生物医疗大数据隐私安全保障机制研究

肖 瑗[1] 卢雅雯[1] 吕智慧[1] 吴 杰[1] 祖立军[2]

[1](复旦大学计算机科学技术学院 上海 200433)

[2](中国银联股份有限公司 上海 201201)

摘 要 在大数据产业发展的背景下,医疗卫生领域也开始探索生物医疗大数据的新用途、新价值。随着生物医疗大数据在临床治疗与科学研究中的应用,相应的数据安全隐患也随之出现,其隐私安全保障问题开始引起人们的重视。为了降低隐私泄露的风险,加强人们对生物医疗数据的保护意识,需要从数据的生命周期角度出发,在采集、存储、访问、应用、共享、销毁这些阶段,对生物医疗大数据的操作、管理行为进行规范,并初步搭建了一个大数据云平台来实现电子数据的安全保障。

关键词 大数据 生物医疗 隐私安全 数据生命周期

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2021.02.051

PRIVACY SECURITY MECHANISM OF BIOMEDICAL BIG DATA

Xiao Ai[1] Lu Yawen[1] Lü Zhihui[1] Wu Jie[1] Zu Lijun[2]

[1](*School of Computer Science, Fudan University, Shanghai 200433, China*)

[2](*China UnionPay Co., Ltd., Shanghai 201201, China*)

Abstract With the development of the big data industry, the medical and health field has also begun to explore new uses and new values of biomedical big data. With the application of biomedical big data in clinical treatment and scientific research, the data security risks have also emerged, and the issue of privacy security has begun to attract people's attention. In order to reduce the risk of privacy leakage and strengthen people's awareness of biomedical data protection, it is necessary to regulate and manage biomedical big data in the stages of data life cycle: data collection, data storage, data access, data application, data sharing, and data destruction. Finally, we built a big data cloud platform to achieve the security protection of digital data.

Keywords Big data Biomedical Privacy security Data life cycle

0 引 言

随着现代社会高速的信息化和网络化,各种应用、服务、网络中产生的数据与信息都在以爆炸式的速度增长。现在关于大数据的研究已是人们耳熟能详的话题,大数据的收集、开发和利用,已经成为当今社会的潮流之一。事实上,大数据的分析应用对于政府或企业的决策有着非常积极的作用。

现在,生物医疗大数据给医疗卫生领域带来了深刻的变革,其被广泛应用在领域内的各个方面,包括电子病历、决策支持系统、远程医疗、个人健康管理、精准医疗[1]等,蕴藏着巨大的医疗价值和科研价值。我国对于生物医疗大数据的发展也很关注,颁布了各类文件支持生物医疗大数据的基础建设。

生物医疗大数据在为人们提供高效、便利服务的同时,也带来了一系列的挑战[2]如隐私安全保障问题。相较于过去,大数据时代下的生物医疗大数据泄露的后果更为严重。例如,个人的身体缺陷、疾病情况,甚至是基因缺陷,都可能会使其在投保险求职时受到不公正对待。并且,随着医疗信息系统的普及,患者就医时被采集到的医疗信息包含了详细的个人信息。数据

收稿日期:2019-08-05。国家重点研发计划项目(2017YFC910101);上海市青年科技英才扬帆计划项目(17YF1425800)。
肖瑗,硕士生,主研领域:计算机应用技术。**卢雅雯**,硕士生。**吕智慧**,副教授。**吴杰**,研究员。**祖立军**,高工。

泄露后,基于个人基本信息,可关联到主体在金融、通信、交通等领域的信息,从而带来严重的经济、精神损失。

为了保障隐私安全,本文梳理了生物医疗大数据研究背景和保护现状,并以生物医疗大数据的生命周期为基础,对生命周期中各个阶段的隐私安全保障行为进行规范。同时,基于 OpenStack 搭建了一个大数据云平台来保障电子数据在云上的安全性。

1　大数据时代下的生物医学

1.1　生物医疗大数据的研究背景

医疗卫生领域每年都会产生海量的生物医疗数据,其数据规模可达到 TB 或 PB 级别[3]。这些生物医疗数据可被简单地分为两类:用于临床医疗的医疗数据和用于科学研究的生物数据。其中用于临床医疗的医疗数据主要为患者的诊疗档案,包括了患者的个人基本信息、诊疗信息、影像报告、治疗方案、药物使用信息、手术记录、住院信息等。而用于科学研究的生物数据则包含了基因数据、生物样本、实验记录等。

