Health Informatics Meets eHealth
G. Schreier and D. Hayn (Eds.)

doi:10.3233/978-1-61499-858-7-80

Effectiveness of Anonymization Methods in Preserving Patients' Privacy: A Systematic Literature Review

Mostafa LANGARIZADEH[a], Azam OROOJI [a,1], Abbas SHEIKHTAHERI[a]
[a] *Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran*

Abstract. Background: An ever growing for application of electronic health records (EHRs) has improved healthcare providers' communications, access to data for secondary use and promoted the quality of services. Patient's privacy has been changed to a great issue today since there are large loads of critical information in EHRs. Therefore, many privacy preservation techniques have been proposed and anonymization is a common one. Objectives: This study aimed to investigate the effectiveness of anonymization in preserving patients' privacy. Methods: The articles published in the 2005-2016 were included. Pubmed, Cochrane, IEEE and ScienceDirect were searched with a variety of related keywords. Finally, 18 articles were included. Results: In the present study, the relevant anonymization issues were investigated in four categories: secondary use of anonymized data, re-identification risk, anonymization effect on information extraction and inadequacy of current methods for different document types. Conclusion: The results revealed that though anonymization cannot reduce the risk of re-identification to zero, if implemented correctly, can manage to help preserve patient's privacy.

Keywords. Anonymization, confidentiality, electronic health record

1. Introduction

Today, healthcare systems have an increasing emphasis on modernizing infrastructures and replacing paper-based medical records with electronic health record systems (EHRs). This transformation makes new opportunities for secondary use of clinical data [1]. An EHR is a system embracing patients' demographic, diagnostic, lab and medication data. These data can be accessed and shared through computer networks among healthcare providers in different organizations [2]. Therefore, EHR is a rich resource for secondary uses such as research, quality assessment and epidemiology [2,3]. However, the high volume of identifiable personal data in EHRs would threaten individuals' privacy [4]. Studies showed that 59% of patients believed that EHR has increased risk of data loss or privacy breach [5]. In this regard, policies and regulations have been developed worldwide to restrict identifiable data sharing and to reduce privacy concerns, such as General Data Protection Regulation (GDPR) was approved by the European Union parliament on April 2016[6]. In the U.S., Health Insurance Portability and Accountability Act (HIPAA) enacted in 1996 and Health Information Technology for Economic and Clinical Health (HITECH) Act enacted in 2009 [7]. According to many regulations, the

[1] Corresponding Author: PhD candidate in medical informatics, Department of Health Information Management, School of Health Management and Information Sciences, No. 6, Rashid Yasemi Av., Vali-e-Asr St., Tehran, Iran. Tel: +98-21-88794301; E-mail: orooji.a@tak.iums.ac.ir

medical team, and those permitted by the patient and those authorized according to law are permitted to access patient's information [8].

Based on many of these policies, researchers and other secondary users can only access identifiable data only if they obtain the required permission from ethical committees and informed consent from patients. This procedure is time-consuming and sometime impossible, especially when the target population is large. However, if the data is anonymized, there is no need for informed consent [9-10]. ISO defines anonymization as the act of eliminating the links between identifiable data and the data subject [11]. In recent years, many techniques have been developed for privacy preservation for healthcare data such as Anonymization, Perturbation, Condensation, Randomization and Fuzzy based methods; among them, anonymization showed to be a promising method [12-13]. Although, anonymization techniques are capable of preserving privacy, they may negatively affect data utility for secondary uses. [12]. In fact, anonymization methods and the resulted anonymized data may result in negative effect on secondary use of data. Hence, many studies have been conducted to investigate these issues and develop solutions for the problems. The aim of this research was to systematically review and categorize the problems of the anonymization techniques and their effect on secondary use of patients' data.

2. Methods

The articles published in the 2005-2016 (conference and peer-reviewed papers) were included. Pubmed, Cochrane, IEEE and ScienceDirect were searched with a variety of related keywords (Table 1). Newspapers, reviews, letter to editor, workshop reports, posters, short reports, books and thesis, articles written in non-English language and also papers related to anonymization on non-health data were excluded. Articles with no access to the full-text were also excluded. In addition, articles were selected if they related to secondary use of anonymized data. In fact, papers related to develop methods for anonymization or those related to problems of these methods were also excluded.

