Chapter 29
Epilogue

Aris Gkoulalas-Divanis and Grigorios Loukides

Abstract This chapter provides a summary of the main topics and methods that have been covered in the book, and it draws inferences about various important aspects of medical data privacy. In particular, it discusses issues and techniques related to preserving privacy in: (1) data sharing, (2) distributed and dynamic settings, and (3) emerging applications. Furthermore, it provides an overview of key legal frameworks for the protection of Personal Health Information (PHI) and of techniques required to comply with these frameworks, such as text de-identification and data governance. Moreover, the chapter discusses some promising directions in the field of medical data privacy.

29.1 Introduction

The collected medical data increase in both quantity and complexity, and are becoming an extremely valuable source for analyses that benefit both medical research and practice. However, the privacy of medical data remains an serious concern. On the one hand, privacy breaches continue to occur. For example, more than 600 privacy breaches that are related to healthcare data were reported only in the last 5 years by the U.S. Department of Health and Human Services [11]. Alarmingly, these incidents affect more than 500 and up to 4.9 million individuals each. On the other hand, attacks to individuals' privacy become more sophisticated and new threats are being continuously discovered. To address these issues and preserve data privacy, significant efforts have been put in order to advance technology, law, and policy.

This handbook aimed to provide a comprehensive coverage of the main research areas in medical data privacy. In the remainder of the chapter, we revisit each of

A. Gkoulalas-Divanis (✉)
Smarter Cities Technology Centre, IBM Research, Dublin, Ireland
e-mail: arisdiva@ie.ibm.com

G. Loukides
School of Computer Science & Informatics, Cardiff University, Cardiff, UK
e-mail: g.loukides@cs.cf.ac.uk

A. Gkoulalas-Divanis, G. Loukides (eds.), *Medical Data Privacy Handbook*,
DOI 10.1007/978-3-319-23633-9_29

these areas and summarize the research challenges and the solutions that have been proposed to offer privacy. In addition, we identify some promising areas for future research in this field.

29.2 Topics and Directions in Privacy Preserving Data Sharing

Protecting the privacy of medical data prior to their sharing is one of the major topics in the field of medical data privacy [4]. One of the pioneering efforts to address privacy in data sharing is the work of Sweeney [9] and Samarati [8]. This work demonstrated that the removal of personal identifiers, such as patient names and phone numbers, is not sufficient to protect the privacy of individuals. This is because other, seemingly innocuous attributes (termed as *quasi-identifiers*) can be used, alone or in combination, to link a patient with their record in the shared dataset. This attack is termed *identity disclosure* and constitutes a serious privacy breach.

Unfortunately, there have been several incidents of medical data publishing, where identity disclosure has transpired. For instance, Sweeney [9] first demonstrated this attack in 2002, by linking a claims database, which contained information of about 135 K patients and was disseminated by the Group Insurance Commission, to the voter list of Cambridge, Massachusetts. The linkage was performed based on patient demographics (date of birth, zip code, and gender) and led to the re-identification of William Weld, then governor of Massachusetts. It was also suggested that more than 87 % of U.S. citizens could be re-identified, based on such (triangulation) attacks. Many other identity disclosure incidents have been reported since then [3]. These include attacks in which (1) students re-identified individuals in the Chicago homicide database by linking the database with the social security death index, (2) an expert witness re-identified most of the individuals represented in a neuroblastoma registry, and (3) a national broadcaster re-identified a patient, who died while taking a drug, by combining the adverse drug event database with public obituaries.

To mitigate the threat of identity disclosure, various data anonymization algorithms have been developed to produce a sanitized counterpart of the original dataset that can be shared with untrusted parties. Chapter 2 surveys the most popular of these algorithms in the context of disseminating patient demographics (relational data) and diagnosis codes (transaction/set-valued data). The chapter explains the objectives of the different methods, as well as the main aspects of their operation. To protect data privacy, anonymization algorithms transform a dataset in order to enforce a formal privacy principle, such as k-anonymity [8, 9], while preserving data utility. There are many algorithms for applying k-anonymity and related privacy principles to various types of data. Some examples of state-of-the-art algorithms, discussed in detail, are those presented in Chaps. 7 and 8.

