International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

A Survey on Publishing Electronic Medical/Health Record with Privacy Preserving

Bhavana A. Khivsara[1], Rajeev Mathur[2], Mahesh R. Sanghavi[3]
[1]*PhD Scholar, PAHER,Udaipur, bhavana.khivsara@gmail.com*
[2] *Principal, LMCST,Jodhpur, rajeev.mathur69@gmail.com*
[3] *Head of Department of Computer Engineering, S.N.J.B's KBJ COE, Chandwad, sanghavi.mahesh@gmail.com*

Abstract- The sharing and integration of electronic medical and health data can be beneficial for a range of medical studies and many research can take the advantages of this data across from clinical experiments to epidemic studies, still it should preserve the privacy of patients. This is because the shared and integrated data need to be protected against several privacy attacks and threats, while utilized for succeeding analysis tasks. In this paper, we present a survey of algorithms that have been proposed for publishing patient health record data, in a privacy-preserving way. We review more number of algorithms, their operation, and highlight their advantages and limitations. We also provide a discussion on directions for future research in this area.

Keywords- Privacy Preserving, Survey, Anonymization, Electronic health records, Algorithms.

I. INTRODUCTION

An electronic health record (EHR), or electronic medical record (EMR), is a systematic collection of electronic health information about an individual patient or population. It is a record in digital format that is theoretically capable of being shared across different health care settings. In some cases this sharing can occur by way of network-connected, enterprise-wide information systems and other information networks or exchanges. EHRs may include a range of data, including demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information. Systems like Electronic Health Record (EHR) are gradually increased the acceptance to collect and store patient data of various types retrieved from multiple different sites, which contain information about patients personal data like zip code, age, gender, education and medical data like medication, allergies, and laboratory test results and diagnosis codes, [4,8]. The EHR systems usage has been increased from 18% in 2001 to 72% in 2012 and is can be exceed to 90% by the end 2020 [6]. Data from EHR systems are increasingly shared for the purposes of improving research [8]. This is because it allows data recipients to perform large-scale, low-cost analytic tasks, which require applying statistical tests (e.g., to study correlations between gender and heart attack), data mining tasks, such as classification and clustering, or query answering. To facilitate the dissemination and reuse of patient- specific data and help the advancement of research, a number of repositories have been established, such as the Database of Genotype and Phenotype (dbGaP) [12], in the U.S., and the U.K. Biobank [17], in the United Kingdom.

II. MOTIVATION

The general scenario for publishing the Electronic Health Records as shown in figure 1, where EHR systems are data providers, which share their data with data publisher. Data publisher publish collect and hold these large volumes of Electronic Health Records. They would like to publish the data for the purposes of data mining which is useful in much research for medical data. But the problem with data publishing is that, it also reveals some information which is considered to be private and sensitive, so the privacy is becoming very important in many data mining applications.

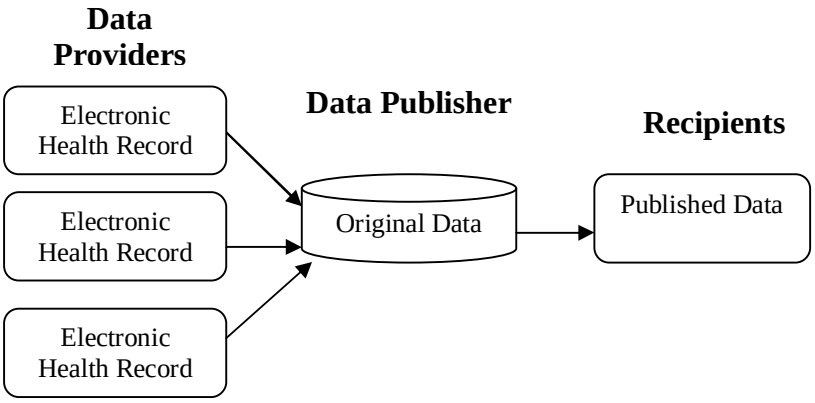

Figure 1: Data Publishing/Sharing Scenario without Privacy Preserving

Following example shows how the released data can be used to re-identify the individual. Suppose, the medical data set as shown in Table I any data base available publically like voter list and Table II is published Electronic Health Record. By linking these two tables, the attacker can easily re-identify that Arjun is suffering from HIV and in this way the privacy of individual is violated. This is happened because the combination of values of Quazi attributes like Zip code, Age and Sex is unique in medical data set.