A total of 659 papers were initially identified. All duplicated (n=58) and non-English (n=13) articles were removed using Endnote software, however, a manual revision was done for verification. Two reviewers independently screened titles (n=588) and abstracts (n=165) and then reviewed the full texts (n=46). Discrepancies resolved by consensus. Finally, 18 publications were included in the review. 28 papers after reviewing full text were excluded mainly because they were related to the problems of the methods not the problems of anonymized data for secondary use.

3. Results

Based on the included papers (aims and results), we classified and discussed the issues of anonymizations in four categories including: (1) data secondary use (SU), (2) re-identification risks (RR), (3) effect on information extraction (IE), and (4) Inadequacy of current methods for heterogeneous documents (IN). Following paragraphs describes each of these categories in detail.

Table 1: The search query used in this study

Science Direct	pub-date > 2005 and pub-date < 2016 and (TITLE-ABSTR-KEY(de-identif*) or TITLE-ABSTR-KEY(deidentif*) or TITLE-ABSTR-KEY(Anonymization) or TITLE-ABSTR-KEY(De-personalization) or TITLE-ABSTR-KEY(Depersonalization) or TITLE-ABSTR-KEY(Pseudonymization)) and ("electronic health record" or "electronic medical record").
Pubmed	(Electronic health record [All Fields] OR Electronic medical record [All Fields]) AND (de-identif*[Title/Abstract] OR deidentif* [Title/Abstract] OR Anonymization[Title/Abstract] OR De-personalization [Title/Abstract] OR Depersonalization [Title/Abstract] OR Pseudonymization [Title/Abstract]) AND ("2006/01/01"[PDAT] : "2016/01/01"[PDAT])
Cochrane	(electronic health record:ti,ab,kw or electronic medical record:ti,ab,kw)and (deidentif*:ti,ab, kw or deidentif*:ti,ab,kw or Pseudonymization:ti,ab,kw or anonymization:ti,ab,kw) Publication year from 2006 to 2016
IEEE	((electronic health record OR electronic medical record) AND (Abstract:deidentif* OR "Abstract":anonymization OR "Abstract":Pseudonymization OR "Abstract":deidentif* OR "Document Title":deidentif* OR "Document Title":anonymization OR "Document Title":Pseudonymization OR "Document Title":deidentif*)) and refined by Year: 2006-2016

3.1. Data secondary use

Secondary usage of health data plays a key role in promoting medical knowledge. In the primary use of EHRs, providing healthcare services, it is necessary to include patient's identification information within the records. In secondary use, however, there is no need for this information [14]. In recent years, there have been many techniques used to preserve patient's privacy. However, one of the most problem of these techniques is possible elimination of much valuable information required for the research purposes and other secondary uses [12]. Therefore, many scientific articles focused on this issue and even proposed a number of methods to strike a balance between privacy preservation and maintaining data value. In this regard, seven papers out of 18 included articles focused on this issue [9, 15-20] (Table 2). For example, Neuberger [18] investigated anonymization approaches and showed that pseudonymization was the best method of striking a balance between data secondary use and privacy preservation. As another example, applications need to be piloted before use in actual environment. Currently, testing is performed with fake data often leads to worse code coverage and fewer uncovered bugs, so testing with real data is important. However, different data privacy laws prevent organizations from sharing these data with test centers because databases contain sensitive information. In this regard, Grechanik [19] proposed a solution for use of anonymized real data in evaluating the effectiveness of such applications.

3.2. Re-identification risk

Re-identification is a process in which attempts taken to find the owner of a record or document which has already been anonymized [21]. Attackers can re-identify data by linking the anonymized data to the other accessible datasets; therefore, anonymization techniques do not guarantee the anonymity of data [16, 22]. Four studies out of included papers focused on this issue [20, 22-24] (Table 2). Some studies introduced methods of estimating the re-identification risk of records [22]. In spite of all efforts to prevent re-identification, the necessity of right legislation for the relevant delinquencies is explored by some researchers [23]. Only one investigation addressed the details of cost-effectiveness evaluation of re-identifying of health data. In this study, the cost of each record was estimated in accordance with the value of each attribute [20]. In addition, El-

Emam [3] found that many attacks succeed due to the inefficiency of the existing anonymization methods.

3.3. Effect on information extraction

Anonymization is a barrier to implementing effective data retrieval mechanisms. Since, it is not possible to do effective query for relevant data using anonymized data [25]. Due to the significance of the issue, many investigations have been conducted to assess the effect of anonymization on different operations such as data extraction or retrieval (nine papers) [2, 9-10, 12-13, 16-17, 25-26]. Some researchers proposed strategies to solve this problem. For instance, an efficient approach was proposed to maintain data appropriateness for data mining purposes even after anonymization [13, 25].

