与隐私权保护［J］．山东科技大学学报（社会科学版），2017，19（4）：9－14.