基于差分隐私的医疗大数据隐私保护模型应用研究

侯梦薇① 卫荣① 兰欣① 邢磊① 那天① 陆亮①

摘 要 目的：随着医疗信息化应用的深入发展，以及医疗大数据挖掘、医疗大数据分析等深层次应用的普及，如何在利用医疗大数据的同时保护好患者的隐私数据，防止其敏感信息泄漏具有十分重要的意义。方法：差分隐私是一种严格且可被证明的隐私保护方法，近年来的研究使其在理论层面不断发展完善，并在数据挖掘、机器学习、推荐系统等领域得到了初步的应用。结果：在对医疗大数据领域的常用隐私保护技术进行综合叙述的基础上，对差分隐私保护技术的基本原理和研究方向进行了阐述。结论：针对不同类型医疗大数据的应用研究做了相应介绍，指出差分隐私技术存在的研究难点，最后展望了其在医学大数据隐私保护领域未来的发展方向。

关键词 医疗大数据 差分隐私 隐私保护 数据发布
Doi:10.3969/j.issn.1673-7571.2019.12.028
[中图分类号] R319 **[文献标识码] A**

Research on Privacy Protection Model and Application of Medical Big Data Based on Differential Privary/ HOU Meng-wei, WEI Rong, LAN Xin, et al//China Digital Medicine.-2019 14(12): 86 to 88

Abstract Objective: With the further development of medical informatization applications, and the popularization of deep-level applications such as medical big data mining and medical big data analysis, how to protect the privacy data of patients while using medical big data to prevent their sensitive data from leaking has important practical significance. Methods: Differential privacy is a strict and proving method of privacy protection. In recent years, research has made it develop and improve at the theoretical level, and has been applied in statistics, machine learning, data mining and other fields. Results: Based on the comprehensive narrative of commonly used privacy preserving technologies in the field of medical big data, this paper describes the basic principles and research direction of differential privacy protection technologies, and presents a corresponding introduction to the application of different types of medical big data. Conclusion: This paper points out the difficulties in the research of differential privacy technology, and finally looks forward to its future development direction in the field of medical big data privacy preserving.

Keywords medical big data, differential privacy, privacy protection, data releasing

Corresponding author Information Technology Department, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, Shaanxi Province, P.R.C.

信息技术的飞速发展使各类数据的发布、采集、分析和存储变得更加方便快捷。在实际的医疗工作中，许多医疗机构因教学、科研等目的，需经常大量收集和发布各类医疗敏感数据。医疗数据的扩散可能会造成患者隐私信息的泄漏，因此仅靠政策法规对医疗机构进行约束远远不够[1]，必须运用面向医疗大数据的隐私保护技术保护患者的隐私，促进我国医疗行业的健康发展。

差分隐私是Dwork[2]在2006年提出的一种针对敏感数据集发布导致的隐私泄露问题的隐私保护模型。基于这一模型，处理后的数据集对任意一个记录的变化是不敏感的，因此一个数据记录在数据集中是否存在对于统计计算结果的影响非常小。攻击者无法通过观察计算结果而获取准确的个体信息，因为一条记录加入数据集而产生的隐私泄露风险被控制在可接受的范围内。

本研究在目前存在的医疗信息隐私保护研究进展和现状进行介绍的基础上，对差分隐私的理论发展、主要研究方向及其在医疗大数据隐私保护问题上的应用进行总结梳理，为未来的研究提供参考。

1 医疗大数据的隐私保护

1.1 医疗大数据中的隐私风险 医疗大数据具有高容量、高速度和多类型的特点，研究人员通过数据挖掘、数据

①西安交通大学第一附属医院网络信息部，710061，陕西省西安市雁塔西路277号

分析等技术对大量医疗数据进行分析和研究，但这也随之带来了隐私泄露的问题。在大数据概念出现之前，大部分隐私保护方法是针对小数据的，而针对小数据的隐私保护方法在被应用到医疗大数据的隐私保护时存在着很大的局限性。因此，隐私保护在医疗大数据时代将面临更大的挑战[3]。

1.2 医疗大数据隐私保护模型 医疗数据集中通常包含着许多患者隐私信息，如医疗诊断结果、处方信息、检验检查报告等。一方面，如果数据持有者不采取适当隐私保护技术而直接将这些数据进行发布，会造成患者的隐私泄漏；另一方面，Netflix用户隐私泄漏等一系列案例[4]表明去除标示符的操作无法保证医疗隐私信息的安全。

如何从医疗数据集中提取有价值信息而又不泄露患者隐私是医学隐私保护的关键问题[5]。针对这一问题，研究者们提出了各类算法保护患者的隐私信息，这些算法和他们的隐私标准被称为隐私保护模型，如图1所示。

