本报和《协和医学杂志》共同策划

医疗大数据 共享和安全的边界在哪里

近年来,《民法典》《网络安全法》《数据安全法》《个人信息保护法》等法律陆续出台,从不同视角对一般数据、个人信息、敏感信息、隐私权等不同分类数据的处理和安全保护进行了规定,明确了大数据智能化建设和应用的根本遵循、基本原则和管理要求,也为大数据智能化实践提供了基本遵循和法律保障。

医疗卫生机构在医疗诊断、卫生防疫、健康管理等实践中,依法采集了大量必要的数据,其中包括相当一部分的个人信息。当今,数字化技术的发展能在多大程度上保障医疗数据在传输、应用中的安全?如何解决智能医学发展过程中出现的数据确权、隐私保护、抵御攻击等一系列问题?前不久,在《协和医学杂志》举办的线上讲座中,专家们就这一话题展开了深入探讨。同时,我们也约请了北京协和医院的相关学者就医疗卫生机构如何保障数据安全分享了心得体会。


RF123供图

保障数据安全,医疗卫生机构从何做起

□北京协和医院 **陈政 张志文 龙笑 吴友武**

法律关注的数据收集、存储、使用、加工、传输、提供、公开等全生命周期的关键节点,正是医疗健康数字化战略规划和大数据智能化建设应用的主要任务和重点适法环节。随着数字技术的发展和赋能,医疗卫生机构亟须在实践中构建数据安全,提升个人信息、敏感信息、隐私信息精准识别、动态标注、精细管理、合理授权、有效鉴权的适法能力,确保好用够用、依法合规。

医疗卫生机构必须加强制度设计,提升技术水平,确保在涉及个人信息的全生命周期管理中,落实法律规定的数据安全和个人信息保护要求,实现全流程可知、可管、可控、可查。

健全相关制度

数据采集感知工作规范和分工机制 按照法定职责任务,形成中央地方、行业社会、网上网下、境内境外依法合规、分工协作的大数据采集感知工作规范,落实数据采集责任,严格执行法律政策规定,利用公开、管理、服务、技术等多种手段,依法获取数据资源。

数据分类分级的制度 严格执行相关法律对数据性质的规定及数据处理的要求,建立医疗健康大数据分类分级制度和重要数据资源目录,对个人信息、敏感信息、隐私信息以及一般数据进行科学、全面、动态的界定,明确数据分级分类的标准规范和管理要求。

与职责任务及使用场景适配的数据使用规则 按照职责分工和任务性质,并统筹不同场景下环境、设备、网络等多种因素,科学授予对应数据资源的使用权限,并实现动态控制,确保好用够用、依法合规。

数据安全的全流程管理制度 制定并执行医疗健康数据收集、使用、存储、加工、传输、提供、公开等全生命周期、全流程的数据安全管理制度,明确各处理环节的数据安全责任和管理要求,对关系国家生物基因安全和重大公共卫生利益等国家核心数据,实行更加严格的管理。建立数据安全应急处置制度,提高数据安全事件感知和控制能力。

数据安全官制度 医疗、卫生机构是重要的数据处理者,应当明确数据安全负责人和管理机构,制订数据安全保护的战略、规划、方案和机构,落实数据安全保护责任,促进医疗大数据更加有序地服务人民健康事业。

建设技术能力体系

建设个人信息、敏感信息、隐私信息识别和标注的能力体系,向收集、存储、使用、加工、公开等关键数据处理环节以及为监测评估、监督检查工作等提供资源化、服务化的识别技术能力,让个人信息识别、标注的能力便捷获得、便利使用。

建设数据分级分类能力体系,积累数据分级分类的知识、条目和算法模型,动态、全面地对数据进行分级分类,既能根据数据项、数据集、数据来源等要素敏感性对数据进行精细化分类,又可根据触及的专控类敏感标识、敏感样本、敏感内容的类型,对数据记录精细化分级。

建设脱敏技术能力体系,对个人信息、敏感信息、隐私信息,进行不同强度的脱敏处理。

建设加解密技术能力体系,对高敏感度的敏感信息、隐私信息等,进行不同强度的加密、解密处理。

建设精细化授权和精准鉴权技术体系,依照法律、行政法规和国家标准的强制性要求,按照职能责任,对使用者按照角色、任务、场景的要求,精细授予对应的级别、类别的数据使用权限。

建设以密码学、联邦学习、可信计算环境等技术为基础的隐私计算技术体系,满足不同强度隐私场景下多种加密计算、分布式计算、安全可信计算等隐私技术动态组合的需求,输出复合的隐私计算技术能力。

