Lee *et al. Hum. Cent. Comput. Inf. Sci.* (2018) 8:19
https://doi.org/10.1186/s13673-018-0142-9

Human-centric Computing and Information Sciences

RESEARCH Open Access

Design of a human-centric de-identification framework for utilizing various clinical research data

Jaedong Lee[1,2], Jipmin Jung[1], Phillip Park[1], Seunghyun Chung[1] and Hyosoung Cha[1*]

*Correspondence:
kkido@ncc.re.kr
[1] National Cancer Center, 323 Ilsanro, Ilsandong-gu, Goyang-si, Gyeonggi-do 10408, Republic of Korea
Full list of author information is available at the end of the article

Abstract

For better life, medical and IT technology are converging and data plays a key role in this convergence. Data in the medical field is information about humans, and these data are sensitive. Because this type of data is frequently accessed by multiple users, a high degree of caution is required during handling. In addition, systematic security precautions must be followed to prevent data from being used outside of the intended purpose, even in clinical research. In this paper, we propose a human-centric framework for clinical environments based on the standards, procedures, and methods outlined in guidelines published in the United States and Korea. This study provides a more balanced approach to the utilization and security of personal information as compared to that in the two previously published guidelines. For the secure clinical studies, this framework makes it possible to provide utility and security in a balanced manner, depending on the forms of provision. As a result, the proposed framework extends the usability of the clinical study, and support secure usage of clinical research data.

Keywords: Privacy, Electronic medical records, Clinical data, De-identification

Introduction

In the fourth industrial revolution, which has seen the rapid development of information and communications technology (ICT), data usage has increased in many fields. Thus, both the number of devices for consuming data and the amount of data consumed are increasing. In the same way that data and information are produced in a variety of fields, the advent of extensive computerization in medical institutions has resulted in the production of many complex and diverse forms of medical data. As a result, in the medical field, requests for access to data are increasing [1, 2].

The collection of clinical data or health information to support healthy living has contributed to the development of medical and life extension technologies. However, negative consequences arise when data is collected, utilized, and interpreted for purposes other than those for which it was collected. Because clinical information often includes sensitive information about a particular individual, e.g., health status, progress of treatments, etc., there can be significant mental, social, and economic damage if it is disclosed improperly. In addition, clinical information collected during treatment processes

Springer Open

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

is then utilized within various departments and uploaded into diverse related institutions. Therefore, as there are various collection methods and many possible external distribution routes, there is a high likelihood that secure processing requirements will be violated. In addition, clinical information can be used for a multitude of purposes, such as primary disease occurrence trend tracking, prescription drug market status, clinical site tracking, and insurance product development via secondary information processing. Therefore, the social cost incurred when clinical data is compromised is increasing when compared with other industries. For these reasons, the importance of securing clinical information has become paramount. Currently, various studies are researching ways to improve the utilization of information while protecting the medical information of subjects [3–7].

Anonymization is a method of converting data into an unrecognizable form without considering re-identification. In contrast, de-identification is a method of converting data into a controlled form that is neither directly or indirectly recognizable, while considering re-identification. De-identification standards have been created for the secure usage and protection of personal identifier information in the United States [8].

In this paper, we propose the implementation of technical and administrative security systems to maintain a balance between utilization and security. Our focus is on the various data that exist in an actual clinical environment, and on the design of a de-identification framework that facilitates the utilization of various clinical research data in a general hospital. This paper consists of five sections including "Introduction": in "Related works" section, we give a survey of most important related works highlighting clinical research data, the de-identification methods and security requirements. In "De-identification framework for the utilization of various clinical research data" section, we describe details regarding our de-identification framework. In "Comparison and analysis" section, we compare and analyze the proposed frameworks. In "Conclusion" section, we present a conclusion which summarizes the paper and describe future research.

Related works

Clinical research data and de-identification

A clinical research data warehouse can be divided into a clinical registry and clinical data. In the case of the clinical registry, this is the data generated from requestor-driven or researcher-driven clinical tests. As this is uniform data in terms of its condition and disclosure, it is used to treat a particular disease. In contrast, clinical data is the data generated during a treatment process that is captured in an electronic medical record (EMR), and consists of lab results, radiology images, diagnoses, etc. Due to its vastness and diversity, general hospitals integrate and manage data in a clinical research data warehouse for researchers to use whenever needed [9].