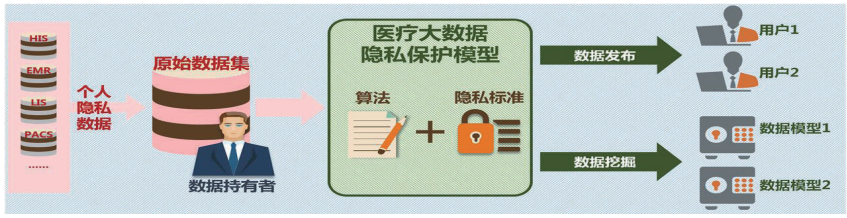

图1 医疗大数据隐私保护模型

2 差分隐私保护

研究人员试图找到一种稳定性高的面对医学大数据的隐私保护模型，能够抵抗拥有最大背景知识的攻击者的攻击。结合医疗大数据本身的特点，差分隐私的提出相当程度上满足了上述所有要求。

差分隐私保护模型的思想源自一个朴素的观察[6]：假设有一数据集D，其中包含David个体，对D进行查询操作f（例如求和、求平均值、求中位数等）所得到的结果为f(D)。如果将David的信息从D中删除后得到数据集D'，对D'进行查询的结果仍为f (D)，则可以得出结论，David的信息并没有因为攻击者反复查询而产生暴露的风险。差分隐私的形式化定义如下。

定义1. 设有随机算法M，P_M为M所有输出构成的集合。对于任意两个邻近数据集D和D'以及P_M的任何子集S_M，若算法M满足：

$$P_r[M(D) \in S_M] \leq \exp(\varepsilon) \times P_r[M(D') \in S_M] \quad (1)$$

则称算法M满足 ε –差分隐私保护，参数 ε 为隐私保护预算。差分隐私的核心就是保证任意一个个体在数据集中存在或不存在对最终的统计查询结果几乎没有影响。具体来说，假设有两个几乎完全相同的数据集（仅有一条记录不相同），分别对这两个数据集进行查询访问时，同一查询语句在两数据集产生同样结果的概率比值近似为 1。

表1 艾滋病诊断数据集示例

姓名	诊断结果
Bob	0
Tom	0
Jane	1
Harry	0
David	1

例如，表1为一个医疗数据集样例D，其中记录为1时代表该患者患有艾滋病，记录为0时代表该患者没有艾滋病。数据集在不泄露具体数据集中记录值的前提下可以为用户提供某些查询统计服务。假设用户输入参数i，调用查询函数f(i)=count(i)获得数据集前i行中所有诊断结果为1的记录行数，并反馈给用户。当攻击者想要推测David是否患有艾滋病，且攻击者已知David位于记录的第5行，则可用count(5)–count(4)推测出结果。

如果DP是一个满足 ε –差分隐私保护算法的查询函数，即DP(i)=f(i)+noise，公式中noise是服从某种随机分布的噪音。假设DP(5)可能的输出来自集合{1, 1.5, 2}，那么DP(5)也会以几乎相同的概率输出{1, 1.5, 2}中任一值，使攻击者不能通过DP(5)–DP(4)得到想要的结果。这种方式使攻击者无法获得查询结果之间的差异，从而保护所有个体的隐私。

3 差分隐私在医疗大数据领域的应用

差分隐私作为近年来的研究热点，理论研究日趋完善[7]。但是由于医疗大数据本身包含许多特殊属性，在引入新的隐私保护机制时必须对这些特殊属性加以考虑，目前的研究与应用多集中于基因组隐私保护、电子健康档案隐私保护以及医疗传感器隐私保护。

3.1 基因组隐私保护 随着DNA测序技术的普及和发展，人们迫切需要对大量DNA序列进行联合比对分析的方法。全基因组关联分析（GWAS）是当前研究遗传信息和疾病之间的关联性时大量使用的一种分析手段，然而由于基因组序列包含敏感信息，基因组序列的发布可能会威胁到个人隐私。Fienberg等人[8]研究了如何在不影响个人隐私的情况下对GWAS进行控制并获得平均的次等位基因频率（MAFs）。作者在文章中针对计算 ε –差分隐私中的χ^2统计和P值并添加Laplace噪音到原始统计信息中，并允许发布加入噪音的统计结果以获取最相近的单核苷酸多态性（SNP）。Raisaro等人[9]针对基因数据的群组探测，提出将同态加密算法和差分隐私