3.4. Inadequacy of current methods for heterogeneous documents

Sometimes, elimination of all identifiers is not even enough to preserve privacy in a special type of document [27]. Studies have shown that different identifiers and documents need to different anonymization approaches. For example, Omran et al. [27] dealt with a key problem through a known anonymization method named k-anonymity. Through k-anonymity, it is not possible to precisely determine which identifiers to be generalized and which to be suppressed. To solve this problem, an ontology-based strategy has been proposed. Moreover, the majority of anonymization methods have been evaluated on a special type of clinical data. Ferrández also indicated that an anonymization method developed for a specific document corpus cannot be appropriate for other types [28]. Among included publications, three papers focused on this issue [20, 27-28] (Table 2).

Table 2: Summary of the research results

Author	**category**	**Aim**	**Related findings**
Neubauer [18]	SU	Assessing existing privacy enhancing methods including anonymization, encryption, depersonalization, role-based access control and pseudonymization.	Pseudonymization supports the privacy of patients and keeps data accuracy intact for secondary usage.
Grechanik [19]	SU	Introducing a new view with which organizations can determine how much test coverage they can lose when using data privacy to database-centric applications.	Using *k*–anonymity (a data privacy approach) leads to serious degradation of test coverage.
Qingming [16]	SU, IE	A utility-based k-anonymity is proposed.	The proposed algorithm has completely less normalized certainty penalty (NCP) cost and it has lower query ratio than other algorithms. $^{+}$
Harada [15]	SU	A new k-anonymity approach in which generalization hierarchies are automatically made by input information.	Automatically establishing generalization hierarchies decreases information loss following k-anonymization *
Loukides [17]	SU, IE	A new anonymization algorithm is suggested, which uses generalization and suppression to choose items, based on data publishers' utility needs, and is leaded by introducing utility criterion measure.	Using an efficient utility measure, this algorithm can be very useful at preserving data utility. It also allows more accurate query answering than other methods.

Author	category	Aim	Related findings
Benitez [22]	RR	Some techniques are introduced to estimate re-identification risk for many de-identification(De-ID) data sharing policies.	The differences in distributing population of U.S. states and their policies for disseminating datasets lead to varying re-identification risks.
Rothstein [24]	RR	Adverse effects of nonconsensual use of anonymized health data in research are included	De-ID was considered an important but insufficient means of keeping health privacy. It is indefensible from technical, ethical, and policy views to go on drawing a regulatory distinction between identifiable and de-identified health data.
Gellman [23]	RR	Risks and dangers to subjects and the research community are highlighted from use of supposedly anonymized information.	There is no ready enforcement for De-ID failures. The use of anonymized information for research goals should be regulated or even forbidden. Although, additional restrictions will make research impossible.
Khokhar [20]	SU, RR, IN	A cost-benefit evaluation is done to test related cost factors associated with the value of anonymized data and the possible damage cost due to privacy breaches.	The analytical cost model is efficient for health information custodians (HICs) to decide better on sharing health data for secondary and commercial uses.
Panackal [13]	IE	Introducing an adaptive utility based anonymization for accessing privacy without compromising the content of data or data mining accuracy	Both original and anonymized data sets are tested for classification accuracy and the conclusions showed that the anonymization process does not provide any important degradation in the accuracy of data mining classification.
Pruski [25]	IE	Introducing an ontology-based approach for efficient information retrieval in encrypted EHRs	The combined use of metadata and ontologies offers exciting features to improve, in terms of relevance, the results of a search. In addition, the uses of standard vocabularies make the construction and interpretation of the queries easier.
Deleger [26]	IE	(1) Evaluating the natural language processing (NLP)-based method to de-identify a large set of diverse clinical notes automatically. (2) Measuring the effect of De-ID on the performance of IE algorithms on the anonymized documents.	The performance of the system was indistinguishable from that of human annotators. The impact of automated De-ID was minimal on the utility of the narrative notes for subsequent IE as measured by the sensitivity and precision of medication name extraction.
Wu [12]	IE	To investigate the issue of health data utility after three anonymization methods with new criterions to assess the data utility (Support Vector Machine (SVM) and Earth Mover's Distance (EMD)$^{\times}$)	The results revealed that there is a significant difference in classification accuracy between evaluations on the original and anonymized data. In EMD experiment, it is shown that privacy preservation methods can significantly jeopardize the data utility due to the highly strict protection principles they impose.
Meystre [9]	SU , IE	The effect of five different De-ID methods was investigated based on clinical text information content (informative and formatting) and clinical information extraction by comparing counts of SNOMED-CT concepts found in the original and anonymized corpus.	The informativeness was only minimally altered by these systems while formatting was only changed by one system. Only about 1.2–3% less SNOMED-CT concepts were identified in anonymized corpus.