An EMR, which refers to a systematized digital bundle of data containing the health information of a patient within a hospital, includes a variety of sensitive information, such as age, weight, medical history, medication information, radiology images, and test results. In addition, EMRs are shared and utilized throughout various interconnected information systems and devices within a hospital. In some countries, the protection of the information on EMRs is legislated, and compliance with the law is strictly enforced

so that individuals cannot be identified when processing data from relevant datasets [10, 11].

In terms of privacy protection, data is classified into direct identifiers (DIDs), quasi-identifiers (QIDs), sensitive attributes (SAs), and non-sensitive attributes (NSAs). DIDs denote information, subjects, or data that is directly related to a patient. This includes social security numbers, patient numbers, mobile phone numbers, etc. Although QIDs cannot be identified as a single item, it may be possible to infer a potential identifier through a combination of several items, including address, blood type, and height. Among QIDs, sensitive attributes are information related to individuals. When combined with other information, it may be possible to identify specific individuals, and thereby cause serious harm [12].

Data can be classified as structured, semi-structured, and unstructured, depending on its form. A standard method for classifying types of data is to determine whether it is in schema form or calculable form. The data is structured data if a form exists and is calculable. If a form exists but is non-calculable, the data is classified as semi-structured data. Unstructured data is data that neither has form nor is calculable. In order to analyze unstructured data, additional formalization work is required [13–15].

International de-identification methods

Since there are differences in the definition and procedures for de-identification and anonymization by country, it is necessary to understand these differences. Therefore, in this section, we describe the key definitions of de-identification and pseudonymization in the European Union, United States, and Korea [16–22].

For de-identification and pseudonymization, the definition of the scope or content is a key factor because it may affect the results of the non-discrimination depending on the definition. In the European Union, pseudonymization is defined as the method by which specific information entities cannot be identified without using additional separately stored information. In contrast, in the United States, pseudonymization is defined as a technique to eliminate linkage among information entities, and is a sub-concept of de-identification. The de-identification process defined in the HIPPA Privacy Rules refers to expert determination, which relies on expert judgment, and the safe harbor method, which deletes 18 identifiers. In Korea, information that can be easily combined with other information and recognizable by the person handling the information is defined as personal information. Among these, de-identification information is defined as the information from which an individual identification element can deleted in whole or in part so that the individual cannot be identified. Since it is a nation-wide standard, individual standards that apply to specific fields of application are not defined in most cases. For example, in the United States, the DID and QID are not classified separately, but are defined using simple examples. As for the DID, it is explained in the safe harbor of HIPPA and the ISO/TS 25237 standard. As for the QID, birthdays, zipcodes, and gender are examples of information with which individuals can be identified when connected with other information. In Korea, the definition of the DID is the same as in the United States, but the QID is categorized according to the characteristics of the attributes into personal characteristics, physical characteristics, credit characteristics, career characteristics, electronic characteristics, and family characteristics, and examples are provided.

There are no separate examples provided in the relevant standards in the European Union [16–19, 22].

For de-identification, the data processing technique has a great effect on the results of de-identification. For this purpose, Korea, the United States, and the European Union use techniques such as masking, transformation, suppression, generalization, and perturbation [16, 18, 19, 22].

An adequacy test that verifies de-identification is the final step. This test analyzes whether the de-identified data has sufficient utility and whether the risk of re-identification is thought to be impossible. In the United States, the re-identification risk is evaluated with sample data after the initial re-identification risk threshold has been established. After comparing the evaluation with the actual re-identification risk, de-identification would be applied if the actual calculated risk is lower than the threshold, while new parameters or transformations are considered if the risk is higher than the threshold. In Korea, an evaluation group that is capable of an objective evaluation of the object data must be established, and an adequacy test of the de-identification level must be executed using basic resources and a k-anonymity model. In the case of the European Union GDPR, an adequacy test is not described [16, 18, 19, 22].

De-identification steps and security requirements

In general, many studies classify structured and unstructured data de-identification into classification, DID processing, QID processing, and SA processing (optional). Table 1 shows the security threats that can occur in each step.

Classification is used to classify data according to the data type (i.e., unstructured data or structured data) using heuristic and artificial intelligence technologies. However, both types can be misclassified by assigning incorrect attributes or applying a biased standard. Since these errors impact the entire process of de-identification, the corresponding data must be re-classified or purged when errors are detected [8, 23, 24].