通过对生物医疗大数据的收集、处理和分析,医疗人员的相关决策获得了海量历史数据的支持,疾病预防和诊疗的效率得到了提升。此外,生物医疗大数据还可用于疾病预防、药物研究、基因分析、疫情监测、人体保健等领域。

但是,随着生物医疗大数据平台和技术的发展,相关隐私泄露事件频发。生物医疗大数据的隐私安全问题面临着重大的挑战。医疗卫生行业的特殊性以及生物医疗数据的敏感性要求人们在快速发展生物医疗大数据的同时,也要加大对生物医疗信息隐私保护的重视。

1.2　国内生物医疗大数据保护现状

我国对于生物医疗大数据的发展也是十分关注,有关生物医疗大数据的文件政策也是层出不穷。2014 年,卫计委颁布了《基于电子病历的医院信息平台技术规范》《基于居民健康档案的区域卫生信息平台技术规范》等文件。2015 年,国务院颁布了《关于城市公立医院综合改革试点的指导意见》,并在《促进大数据发展行动纲要》中指出要在健康医疗领域全面推广大数据应用,构建以人为本、惠及全民的民生服务新体系[4]。2016 年,国务院颁布了《关于促进和规范健康医疗大数据应用发展的指导意见》;中国信息通信研究院颁布了《大数据白皮书(2016)》,并在其中描述了医疗领域大数据应用的进展情况及发展趋势。2017 年,国家开始施行《中华人民共和国网络安全法》,将对信息安全的保护由行政法规层面逐步上升到了法律层面。2018 年,国家卫健委颁布了《国家健康医疗大数据标准、安全和服务管理办法(试行)》,对生物医疗数据的标准管理、安全管理、服务管理、管理监督四个方面进行了规范。2019 年,《互联网个人信息安全保护指南》正式发布,明确规定了个人信息的管理机制、技术措施、业务流程和应急处置办法,进一步加强了个人信息的安全保护。这些文件内容覆盖了医院信息化、医药信息化、数据融合等领域,为生物医疗大数据的建设提供了强有力的支持。

2　生物医疗大数据生命周期下的数据安全保障

生物医疗大数据中涵盖了大量的个人隐私信息,为了降低隐私泄露的风险,需要对数据使用者和管理者的数据操作行为进行规范。本节以生物医疗大数据的生命周期为线索,对生命周期各个阶段的数据安全保障进行研究并给出建议。

从生物医疗视角出发,基于张静[5]对于大数据生命周期的定义,将生物医疗大数据的生命周期分为数据采集、数据存储、数据访问、数据应用、数据共享、数据销毁这六个阶段。

2.1　数据采集阶段

数据采集阶段是大数据生命周期的第一个阶段。在这一阶段,个人的生物医疗数据被采集,为未来的数据处理和分析奠定了基础。

个人生物医疗数据的采集手段繁多,包括个人资料填写、医生就诊问询、医疗设备收集、医学研究志愿者自愿提供等。获取的内容主要有个人基本信息、个人医疗信息和生物数据样本。其中:个人基本信息包括姓名、电话号码、家庭住址、婚姻状况等信息;个人医疗信息包括病情、药方、过敏史、患病史等信息;生物数据样本包括血液样本、基因样本、生物组织样本等。收集到的数据和样本会被用于数据主体的临床治疗或医疗相关的科学研究。

收集数据时需要获得数据主体的知情和同意。在获取数据或生物样本时需要以文字形式告知数据主体获取的方式、内容和用途。若在获取时不确定数据是否具有后续用途,需要获得数据主体的动态知情同意,即每次数据用于新的用途之前,就要向数据主体说明,再次获得数据主体的同意。必须要在获得数据主体知情和同意的前提下才可以进行数据的采集工作,在数

据主体不同意的情况下不应当采集数据或生物样本。

在数据的采集过程中应当遵循最小化原则,避免收集无关目的的隐私数据,即收集的数据的类型和数量应与获取目的有直接关联。同时,收集隐私数据应有特别提示,在以书面或网络形式获取数据时需标明是否为隐私数据,以及必填/非必填项。

数据采集时也需要对数据进行简单的处理,包括对采集到的数据进行核对与矫正。对于生物样本,则需要及时贴好标签,做好相应的标识以便与其记录进行关联。

进行数据采集的人员需要进行管理。其中涉及隐私数据采集的人员需要经过隐私数据安全培训,并签订安全保密协议。接触数据的人员不得篡改或记录数据,不得保留数据备份、部分或全部生物样本。