The effectiveness of anonymization algorithms in terms of offering data utility as well as achieving efficiency, varies significantly. Thus, it is important for a data owner, such as a healthcare institution, to be able to evaluate a range of algorithms and select the best one to apply to a given dataset. Until recently, this was a tedious task, requiring substantial domain knowledge and computer science expertise. However, recent software systems such as those discussed in Chaps. 5 and 6, have simplified the process of producing anonymous datasets by supporting a large number of popular anonymization methods and configurations. In addition, these systems allow, or can be easily extended to allow, the prevention of attacks beyond identity disclosure [5, 6].

Furthermore, there are cases in which protection from a wide class of attacks is necessary. This is possible using the principle of differential privacy [2], which is largely independent of the background knowledge of attackers. Differential privacy ensures that the addition or removal of any record to/from a dataset has little effect on the outcome of a calculation. This is achieved through noise addition (of appropriate magnitude). In several applications that can be supported with the release of aggregate statistics and do not require publishing of (truthful) data at a record level, differential privacy offers a good solution. As discussed in Chap. 3, there are many algorithms for enforcing differential privacy on relational, set-valued, as well as dynamic stream data (e.g., events produced by health monitoring applications). The majority of these algorithms add noise to the data or to the answers of queries that are applied to the data. Minimizing the impact of noise is essential in order to enhance the data utility, but it is not straightforward, particularly for high-dimensional data, as explained in Chap. 4.

While significant progress for enhancing privacy-protection in data sharing has been made, there are still many important research issues that warrant further investigation. First, sophisticated attacks that are possible by combining independently released datasets must be prevented. This should be done under the realistic assumption that coordination between the institutions that have released the datasets is prohibited. An interesting method to prevent a class of such attacks for demographics has been presented in Chap. 8.Further research is needed to prevent these attacks in more complex data, such as data containing both demographics and diagnosis codes [7], or longitudinal data [10]. In addition, it is important to design methods that guard against these attacks, while providing guarantees for the utility of the data in intended query answering and mining tasks. Another class of attacks that need to be thwarted are those performed on aggregated data. These attacks have been the focus of the statistical disclosure control community for years and have led to various methods, which are surveyed in Chap. 9. To increase the adoption of these methods on medical data, it is important to strengthen the protection they offer and to adapt them to operate in web-based data sharing platforms.

Furthermore, to increase the use of privacy-preserving data sharing methods, it is important to address scalability issues. In fact, most of the existing anonymization methods are applicable to datasets that fit into the main memory. Thus, they cannot be used to protect datasets with sizes of several GBs or even TBs. The development of scalable anonymization methods that potentially take advantage of

parallel architectures to solve this problem is therefore worthwhile. Recent steps towards this direction are the works of [12, 13], which are based on the Map/Reduce computing paradigm [1].

29.3 Topics and Directions in Privacy Preservation for Distributed and Dynamic Settings

Medical data are often distributed among multiple parties, such as collaborating healthcare researchers or healthcare organizations that form a consortium. Due to the privacy considerations that these parties typically have, simply sharing raw data among them is not feasible. This has led to the development of privacy-preserving record linkage methods, which allow a set of parties to construct a "global" data view without sharing their raw data. There are two main directions in the development of privacy-preserving record linkage methods, as discussed in Chap. 10. Most of these methods construct the "global" view by employing secure multiparty computation protocols. That is, they perform the operations needed for the record linkage using encrypted data. An alternative way is based on data transformations, and trades-offs privacy for efficiency. That is, each party transforms their data to enforce k-anonymity or differential privacy, and then the record linkage (matching) is performed on the transformed data.