Table 1 Voter List

NAME	ZIPCODE	AGE	SEX
Mohit	423065	29	M
Sunil	422036	32	F
Rohini	422035	27	F
Arjun	423012	47	M

Table 2 Electronic Health Record

ID	ZIPCODE	AGE	SEX	DIAGNOSIS
1	423065	29	M	Heart Disease
2	422036	32	F	Flu
3	423245	38	M	Headache
4	422035	27	F	Cancer
5	423012	47	M	HIV

While the publishing and sharing of patient health record is greatly useful for research, it must be performed in a way that Way that should preserve the patients' privacy. Many approaches has been proposed to achieve this, by applying various techniques [2], such as cryptography (e.g., [3]) and access control .However, these approaches are not able to offer anonymity to individual patient(i.e., that patients' private and confidential information will not be disclosed) when data about patients are published . This is because the data need to be published and shared extensively to the number of recipients sometimes potentially unknown.

Towards preserving identity, there are number of policies that limit the sharing of patient medical data are emerging worldwide. For example, in the U.S., the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) [22] provides two policies for protecting anonymity, the first policies listed eighteen identifiers that must be removed from data, prior to their publishing or sharing, while in second policy an expert needs to certify that the data to be shared create a low privacy risk before the data can be shared with outsiders. Similar policies are placed in number of countries like the U.K., Canada and the European Union [1]. These policies focus on preventing the privacy threat of identity disclosure (also referred to as re-identification).

To tackle re-identification, and other privacy fear various techniques have been developed. Most of which aim at publishing a dataset of patient records, while satisfying certain privacy and data utility objectives. The privacy is achieved by using privacy models, and enforced algorithms that modify a given dataset. The majority of the proposed algorithms are applicable to data containing

demographics, focus on preventing the threats of identity, attribute, and/or membership disclosure and work by transforming the data using generalization and/or suppression techniques. The Electronic Health Record publishing with privacy preserving algorithm scenario is as shown in fig. 3.

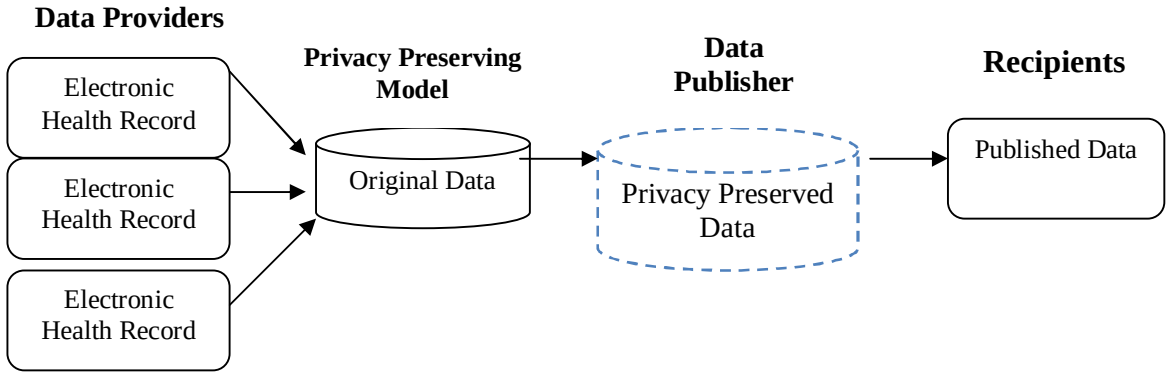

Figure 2: Data Publishing/Sharing Scenario with Privacy Preserving Model

III. LITERATURE REVIEW

In this section, we first discuss the major privacy threats related to the disclosure of sensitive information. Then privacy models are discussed that can be used to protect against different types of threats. The importance of discussing privacy models is that the privacy models can be used to evaluate data safety before to make it public and privacy models can be integrated into algorithms to guarantee that the data can be changed in a way that preserves privacy.

