DOI: 10.13376/j.cbls/2022001
文章编号：1004-0374(2022)01-0001-04

睿●观●察

寻找生命健康大数据在安全保护与开放共享之间的平衡——对《中华人民共和国个人信息保护法》的思考

吴家睿
(中国科学院分子细胞科学卓越创新中心，上海 200031)

随着21世纪初人类基因组计划的完成，生命健康科学迈入了大数据时代。不久前有统计指出，全世界在2013年产生的医疗健康数据大约在153 EB (1 EB = 10^{18} Byte)，而在2020年则增长到了2 314 EB[1]。大数据是生命健康研究领域的一个“新物种”，其获取与管理、保护和利用等各个方面都有着不同于传统科学的特征，也引发了一系列挑战；如何在保护数据安全和数据开放共享之间构建平衡就是一个急需解决的关键问题。为此，中国政府最近颁布了两部相关的法律：《中华人民共和国数据安全法》(自2021年9月1日起实施；以下简称“安全法”)和《中华人民共和国个人信息保护法》(自2021年11月1日起实施；以下简称“保护法”)。笔者认为，这两部法律为合理合规地保护和利用数据奠定了重要的基础，但仍然存在着许多值得探讨的地方。

1 如何区分信息与数据

信息和数据是两个不同的概念，但有着紧密的关系。按照“保护法”的规定，“个人信息是以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息”。而根据“安全法”对“数据”的界定：“本法所称数据，是指任何以电子或者其他方式对信息的记录”。显然“保护法”所指的“个人信息”实际上就是“个人数据”。换句话说，个人所拥有的各种生物学的和非生物学的信息，只有被记录下来成为数据，才属于“保护法”的保护对象。例如，个人的行踪轨迹属于个人信息，只有被手机或者可穿戴设备记录下来才成为个人数据。美国政府在1996年颁布了著名的《健康保险携带和责任法案》(Health Insurance Portability and Accountability Act, HIPAA)，涉及到个人数据和个人信息的保护。该法案中有这样一个规定：医生与患者交流时需将电脑屏幕调整到适当的角度以避免他人的观看。这个规定显然强调的是保护非数据的个人信息。

记录信息的本质就是要对信息进行“处理”。“保护法”给出了一个明确的定义：“个人信息的处理包括个人信息的收集、存储、使用、加工、传输、提供、公开、删除等”。因此，“保护法”的目标是规范个人信息的处理，主要针对的是个人信息的处理者，包括各种社会组织和个人。“保护法”共有八章，内容完全是围绕着个人信息处理展开。仅从该法案各章的标题就能够很好地反映这一点。除了第一章“总则”和第八章“附则”外，第二章标题是“个人信息处理规则”；第三章是“个人信息跨境提供的规则”；第四章是“个人在个人信息处理活动中的权利”；第五章是“个人信息处理者的义务”；第六章是“履行个人信息保护职责的部门”；第七章是“法律责任”。显然，“保护法”可以视为“个人信息处理法”。

2 如何处理个人信息

2021年1月实施的《中华人民共和国民法典》(以下简称“民法典”)第一千零三十二条明确规定：“自然人享有隐私权。……隐私是自然人的私人生活安宁和不愿为他人知晓的私密空间、私密活动、私密信息”。可以说，“保护法”的首要任务是保护数据领域的“个人隐私”。正是基于保护自然人隐私的考虑，“保护法”把处理过的个人信息即个人数据划分为可识别的和不可识别的(匿名化的)：“个人信息是以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息，不包括匿名化处理后的信息”。“保护法”专门给“匿名化”做了这样一个定义：“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”。也就是说，“保护法”关注的是个人数据与特定自然人的关系，如果从个人数据再也不能直接或者间接地识别出特定自然人，自然人的隐私就不会受到侵犯；因此，这种“匿名化”处理后形成不能识别特定自然人的个人数据就不被纳入“保护法”的保护对象。