In the DID, QID, and SA processes, the generalization, randomization and elimination methods are primarily used for structured data while the cryptography, replacement, and elimination methods are used for unstructured data due to the difficulty of detection and processing [25].

Table 1 De-identification threats of structured data and unstructured data

Type	1. Classification	2. Processing of DID	3. Processing of QID	4. Processing of SA (optional)
Structured data				
Methods	Manual (human) Heuristic Artificial Intelligence	Generalization, Randomization, Elimination		
Threats	Mis-classification	Misprocess, Single out, Likability, Inference	Homogeneity attack, Background knowledge attack	Similarity attack, Skewness attack
Unstructured data				
Methods	Manual (human) Heuristic Artificial Intelligent	Cryptography, Replacement, Elimination		
Threats	Mis-classification	Misprocess, Single out, Likability, Inference		

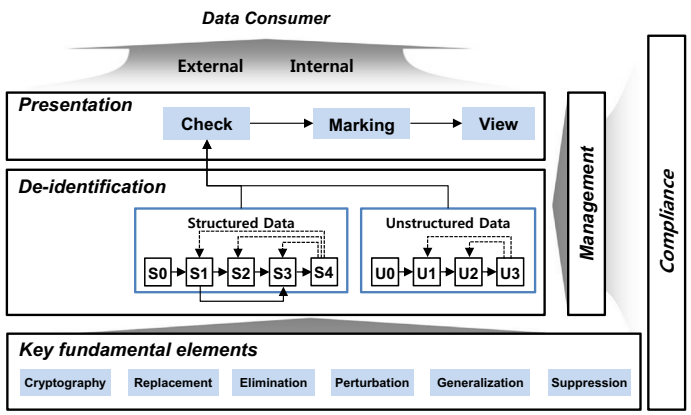

Fig. 1 Architecture of proposed framework

In the DID processing step, the process is performed on a value that directly refers to an individual. The following are common security threats in data de-identification. (A) Errors due to the choice of weak criteria and/or an out of range value (misprocess), (B) when an individual item is identifiable (singled out), (C) linkable if the individual item is directly identifiable through association with other identifiable items, and (D) it is possible to infer a specific person through the meaning of the attribute of the value (Inference). These four types of threat are the same as those that occur when de-identifying unstructured data. In addition, the rigorous application of the de-identification standard and management are required because the connectivity between data subjects is strong in the case of DID [26–31].

In the QID processing step, processing is performed on values that are potentially identifiable in combination with other information. Even though the QID is indirect information related to a specific information subject, if it does not account for sensitive attributes belonging to an equivalence class that are identical after de-identification, there is a threat that specific information can be inferred about subjects through homogeneity and background knowledge [26, 27, 32].

If sensitive content is contained even during the processing of the SA and QID, it must be classified separately and processed uniquely. It is safe to ensure that the number of different sensitive attributes is sufficient so as to be difficult to identify. At the time, there is a threat of re-identification if the distribution or ratio of the specific value in the QID group and the semantic closeness after de-identification are not taken into consideration [23, 27, 33].

De-identification framework for the utilization of various clinical research data

Architecture

In this section, the structure of the framework used to de-identify various types of clinical research data is described. As shown in Fig. 1, the proposed framework is divided into six elements, each of which have several processes.

Key fundamental elements is a component that contains basic and core technology for de-identification. This component consists of six components: cryptography,

replacement, elimination, perturbation, generalization and suppression. At the bottom, *De-identification* component is supported, and *Management* component is controlled.

De-identification component is located between *Key fundamental elements* and *Presentation* component, and performs primary de-identification processing of input data according to data type. During this process, the *Key fundamental elements* are supported, and if the primary de-identified data is requested, *Presentation* component provides the requested data.

Presentation is a component that performs secondary de-identification before providing data to the data consumer. In addition, it receives first de-identification data from the *De-identification* component, and is responsible for data transmission and display to the data consumer.

Data consumer is the topmost and refers to the subject that receives the de-identified data. *Data consumer* is divided into internal and external and different compliances are applied.

Management component is responsible for controlling each component in the framework to work organically. *Management* component is also located between each component and compliance. It coordinates the compliance of laws and regulations at this position.

Compliance refers to the legal regulations or guidelines that must be complied during de-identification. It also includes the ethics, which refers to the common values held by the majority, in order to prepare for the possibility that the rights of the information subjects may be invaded even though the developed technology complies with all legal regulations or guidelines.

