生物医学大数据中的隐私计算

辛均益 陈如梵 王 林

（杭州医学院 杭州 310059）（杭州锘崴信息科技有限公司 杭州 310053）（济南大学 济南 250022）

唐丹叶 孙 琪 沈 涛 王 爽

（杭州锘崴信息科技有限公司 杭州 310053） （检科测试集团有限公司 北京 100176） （1 杭州锘崴信息科技有限公司 杭州 310053 2 四川大学华西医院 成都 610041）

〔摘要〕 分析生物医学数据流通难点、传统隐私保护手段缺陷，介绍隐私计算优势及其主流技术路线，阐述生物医学中隐私计算的典型应用场景和有关案例，为该领域相关研究提供参考。

〔关键词〕 隐私计算；生物医学数据；数据流通；联邦学习；可信执行环境

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.10.001

Privacy－preserving Computing in Biomedical Big Data *XIN Junyi, Hangzhou Medical College, Hangzhou 310059, China; CHEN Rufan, Hangzhou Nuowei Information Technology Co. Ltd., Hangzhou 310053, China; WANG Lin, Jinan University, Jinan 250022, China; TANG Danye, SUN Qi, Hangzhou Nuowei Information Technology Co. Ltd., Hangzhou 310053, China; SHEN Tao, CAIQTEST Group, Beijing 100176, China; WANG Shuang, 1Hangzhou Nuowei Information Technology Co. Ltd., Hangzhou 310053, 2West China Hospital, Sichuan University, Chengdu 610041, China*

〔**Abstract**〕 The paper analyzes the difficulties of biomedical data sharing and the defects of traditional privacy protection methods, introduces the advantages of privacy－preserving computing and its mainstream technology route, and expounds the typical application scenarios and related cases of privacy－preserving computing in the biomedical field, so as to provide references for related research in this field.

〔**Keywords**〕 privacy－preserving computing; biomedical data; data sharing; federated learning; trusted execution environment

1 引言

近年来随着新技术不断发展与深度融合应用，生命科学和医学领域数据规模正在迅速扩增。以新一代测序技术应用为例，每台高通量的测序仪每天可产生约 100 GB 的基因组测序数据。在此背景下，生命科学与医学基础研究正在从实验科学向数据、人工智能驱动的新方向发展。生物医学数据的分散分布与存储为大范围数据应用带来全新挑战。如果说大数据是生命科学和医学研究的重要基础，那么数据共享便是形成大数据的必要手段。医学领域积累了海量数据，但存储相对分散，且敏感度高、类型多样，导致较为严重的“数据孤岛”问题。同时，不同医院、不同机构之间存储的数据也存在异构、非独立同分布等问题。为了最大程度利用相对

〔收稿日期〕 2022－08－28

〔作者简介〕 辛均益，副院长，高级工程师；通信作者：沈涛，高级工程师。

分散的数据以服务各项生物医学应用，进行数据共享十分必要，同时需将数据资源有效且安全地连接起来。

生物医学研究中患者信息的隐私问题是不容回避的现实挑战。医疗数据涉及患者隐私，具有特殊性和敏感性。2021 年我国颁布的《中华人民共和国个人信息保护法》[1] 中将医疗健康信息视为敏感信息，医疗行业中大量患者相关信息均会进入敏感个人信息的范畴，此类信息包含诊疗过程中的病历信息、不良反应报告信息、临床试验数据信息等。因此，在大数据应用环境中保护好个人医疗相关隐私数据至关重要。尽管有相应监管要求，医疗数据的合规有序流通问题依然严峻。患者级明文数据过于敏感，不能直接共享，但生物医学数据开放合作势必为未来疾病防治带来重大影响和突破，如何解决这两者之间的冲突一直以来都是生物医学数据开放方面的一个难点。针对这种两难的境地，本文探讨采用隐私计算安全技术对相关数据进行处理，在共享数据的可实施性和数据隐私安全之间进行平衡。