Despite the significant research efforts to develop privacy-preserving record linkage methods, there are some open issues that require further investigation. In terms of privacy, it is important to design flexible methods, which can deal with attributes of different sensitivity as well as with malicious parties. In terms of linked data quality, it is important to design methods that can deal with complex data and offer high accuracy of linkage. An important step towards advancing the research in this field is the establishment of dedicated record linkage centers, as discussed in Chap. 11. The principle behind the operation of these centers is that human intervention can increase the quality and privacy of linked data.

In addition, privacy-preserving methods need to deal with the dynamic aspects of healthcare information management. This is because information is constantly exchanged between different parties, in Health Information Exchange (HIE) systems. These systems are increasingly used for purposes ranging from patient diagnosis and treatment to detection of medical identity theft and financial fraud. However, they pose privacy risks, including potential misuse and unwanted sharing of information, medical identity theft and financial fraud, as explained in Chap. 12. These issues are amplified with the use of large quantities of complex medical data, which are generated by mobile and large-scale networks, and health monitoring medical devices. Tackling these issues while preserving privacy in HIE systems is a promising direction for future research. For example, the security architecture of HIE systems needs to protect data residing in mobile platforms and to account for potentially vulnerable medical devices.

Another important area for preserving the privacy of health information is the design of access control methods. These methods determine what information should be accessed by each party, as well as when, how, and why access to information is performed. To control access to healthcare information in a flexible way, the Role Based Access Control (RBAC) model can be used. This model assigns permissions to users, based on their roles in organizations to denote specific job functions and the associated authorities and responsibilities. RBAC can also serve as basis for solutions that support team collaboration and workflow management, as discussed in Chap. 13. Another way to control access to information is based on patient consent management (i.e., give patients the ability to grant and revoke access to their data). An approach for patient consent management that is designed to deal with the changes to the context of data over time was presented in Chap. 14. Future research directions in the area of consent management include the design of methods that are suitable for mobile devices. In particular, it is interesting to investigate how patients can preserve the control over their data and consent policies, when it is not possible for the system to interact with the device (e.g., after the mobile device has been stolen, lost, or destroyed). Another important direction is to design cross-domain policies for consent management that are transferable between different systems. Such policies can greatly simplify consent management in practice.

Furthermore, there is an increasing interest towards the use of cloud infrastructures for the storage and processing of medical data. This poses certain privacy threats, including malicious data access, intentional data modification, and identity spoofing, which were surveyed in Chap. 15. The majority of existing methods to mitigate these threats are based on cryptographic primitives and are not sufficiently scalable. Promising research directions in this field involve the improvement of the scalability of these methods, as well as the design of techniques to automatically determine whether or not some data must be encrypted in the cloud environment. Another research direction is to improve the auditing and accountability capabilities of methods for managing medical data in the cloud. This is important to warrant the integrity and confidentiality of medical data.

29.4 Topics and Directions in Privacy Preservation for Emerging Applications

The privacy-preserving management, sharing, and analysis of medical data is necessary to support an increasing number of emerging applications. For example, applications featuring genomic, medical image, and Radio Frequency IDentification (RFID) data were discussed in the book, along with applications requiring biomedical signals and health social network data.

Genomic data are essential to realize the vision of personalized medicine (i.e., develop drug and treatments that are tailored for a particular patient). Therefore, they are increasingly shared among healthcare organizations. The sharing, however,

of genomic data poses serious privacy threats, as discussed in Chap. 16. These threats may result in privacy breaches that affect the patient, whose DNA data are shared, as well as the patient's ancestors and descendants. To guard against these threats, it is possible to apply differential privacy [2] to genomic data. An effective algorithm for enforcing differential privacy on genomic data was presented in Chap. 17. The development of differentially private algorithms for genomic data that optimize the utility of the data for a given analysis task, such as performing a statistical association test or building a mining model, is very important. Another approach for preserving the privacy of genomic data is based on cryptography. This approach is applicable when parties are interested in obtaining the result of a particular analysis task, as explained in Chap. 18. Examples of such tasks are genetic association studies and medical tests. To increase the practicality of this approach, it is essential to design methods that are appropriate for the range and complexity of future analysis tasks (e.g., those involving multiple genomes). Furthermore, methods that allow non-expert end-users to run computational genomic tests would be very useful. However, the development of such methods raises questions related to the information that needs to be presented to end-users. These questions include how the presented information can match the expectations of the end-users and the scientific community.