Service scenario

The de-identification framework for medical institutions must be applicable to actual medical institutions and accommodate a variety of regulations and technical measures. Figure 2 depicts the scenarios in which the practical elements of compliance and supporting components are applied to the proposed framework. Likewise, the process,

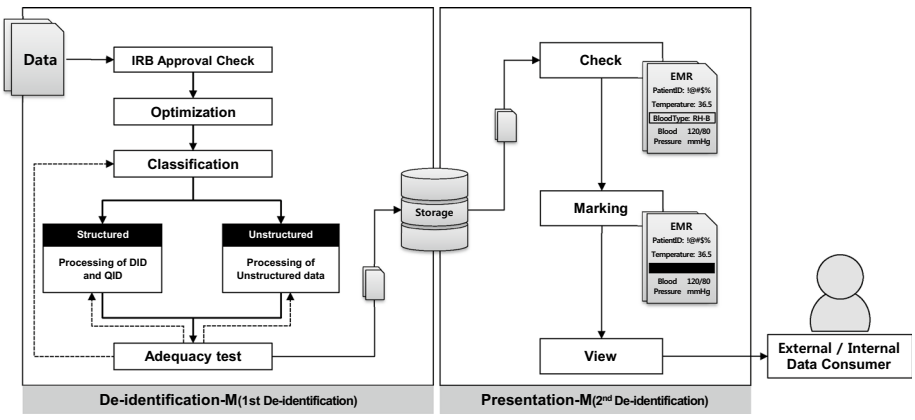

Fig. 2 Service scenario

Table 2 Data optimization and classification in structured data and unstructured data

Structured data	Unstructured data

//S0. Data Optimization
OptimizeID(inputdata){
Auth = Authentication(irbapproval, projectid)
if(Auth == CORRECT){//equal?
ChkResult = CheckValidation(inputdata)
if(ChkResult == ERROR)
exit
else
ClassifyID(ChkResult)
}
exit
}
	//U0. File Optimization
	SelectOptiFILE(){
	Auth = Authentication(irbapproval, projectid)
	if(Auth == CORRECT){//equal?
	SelectedFile = FileSelection(inputfile)
	ChkFile = CheckValidation(SelectedFile)
	if(ChkFile == ERROR)
	exit
	else
	ClassifyFILE(ChkFile)
	}
	exit
	}

//S1. Data Classification
ClassifyID(ChkResult) {
classifiedID[DID|QID|SA] = locator(ChkResult)
if (classifiedID == DID)
goto S2
else if (classifiedID == QID){
if (classifiedID == SA) {
{checkSA = 1}
goto S3
}
else//NSA
goto S2
}
	//U1. File Classification
	ClassifyFILE(ChkFile) {
	ClfFile[C|U] = CheckChangeable(ChkFile)
	if(ClfFile == C)//Changeable File
	goto U2
	else
	goto U2//Unchangeable File
	}

technique, and evaluation of de-identification are proposed as key points. Each step is then illustrated using pseudocode and descriptions thereof.

Data optimization and classification

In Table 2, data optimization and classification refers to the preprocessing steps performed before de-identification to improve the quality of the de-identification results. Through this process, de-identification errors that occur because of data entry errors and the absence of data can be reduced. In the proposed framework, the IRB approval and research project numbers are identified before optimizing the structured and unstructured data to prevent erroneous inputs and processes. For the structured data, all columns are classified into DID, QID, SA, and NSA after checking for errors in the input data. Next, the unstructured data is classified into changeable and unchangeable files, and, depending on whether they are changeable, check for errors after processing.

Steps S2 and S3, which process the DID and QID of the structured data, and U2, which is used to process files separated by unstructured data, are described separately.

Processing of structured data (DID and QID)

In Table 3, the DID and QID processes are performed in the order of the initial processes: standard setting, non-identification, and risk identification. Detailed de-identification techniques are classified according to the features of the target data. DID requires cryptography, replacement, and elimination techniques due to the strong connection to the information subject. In addition, although the QID does not have as strong a connection as the DID, it is potentially identifiable when combined with other information.

