健康医疗大数据时代的隐私保护探析

王 爽 尹聪颖

（美国印第安纳大学信息计算工程学院 美国布卢明顿 47408） （HC3i 中国数字医疗网 北京 100190）

〔摘要〕 以基因数据为例，全面分析健康大数据隐私面临的挑战，从联盟数据分析、同态加密、硬件加密、差分隐私几方面探讨隐私数据保护策略，阐述有关数据安全和隐私保护法律建设，以期为相关研究提供参考。

〔关键词〕 健康医疗大数据；数据隐私；隐私保护

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.01.001

Discussion and Analysis on the Privacy Protection in the Age of Big Data in Healthcare *WANG Shuang*, *School of Informatics*, *Department of Computing and Engineering*, *Indiana University*, *Bloomington*, *IN*, *U S A*, *47408*; *YIN Congying*, *HC3i. cn*, *Beijing 100190*, *China*

〔**Abstract**〕 The paper, by taking genetic data as an example, conducts a comprehensive analysis on the challenges confronting the privacy of healthcare big data, discusses the protection strategies of private data from such aspects of analysis on alliance data, homomorphic encryption, hardware encryption and differential privacy. It also elaborates on the building of laws related to data security and privacy protection, and provides references for relevant study.

〔**Keywords**〕 healthcare big data; data privacy; privacy protection

1 引言

随着大数据时代来临，人口健康与医疗数据面临着日益严峻的安全挑战。卡内基梅隆大学 Latanya Sweeney 教授于 2000 年发表的《简单的人口统计往往能识别出人的独特性》[1]报告指出少数特征的组合常常结合在一起即可唯一地识别某些个体。美国选举人公共注册信息中 87% 的基于 5 位邮编、性别、出生日期即有可能被唯一识别出个人身份；53% 通过地点、性别、出生日期可能被唯一识别出个人身份；在县一级，18% 通过县、性别、出生日期可能被唯一识别出个人身份。显然上述个人信息的数据字段是不应该被公开的，因为这有可能泄露个人隐私。如该研究曾使用麻省总医院的出院数据和选举投票的注册数据进行匹配，最终链接出某麻省议员的住院信息。

〔收稿日期〕 2019－01－16

〔作者简介〕 王爽，博士，助理教授，主要研究方向：生物医疗信息、数据隐私与安全、大数据分析、机器学习、高性能计算、数据压缩，先后在国内外发表论文 90 篇，据 2019 年 1 月 Google Scholar 的搜索结果论文被累计引用 1 262次，领导开发的独特的基于 SGX 硬件技术的大规模基因安全数据分析模型于 2016 年获得 Intel 杰出贡献奖；尹聪颖，硕士，HC3i 中国数字医疗网副主编。

2 健康大数据隐私挑战——以基因数据为例

2.1 概述

健康医疗大数据在全球快速发展，越来越多的个人数据被脱敏后公开，用于精准医学等各类大数据研究。然而如上述报告所述健康医疗数据的公开

或将引出一系列隐私安全问题。

2.2 脱敏分享的隐私安全问题

健康医疗大数据时代，大量医疗数据被不断采集。人们往往认为一组医疗数据将名字、身份证信息去掉后便安全，可以公开使用。然而当这组数据跟另一组数据连在一起时可能会完全暴露个人隐私。如果加入基因数据，隐私安全威胁会更加明显。随着基因检测技术发展，只需大概 75 个统计上独立的单核苷酸多态性（Single Nucleotide Polymorphism，SNP）位点即可唯一确定一个人[2]，所以说基因数据比指纹数据更敏感。当基因检测数据与一些病理数据相结合时很容易匹配到具体个人，这种确认会侵犯人类隐私。数据脱敏是指对数据中包含的秘密或隐私信息（如个人身份识别信息、商业机密数据等）进行数据变形处理，使得恶意攻击者无法从经过脱敏处理的数据中直接获取敏感信息，从而实现对机密及隐私的防护。

2.3 个人隐私与基因联想

基因与个人隐私之间的关系十分微妙。2018 年美国警方通过一家名为 GEDMatch 的家谱网站上一名亲戚的遗传信息确认到 40 年前的金州杀手案罪犯身份[3]。这一手段运用到医学信息上，如果已知某人的基因就能知道此人是否得过某种疾病。如艾滋病人去参加癌症或糖尿病的研究，只提供自身基因信息不公开其他信息，获得信息的人对患者基因在公共数据库中进行比对就能够获得其个人信息，进而获得其患有艾滋病的信息，造成个人隐私风险，可能损害个人权益。如果雇主知道雇员是糖尿病患者，可能会因怀疑其能否参加重体力劳动而解雇该雇员。如果保险公司通过基因检测知道参保人有较大的重疾可能性，就会降低保额，增加保费，甚至拒绝提供保险服务。