2 医疗数据流通合规难点

2.1 监管政策

针对医疗数据合规性的监管相对分散，缺乏统一标准。这一点在跨境（域）医疗数据流动上尤为明显。我国针对医疗数据跨境（域）监管要求分散在不同监管部门和法律法规中，这些法律法规在不少环节上都有重叠。这意味着对于同一数据处理主体，当涉及医疗数据跨境（域）时，很可能同时受到多种法规监管且这些法规之间各不相同，很难同时满足所有要求。此外，国内很多法律法规都提到当涉及医疗数据跨境（域）时，需要进行安全评估，然而针对医疗场景的跨境（域）数据安全评估的具体量化标准仍有待明确，这无疑增加了医疗数据跨境（域）流动难度。

2.2 技术实现

传统的隐私保护手段不够完善，很难满足现行法律对于隐私安全的要求，导致医疗数据流动困难。以往在科研过程中会利用数据脱敏手段对需要共享的数据进行处理，以保证科研过程中数据及隐私的安全性。然而数据脱敏已被证实存在一定漏洞。1997 年的一项研究显示[2]，尽管通过传统数据脱敏技术，剔除了可以直接识别患者身份信息的识别符，然而脱敏后的数据结合公开的数据库仍然可以重新识别某些个体的身份信息。为证明这并非个例，Sweeney 团队进行一项范围更广的研究[3]。结果显示通过邮编、生日、性别的信息求交集，63%～87% 美国人可被唯一确认。此后 Emam 团队的研究则进一步证实，即便依照美国《健康保险携带和责任法案》（Health Insurance Portability and Accountability Act，HIPAA）中安全港（Safe Harbor）脱敏方法[4]，通过类似的背景信息求交集，大概每 15 000名个体中就有两名可以被重新识别[5]。类似的风险也存在于我国医疗数据中，一项有关我国患者数据隐私风险评估的研究显示，我国患者脱敏后数据的重识别风险约为 0.01%，与之前国外研究结果中的比例相当。这项研究也是目前我国最大规模的患者数据隐私风险评估。这意味着，数据脱敏不等于匿名化，尽管经过处理的数据隐私泄露风险降低，但仍具有一定的泄露可能性。隐私泄露导致多方面的负面影响，如果所泄露的是基因数据，其伤害还会蔓延至患者的血亲及整个家族。《中华人民共和国数据安全法》[7]《中华人民共和国个人信息保护法》[1] 等法律法规严格规范了数据的安全流动。其中《中华人民共和国个人信息保护法》规定“匿名化”是指个人信息经过处理无法识别特定自然人且不能复原的过程。因此数据脱敏并不等价于数据匿名化。《中华人民共和国个人信息保护法》将匿名化的信息排除在“个人信息”概念以外，以便促进数据利用与流通。此外，数据脱敏这类隐私保护手段还具有一定的使用场景局限性，无法处理特定类型数据，如基因数据。基因数据具有唯一性，通过十几个或几十个基因位点就能够识别出个体的身份[8]，然而这些位点又有很重要的研究价值，简单地剔除这些位点虽然能够实现隐私安全保护，但却破坏了基因数据完整性，研究价值大幅降低。

3 隐私保护的3 个阶段

3.1 数据脱敏

数据脱敏是通过数据消隐、泛化、置换、扰动等方法对原数据中的隐私信息进行处理，以便降低数据中敏感信息的过程。数据脱敏的主要目的是保护个人身份信息、敏感的个人数据或业务数据。在美国 HIPAA 法案下，数据脱敏主要包括两种方法，一种是专家模式，具有数学或者统计学背景的专业人士，通过统计学的方法来移除数据中的敏感信息。但这种方法无法提供明确的数据脱敏规则，在实际应用中具有很大局限性。另一种被广泛采用的方法是安全港，其规定 18 种识别符，通过移除这 18 种识别符号来降低身份及敏感信息泄露风险。数据脱敏的难点是脱敏后既要保护数据隐私，又要保持数据可用性。数据脱敏处理过程主要包括敏感数据识别、制定脱敏策略、敏感数据处理、审计评估等阶段。除之前所述的脱敏数据被重新识别风险以外，脱敏数据在共享过程中仍以明文传输，存在被复制后转卖或超范围使用甚至滥用的风险，可见脱敏数据无法满足生命科学和医学研究发展要求。