Table 3 Processing of identifier and quasi-identifier

```
//S2. Processing of identifier
HandleID(id, rule){
  setRule = basicRule
  CheckedRisk = CheckRisk(value1)
  PermittedRisk = SetRisk(value2)
//The Rule, ChkRisk, PermittedRisk can be customized.
  For column = 1 to Number of columns{
   if(CheckedRisk < PermittedRisk){
    goto S3
   }
   if(classifiedID == DID){
    returnedId = DeidentificationID(id, setRule)
//Cryptography, Replacement, Elimination
    CheckedRisk = CheckRisk(returnedId)
//Risk evaluation (by each column)
   } Next column
  }
  goto S3
}
```

```
//S3. Processing of quasi-identifier
HandleQID(qid, level){
  setLevel = basicLevel
  CheckedRisk = CheckRisk(value1)
  PermittedRisk = SetRisk(value2)
//The Level, ChkRisk, PermittedRisk can be customized.
  For column = 1 to Number of columns{
   if(CheckedRisk < PermittedRisk){
    goto S4
   }
   if(classifiedID == QID){
    returnedQid = DeidentificationQID(qid, setLevel)
//Suppression, Generalization, Perturbation,
    CheckedRisk = CheckRisk(returnedQid)
//Risk evaluation (by each column)
   } Next column
  }
  goto S4
}
```

Table 4 Processing of unstructured data

```
//U2. File processing
HandleFile(inputfile, Rule){
  filetype = CheckType(inputfile)
  if (filetype == C){//Changeable File
   duplicatedFile = CopyFile(inputfile)
   OriginalFileEncryption(inputfile, key, currentTime)
   result1 = DeidentificationFile(duplicatedFile, rule)
   //Cryptography, Replacement, Elimination
  } else {//Unchangeable File
    UnchangeableFileEncryption(inputfile, key, currentTime)
  }
  goto U3
}
```

Therefore, techniques such as suppression, generalization, perturbation, swapping, and sub-sampling are used.

Processing of unstructured data

As shown in Table 4, this step involves the processing of unstructured data. The data is processed by de-identification or encryption, depending on whether it is changeable. If de-identification is required, then a copy is processed instead of the original. At this time, the de-identification techniques involve cryptography, replacement, and elimination.

Adequacy test

As shown in Table 5, the adequacy test evaluates whether the de-identification procedures and methods, and the re-identification risk of the data provided, are appropriate. An adequacy test is performed for both the structured and unstructured data from a single clinical study. Priority is given to the classification and adequacy of the results of the de-identification. Finally, after a review by the honesty broker and adequacy test

Table 5 Adequacy test of structured data and unstructured data

```
//S4, U3. Adequacy test
Adequacy(S3|U3){
  CheckedClassification = CheckClassification(){
   if(CheckedClassification == INVALID){
    goto S1 | U1//Re-classification
   }
  }
  CheckedDeidentification = CheckDeidentification(){
   if(CheckedDeidentification == INVALID){
    if(classifiedID == DID){
     goto S2
    } else if(classifiedID == QID) {
     goto S3
    } else {
     goto U2
    }
   }
  }
  Call Dataprovider()
}
```

committee along with the profile of the research data, data is provided according to the form and purpose of the request.

Comparison and analysis

When performing de-identification, its purposes and targets play critical roles. In this chapter, we compare the proposed Korean Guideline for De-identification of Personal Information (published by various South Korean government agencies), the US National Institute of Standards and Technology NISTIR 8053 (2015) and the proposed framework.

In the Korean guidelines, techniques including pseudonymization, aggregation, data reduction, data suppression, and data masking are used once the DID and QID are selected and classified in the initial dataset. Then, the de-identification adequacy of the data is evaluated. The safe harbor defined in HIPPA was selected as the standard for classifying the DID. For the QID, de-identification is performed by considering the individual characteristics of physical, credit, career, electronic, and family. On this basis, only data that has been determined as appropriate will be used [8, 22].

Although there are no separate criteria defining the DID and QID in NISTIR 8053, the standard indicates that the selected DID in a dataset should be masked or transformed into other content that is difficult to directly identify. The standard goes on to define what information an attacker can access and determines if the QID can be re-identified. To minimize the disclosure risk, the standard then defines which fields are adequate for the purposes of usage and disclosure, and determines the degree of maximum de-identification. After de-identification within the maximum level of the definition, an adequacy test is performed. At that point, the data is provided if it is below the acceptable risk level [6, 20].