在赋能与安全中找到平衡

强化使用者的多维身份认证和持续信任评估,确保主体可信。

强化数据赋能关键节点处理逻辑的闭环控制,确保行为合规。强化存储、传输、使用关键环节数据安全保护,确保实体安全。强化大数据动态审计、安全分析和预警研判,确保风险可知。强化个人信息保护和数据安全红线责任,确保全局可控。强化数据处理全流程重点行为日志记录和保护,确保操作可查。

构筑三道防线

数据采集单位第一道处理 由数据采集单位,通过数据分类分级和过滤技术措施,将特定目标的高敏感且不可复用的数据进行滤除,对隐私度、敏感度高但需要用于医疗或公共卫生管理的数据,直接去标识化处理后,向大数据平台提供,同时提供全局性唯一记录ID,便于严格事由审批后,由指定部门回溯标识信息并落地调查。

大数据平台第二道处理 由大数据平台,通过数据分级分类和脱敏技术措施,对敏感度高且不可复用的数据进行滤除或者高强度加密处理;对隐私度、敏感度高但需要用于医疗或公共卫生管理的数据,直接去标识化处理,并为后续经审批后回溯提供条件;对一般敏感度的个人数据,在分级分类后,对数据进行变换、断开标识并标识记录关系等脱敏处理,但具备查询、比对、分析统计等计算条件。

平台数据服务第三道处理 由大数据平台服务技术,基于访问主体的职责分工和任务事由的审批情况,以及主体所在的网络、设备、APP、物理环境等实际因素,动态适配和反馈不同强度的脱敏数据或原始数据。

保障AI模型安全,“防”要走在“攻”前面

□周少华

医学影像是构成医疗大数据的重要部分。我先简单介绍一下医学影像的一些特点。

影像:多模高精 医学影像包括了CT、超声、核磁等等多种模态。随着技术的进步,很多医学影像已越来越接近真实图像,具备高精度的特点。

数据:非标孤立 影像数据的采集在不同中心没有统一的标准,同样的胸片,在不同医院拍可能都不一样。而在这些数据之间,也存在彼此孤立的状态,不同的医院,甚至同一家医院内部不同科室之间也没有形成互联。

疾病:长尾突发 很多影像数据属于典型的长尾分布,大量的数据分属于有限几个类别的常见病;罕见病品类众多,但每一类却只有极少量的数据。此外,像新冠肺炎这样的突发疾病,早期的数据量也是非常稀少的。在这种情况下,数据的采集,以及在此数据基础上构建AI系统是很有难度的。

标注:稀疏有噪 医学影像形成数据集之后,就要依据需要进行高质量的标注,比如,某一个器官的边缘在哪里。但是,我们目前拥有的标注量很少,而且不同的医生对同一个影像有时会产生不同的解读,造成标注的噪声。

样本:各异不均 比如肺结节,正样本和负样本之间差异很小,且正负样本之间的数量非常不均衡,通常负样本的数据量大得多。

任务:复杂多样 智能医学为我们提出的工作任务是多样化的,而面临的数据集又是极为差异化的。

安全:脆弱不稳 假设我们现在针对某一类影像构建了一个AI模型,要根据影像检测相关疾病的特征点。而一旦对这个模型进行攻击,比如,加一点点“噪声”,且这个“噪声”又是肉眼看不见的,那么这个AI模型的输出就会完全发生改变。

业界有个实验,输入熊猫的图片,用AI系统进行识别,它以约60%的自信度认为这是熊猫。然后,再在图上加上很小的“噪声”,这时候AI系统就输出了另外一个结果:猴子,并以接近100%的自信度否认了这是一只熊猫。

我们需要注意的是,攻击本身的形式是非常多样的,而且对人类肉眼来说,它能做到完全不被察觉。同时,我们最近做的实验表明,医学影像AI模型从安全性角度来说比自然图像更加脆弱,更容易受到攻击。

为了抵御攻击,我们必须构建防守的模型,从而成功识别某个图像是攻击图像还是原始图像。而一旦一个更有力的攻击又加入进来时,我们就必须再去找到这个新的变化,一步一步建立起整个防守系统。这也是我们目前正在积极推进的工作。

技术是在不断进步的,对于医工人员来说,“防”的工作要走在“攻”的前面,有了防守系统之后,就能主动去寻找攻击,更好地保障数据安全。

(作者系中国科学技术大学讲席教授兼生物医学工程学院执行院长,博士生导师,中科院计算所客座研究员,香港中文大学(深圳)客座教授)

用好区块链,实现医疗大数据可溯源

□彭绍亮

医疗大数据主要分为两大类:电子病历和医学影像,当然还包括一些其他病理的数据。而智能医疗,就是希望能够对这些数据进行二次挖掘,在诊断和治疗上进行辅助。

广义的电子病历概念很广,包括患者的基本信息、账单数据、药物史、临床诊断历史、生化检测、用药、理化甚至是医疗保险的数据。在这些基础上,结合人工智能,我们可以进行临床的预测、精准医学、药物设计等,在很大程度上帮助医生和患者共同制定临床诊疗路径,分析、提高诊断的精准度和效率。