2.2 数据存储阶段

数据被采集后需要根据相应的要求进行存储。海量的数据被集中存储和管理,这要求我们保障数据存储环境的安全性。

首先,需要明确存储的对象。存储的数据对象包括以纸质、网络、医疗器械等方式采集到的生物数据和医院、医疗相关研究机构获取的生物样本。存储的目的是为数据主体后续治疗或后续患病治疗提供参考,也会为相似病例治疗提供参考,部分会成为科研病例的素材。

其次,不同介质的数据会有不同的存储手段。重要纸质材料应有专门房间妥善保存。经过采集和录入的数据应存储在数据库中,存储数据的服务器及其备份服务器等应放置在可靠安全的环境里。生物样本应存放在适宜的环境下。

然后,存储的数据也需要进行一定的处理。在隐私保护方面,需要对姓名、身份证号等关键追溯性信息做脱敏处理,对隐私数据设置隐私标记。信息安全技术个人信息安全规范中有规定:收集个人信息后,需要立即进行去标识化处理,并采取技术和管理两方面的措施,将去标识化后的数据与可用于恢复识别个人的信息分开存储,并确保在后续的个人信息处理中不能重新识别个人[6]。同时,数据保存应遵从时间最小化原则,即个人信息保存期限应为实现目的所必须的最短时间,超出个人信息保存期限后,应对个人信息进行删除或匿名化处理。

另外,从数据管理的角度来看,数据存储方应建立专门的数据管理系统来对获取的生物数据进行管理。为数据管理系统所处网络划分不同的网络区域,并按照方便管理和控制的原则为各网络区域分配地址[7];对存储数据的数据库网络进行防火墙等隔离手段,保证网络隔离;定期进行数据备份(本地及异地),并做好容灾方案。

最后,在管理人员方面,数据存储方应为数据管理系统、存储数据的机房、保存生物样本的房间分配相应管理者,并明确其责任范围。管理者需要经过隐私数据保护培训并签订数据保护协议。同时需要建立数据管理制度体系,其中包括安全策略、管理制度、操作规程等[7]。

2.3 数据访问阶段

经过采集、存储阶段后,生物医疗数据已经可以支持简单的医疗诊断行为,例如患者的医疗数据被医护人员访问查询以便于治疗方案的确定和实行。为了保障患者的隐私,降低信息泄露风险,访问行为需要被约束和控制。

在访问手段上,由于生物医疗数据在物理意义上可分为电子数据、纸质文档、生物样本,所以访问手段也相对多样。电子数据可以通过数据管理系统访问,也可以直接访问数据库;纸质文档和生物样本则需要直接接触和翻阅。

对于电子数据,应对访问人员进行访问控制和安全审计。访问数据管理系统需要有合法身份,通过其身份对应权限进行访问。访问系统的合法身份在获取其身份及对应权限前需要了解涉及隐私数据类型,并签订协议。非系统内人员如有正当理由需要访问系统,需要进行审批,获得临时身份后访问。

对于物理数据,访问机房或存储纸质文档、生物样本的房间需要进行审批,并对人员进出进行记录。

2.4 数据应用阶段

应用阶段是生物医疗大数据产生价值的重要阶段。在这一阶段,海量的数据被处理、分析和解释,能有效地辅助医疗领域的决策制定。

应用数据的整个过程中都要获得数据所有者的明示同意,即数据所有者对其个人数据的处理做出明确授权的行为,包括书面声明等。在应用数据前,应获得数据主体的明示同意。在应用数据的过程中,应用范围不得超出数据收集过程中所声称的范围,若超出上述范围,需再次征得数据主体的明示同意。对收集的数据进行加工处理后产生的新数据应被认为是数据主体的生物数据,所以对新数据的使用也应获得数据主

体的明示同意。

在数据应用的过程中,需要对数据的操作、管理行为进行约束,有专人基于数据应用相关规章制度监管数据应用过程,负责数据使用的申请和审批。应消除数据中与研究目的无关的信息,使数据无法追溯到主体;应采取权限控制技术,使不同领域的人员仅获取其领域所需的生物医疗数据,降低数据窃取的可能性;应保障对数据进行分析、挖掘后产生的新数据[8]的安全性;应控制数据的流动,限制数据的使用范围,使数据不进入保险、保健等盈利行业;应保障数据可视化过程中的安全性,使个人的信息不被公开泄露。