In the proposed framework, the initial data and files are separately processed as unstructured and structured data. In the case of unstructured data, encryption or non-identification processing is performed according to the change and processing

Table 6 Comparison of de-identification framework

Item	Proposed Framework	NISTIR 8053 [8]	Korean Guideline [22]
Handling multiple data types in one framework.	Yes	No	No
Does it support the data processing for clinical research?	Yes	A little	No
De-identification Adequacy test	Yes	Yes	Yes
Distinguish between data consumers?	Yes	No	No

possibilities. For structured data, dataset optimization is first performed to prevent errors, such as range errors, blank spaces, etc. After optimization, the classification of DID, QID, and SA proceeds. Then, rule setting, de-identification, and risk assessment is performed until the number of columns is achieved. Cryptography, replacement, and elimination techniques are then used to ensure the non-identification of DID, and suppression, generalization, and perturbation techniques are used for the QID. The next steps proceed differently, depending on whether the provided object passes the adequacy test by being below the level of acceptable risk [6, 20].

Finally, Table 6 compares the proposed framework with the other two schemes described above. In particular, it is focused on differences in the de-identification process or function. And the proposed framework is designed to support the use of various clinical research data in general hospitals and meets all the items in Table 6.

Conclusion

Information that cannot be specified can be identified if the amount and type are increased. In particular, data in the medical field is very sensitive. Since data is frequently accessed by multiple users, substantial caution is required when handling data. In addition, systematic security protocols must be followed to prevent data from being used outside its intended purpose, even in clinical research. In this regard, a system for de-identification should be provided in clinical research that requires periodic review and modification, rather than one-time actions.

In this paper, we provided an overview of the latest United States and Korean de-identification guidelines, de-identification steps, and phased security threats. On this basis, we proposed a de-identification framework for clinical research. The proposed framework processes clinical data in the order of refinement, classification, de-identification, and adequacy tests. The proposed framework provides a balanced approach to efficient utilization and effective security to ensure safe clinical research.

Note that we did not undertake a detailed analysis of the optimization and classification of data to improve the de-identification results. Accordingly, in future research, we will conduct in-depth research on the optimization and classification of both DID and QID prior to de-identification.

Authors' contributions

JDL carried out design of the proposed framework. JMJ and PLP analyzed related research. SHC performed comparative analysis. HSC mainly managed and supervised this paper. All authors read and approved the final manuscript.

Author details
[1] National Cancer Center, 323 Ilsanro, Ilsandong-gu, Goyang-si, Gyeonggi-do 10408, Republic of Korea. [2] SeoulTech, Seoul, Republic of Korea.

Acknowledgements
This study was supported by a grant from the National R&D Program for Cancer Control, Ministry of Health and Welfare, Republic of Korea (1631180).

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 16 April 2018 Accepted: 12 June 2018
Published online: 28 June 2018