3.2 数据安全沙箱

数据安全沙箱是一个隔离的数据处理环境，用户在一个受控的环境内使用数据进行分析，并在权限范围内获得数据分析的相关结果，而非原始数据，从而一定程度上解决数据共享过程中泄露和权限控制的问题。此外，安全沙箱提供隔离运行环境，对于恶意代码有防护作用，可以一定程度上避免对数据源方其他系统的影响。但是在多中心合作过程中，沙箱也面临着新的挑战，例如怎样找到一个多中心共信的沙箱进行数据汇聚计算。

3.3 隐私计算

隐私计算是融合密码学、系统安全、机器学习和硬件安全技术来有效解决数据共享过程中隐私安全保护的新兴技术。隐私计算可以支持多个参与方在不泄露各自数据和机密信息的前提下，联合参与计算和数据共享，以实现联合建模和分析、隐私查询、隐私求交等目的[9]。随着人工智能和大数据的普及，隐私计算已经成为数据安全和数据价值之间的桥梁和安全底座。利用隐私计算技术，数据变得“可用而不可见”，实现原始数据无需共享下的多中心合作，多主体间的数据安全合规交换，平台内数据分级授权使用，确保数据安全可控。

4 隐私计算的主流技术路线

4.1 联邦学习

机器学习和统计分析模型的效果跟训练数据集的数据量、数据特征和数据质量有着密切关系。较大的数据样本量、丰富的数据维度和高质量的数据往往是生物医学研究中的前提条件。但由于生物医学数据存储往往比较分散，分布在不同机构的不同系统中，且受限于数据共享法律和伦理限制，存在“数据孤岛”问题。联邦学习即是为了解决这个问题而提出的，其主要思想是多个数据源（如医院、医疗研究机构等）共同参与 1 个全局模型的构建，每个数据源使用本地数据进行计算并生成相应的本地模型统计信息（如模型的本地梯度），全局节点通过收集到的各数据源本地梯度进行全局模型的更新并返回更新后的模型到各数据源，通过反复迭代直到全局模型收敛。联邦学习使得全局模型的效果能够与将数据共享集中起来训练的模型效果基本一致，甚至在某些场景下具有更强的鲁棒性。有学者[10]在 2012 年提出医疗在线安全联邦学习构架，有效解决了多中心生物医疗数据协作过程中“数据可用不可见”和隐私保护问题。在联邦学习执行过程中数据仍保留在各数据源，各方仅提供梯度或模型中间结果，即便如此，该类中间信息仍可通过一定计算反推获得部分原始数据信息，存在隐私信息泄露风险。因此有的场景会综合利用同态加密、多方安全计算、可信执行环境等多种隐私计算技术融合的方法来解决此类问题。

4.2 多方安全计算

多方安全计算[11]是多个数据拥有者根据各自私

有数据联合计算，确保每方获取联合计算结果，且不泄露己方数据。多方安全计算为解决多中心环境下的信息安全协作共享提供一种新的计算模式，对多中心数据安全具有重要价值。但多方安全计算在执行过程中的计算量和通信量十分庞大，对于网络带宽有限、算法复杂或数据量较大的任务场景具有一定局限性。