Another type of complex medical data that is becoming increasingly important in applications is images. In fact, medical image data play a central role in healthcare applications related to diagnosis, therapy, follow-up, and clinical research. At the same time, medical image data must be protected from unauthorized access, modification, and copying. This is possible using a combination of encryption and watermarking techniques, as explained in Chap. 19. In addition, it is a legal and ethical requirement to prevent patient re-identification from medical image data and metadata. This requirement is particularly important for neuroimage data and metadata, which can reveal sensitive information about individuals, including serious medical conditions, as it was discussed in Chap. 20. While there are significant efforts for designing frameworks and platforms for the management of neuroimage data, privacy-preserving solutions for such data are lacking. In particular, techniques that are based on objective performance metrics, comply with key regulations, and can be easily integrated in large-scale platforms are needed.

Emerging applications, such as the prevention of infections in hospital environments as well as the monitoring of patients' locations, are based on RFID data. This type of data, however, creates two privacy issues, namely tracking of the location of a patient, and preserving the confidentiality of data that is stored in the RFID tag. Both issues were discussed extensively in Chap. 21. Eliminating tracking is very important, but also challenging. This is because most existing RFID protocols are based on the ability of the reader to retrieve the tag's ID number, which directly allows tracking. On the other hand, preserving the confidentiality of data stored in the RFID tag can be achieved using encryption. The main challenge is to develop scalable methods that are able to deal with the increasing number of RFID tags.

Healthcare applications in hospital environments also feature biomedical signals. A very important type of signal in the context of the diagnosis of heart-related

disorders is the ElectroCardioGram (ECG). More specifically, ECG classification algorithms are central to the diagnosis of ventricular contractions, fibrillation, and tachycardia. However, these algorithms need to be executed on encrypted data, to address privacy considerations, as discussed in Chap. 22. The main challenge here is to preserve both the representation of the ECG data and the intermediate computation results, which are needed by the classification algorithms. Another challenge, which becomes increasingly important, is to be able to execute these algorithms in real-time. Both challenges warrant further research to be fully addressed.

A different type of emerging medical applications are based on health social network data. These applications focus on issues of patient health management, which include providing emotional support, as well as sharing advice and medical information. Health social network data involve personal and sensitive information and need to be protected from attacks, including re-identification, discrimination, profiling and sensitive information disclosure, as explained in Chap. 23. In addition, it is important to protect the communication and interactions between the users of health social networks, which include patients, doctors, and caregivers. Preserving privacy in health social network data is a multifaceted problem, which needs further research. In particular, it is interesting to develop methods that allow the privacy-preserving sharing of health network data, in accordance with user-specified privacy policies and access controls.

29.5 Topics and Directions in Privacy Preservation Through Policy, Data De-identification, and Data Governance

The preservation of medical data is a legal requirement, posed by various legal frameworks. These include laws in the United States, United Kingdom, and Canada, as well as data sharing policies and regulations that have been developed by major research funding organizations in these countries. These frameworks are a first, important step to protecting medical data but may also act as barriers to the global sharing of data, as discussed in Chap. 24. In addition, the breach of privacy laws and data sharing regulations incurs significant financial costs to healthcare organizations. Interestingly, the cause of such privacy breaches is often human error. Therefore, an analysis of the types of human errors is important to improve our understanding of how these errors lead to privacy breaches and how they can be avoided. Such an analysis was performed in Chap. 25, for the case of the HIPAA privacy rule.