2.4 基因数据安全威胁家族

包括基因在内的健康医疗数据快速增长，随着相关应用的不断展开，人类隐私安全威胁日益严峻。其中基因数据关系到的不只是一个人，而是整个家族。而且基因数据十分“强健”，即便将基因上某个位点去掉，还是可以通过其他基因来确认。而用户到商业化的基因测序公司进行测序服务后，公司有可能将数据卖给药厂或其他公司，用于药品研发或其他用途。如 2018 年知名医药企业葛兰素史克（GSK）与商业化基因检测公司 23andMe 达成 3 亿美元的股权投资交易，后者将利用已有的 500 万名用户数据为 GSK 提供 4 年独家合作[4]。这种做法不仅会暴露个人隐私，还可能导致家族隐私暴露。哈佛大学做过一项调查，称 92% 的美国人不愿意公开基因数据，因为子孙后代的信息都有可能会被公开。

2.5 云共享安全隐患多

原始的基因数据非常大，一个人的基因测序数据约有 300GB，精准医学要做上百万人的基因数据分析，量非常大。不可能在每个机构或医院都建立超级计算中心，因此美国医疗机构或科研院校将数据放在公有云上，但问题较多，存在很多隐私安全风险，因为公有云中的计算资源是被很多用户共享的，数据在计算和存储的过程中还会存在备份操作，不加以保护的数据安全无法得到有效控制。

3 隐私数据保护策略

3.1 概述

健康医疗大数据的巨大潜力吸引无数医疗机构、科研团体积极探索，一边是数据带来的隐私安全“黑洞”，一边是精准医学打开的未来医疗世界大门，隐私安全保护与数据公开应用能否兼得？医疗大数据隐私保护的基本方法，见图 1。

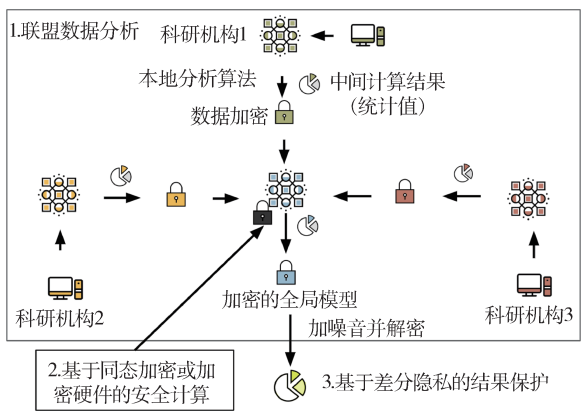

图 1 医疗大数据隐私保护的基本方法

3.2 联盟数据分析[5]

如果多个医疗机构联合进行医疗或者基因数据的分析，其实是不需要直接交换个体数据的，只需要通过交换统计值就能得到想要的结果。分享统计值可以有效降低数据分享过程中的隐私风险。如学习一个模型需要看某些参数，这些参数代表疾病的高风险性与年龄、性别或其他特征的关系，可以将从每个医疗机构内部个体数据得到的统计值整合成一个全局的模型参数，然后将对应的模型分享给研究人员，训练模型的原始数据并不需要分享给其他医疗机构或研究人员。

3.3 同态加密[6]

如果是基于公有云做数据运算，为提高安全等级可以选择同态加密。同态加密是级别非常高的一项加密手段，可以在加密数据上做加密运算，得到的结果也是加密的，只有授权的用户才能拿到加密的结果。这样既能使用户放心使用公有云资源，又能保护个人隐私等数据安全。目前基于正则长度方程（Regularized Long Wave Equation，RLWE）的同态加密安全等级非常高，根据已知的研究成果，即使是量子计算实现后都不能破解。在可预见的未来医疗数据应用中都是非常安全的。

3.4 硬件加密[7]

硬件加密是利用英特尔第 6 代之后的 CPU 芯片的一个加密区域，有授权的用户可以访问。所有数据在硬件外都是加密的，非授权用户看不到。目前在圣地亚哥 Rady 儿童医院、伦敦帝国学院、新加坡基因研究所 3 国联合开展的川崎病研究项目中，由于 3 个国家对于基因数据隐私保护的要求不同，项目数据传输、分析是通过硬件加密的方式实现。两位教授领导的团队在世界范围内最先实现在加密硬件上大规模进行带有隐私保护的基因计算研究，基于 Intel 芯片进行硬件加密满足不同机构、国家对于基因数据隐私保护的要求。该项工作获得 Intel 杰出成就奖，被多个权威学术期刊（包括《科学》杂志）引用。

3.5 差分隐私[8]

如果只是做一些前期探索性研究，研究者并不需要原始数据，只需要与原始数据相似的数据信息即可。具体做法是在原始的数据上添加噪音，或者先在原始数据上拟合出一个分布，然后在这个分布的空间中再抽象出数据。这个数据会与原始数据很相像，但是没有任何点能够对应到原始数据。使用这种数据去开展研究，无从得知数据具体来源。

4 各国数据安全和隐私保护法律建设

4.1 概述

随着健康医疗大数据应用的深入，更多隐私安全挑战正在涌现，需要更加先进的隐私安全保护技术和方法帮助应对大数据可能带来的困扰。因此美国和欧盟一方面加强相关数据安全法律建设，另一方面也在积极鼓励细分领域的科技创新。