4.3 同态加密

传统数据运算需要在明文上进行，在数据拥有者需要将计算任务外包时，相应数据也需要交给任务执行者，这带来巨大的安全隐患。与一般关注数据存储安全方案不同的是，同态加密主要解决的是数据处理安全问题。同态加密可以简单概括为实现密文上的特定运算，且运算结果通过解密后与明文计算的结果相同。这使得许多外包或多方协作场景对数据安全有了进一步的保护，通过传输加密数据、在加密数据上执行运算得到加密结果、传输加密结果、解密获取结果这一流程实现不泄露数据隐私的协作运算。尽管这一思想非常直观简洁，但是目前同态加密算法只能执行部分操作[12-13]，对一些复杂的非线性计算仍需要消耗巨大的计算资源，离实际应用落地还有很长的路要走。

4.4 可信执行环境

可信执行环境（Trusted Execution Environment，TEE）通常指存在于计算机硬件上（如 CPU）的一块特定隔离区域（如 Enclave），这块区域可以给数据和代码的执行提供一个安全的空间，以保证机密性和完整性。可信执行环境可以在不信任操作系统、虚拟机、基本输入输出系统（Basic Input Output System，BIOS）等底层资源的支持下，为特定区域内的数据和技术提供安全保证。可信执行环境实现了安全性和可用性之间较好的平衡，然而其安全性依然在一定程度上依赖硬件设计的信任。一些研究表明可信执行环境也面临着例如测信道攻击等相关问题[14]。

5 医疗大数据跨境安全共享实现思路

医疗大数据包含海量个人隐私信息，在进行医疗数据合规共享时需要保证其全生命周期数据安全，需要对数据使用者和管理者操作行为进行规范[15]。目前国际上对于个人健康医疗数据跨境流动的专门标准并不多，2004 年国际标准化组织发布的《健康信息学推动个人健康信息跨国流动的数据保护指南》提出，除保护数据主体切身利益所必要的传输之外，个人健康数据不应传输，除非得到数据主体明确的授权。因此在使用医疗数据过程中需理清数据种类是什么，数据在哪，谁在用，如何用，实现数据保护、数据形式合规与实质合规。对我国而言，一方面，可通过完善医疗数据合规跨境的制度体系，推动面向国际的数据流动；另一方面，还可探索数据跨境试点。此外，可参考欧盟及其他国家经验，设立符合我国国情的多样化合法流动机制，以及指引性的数据跨境流动协议范本。在形式合规方面，医疗数据安全合规体系的建立必须围绕医疗数据全周期的运行开展。相关企业和机构应首先确保根据本国法律法规要求，进行数据分类分级。此外，建设完备的医疗数据跨境人员保障制度同样必不可少，应形成由决策层、管理层、执行层、监督层及协同层构成的组织结构，同时加强人员培训，通过配置检查和旁站式验证确保相关制度落实。在实质合规方面，相关方必须通过采取必要的技术措施，如隐私计算技术，确保形式合规的充分落实和执行，使跨境医疗数据在“可用不可见”模式下，实现可管、可控、可计量的共享。通过结合区块链技术，实现数据使用全流程可溯源、可追踪，确保医疗数据跨境过程中的有效保护和合法利用，并使其处于持续安全的状态，避免数据出境及再转移后被泄露、毁损、篡改、滥用等。一个完善的医疗数据隐私保护平台应具备 3 项核心功能，包括隐私合规多维的检测、智能数据的分级分类与精细化安全管控，以及隐私计算与区块链结合。