A. Privacy threats

Data set which has to be prone to privacy treat are categorized into three different types of sets, direct identifiers, quasi-identifiers, and sensitive attributes. Direct identifiers attributes can explicitly re-identify individuals, such as individual name, mobile number, mailing address, social security number (Like PAN card or Adhar card in India) or any other national IDs. On the other hand, quasi-identifiers are attributes which in combination can lead to identity disclosure, such as gender, date of birth, age and zip code [18]. Last, sensitive attributes are those that patients are not willing to be associated with. Examples of these attributes are specific diagnosis codes (e.g., psychiatric diseases, HIV, cancer, etc.) and genomic information. In Table Table1, we present an example dataset, in which Name is key identifier, Gender,Age and Zip Code are quasi-identifiers, and Diagnosis is a sensitive attribute.

Based on the above-mentioned types of attributes, we can consider the following classes of privacy threats:

1> Identity disclosure (or re-identification) [19]. It occurs when an attacker can correlate a patient with their record in a published dataset. For example, an attacker may re-identify Bob in Table 1, even if the table is published depressed of the direct identifiers (i.e., Name). This is because Bob is the only person in the table with age 22 and also lives in zip code 423508.

Table 3: Classification of Attributes

Key attribute	Quasi_identifier			Sensitive attributes
Name	Gender	Age	Zip	Diagnosis

2>Membership disclosure [14]: This threat occurs when an attacker can recognize with high probability that an individual's record is enclosed in the published data. For example, consider a

dataset which contains information on only HIV-positive patients. The fact that a patient's record is contained in the dataset allows suppose that the patient is HIV-positive, and thus causes a threat to privacy. Note that membership disclosure may occur even when the data are protected from identity disclosure, and that there are several real-world scenarios where protection against membership disclosure is required. Such interesting scenarios were discussed in detail in [14, 15].

3> Attribute disclosure (or sensitive information disclosure) [11]: This threat occurs when an individual is associated with information about their sensitive attributes. Attribute disclosure occurs when confidential information about an individual is revealed and can be attributed to the individual .This information can be, for example, the individual's value for the sensitive attribute (e.g., the value in Diagnosis in Table1, or a range of values which contain an individual's sensitive value . There have been several incidents of patient data publishing, where identity disclosure has been revealed.

Sweeney [19] first demonstrated the problem in 2002, by linking a claims database, which contains 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 (e.g., Date of birth, Zip code, and Gender) and led to the re-identification of, William Weld, then gov- ernor of Massachusetts. It was also suggested that more than 87% of U.S. citizens could be re-identified, based on such attacks. Many other identity disclosure incidents have been reported since [3].

B. Privacy models

In this section, we present some well-established privacy model that against the aforementioned threats. These privacy models: (i) model what leads to one or more privacy threats and (ii) describe a computational strategy to enforce protection against the threat. Privacy models are subsequently categorized according to the privacy threats they protect from, as also presented in Table 2.

Table 4: Privacy models to guard against different attacks.

Attack Type	Privacy Model
Identity disclosure	k-anonymity
	(1,k)anonymity
	(k,1)anonymity
	(k,k)anonymity
Membership Disclosure	δ-Presence
	c- confidence δ-Presence
Attribute Disclosure	l-diversity
	t-closeness
	Range based
	Variance based

Models against identity disclosure

A plethora of privacy models have been proposed to prevent identity disclosure in medical data publishing. These models can be grouped, based on they type of data to which they are applied, into two major categories: (i) models for demographics and (ii) models for diagnosis codes.