Author	category	Aim	Related findings
Gkoulalas-Divanis [2]	IE	Introducing a novel anonymization methods which is able to anonymize data with a desired balance between utility and privacy[α]	The results showed the relative error in query answering.
Liu [10]	IE	Evaluating performance of four De-ID methods that may be used to ensure regulatory compliance while also making practical database updating and querying easier.	Different De-ID methods have different effects on database operations such as time needed for data insertion, initial data loading and query on the database. Overall, De-ID has an undesirable effect on longitudinal study prevention.
Omran [27]	IN	A new ontology-base k-anonymization is proposed to determine which information can be generalized and which information needs to be suppressed.	The method could play an important role in protecting the privacy of personal health records without sacrificing the value of information for primary and secondary usages.
Ferrández [28]	IN	Various De-ID methods are comparing according to the generalizability and portability on different document sources as train and test sets.	There is no good report and results for these three systems as generalizability experiment.

[+] *Normalized certainty penalty (NCP) and Query answerability are two metrics that measure the utility of the data.*

[*] *Information loss is measured in terms of information entropy using a frequency distribution.*

[×] *Classification by SVM and evaluating the similarity between anonymized and original tables based on EDM are two approaches for investigating the utility loss of privacy preservation techniques.*

[α] *The utility policy constructed by Utility Policy Extraction (UPE) leads to the production of anonymized data that allows accurately computing the number of patients with the selected diseases.*

4. Discussion

The present study explored the efficiency of and the issues related to the anonymization of EHRs. It revealed that anonymization, though appropriate for preserving patients' privacy of healthcare data, cannot dispose of data re-identification risk altogether. On the other hand, when through the anonymization process a great portion of identifiable data is removed, the data will not be appropriate for a secondary use. This issue has been pinpointed in Meystre's study [29] under the title of over-scrubbing. Many anonymization methods have been suggested which mainly addressed data utility after anonymization. Such methods were explored in the first category i.e. data secondary use. The findings of the second category showed that if standard methods are followed for anonymization, there will be a lower risk of re-identification. Moreover, an accurate and detailed analysis of different types of re-identification risks and their effects could be helpful to data disclosure policy making and the right implementation of anonymization methods.

The main goal of data aggregation is analysis and use of the extracted information. The effect of anonymization on IE and database operations has been explored separately in the third category. Recent investigations proved that new anonymization methods have inconsiderable impacts on database operations, IE-based applications and text mining. Another issue taken into account was the text type. Unfortunately, the majority of anonymization methods have been evaluated on a specific type of clinical notes. Furthermore, they are mostly focused on English-language texts. However, the body of research selected in the fourth category showed that an anonymization method, once designed for a certain type of clinical texts, will not produce desirable results on other text types.

All issues, covered here, addressed the significant issues related to anonymization domain which has got to be resolved through efficient approaches. However, this does not imply that anonymization is improper for privacy preservation.

The present study systematically reviewed the recent published researches about patient information anonymization. The effectiveness of this anonymization procedure was investigated in four categories: secondary use of anonymized data, re-identification risk, effect of anonymization on IE, and inadequacy of current methods for different text types. Although anonymization does not reduce the risk of re-identification to zero, if implemented correctly, could be useful in preserving patients' privacy. Moreover, a comprehensive analysis of different types of re-identification attacks plays a key role in data exposure policy making and developing anonymization algorithms.

References

[1] I. Anciu, J.D. Cowan, M. Basford, X. Wang, A. Saip, S. Osgood, et al, Secondary use of clinical data: the Vanderbilt approach, *J Biomed Inform* **52** (2014), 28-35.

[2] A. Gkoulalas-Divanis, G. Loukides, J. Sun, Toward smarter healthcare: Anonymizing medical data to support research studies, *IBM J Res Dev* **58**(1) (2014), 9:11.

[3] E.l. Emam, E. Jonker, L. Arbuckle, B. Malin, A systematic review of re-identification attacks on health data. *PloS one* **6**(12) (2011), e28071.

[4] P.C Kaur, T. Ghorpade, V, Mane, Analysis of data security by using anonymization techniques. 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence); 14-15 Jan (2016).

