Proceedings of the 2nd International Conference on Biological Engineering and Medical Science (ICBioMed 2022), Part I
DOI: 10.54254/2753-8818/3/20220443

Privacy Issues in Big Data in Medicine

Jiazhen Lin

College of Medicine and Biomedical Information Engineering, Northeastern University, Shenyang, China, 110819

2441382@dundee.ac.uk

Abstract. With the rapid development of machine learning and artificial intelligence, the digital world is generating data at an alarming rate. This big data is bringing life to the development of medicine, from monitoring the safety of drugs, to the allocation of medical resources, to the diagnosis and treatment of complex diseases. But the development of big data also brings many risks and challenges, and it is particularly important to deal with the privacy of the data. Big data in medicine includes biological data, biometric data and e-health data, which are used by multiple organizations such as academia, medical institutions, and government departments. On the one hand, the data is widely used, and on the other hand, the public's idea of data protection is getting stronger and stronger, and many problems have arisen over time. This article concludes from the previous literature that more and more data are being used in an unregulated way and more people think they have lost control over data because of the incomplete aspects of the law and the difficulty for the government to control the level of privacy protection in some cases.

Keywords: big data, medicine, privacy, data security

1. Introduction

The concept of Big Data was introduced by McKinsey, a leading global consulting firm, back in the 1990s, when McKinsey stated, "Data, which has permeated every industry and business function today, has become an important production factor. The mining and use of massive amounts of data by people heralds a new wave of productivity growth and consumer surplus". Since then, big data has been used in various fields such as physics, biology, environmental ecology and even military, financial and communication industries. With the rapid development of machine learning and artificial intelligence, big data has once again aroused attention and has been frequently used in the medical field. From monitoring the safety of drugs, to the allocation of medical resources, to the diagnosis and treatment of complex diseases, big data has been creating value for the development of medicine [1], but the rapid development has also raised concerns about data privacy issues. This review introduces the concepts of big data and privacy separately to give readers a basic idea of both, followed by the problems and their causes at this stage, and also gives a feasible way forward. By summarizing the medical privacy issues in big data in recent decades and suggesting some feasible ways forward, this paper enables the reader to have a detailed understanding of this area.

2. What is big data and what is privacy?

Since the 1980s, the volume of data has generally increased, with collection rates doubling every 40

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/3/20220443

months [2], and by 1990, McKinsey introduced the term "Big Data", which is defined by the McKinsey Global Institute as a collection of data so large that it is significantly beyond the capabilities of traditional databases in terms of acquisition, storage, management, and analysis. It is a collection of data that is so large that it exceeds the capabilities of traditional databases in terms of acquisition, storage, management, and analysis. It has four major characteristics: 1): unlike data, the starting unit of big data is at least PB. China Mobile has over 800 million users, 14 TB of new data acquired every day, and a cumulative stock of 300 PB. Amazon deals with trillions of bytes of data every day. 2): a wide range of data types, including letters, numbers, text, pictures, audio and other forms. 3): the fast processing of data; In the face of massive data, big data is different from the traditional data mining, it has a higher processing efficiency. 4): the value of data density is low; an 8Mbps camera an hour to produce data volume of about 3.6GB, a month 2.59TB, many cities cameras up to hundreds of thousands, a month of data volume reached hundreds of petabytes, save too long storage volume can even reach EB magnitude. The volume of data is so huge, but often only a small fraction of it is valuable enough to characterise the low value density of the data. With the rapid development of machine learning and artificial intelligence, the application field of big data has become more extensive. For example, the decision of prevention and control by analyzing the data of COVID-19 [3], the shopping platform will infer your preferences and push the corresponding goods according to your browsing records, the navigation software deduces the real-time road condition and provides the best route with the help of big data, and the electronic medical record launched by the hospital has greatly improved the The electronic medical records introduced by hospitals have greatly improved the efficiency of work. Although Big Data has effectively changed our lives, it has also raised concerns about privacy issues.

According to Moore, because humans need to cooperate to survive, an assembled society requires a governing order to maintain stability, and privacy describes the extent to which this governing order can interfere with the lives of individuals at most. The modern discussion of privacy dates back to the early 17th century when Sir Edward Coke, the British Attorney General, proposed that "every man's house is to him his castle and fortress, and his defense against injury and violence, and his repose. This idea was amended in 1791 to "The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures" and incorporated into the U.S. Constitution. The Universal Declaration of Human Rights, subsequently adopted in 1948, established privacy as a fundamental human right. With the advent of the digital communication revolution, privacy is defined as the right to control personal information, which includes information about one's property, correspondence, and various actions and transactions [4]. At a time when big data tells the development, the coexistence of the demand for large amounts of data in various fields on the one hand, and the protection of data privacy by law on the other hand, is bound to raise many issues. In the following paper, we will discuss the current situation of the problem and the reasons for its formation.