Models for demographics. The most popular privacy model for protecting demographics is k-anonymity [18,19]. k-anonym- ity requires each record in a dataset D to contain the same values in the set of Quasi-IDentifier attributes (QIDs) with at least *k*-1 other tuples in D. Recall that quasi-identifiers are typically innocuous attributes that can be used in combination to link external data sources with the published dataset. Satisfying k-anonymity offers protection against identity disclosure, because it limits the proba bility of linking an individual to their record, based on QIDs,

to 1=*k*. The parameter k controls the level of offered privacy and is set by data publishers, usually to 5 in the context of patient demographics [13].

Another privacy model that has been proposed for demographics is k-map [20]. This model is similar to k-anonymity but considers that the linking is performed based on larger datasets (called population tables), from which the published dataset has been derived. Thus, k-map is less restrictive than k-anonymity, typically allowing the publishing of more detailed patient information, which helps data utility preservation. On the negative side, however, the k-map privacy model is weaker (in terms of offered privacy protection) than k-anonymity because it assumes that: (i) attackers do not know whether a record is included in the published dataset and (ii) data publishers have access to the population table.

CONCLUSION

In this work, we presented a survey of privacy algorithms that have been proposed for publishing structured patient data. We reviewed no of privacy algorithms, derived insights on their operation, and highlighted their advantages and disadvantages. Subsequently, we provided a discussion of some promising directions for future research in this area.

REFERENCES

[1] EU Data Protection Directive 95/46/ECK; 1995.
[2] Aggarwal CC, Yu PS. Privacy-preserving data mining: models and algorithms. Springer; 2008.
[3] Berchtold S, Keim DA, Kriegel H. The x-tree: an index structure for high- dimensional data. In: VLDB; 1996. p. 28–39.
[4] Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Use of electronic medical records for health outcomes research: a literature review. Med Care Res Rev 2010;66(6):611–38.
[5] Gionis A, Mazza A, Tassa T. k-Anonymization revisited. In: ICDE; 2008. p.744–53.
[6] Hsiao CJ, Hing E. Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001–2012. NCHS data brief; 2012. p. 1–8.
[7] Koudas N, Zhang Q, Srivastava D, Yu T. Aggregate query answering on anonymized tables. In: ICDE '07; 2007. p. 116–25.
[8] Lau EC, Mowat FS, Kelsh MA, Legg JC, Engel-Nitz NM, Watson HN, et al. Use of electronic medical records (EMR)for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. Clin Epidemiol 2011;3(1):259–72.
[9] Li N, Li T, Venkatasubramanian S. t-Closeness: privacy beyond k-anonymity and l-diversity. In: ICDE; 2007. p. 106–15.
[10] Loukides G, Shao J. Capturing data usefulness and privacy protection in k-anonymisation. In: SAC; 2007. p. 370-4.
[11] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-Diversity: privacy beyond k-anonymity. In: ICDE; 2006. p. 24.
[12] Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007;39:1181–6.
[13] Malin B, Loukides G, Benitez K, Clayton EW. Identifiability in biobanks: models, measures, and mitigation strategies. Hum Genet 2011;130(3):383–92.
[14] Nergiz ME, Atzori M, Clifton C. Hiding the presence of individuals from shared databases. In: SIGMOD '07; 2007. p. 665–676.
[15] Nergiz ME, Clifton C. d-presence without complete world knowledge. Tkde 2010;22(6):868–83.
[16] Nergiz ME, Clifton C. d-presence without complete world knowledge. IEEE Trans Knowl Data Eng 2010;22(6):868–83.
[17] Ollier WER, Sprosen T, Peakman T. UK biobank: from concept to reality.Pharmacogenomics 2005;6(6): 639–46.
[18] Samarati P. Protecting respondents identities in microdata release. Tkde2001;13(9):1010–27.
[19] Sweeney L. k-anonymity: a model for protecting privacy. Ijufks2002;10:557–70.
[20] Sweeney L. Computational disclosure control: a primer on data privacy protection. PhD thesis, AAI0803469; 2001.