[5] National Partnership foe women and Families, Making IT meaningful: How consumers value and trust health IT, Washington, DC, USA (2012).

[6] EU General Data Protection Regulation, https://www.eugdpr.org, last access: 3.3.2018.

[7] K.A. Wager, F.W. Lee, J.P. Glaser, *Health Care Information Regulations, Laws, and Standards*, in: Health Care Information Systems: A Practical Approach for Health Care Management. Wiley, San Francisco, CA., 2013. pp. 85-90.

[8] E. Khorshidi, Patient's Bill of Rights, http://www.e-khorshidi-lawyer.ir/index.php?ToDo=ShowArticles&AID=13145, last access: 3.3.2018.

[9] S.M. Meystre, O. Ferrandez, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore. Text de-identification for privacy protection: a study of its impact on clinical text information content, *J Biomed Inform* **59** (2014), 142-150.

[10] J. Liu, S. Erdal, S.A. Silvey, J. Ding, J.D. Riedel, C.B. Marsh, et al, Toward a fully de-identified biomedical information warehouse. AMIA Annu Symp proc AMIA Symposium, (2009), 370-374.

[11] ISO, Health informatics – Pseudonymization, (2017).

[12] L. Wu, H. He, O.R. Zaïane, Utility of privacy preservation for health data publishing. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, (2013), 510-511.

[13] J.J Panackal, A.S. Pillai, Adaptive Utility-based Anonymization Model: Performance Evaluation on Big Data Sets, *Procedia Comput Sci* **50** (2015), 347-352.

[14] B.S. Elger, J. Iavindrasana, L.L. Iacono, H. Müller, N. Roduit, P. Summers, et al. Strategies for health data exchange for secondary, cross-institutional clinical research, *Comput Methods Programs Biomed* **99**(3) 2010 ,230-251.

[15] K. Harada, Y. Sato, Y. Togashi, Reducing Amount of Information Loss in k-Anonymization for Secondary Use of Collected Personal Information. 2012 Annual SRII Global Conference, (2012).

[16] T. Qingming, W. Yinjie, L. Shangbin, W. Xiaodong, Utility-based k-anonymization, The 6th International Conference on Networked Computing and Advanced Information Management, (2010).

[17] G. Loukides, A. Gkoulalas-Divanis, Utility-preserving transaction data anonymization with low information loss, *Expert Syst Appl*, **39**(10) 2012, 9764-9777.

[18] T. Neubauer, B. Riedl, Improving patients privacy with Pseudonymization, *Stud health technol inform* **136** (2008), 691-696.

[19] M. Grechanik, C. Csallner, C. Fu, Q. Xie, Is Data Privacy Always Good for Software Testing? 2010 IEEE 21st International Symposium on Software Reliability Engineering, (2010).

[20] R.H. Khokhar, R. Chen, B.C. Fung, S.M. Lui, Quantifying the costs and benefits of privacy-preserving health data publishing, *J Biomed Inform*, **50** (2014), 107-121.

[21] H. Kikuchi, T. Yamaguchi, K. Hamada, Y. Yamaoka, H. Oguri, J. Sakuma, Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization, 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), (2016).

[22] K. Benitez, B. Malin, Evaluating re-identification risks with respect to the HIPAA privacy rule, *J Am Med Inform Assoc*, **17**(2) 2010, 169-177.

[23] R. Gellman, Why deidentification fails research subjects and researchers, *Am J Bioeth* **10**(9) (2010), 28-30.

[24] M.A. Rothstein, Is deidentification sufficient to protect health privacy in research?, *Am J Bioeth* **10**(9) (2010), 3-11.

[25] C. Pruski, F. Wisniewski, Efficient medical information retrieval in encrypted Electronic Health Records, *Stud Health Technol Inform* **180** (2012), 225-229.

[26] L. Deleger, K. Molnar, G. Savova, F. Xia, T. Lingren, Q. Li, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction, *J Am Med Inform Assoc*, **20**(1) (2013), 84-94.

[27] E. Omran, A. Bokma, S. Abu-Almaati, A K-anonymity Based Semantic Model For Protecting Personal Information and Privacy, IEEE International Advance Computing Conference, (2009).

[28] O. Ferrandez, B.R. South, S. Shen, F.J. Friedlin, M.H. Samore, S.M. Meystre, Generalizability and comparison of automatic clinical text de-identification methods and resources, AMIA Ann Symp proc, (2012), 199-208.

[29] S.M. Meystre, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, *BMC Med Res Methodol* **10** (2010).