相结合的方法，使研究人员可以使用基因组数据进行研究同时保护了患者的个人隐私。

3.2 电子健康档案隐私保护 差分隐私在电子健康档案数据中的应用主要是面向人口统计学信息或诊断信息。Mohammed等人[10]提出一种针对人口统计学信息的非交互式差分隐私方法，实验结果证明该算法保持了分类的准确性，且可扩展性和性能都优于现有的分类算法。Chen等人[11]展示了如何对涉及一系列有相互关系的诊断数据的计数查询进行隐私保护操作。针对差分隐私在关系型数据上可伸缩性不足的缺陷，文章提出一种基于概率的自顶向下分割算法，实验结果表明此方法在大型关系数据集上保持了很高的应用价值。此外，针对那些允许从电子健康档案中对汇总信息进行不同的数据发布的系统，Gardner等人[12]提出了一种满足差分隐私的新型统计健康信息发布系统，并在真实的电子健康档案数据集中证明了系统的可行性和实用性。

3.3 医疗传感器隐私保护 可穿戴传感器采集的大数据通常包含患者敏感信息，如物理环境、位置信息等，必须得到保护。Lin等人[13]提出一种针对医疗传感器大数据的差分隐私保护方案，引入动态噪音阈值，使该方案更加适合大数据的隐私保护。针对用户隐私保护的计算开销，Lin[14]应用哈尔小波转换方法将直方图转换为完整的二叉树，实验结果表明树形结构大大降低了用户隐私保护的计算开销。

4 总结与展望

随着电子健康档案、可穿戴式监测设备、转化医学和基因研究的兴起，越来越多的个人敏感信息连入互联网，其利益影响之重大，面临威胁之严峻，已经超过了传统的信息安全和隐私保护研究范畴。越来越多研究者开始关注如何有效地保护医疗大数据中的个人隐私[15]。差分隐私作为一种新型隐私保护框架，逐渐应用于医学大数据隐私保护领域。与此同时，差分隐私是一个相对年轻的研究领域，在理论和应用上都还存在着一些需要解决的问题，归纳为以下三点：①差分隐私在保证患者隐私数据安全的同时，也会制约医疗数据可用性，因此如何同时兼顾隐私安全和数据可用性一直是差分隐私的主要研究方向；②隐私预算是差分隐私实现机制的关键要素，如何针对不同应用的隐私保护要求合理地分配隐私保护预算也是一个研究热点；③医疗数据集拥有很大的数据量，加入噪音之后结果集必然非常巨大，如何提供合适的结果集后置处理函数或过滤函数，减少数据量并提升数据的可用性将是一个非常重要的研究方向。

参考文献

[1] 赵新蓉，赵韡.大数据背景下的患者隐私安全挑战[J].中国数字医学，2016，11(8)：13-15.

[2] Dwork C.Differential Privacy[J].Lecture Notes in Computer Science，2006，26(2)：1-12.

[3] Blond SL，Zhang C，Legout A，et al.I know where you are and what you are sharing:exploiting P2P communications to invade users'privacy[C]//Berlin，Germany，2011-11.

[4] 陈敏，牟海燕，秦健.健康医疗大数据标准体系框架研究[J].中国数字医学，2018，13(4)：14-16，33.

[5] 许培海，黄匡时.我国健康医疗大数据的现状、问题及对策[J].中国数字医学，2017，12(5)：24-26.

[6] 朱天清，何木青，邹德清.基于差分隐私的大数据隐私保护[J].信息安全研究，2015，1(3)：224-229.

[7] 康海燕，马跃雷.差分隐私保护在数据挖掘中应用综述[J].山东大学学报：理学版，2017，52(3)：16-23，31.

[8] Fienberg SE，Slavkovic A，Uhler C.Privacy Preserving GWAS Data Sharing[C]//IEEE，International Conference on Data Mining Workshops.IEEE Computer Society，2011：628-635.

[9] Raisaro JL，Choi G，Pradervand S，et al.Protecting Privacy and Security of Genomic Data in i2b2[R].Institute of Electrical and Electronics Engineers，2017.

[10] Mohammed N，Chen R，Fung BCM，et al.Differentially private data release for data mining[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.ACM，2011：493-501.

[11] Chen R，Mohammed N，Fung BCM，et al.Publishing SetValued Data via Differential Privacy[J].Vldb，2012，4(4)：1087-1098.

[12] Gardner J，Xiong L，Xiao Y，et al.SHARE：system design and case studies for statistical health information release[J].Journal of American Medical Informatics Association Jamia，2013，20(1)：109-116.

[13] Lin C，Song Z，Song H，et al.Differential Privacy Preserving in Big Data Analytics for Connected Health[J].Journal of Medical Systems，2016，40(4)：97.

[14] Lin C，Wang P，Song H，et al.A differential privacy protection scheme for sensitive big data in body sensor networks[J].Annals of Telecommunications，2016，71(9-10)：465-475.

[15] 杭长山.大数据安全与隐私保护[J].电子技术与软件工程，2017(10)：205.

【收稿日期：2018-09-06】
【修回日期：2018-11-09】
（责任编辑：郑艳）