References

1. Karystianis G, Sheppard T, Dixon WG, Nenadic G (2016) Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database. BMC Med Inform Decis Making 16(1):1–18
2. Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P (2013) Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiother Oncol 108(1):174–179
3. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ (2014) How the provenance of electronic health recorddata matters for research: a case example using system mapping. EGEMS (Wash DC) 2(1):1058. https://doi.org/10.13063/2327-9214.1058
4. Fernández-Alemán JL, Señor IC, Lozoya PÁO, Toval A (2013) Security and privacy in electronic health records: a systematic literature review. J Biomed Inform 46(3):541–562
5. Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: 2008 IEEE symposium on security and privacy (sp 2008), Oakland, CA, pp 111–125
6. Abdelhak Mansoul, Baghdad Atmani (2017) Combining multi-criteria analysis with CBR for medical decision support. J Inf Process Syst 13(6):1496–1515
7. Kar Jayaprakash, Mishra Manoj Ranjan (2016) Mitigating threats and security metrics in Cloud Computing. J Inf Process Syst 12(2):226–233
8. Garfinkel SL (2015) De-identification of personal information (NISTIR 8053), NIST, http://dx.doi.org/10.6028/NIST.IR.8053. Accessed 10 Apr 2018
9. George J, Kumar V, Kumar S (2015) Data warehouse design considerations for a healthcare business intelligence system. In: Proceedings of the World Congress on Engineering, vol 1, July 2015
10. West VL, Borland D, Hammond WE (2014) Innovative information visualization of electronic health record data: a systematic review. J Am Med Inform Assoc 22(2):330–339
11. Weiskopf NG, Weng C (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 20(1):144–151
12. Guido Z, Daniel K, Anthony N, Anton B (2014) De-identification of health records using Anonym: effectiveness and robustness across datasets. Artif Intell Med 61(3):145–151
13. Shin SY, Lyu Y, Shin Y, Choi HJ, Park J, Kim WS, Lee JH (2013) Lessons learned from development of de-identification system for biomedical research in a Korean Tertiary Hospital. Healthc Inform Res 19(2):102–109
14. Shin SY, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, Lee MS, Choi CM, Kim WS, Lee JH (2015) A de-identification method for bilingual clinical texts of various note types. J Korean Med Sci 30(1):7–15
15. Muqun L, David C, John A, Lynette H, Bradley AM (2014) De-identification of clinical narratives through writing complexity measures. Int J Med Inform 83(10):750–767
16. Garfinkel SL (2016) NIST SP 800-188 De-Identifying Government Datasets (2nd Draft). NIST, Gaithersburg
17. ISO 25237:2017 Health informatics Pseudonymization (2017), ISO/TC 215 Health informatics
18. Graham C (2012) Anonymization: managing data protection risk code of practice. Information Commissioner's office, Wilmslow
19. Opinion 05/2014 on Anonymization Techniques (2014), Article 29 Working Party, European Union (EU)
20. El Emam K, Jonker E, Sams S, Neri E, Neisa A, Gao T, Chowdhury S (2007) Pan-Canadian de-identification guidelines for personal health information. Children's Hospital of Eastern Ontario Research Institute, Ottawa
21. Office of the Australian Information Commissioner (2014) Privacy business resource 4: de-identification of data and information. Australian Government, Australia. https://www.oaic.gov.au/resources/privacy-law/privacy-archive/privacy-resources-archive/privacy-business-resource-4-de-identification-of-data-and-information.pdf
22. Korean government interdepartmental Joint (2016) Guidelines for De-identification of Personal Data. Korean Government, Korea. https://www.privacy.go.kr/cmm/fms/FileDown.do?atchFileId=FILE_000000000830764&fileSn=0
23. Prasser F, Kohlmayer F, Kuhn KA (2016) Efficient and effective pruning strategies for health data de-identification. BMC Med Inform Decis Making. https://doi.org/10.1186/s12911-016-0287-2
24. Mark E, Elaine M, Kieron O, Caroline T (2016) The anonymisation decision-making framework. UKAN (UK Anonymisation Network), Manchester
25. Lee YR, Chung YC, Kim JS, Park HK (2016) Personal health information de-identified performing methods in Big Data Environments. Int J Softw Eng Appl 10(8):127–138

26. Lee YJ, Lee KH (2017) Re-identification of medical records by optimum quasi-identifiers. In: 2017 19th international conference on advanced communication technology (ICACT), 19–22 Feb 2017
27. Merener MM (2012) Theoretical results on de-anonymization via linkage attacks. Trans Data Priv 5(2):377–402
28. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G (2015) Combining knowledge and data-driven methods for de-identification of clinical narratives. J Biomed Inform. https://doi.org/10.1016/j.jbi.2015.06.029
29. Jiang Zhipeng, Zhao Chao, He Bin, Guan Yi, Jiang Jingchi (2017) De-identification of medical records using conditional random fields and long short-term memory networks. J Biomed Inform 75:s43–s53
30. Menger V, Scheepers F, van Wijk LM, Spruit M (2018) DEDUCE: a pattern matching method for automatic de-identification of Dutch medical text. Telemat Inform 35(4):727–736
31. Phuong ND, Chau VTN (2016) Automatic de-identification of medical records with a multilevel hybrid semi-supervised learning approach. In: 2016 IEEE RIVF international conference on computing & communication technologies, research, innovation, and vision for the future (RIVF), Hanoi, pp 43–48
32. Acharya S, Patel A (2017) Towards the design of a comprehensive data de-identification solution. In: 2017 IEEE international conference on bioinformatics and biomedicine (BIBM), Kansas City, MO, pp 1–8
33. Prasser F, Eicher J, Bild R, Spengler H, Kuhn KA (2017) A tool for optimizing de-identified health data for use in statistical classification. In: 2017 IEEE 30th international symposium on computer-based medical systems (CBMS), Thessaloniki, pp 169–174

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com