“保护法”为了保护数据领域的个人隐私，在第五十一条规定中明确要求个人信息处理者要“采取相应的加密、去标识化等安全技术措施”来保护个人信息。“保护法”也对“去标识化”给予明确的定义：“是指个人信息经过处理，使其在不借助额外信息的情况下无法识别特定自然人的过程”。也就是说，“去标识化”处理过程是保护个人隐私的必要措施，是实现个人数据安全的基本要求。

这里引出了一个重要的问题：如何区别去标识化的个人信息和匿名化的个人信息？匿名化的个人信息处理过去是采用删除个人身份信息，如姓名、年龄、性别和住址等来防止对特定自然人的识别。但在大数据时代，不同的数据类型之间往往有着许多直接和间接的联系，通过数据之间的分析，能够挖掘出被隐藏很深的个人信息，使得这种传统的匿名化方法在面对大数据分析和搜寻时不再有效[2]。例如，美国著名罪犯“金州杀手”逃匿了 30 多年，该罪犯留下的一段 DNA 序列对警方一直没有什么帮助，这段序列可以视为匿名化的。2018 年初，警方将该序列与一个公开的基因组数据库 GEDmatch 里的 DNA 数据进行比对，发现了与罪犯有亲缘关系的人，最终抓住了这个罪犯。显然，该序列就应该属于去标识化的。由此可见，判断个人信息的处理属于去标识化的还是匿名化的，在很大程度上取决于信息处理者的技术能力和现实条件。

从处理的个人信息种类来看，“保护法”把一类个人信息定义为“敏感个人信息”——“包括生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹等信息”。可以说，生命健康领域内的数据基本上都属于敏感个人信息。“保护法”对敏感个人信息的处理有更为严格的保护要求，在第二十八条中规定：“只有在具有特定的目的和充分的必要性，并采取严格保护措施的情形下，个人信息处理者方可处理敏感个人信息”。但是，对这样规定如何理解和解释？例如，我国研究者为生物样本提供者拟定的知情同意书中是这样写的：“我同意所捐献样本和信息用于所有医学研究，为早日攻克疾病和病患医治作贡献”[3]。这种说法符合“具有特定的目的和充分的必要性”的规定了吗？显然，这种“特定的目的”和“充分的必要性”等的界定如果不是很明晰的话，个人生物学信息和医疗健康信息的处理者将面临着不确定的法律风险。

“保护法”在第二十九条中特别强调：“处理敏感个人信息应当取得个人的单独同意”。显然，这一规定是要确保个体的“知情权”。但是，在现实复杂情况中，尤其在大数据时代，要取得每个人的单独同意并非易事。在医学伦理的国际“基本法”——《赫尔辛基宣言》关于“知情同意”的规定中能看到这种复杂性：“针对使用可识别身份的人体材料或数据进行的医学研究，例如针对生物样本库或类似储存库中的材料或数据进行的研究，医生必须征得材料或数据采集、储存和 / 或再使用的知情同意。可能存在特殊情况使得获取这类研究同意不可能或不现实，在这种情形下，只有经过研究伦理委员会考量和批准后研究才可进行”。此外，由于“保护法”规定的个人信息处理方式涉及范围很广，包括“收集、存储、使用、加工、传输、提供、公开、删除等”，因此需要考虑不同的处理方式在执法过程中要有所区别。也就是说，在生命健康领域执行“保护法”处理敏感个人信息规定时，需要出台能够适应复杂现实情况的配套政策，使法律执行更加清晰，并降低执行难度。

3 如何进行个人数据的确权

个人信息一旦被记录下来成为个人数据，就产生了数据确权的问题，即个人数据的权属如何确定——是个人信息的提供者本人还是处理者一方？例如，美国最大的个人基因组信息分析公司“23andMe”为超过百万的客户提供了全基因组测序服务，但这些所测的个体基因组数据全部属于公司。尽管“安全法”在第三条里给出了数据的“全生命周期”——“数据处理，包括数据的收集、存储、使用、加工、传输、提供、公开等”，但却没有提及数据处理最重要的一环——“确权”。