4.2 美国

美国在数据安全方面的法律建设起步较早，1996 年发布《健康保险流通与责任法案》（Health Insurance Portability and Accountability Act，HIPAA/1996，Public Law 104 - 19），公布个人健康信息的隐私保护标准和实施指南，明确要求医疗数据的安全等级和脱密方式。美国数据安全研究组织还在推进基因研究、数据安全两大领域人才的跨界交流，以探索更加先进的基因安全保护技术，如组织全球基因安全保护竞赛。作为竞赛联合发起人，笔者对于全球基因安全保护技术发展深有体会。最初参赛队伍的数据模型因为尺寸不合适，不能用到基因上，现在各参赛队伍已经能够在成熟的模型上不断提高。自 2014 年开展至今全球对于基因安全的意识都在提升，据悉目前全球有超过 100 个队伍参与其中，包括斯坦福大学、麻省理工大学、微软公司、IBM 公司等。该项竞赛多次被国际权威媒体报道，包括 Nature News 和 GenomeWeb 等。

4.3 欧盟

2018 年 5 月欧盟正式开始实施《一般数据保护条例》（Genral Data Protection Regulation，GDPR），旨在加强对欧盟境内居民的个人数据和隐私保护。该法律加大数据隐私泄露的处罚力度，其中最高达 2 000 万欧元，或企业 1 年营业额的 4% 的罚款。可见各国都在不断加强对于医疗数据隐私保护的重视程度。

4.4 中国

中国也在不断加强对于隐私保护的力度，如 2017 年 6 月颁布并实施的《中华人民共和国网络安全法》中明确规定在未获得知情同意前，网络运营者不得向第 3 方提供个人信息，也不得擅自泄露、篡改、毁损其收集的个人信息。同时该法律也提到经过脱敏处理的数据，如果无法被用来识别特定个人信息的情况除外。但是该法律并没有像美国的 HIPAA 法案一样提供详细的规定，以指导数据收集方如何生成可以满足条件的脱敏数据。作为法律的补充，2018 年 5 月颁布的《信息安全技术个人信息安全规范》对个人信息收集、保存、使用等各个环节提出具体要求。但是其中并没有提出专门针对医疗大数据标识化处理的条款。在该规范中与医疗相关的数据都被定义为个人敏感数据，收集和使用前需要获得个人知情同意，除非以下 3 种情况：一是与公共安全、卫生、重大公共利益直接相关的数据；二是出于维护个人信息主体或其他个人的生命、财产等重大合法权益，但又很难得到本人同意；三是个人信息控制者为学术研究机构，出于公共利益开展统计或学术研究所必要，且其对外提供学术研究或描述的结果时，对结果中所包含的个人信息进行去标识化处理的。关于标识化处理，中国于 2017 年 8 月发布《信息安全技术个人信息去标识化指南》，其中描述个人信息去标识化的目标和原则，提出去标识化过程和管理措施，对常用的脱敏方法进行介绍。

5 结语

健康医疗大数据时代，单纯依赖政策的保护、技术的革新实现个人隐私保护是不够的。未来医疗将是全民主动参与的时代，每个人都是数据的提供者、使用者和受益者。只有主动提升隐私安全保护意识才能更有效地保护个人权益，在健康医疗大数据背景下获得数据赋予的健康收益，真正实现个人对隐私的掌控。

参考文献

1 L Sweeney, Simple Demographics Often Identify People Uniquely [J]. Health, 2007 (671): 1 - 34.

2 Z Lin, A B Owen, R B Altman, Genomic Research and Human Subject Privacy [J]. Science, 2004, 305 (5681) 183.

3 A Selk. The Ingenious and "Dystopian" DNA Technique Police Used to Hunt the "Golden State Killer" Suspect [EB/OL]. [2018 - 04 - 20]. https://www.sfgate.com/news/article/The - ingenious - and - dystopian - DNA - technique - 12871539.php.

4 药明康德. GSK 投资 3 亿美元与 23andMe 合作研发创新疗法 [EB/OL]. [2019 - 01 - 14]. http://med.sina.com/article_detail_109_2_49495.html.

5 C - L Lu, S Wang, Z Ji, et al. WebDISCO: a web service for distributed cox model learning without patient - level data sharing [J]. J. Am. Med. Inform. Assoc., 2015, 32 (2): 1212 - 1219.

6 S Wang, Y Zhang, W Dai, et al. HEALER: homomorphic computation of ExAct Logistic rEgRession for secure rare disease variants analysis in GWAS [J]. Bioinformatics, 2016, 33 (6): 211 - 218.

7 F Chen, S Wang, W Dai, et al. PRINCESS: privacy - protecting rare disease international network collaboration via encryption through software guard extensionS, Bioinformatics, 2017, 33 (6): 871.

8 M Wang, Z Ji, H Kim, et al. Selecting Optimal Subset to Release Under Differentially Private M - estimators from Hybrid Datasets [J]. IEEE Trans. Knowl. Data Eng., 2017 (99): 1.