3. Privacy issues arising from big data

3.1. Current status of the problem

With the development of the online world, people's lives are gradually "data-driven" and a large amount of personal information is stored on the Internet. In the medical field, although most countries consider medical data to be legally owned by patients [5], people have expressed their loss of control over the data, and this situation is particularly evident in the medical field. particularly evident. According to a recent Eurobarometer survey, 15% believe they have full control over their personal information and as many as 31% believe they have no control over their information at all, in addition to the 69% who believe their data is being used in areas they did not authorize in the first place. And most people want to have control over their information themselves. According to a recent survey on Americans' attitudes toward privacy, security, and surveillance, 93% of respondents agreed that it is important or very important to control who has access to information about them [6].

Medical data, as patient privacy, has legal protection on the one hand, and patients themselves want

DOI: 10.54254/2753-8818/3/20220443

to be able to control their own data. So why do only a minority of people think they can control their own data?

3.2. Reasons for problem formation

*3.2.1. Incompleteness of the law. I*n 1996, Congress enacted the Health Insurance Circulation and Accountability Act, and in 1999, the Secretary of Health and Human Services, HHS, proposed an initial version of the Privacy Rule, which now incorporates amendments to the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2008 and the "Genetic Information Nondiscrimination (GINA) Act of 2008 [7] and the EMR added in 2013 [8]. As the primary governing law for health data privacy, although HIPAA has undergone several amendments since its introduction, it still seems to have many problems today.

The first is that, in the face of the rise of artificial intelligence, the previous way of encrypting information is no longer applicable today. Previously, HIPPA protected patients from privacy violations by removing a set of 18 specified identifiers while sharing data out to prevent patient identification, but nowadays, de-identified data can be used to identify patients by measuring other data sets in comparison [9]. Using artificial intelligence and big data techniques, researchers at the University of Zurich mined 120,000 legal records and compared them to publicly available datasets, demonstrating that big data combined with artificial intelligence can identify individuals in confidential cases, and another researcher in the study noted that the procedure could in principle be applied to any publicly available database [10]. This shows that there is a relatively large gap in the protection of patient privacy by removing specified identifiers.

Then there is the difficulty of defining whether there is a privacy violation in today's law. There are cases where citizens claim that their sensitive information was not protected due to corporate negligence, but the courts dismiss the claims as not having suffered actual harm [11]. In addition, in the face of the current era of data mining, it is difficult to define whether indirectly generating data about others through data mining is a privacy violation. knowledge, and that there are legal and normative reasons to reject the notion that inference may violate privacy [9].

Secondly, the scope of HIPPA is too small. HIPAA proposes that the regulations apply to all "covered entities", which include most healthcare providers, insurance companies and hospitals [9], but these entities do not include Many large technology companies, such as Apple and Google, are excluded. With technology advances and the pursuit of health, wearable devices have grown rapidly, but the law has not kept pace with the times. When people use wearables or other mobile health apps, they first have to sign a privacy consent form. Not signing it means they can't receive their services, but once they do, it means the user's information becomes part of someone else's data set, and that data that people don't care about is often important. For example, a company program called Strava has a location feature, and in the Syrian war, American soldiers using Strava exposed their location and movements [12]. Most people today do not understand how valuable their ownership is, especially when faced with consent forms that are voluminous and contain a lot of jargon, and people tend to sign them without being familiar with the security policies of the application [13], which is an urgent problem.

Finally, the GINA (Genetic Information Non-Discrimination Act) also has limitations. Although it prohibits insurance companies and businesses from discriminating against genetic information, its focus is not on limiting access to personal information but on protecting the person from discrimination if the information is accessed, and only from a small percentage of discrimination. It is difficult for the law to limit the way other people look at you and talk about you in private once they know your information. Similarly, with the use of big data, many disease risks are predictable, but current laws cannot limit how companies or insurance companies treat this category of people who are currently healthy but at high risk of disease.

DOI: 10.54254/2753-8818/3/20220443

3.2.2. It is difficult to define the level of protection of data. The time and cost involved would be enormous if each use of someone else's information had to be communicated and consented to, and the use and scope of the data would need to be communicated, and the need for data providers to approve consent forms one by one would also be a problem for them, which would be a huge blow to both researcher motivation and academic development. Similarly, over-protection of data can also have an impact on research results. For example, when data from multiple regions is needed for comparison, it is difficult for researchers to link highly de-identified data [14] and the data will become fragmented, which will be a serious obstacle to the development of the technique. But too much open information will violate the privacy of the general public, so it is difficult to define the level of protection for data.

4. A viable way forward

Facing the risks and problems brought by the era of big data, it is urgent to propose feasible solutions for the rational use of big data in the medical field.