2.5　数据共享阶段

随着共享信息平台的建立,各行各业都开始尝试进行数据的共享,医疗行业也不例外。特别在临床医疗领域内,患者通常会到不同的医院治疗疾病,这时,个人生物医疗数据就可以在不同的医院中进行共享。此举消除了数据的孤岛,让医生的诊断决策有更坚实的基础。数据共享在为医疗领域带来便利的同时也增加了隐私泄露的可能性。数据共享双方都应严格规范自身的数据共享行为,防止恶意人员获取共享数据。

共享的数据内容主要包括各医疗机构之间相互协作进行临床治疗所需的医疗数据和各科研机构用于医学研究所需的生物数据。数据共享的方式主要包括线下方式的数据共享;基于共享数据库的在线数据共享;基于请求和反馈的数据共享。

为了保障传输过程中的数据安全,出于研究目的传输的数据应进行匿名化、去标识化处理,让数据无法追溯到个体;线下共享数据时应采取措施保证传输过程的安全性;线上传输数据时应采用文本、图像加密等技术保证数据的完整性和保密性。

在使用共享数据时,共享数据接收方应将共享数据与接收方原有数据隔离存储,并基于最小授权原则对接收的数据进行访问控制、提供身份鉴别服务,也应对共享数据的操作行为进行安全审计,并保留审计记录。

需要有第三方机构对共享数据发送方和共享数据接收方的数据共享行为进行管理。第三方机构应制定数据共享相关规章制度和文件,执行并落实相关的管理制度,监管数据共享行为。

共享数据双方应配合第三方机构的指导和监管,遵循数据共享的相关流程规定,不应私自进行数据共享。共享数据发送方应保证发送数据的真实性,不得篡改数据;当共享时限到达后,共享数据发送方应检验共享数据接受方是否归还共享样本、是否删除共享数据。共享数据接收方人员不得私自获取、复制、更改、存储共享数据;当共享时限到达时,共享数据接收方应归还共享样本并删除共享数据。

2.6　数据销毁阶段

数据销毁阶段是数据生命周期的最后一个阶段。所有的数据都有时效性,收集到的生物数据在患者康复、研究结束或数据到达保存期限后应被销毁。

数据的销毁可分为数据删除和实物销毁。数据删除对应于电子数据的删除,需要采取一定的措施防止他人通过技术手段恢复存储设备中的生物数据,例如乱码数据覆盖、设备格式化等。实物销毁对应于纸质文档、存储设备、生物样本的销毁,其中:存储设备应采取永久消磁或彻底销毁手段进行处理;纸质报告应进行粉碎处理;样本应按照相应规章制度进行处理。

各机构需要在数据保存期限到达后销毁数据。其中:对生物数据进行处理、计算后的衍生数据应设有保存期限,到期后应删除;共享数据到期后应删除;生物样本等实物到达保存期限或不能使用后应销毁;数据主体在机构违反法律法规或与数据主体的约定时,要求机构销毁个人数据时,机构应删除数据并销毁对应实物;机构停止运营后应删除所有生物数据并销毁对应实物。

各机构的数据销毁人员也应遵循相关的规定,不应保留、复制销毁数据,应检查销毁结果,若有遗漏则再次销毁。

3　平台实现

本文基于 OpenStack 初步建立了一个大数据平台,提供了电子数据的安全保障管理环境。

本平台为每个业务系统建立专用的虚拟资源空间,使之在相对隔离的环境中可信、高效地运行,并可按需灵活调整。具体功能包括:平台物理资源调度和管理、虚拟运行环境的自动化配置和交付、平台性能监测和优化等。

除此之外,为了保障平台和平台中各系统的信息安全,如图 1 所示,本平台采用基础平台安全防护,基于虚拟化的安全隔离、安全初始化及交付、接入控制、安全审计监控、多粒度访问控制等机制,建立安全保障体系,为平台和系统的运行提供安全服务。