To comply with the aforementioned legal frameworks, the shared data must be de-identified (i.e., be devoid from PHI). This is challenging to perform, particularly in the case of unstructured (text) data, such as doctors' clinical notes and prescriptions. Both the detection and the protection of PHI entails significant computational challenges, which are explained in Chap. 26. The detection of PHI can be performed

based on pattern matching (e.g., based on rules provided by domain experts) and machine learning (e.g., classification) algorithms, while privacy protection is typically achieved by the removal or transformation of PHI. An interesting family of transformation-based de-identification methods were reviewed in Chap. 27. These methods replace PHI with surrogates (i.e., synthetic terms that are realistic, in the sense that they can be read naturally) and offer high utility for certain scenarios, such as those involving correspondence between doctors. Despite the progress in de-identification methods, various challenges remain and offer opportunities for future research. These include increasing the generalizability and accuracy of de-identification methods for specific types of PHI, as well as the development of a solid methodology for quantifying the risk of re-identification attacks based on de-identified text data.

Another requirement that is becoming increasingly important to comply with legal privacy frameworks, is medical data governance. This is a procedural requirement, which assists multiple parties (e.g., a consortium of healthcare organizations) to achieve a common goal, such as perform a research study. Chapter 28 provided a comprehensive discussion of medical data governance. To improve the state-of-the-art in this area, more research is needed to deal with the fact that parties often have different privacy and data analysis requirements, as well as with the increased complexity of data. For instance, it is interesting to extend medical data governance solutions to datasets derived from both EHR systems and health social networks.

29.6 Conclusion

Medical data privacy is an area with a broad spectrum of applications, ranging from genomics to patient monitoring and health social networks. As a result, it has attracted significant research interest from the computer science, medical informatics, and statistics communities. This has resulted in various important and practically useful approaches for preserving privacy that were categorized in four areas. The chapter summarized the topics and methods in each of these areas. In addition, it highlighted some possible directions for future work.

References

1. Dean, J., Ghemawat, S.: Mapreduce: a flexible data processing tool. Commun. ACM **53**(1), 72–77 (2010)
2. Dwork, C.: Differential privacy. In: ICALP, pp. 1–12 (2006)
3. Emam, K.E., Jonker, E., Arbuckle, L., Malin, B.: A systematic review of re-identification attacks on health data. PLoS ONE **6**(12), e28071 (2011)
4. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. J. Biomed. Inform. **50**, 4–19 (2014)

5. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l-diversity. In: ICDE, pp. 106–115 (2007)
6. Nergiz, M.E., Atzori, M., Clifton, C.: Hiding the presence of individuals from shared databases. In: SIGMOD, pp. 665–676 (2007)
7. Poulis, G., Loukides, G., Gkoulalas-Divanis, A., Skiadopoulos, S.: Anonymizing data with relational and transaction attributes. In: ECML/PKDD, pp. 353–369 (2013)
8. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
9. Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **10**(5), 557–570 (2002)
10. Tamersoy, A., Loukides, G., Nergiz, M.E., Saygin, Y., Malin, B.: Anonymization of longitudinal electronic medical records. IEEE Trans. Inf. Technol. Biomed. **16**(3), 413–423 (2012)
11. U.S. Department of Health and Human Services.: Breaches affecting 500 or more individuals. http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/index.html (2015). Accessed 6 Sept 2015
12. Zhang, X., Yang, C., Nepal, S., Chang, L., Dou, W., Chen, J.: A mapreduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud. In: Cloud and Green Computing, pp. 105–112 (2013)
13. Zhang, X., Liu, C., Nepal, S., Yang, C., Dou, W., Chen, J.: A hybrid approach for scalable sub-tree anonymization over big data using mapreduce on cloud. J. Comput. Syst. Sci. **80**(5), 1008–1020 (2014)