First of all, it is extremely important to improve the legal system; laws are not static and should be an ongoing effort to change with the development of society [15], especially laws like HIPAA that protect medical data should be kept up to date with the rapid development of technology. In today's world, where technology companies like Apple and Google continue to squeeze the coverage of the law and the scope of what HIPPA actually protects is decreasing, it is important to establish a system of corporate accountability. In fact, as early as 2016, Europe enacted the European General Data Protection Regulation (GDPR), which proposed that any company and organization would be subject to privacy laws, and the scope of protected information was expanded to include any information related to individuals [16], which greatly improved the protectiveness of the law for privacy. The second is to face the problem of re-identification of desensitized medical data. Big data analytics software can be the use of more advanced encryption algorithms, and studies have shown that the use of pseudonymized enhanced de-identified input patterns can significantly improve the accuracy of results [17]. Finally, it is important to develop public awareness of ownership, as many people do not yet have a clear understanding of the importance of the medical data they generate and the value of data ownership, and knowledge of this can help them better protect their privacy. Of course, making lengthy information sheets shorter and easier to understand is also a possible approach.

5. Conclusion

Many fields have flourished in the era of big data, and at the same time, many challenges have been raised. This paper discusses the privacy issues of big data in medicine from the basic concepts, existing problems and causes of the problems, and concludes with a possible way forward, but some problems remain, such as it is difficult for the law to define whether there is privacy in the face of indirect information generation. infringement, and can only be reluctantly addressed through increased corporate accountability. This paper is also limited by the legal protection of privacy on the one hand, and the need for a large amount of data for academic development on the other. In future research, we can explore what the appropriate scope of legal protection should be, and whether there are better solutions to face the problem of indirect information generation.

References

[1] Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. Health affairs (Project Hope), 33(7), 1139–1147. DOI: https://doi.org/10.1377/hlthaff.2014.0048

[2] Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. Science (New York, N.Y.), 332(6025), 60–65. DOI: https://doi.org/10.1126/science.1200970

[3] Alsunaidi, S. J., Almuhaideb, A. M., Ibrahim, N. M., Shaikh, F. S., Alqudaihi, K. S., Alhaidari, F. A., Khan, I. U., Aslam, N., & Alshahrani, M. S. (2021). Applications of Big Data Analytics to Control COVID-19 Pandemic. Sensors (Basel, Switzerland), 21(7), 2282. DOI:

https://doi.org/10.3390/s21072282
[4] Kayaalp M. (2018). Patient Privacy in the Era of Big Data. Balkan medical journal, 35(1), 8–17. DOI: https://doi.org/10.4274/balkanmedj.2017.0966
[5] Viceconti, M., Hunter, P., & Hose, R. (2015). Big data, big knowledge: big data for personalized healthcare. IEEE journal of biomedical and health informatics, 19(4), 1209–1215. DOI: https://doi.org/10.1109/JBHI.2015.2406883
[6] Vayena, E., & Blasimme, A. (2017). Biomedical Big Data: New Models of Control Over Access, Use and Governance. Journal of bioethical inquiry, 14(4), 501–513. DOI: https://doi.org/10.1007/s11673-017-9809-6
[7] Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification rules under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; other modifications to the HIPAA rules. (2013). Federal register, 78(17), 5565–5702.
[8] Kels C. G. (2020). HIPAA in the Era of Data Sharing. JAMA, 323(5), 476–477. DOI: https://doi.org/10.1001/jama.2019.19645
[9] Price, W. N., 2nd, & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature medicine, 25(1), 37–43. DOI: https://doi.org/10.1038/s41591-018-0272-7
[10] Martin J. (2020). Big data, big future. BioTechniques, 68(4), 166–168. DOI: https://doi.org/10.2144/btn-2020-0027
[11] Nissim K, Wood A. Is privacy privacy?. Philos Trans A Math Phys Eng Sci. 2018;376(2128):20170358. DOI: 10.1098/rsta.2017.0358
[12] Belani, S., Tiarks, G. C., Mookerjee, N., & Rajput, V. (2021). "I Agree to Disagree": Comparative Ethical and Legal Analysis of Big Data and Genomics for Privacy, Consent, and Ownership. Cureus, 13(10), e18736. DOI: https://doi.org/10.7759/cureus.18736
[13] Cilliers L. (2020). Wearable devices in healthcare: Privacy and information security issues. Health information management: journal of the Health Information Management Association of Australia, 49(2-3), 150–156. DOI: https://doi.org/10.1177/1833358319851684
[14] Eisenberg, R. S., & Price, W. N., 2nd (2017). Promoting healthcare innovation on the demand side. Journal of law and the biosciences, 4(1), 3–49. DOI: https://doi.org/10.1093/jlb/lsw062
[15] Allen A. L. (2021). HIPAA at 25 - A Work in Progress. The New England journal of medicine, 384(23), 2169–2171. DOI: https://doi.org/10.1056/NEJMp2100900
[16] Galvin, H. K., & DeMuro, P. R. (2020). Developments in Privacy and Data Ownership in Mobile Health Technologies, 2016-2019. Yearbook of medical informatics, 29(1), 32–43. DOI: https://doi.org/10.1055/s-0040-1701987
[17] Kayaalp, M., Browne, A. C., Callaghan, F. M., Dodd, Z. A., Divita, G., Ozturk, S., & McDonald, C. J. (2014). The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them. Journal of the American Medical Informatics Association : JAMIA, 21(3), 423–431. DOI: https://doi.org/10.1136/amiajnl-2013-001689