6 医疗场景下隐私计算典型案例

6.1 多中心隐私数据分析、建模

在实践中，单一中心样本量和样本维度往往难以支撑一项研究的进行，因此需要多家机构、中心

合作以增加样本量、丰富数据维度。为了保护患者隐私，相关法律法规严格限制生物医学数据的不安全流动，这也意味着需要将明文数据物理聚合的传统集中式计算不再适用于医疗场景。联邦学习等隐私计算技术可以在明文数据不出域的前提下，实现带有隐私保护的跨中心数据协作、分析和建模，既满足日益严格的隐私保护要求，又能促进多方数据协作。有研究团队运用联邦学习等隐私计算技术开发带有隐私保护的多中心全基因组关联分析技术框架——iPRIVATES，用于强直性脊柱炎的研究[16]。利用 iPRIVATES，通过分布式、联邦式计算方式，实现在不分享明文数据（个体级基因数据）前提下，构建多中心临床辅助诊断模型。研究结果显示，分布式计算下算法耗时和靶点特征等价于物理聚合数据的集中式计算。同时，由于分布式计算可以触达更多数据源，该项研究中包含跨省的多家三甲医院、高校数据，其样本量和数据维度相比于集中式计算更高，因此研究得出的结论更可靠、准确且更具有普适性。另一个关于多中心隐私数据建模的案例则聚焦于罕见病。由于罕见病的特性，在临床诊断中往往容易被误诊，例如该案例中研究的儿童川崎病，其早期症状并不典型，很容易与普通的呼吸道感染等疾病混淆，导致错失最佳治疗时机。为解决这一问题，需要联合更多数据对疾病进行分析，找出致病的基因位点，提供更多诊断依据，使医生能够在临床诊断中及早发现。该研究团队联合国际川崎病联盟，实现跨 3 国的多中心儿童川崎病数据联合分析[17]。这项研究中仍然使用安全联邦式计算结合可信执行环境，在各数据源处部署本地计算节点，完成本地计算后，通过交换加密的中间统计信息，构建更为精确的全局模型。由于明文数据不出域，在保证患者隐私信息的同时也满足各国不同的隐私政策要求。研究结果显示其所使用的技术框架可以在分钟级完成跨多个国家的联合数据分析，且结果等价于集中式计算。

6.2 多中心目标条件隐私查询

随着医院信息化程度的加深，临床辅助诊断系统的普及率也越来越高。然而这其中存在一些隐私安全隐患。在临床辅助诊断系统中，查询条件、被查询数据源、匹配信息、匹配结果等都是高敏感度数据，不带隐私保护的明文查询很有可能导致患者身份及敏感信息泄露。一项由复旦大学附属儿科医院发起的研究项目中，将隐私计算应用到儿童罕见病查询网络中，实现目标条件隐私查询。研究团队利用 TEE 等多种隐私计算技术保证查询过程中的数据隐私安全。查询过程中，通过自然语言处理，提取出查询条件，在儿童罕见病网络中进行相似患者匹配，返回患者可能潜在存在的罕见病报告。其中，用户输入的查询条件在加密状态下与加密数据库进行匹配，生成加密的报告，为所有敏感数据和信息提供全流程隐私安全保护。

6.3 带有隐私保护的医疗临床数据库

医疗临床数据库或专病数据网络的构建有助于提高科研效率，进一步挖掘临床数据价值，为临床医生提供更多真实世界证据、辅助临床决策。通过隐私计算构建带有隐私保护的医疗临床数据库或专病数据网络能够解决其中存在的隐私安全隐患。在全球范围内，结直肠癌（Colorectal Cancer，CRC）已成为仅次于肺癌和前列腺癌的男性第 3 大常见恶性肿瘤，也是女性中仅次于乳腺癌的第 2 大常见恶性肿瘤。人工智能的发展为 CTC 临床信息预测、避免过度治疗或治疗不足带来新机遇。临床信息可以是肿瘤亚型、复发、死亡率和疾病进展。中华医学会消化外科结直肠癌学组将隐私计算技术应用到类似医疗临床数据库的搭建中，实现全国范围内带有隐私保护的结直肠癌数据共享。该项目支持每家医院单独管理各自数据及密钥，但不同医院之间能在隐私计算技术的支持下进行跨院的联合数据统计、分析等，兼顾隐私保护和数据共享的双重目标。同时，系统使用的 TEE 环境为英特尔软件防护扩展技术（Intel SGX）。该技术是一种基于硬件的隔离和内存加密机制，对执行的代码和数据提供针对软件和物理攻击的额外保护，其可信计算基仅包括处理器硬件和应用程序代码本身。通过提供一组扩展指令集，可以将敏感信息及其处理逻辑放置在安全的