“数据确权”是一个复杂而又敏感的问题，在个人数据领域主要涉及到个人的权益和信息处理者的利益。虽然“保护法”在确定个人信息处理的各个环节时没有提到“确权”，但在第三十条中有这样一个规定：“个人信息处理者处理敏感个人信息的，……还应当向个人告知处理敏感个人信息的必要性以及对个人权益的影响”。实际上，研究者也很关注这一问题，如上文提到的知情同意书示范样

本是这样告知样本提供者：“研究结果若衍生任何专利权或商业利益时，所有权益将与您无关。……您和其他捐献者的贡献将会推动医学技术进步，从而获得更有效的疾病诊断、治疗方法，这将惠及您以及相似疾病的其他患者，这是您和其他捐献者的共同利益”[3]。

数据，尤其是大数据，已经成为一种重要的社会经济资源，目前国内外最富有的公司大多是涉及大数据的公司。大数据在生命健康领域同样也成为了重要的资源，不仅对科学研究具有重要的价值，而且也有可能带来巨大的经济利益。例如，美国23andMe公司将3 000名帕金森症患者的全基因组信息去标识化以后以6 000万美元的价格卖给了Genentech公司。2018年，Roche制药公司用43亿美元分别收购了收集癌症患者临床信息的Flatiron Health公司以及收集癌症患者的样本和基因组测序数据的Foundation Medicine公司；这两家公司最有价值的就是临床大数据。显然，个人数据权属问题的解决方案要处理好提供信息的个人的权益和信息处理者的利益，从而才能够有助于数据生态和数据产业的健康发展。

“保护法”对个人信息是否进行处理规定了明确的个人自决权：“个人对其个人信息的处理享有知情权、决定权，有权限制或者拒绝他人对其个人信息进行处理”。但是，“保护法”对处理后形成的个人数据在确权方面却没有规定。个人数据确权的关键是对其所有权的判定。按照“民法典”第一百一十四条规定，所有权是一种“物权”，即“物权是权利人依法对特定的物享有直接支配和排他的权利，包括所有权、用益物权和担保物权”。由于个人数据作为“特定的物”是在提供信息的个人和信息处理者两种“民事主体”的共同参与下形成的，所以如何确定个人数据的权利人就成为必须要解决的问题。目前我国的“民法典”、“安全法”和“保护法”等与个人数据相关的法律均没有对此给予明确的回应。显然，解决数据确权这一新生事物不仅需要国家有关部门制定相关的法规和办法，而且需要学术界进行深入的理论探讨。

4　如何协调大数据的安全与共享

处理单个或少量个人信息形成的数据的“价值密度”远低于处理大规模人群信息形成的大数据集合的“价值密度”，而且后者通常还具有巨大的增值潜力。这不仅表现在基于互联网的大数据，同样也表现在人口健康领域的大数据。例如，2012年建成的UK Biobank收集了50万英国人的生物学样本以及基因组数据和各种表型数据[4]；在不到10年的时间内，世界各国众多研究者利用这些样本和数据进行了各种健康问题的研究，并发表了上千篇研究论文。当然，UK Biobank的价值远不止于此。美国政府同样也高度重视大数据在健康领域的价值，于2017年正式启动了为期10年的“All of Us Research Program”，计划收集100万美国志愿者的生物学样本和健康相关的大数据。

为了确保高价值的规模化个人数据的国家安全，“保护法”第四十条专门规定：“关键信息基础设施运营者和处理个人信息达到国家网信部门规定数量的个人信息处理者，应当将在中华人民共和国境内收集和产生的个人信息存储在境内。确需向境外提供的，应当通过国家网信部门组织的安全评估”。按照2021年10月29日国家互联网信息办公室发布的《数据出境安全评估办法(征求意见稿)》，凡处理个人信息达到100万人的个人信息处理者向境外提供个人信息，或累计向境外提供的个人信息超过10万人的数据出境均需要进行安全评估。换句话说，国家把拥有百万个体数据的个人信息处理者(关键信息基础设施运营者)定为数据安全的重点关注对象，同时对10万人规模的个人数据出境进行安全管控。值得注意的是，个体生物学信息和健康信息等“敏感个人信息”受到了更严格的安全管控，只要累计向境外提供超过1万人以上敏感个人信息就需要进行安全评估。