但是早些年,我们的研究只能基于单中心的医疗大数据,没有解决数据孤岛的问题。而近些年,随着区块链技术的发展和应用,我们已经能越来越好地对数据的可信度和分享进行保障。

简单说来,区块链就是一个去中心化的、分布式的账本数据库,具有隐私保护、不可篡改的特性。它由一个块身和块头组成,和最早的比特币很相似,分为数据层、网络层、共识层、激励层、应用层。区块链解决了从信息互联互通到价值传递,从信息互联网到价值互联网的问题。

在以往,互联网只能传递信息,而现在通过区块链,我们能够传递价值,就像比特币一样,在信息的基础上,还能把赋予在它身上的金融属性发挥出来,也叫作资产的数字化。我们国家是最早推出数字货币的国家,原因就是看中了区块链的去中心化、匿名性、便捷性、不可篡改性和可编程性。

而我们正在做的,就是让区块链在医疗健康等多个领域进行多元化的渗透,形成一个可信的医疗大数据平台。

在这个平台上,相关的用户及机构可以进行资质的认证和注册,将原本分属不同中心的数据灌入,打造成一个可信的数据池,除了医疗本身的信息,还包含了企业项目、区域建设、地理信息、人文社会以及一些政府开放的数据。在此基础上,由政府部门、医疗机构等对数据池的有效性、分界性进行监督和审定,而后,平台上的用户即可根据不同的访问权限,在平台保障数据安全的前提下,对数据进行授权、分享、查询等。

我们曾研发过一个药品和疫苗溯源的区块链,以防止假疫苗和假药事件的发生。在药品生产过程中,我们就开始对其进行监管。通过手机端,我们设计了一个很简单的应用,让用户得以将在售药品、在库药品、已发药品进行全流程的记录。这样,一株药就是一个区块链,手机一扫,就可以看到这株药的来源。

当然,这个链条也在不断增长,我们也可以随着它的变化,看到这株药物或疫苗的生产、出库、入库、打包的时间,保证它的生产及流通安全。尤其是在药品或疫苗生产前期的细胞培养阶段,我们也会进行持续录入,直到后面的灭活、纯化、制备等步骤,保证数据的可信、可用。在使用过程中,出现任何问题,都可以做到全流程可溯源。

与此同时,我们也要认识到,解决数字孤岛问题,进行更深入的智能学习,数据的移动势必会产生安全隐患。比如被篡改,或是造成隐私泄露等。所以,我们希望,能把智能学习的模型进行压缩,使它在不同的数据之间移动,让数据“可用不可见”。这样,对于医疗机构来说,医疗数据便无须进行脱敏处理,不用离开医院,便可以套在人工智能的模型上,简单应用,发挥价值。

(作者系湖南大学国家超级计算长沙中心副主任,博士生导师,教育部2020“长江学者奖励计划”特聘教授)

对话——

问:作为临床医生,大家都知道高质量的临床研究一定需要基于多中心的持续对照研究。但在这个过程中,患者隐私会否泄露、数据传输能否安全,目前相关技术的发展水平能否消除这些隐患?

答:数据集的应用和安全性应该找到一个平衡点。

一方面,数据保护太严,会严重影响AI模型的学习效率,也会影响准确性,导致因噎废食,这与我们利用智能医学进行赋能的初衷是相违背的。医学太复杂了,只有多中心的数据,才能提供多角度的信息,从而产生新的火花。

目前,我们的技术已经基本能保证在数据不出医院的前提下,让大家一起使用数据,共同训练模型。同时,为了满足需求,也解决大家的顾虑,很多团队也在研究如何使相关的硬件加速,或者不用同态加密技术也可以保护隐私的算法。

从根本上说,开放是第一步,这是一个认知问题。只有实现了数据共享,才能在这个基础上找到共同点、训练模型,推进智能医学的发展。

但另一方面,我们还必须警惕两条红线:绝对不要把数据流到境外去,绝对不要进行数据交易。对此,业界也一直在呼吁政府部门和相关机构出台数据共享的标准和方式,我们也在编写一些共识,探讨可行性,希望加强对相关人员的约束,确保数据安全。

(本段文字摘自线上讲座问答环节)

主办单位:健康报社有限公司 出版单位:健康报社有限公司 社长:邓海华 总编辑:孙伟 社址:北京市东直门外小街甲6号 邮编:100027 电话:010-64672395 新闻热线:xwrx@jkb.com.cn 广告业务:010-64622318 发行业务:010-64622910 周一至周五对开八版 年价:456元 零售价:1.9元 中青印刷厂印刷(地址:北京市东直门海运仓2号)