飞地环境，防止系统管理员、操作系统等对信息的篡改和访问。目前该数据库已覆盖全国 24 个省、60 余家三甲医院，为相关领域的科研项目提供高质量的数据样本支持，也进一步推动相关防治、干预措施的发展进程。

7 结语

生物医学大数据的共享和联合分析已经成为现代医学研究中必不可少的环节，隐私计算通过技术手段为医疗数据价值在安全可控前提下的共享提供有效的安全保障，推动医疗数据价值最大化利用。此外，在医疗领域广泛推动隐私计算应用和数据安全共享，有助于提升医疗资源利用率。展望未来，隐私计算也将推动精准医学加速发展，助力医疗健康行业迈入全新发展阶段。

参考文献

1 中国人大网. 中华人民共和国个人信息保护法 [EB/OL]. [2022 - 07 - 30]. http://www.npc.gov.cn/npc/c30834/202108/a8c4e3672c74491a80b53a172bb753fe.shtml.

2 Sweeney L. Weaving Technology and Policy Together to Maintain Confidentiality [J]. The Journal of Law, Medicine & Ethics, 1997, 25 (2-3): 98-110.

3 Sweeney L. Simple Demographics Often Identify People Uniquely [J]. Health (San Francisco), 2000 (671): 1-34.

4 Dwyer S J, Weaver A C, Hughes K K. Health Insurance Portability and Accountability Act [J]. Security Issues in the Digital Medical Enterprise, 2004, 72 (2): 9-18.

5 El Emam K, Jonker E, Arbuckle L, et al. A Systematic Review of Re-identification Attacks on Health Data [J]. PloS One, 2011, 6 (12): e28071.

6 Gong M, Wang S, Wang L, et al. Evaluation of Privacy Risks of Patients' Data in China: Case Study [J]. JMIR Medical Informatics, 2020, 8 (2): e13046.

7 中国人大网. 中华人民共和国数据安全法 [EB/OL]. [2022 - 07 - 30]. http://www.npc.gov.cn/npc/c30834/202106/7c9af12f51334a73b56d7938f99a788a.shtml.

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

9 魏明月，陈敏，胥婷，等. 临床科研场景下医疗数据安全开放共享平台设计 [J]. 中国数字医学，2021，16 (7)：27-32.

10 Wang S, Jiang X, Wu Y, et al. Expectation Propagation Logistic Regression (EXPLORER): Distributed Privacy-Preserving Online Model Learning [J]. Journal of Biomedical Informatics, 2013, 46 (3): 480-496.

11 Micali S, Goldreich O, Wigderson A. How to Play Any Mental Game [C]. New York: Proceedings of the Nineteenth ACM Symp. on Theory of Computing, ACM, 1987.

12 Gentry C. A Fully Homomorphic Encryption Scheme [M]. Palo Alto: Stanford University, 2009.

13 Wang S, Zhang Y, Dai W, et al. HEALER: Homomorphic Computation of Exact Logistic Regression for Secure Rare Disease Variants Analysis in GWAS [J]. Bioinformatics, 2009, 32 (2): 211-218.

14 Wang W, Chen G, Pan X, et al. Leaky Cauldron on the Dark Land: Understanding Memory Side-Channel Hazards in SGX [C]. Dallas: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017.

15 肖瑗，卢雅雯，吕智慧，等. 生物医疗大数据隐私安全保障机制研究 [J]. 计算机应用与软件，2021，38 (2)：2.

16 Wu X, Zheng H, Dou Z, et al. A Novel Privacy-preserving Federated Genome-wide Association Study Framework and Its Application in Identifying Potential Risk Variants in Ankylosing Spondylitis [J]. Briefings in Bioinformatics, 2021, 22 (3): 1-10.

17 Chen F, Wang S, Jiang X, et al. Princess: Privacy-protecting Rare Disease International Network Collaboration via Encryption through Software Guard Extensions [J]. Bioinformatics, 2017, 33 (6): 871-878.