大数据的开放与流动是实现其内在价值的基础，也是大数据时代的基本准则。2019年11月，国际科学理事会数据委员会(CODATA)发布了《科研数据北京宣言》，其原则之一就是鼓励国家间数据的开放与共享。在人口健康领域，数据的开放与共享尤为重要，不同人种、不同疾病谱和不同环境之间数据的比较将更有利于我们认识人体生理和病理的活动规律。不久前，美国国立卫生研究院(NIH)牵头组建了一个“国际十万人队列联盟”(International Hundred Thousand Plus Cohort Consortium, IHCC)，把43个国家100多个人群研究队列汇集在一起，参与人数超过5 000万[5]。显然，这种跨国研究活动的前提是这些国家的研究数据可以跨境流动与共享。

然而，数据的开放与流动又需要保证其安全性，二者必须兼顾。正如“安全法”第十一条所说：“国

家积极开展数据安全治理、数据开发利用等领域的国际交流与合作，参与数据安全相关国际规则和标准的制定，促进数据跨境安全、自由流动”。但是，如何落实数据在安全前提下的流动还有许多法律法规方面的实施细则需要完善。例如，“保护法”第三十九条规定：“个人信息处理者向中华人民共和国境外提供个人信息的，应当向个人告知境外接收方的名称或者姓名、联系方式、处理目的、处理方式、个人信息的种类以及个人向境外接收方行使本法规定权利的方式和程序等事项，并取得个人的单独同意”。假如我国研究者考虑加入“国际十万人队列联盟”这一研究项目，就需要先考虑如何满足该规定以便人群队列大数据的跨境流动与共享；可以想见，这绝非易事。

5 结语：如何在大健康时代保护和利用大数据?

20 世纪属于基于小数据的传统临床医学时代，在处理个人信息时要有特定的目的 (如临床试验)；相应的医学伦理学的核心是保护个人隐私，信息处理者需要与个人签订与信息处理目的直接相关的“具体知情同意”(Specific consent)。虽然“保护法”是 2021 年 11 月才开始实施，但看上去却更适合这样的时代。该法不仅规定个人信息处理者处理敏感个人信息的基本条件是“只有在具有特定的目的和充分的必要性”(第二十八条)，而且还强调“个人信息的保存期限应当为实现处理目的所必要的最短时间”(第十九条)。21 世纪则是针对全人群和全生命周期的健康维护和促进的“大健康时代”，生命健康大数据是实现“大健康时代”目标的必由之路。英国的 UK Biobank 服务于这个目标；美国的“All of Us Research Program”同样服务于这个目标，“即它不聚焦在某一种疾病、某一种风险因子，或者是某一类人群；反之，它使得研究者可以评估涉及到各种疾病的多种风险因子”[6]。因此，信息处理者需要与个人签订的应该是与具体目的没有直接关系的“广泛知情同意”(Broad consent)：“我同意所捐献样本和信息用于所有医学研究，为早日攻克疾病和病患医治作贡献”[3]。显然，“保护法”需要针对大健康时代的目标来考虑大数据的保护和利用。

[参 考 文 献]

[1] Banks MA. Sizing up big data. Nat Med, 2020, 26: 5-6
[2] Price WN 2nd, Cohen IG. Privacy in the age of medical big data. Nat Med, 2019, 25: 37-43
[3] 中国医药生物技术协会组织生物样本库分会中国研究型医院学会临床数据与样本资源库专业委员会. 医疗卫生机构生物样本库通用样本采集知情同意书示范范本. 中国医药生物技术, 2019, 14: 477-80
[4] Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature, 2018, 562: 203-9
[5] Manolio TA, Goodhand P, Ginsburg G. The International Hundred Thousand Plus Cohort Consortium: integrating large-scale cohorts to address global scientific challenges. Lancet Digit Health, 2020, 2: e567-8
[6] NIH-Wide Strategic Plan for Fiscal Years 2021–2025[EB/OL]. https://www.nih.gov/sites/default/files/about-nih/strategic-plan-fy2021-2025-508.pdf