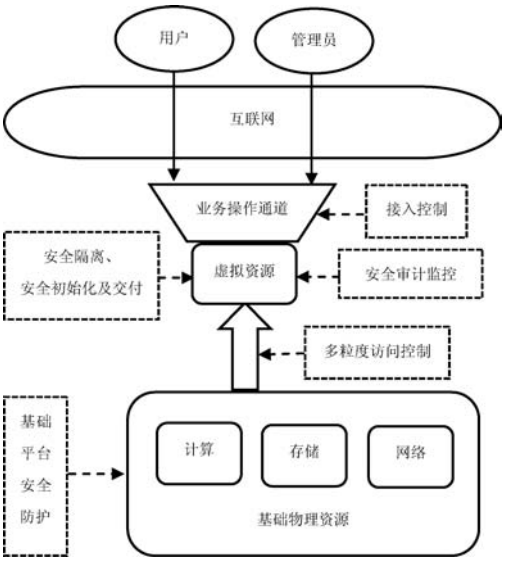

图 1 平台安全服务部署情况

本平台的基础平台安全防护采用了控制平台和业务平台相对分离的思路。控制平台完全对外隔离,仅连接平台的物理资源;业务平台是构筑于控制平台之上的虚拟化平台,可对外连接。控制平台与业务平台间的连接将受到严格控制,安全防护的重心将放在控制平台。针对控制平台,我们根据网络、主机、存储设备的具体规划,采取相应的安全防护机制,包括外网部署防火墙、内网划分独立网段、采用统一身份验证和授权管理、关键通道入侵检测设施等。

基于虚拟化的安全隔离为每个虚拟机分配专用的计算、存储和网络资源,可防止残余数据的利用,消除侧信道。安全初始化及交付阶段采用随机化因素改变虚拟空间的缺省安全机制配置,并通过安全的通道和环节将相关的认证因子进行交付。

接入控制部分提供身份认证和授权管理服务,采用安全接入机制使授权用户进入系统,防止非授权用户对平台和系统造成损害。

安全审计监控实现多层次监测数据的关联分析,并向用户提供对监测数据的查询和分析服务,实现安全审计。用户通过安全审计和监测服务,可实现对相关事件的溯源。

多粒度访问控制机制能实现对多样化数据资源的保护。生物医疗大数据类型驳杂、体量巨大,难以采用统一的数据访问控制机制,因此本平台采用了多粒度的访问控制,对资源类型和资源实例进行权限管理。

业务系统在上线运行后会面临各式各样的数据安全挑战,相应的安全防护措施是必不可少的。本平台为电子数据安全保障提供了必要的安全服务,降低了信息安全风险。但对数据的攻击手段是不断变化、不断发展的,因此,本平台会在未来继续完善数据保护措施,使电子数据保护方案更加完备。

4 结 语

当下我国医疗卫生领域在生物医疗大数据的使用方面尚未形成标准的规范,这导致数据的安全保障管理方面存在很多风险。本文则以数据的生命周期为基础,面向数据使用者和管理者,给出在数据采集、存储、访问、应用、共享、销毁等阶段的隐私安全保障建议。希望能以此为基础形成生物医疗大数据监管规范框架并撰写数据共享的保障监管规范。

同时,本文建立了一个基于 OpenStack 的大数据平台,为数据系统的运行提供了安全的防护,保障了电子数据的安全。

参 考 文 献

[1] 姬博歆. 医疗大数据应用中的隐私权保护研究[D]. 哈尔滨:黑龙江大学,2017.

[2] 张国庆,李亦学,王泽峰,等. 生物医学大数据发展的新挑战与趋势[J]. 中国科学院院刊,2018,33(8):853 - 860.

[3] 中国信息通信研究院. 大数据白皮书[R/OL]. (2016 - 12 - 28)[2019 - 07 - 29]. http://www.cac.gov.cn/2016 - 12/28/c_1121534609.htm.

[4] 国务院. 促进大数据发展行动纲要[R/OL]. (2015 - 08 - 31)[2019 - 07 - 29]. http://www.gov.cn/zhengce/content/2015 - 09/05/content_10137.htm.

[5] 张静. 云环境下医疗大数据隐私安全风险评估[D]. 昆明:云南财经大学,2018.

[6] 全国信息安全标准化技术委员会. 信息安全技术个人信息安全规范:GB/T 35273-2017 [S]. (2018 - 01 - 24)[2019 - 07 - 29].

[7] 公安部网络安全保卫局. 互联网个人信息安全保护指南[Z]. [2019 - 07 - 28].

[8] 陈文捷,蔡立志. 大数据安全及其评估[J]. 计算机应用与软件,2016,33(4):34 - 38,71.

(上接第 226 页)

[10] 陈琦,潘伟民. 基于自编码器的图像去噪设计与实现[J]. 新疆师范大学学报(自然科学版),2018,37(2):80 - 85.

[11] 李传朋,秦品乐,张晋京. 基于深度卷积神经网络的图像去噪研究[J]. 计算机工程,2017,43(3):253 - 260.

[12] Zhang K, Zuo W M, Chen Y J, et al. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising [J]. IEEE Transactions on Image Processing, 2017, 26(7): 3142 - 3155.

[13] 李敏,章国豪,曾建伟,等. 结合 Inception 模型的卷积神经网络图像去噪方法[J]. 计算机工程与应用,2019,55(20):139 - 144.