J Korean Med Sci. 2018 Jan 29;33(5):e41
https://doi.org/10.3346/jkms.2018.33.e41
eISSN 1598-6357·pISSN 1011-8934

JKMS

Opinion

Medical Informatics

Issues and Solutions of Healthcare Data De-identification: the Case of South Korea

Soo-Yong Shin

Department of Computer Science and Engineering, Kyung Hee University, Yongin, Korea

Received: Aug 3, 2017
Accepted: Nov 7, 2017

Address for Correspondence:
Soo-Yong Shin, PhD
Department of Computer Science and Engineering, Kyung Hee University, 1732 Deogyeong-daero, Giheung-gu, Yongin 17104, Republic of Korea.
E-mail: sooyong.shin@khu.ac.kr

ORCID iDs
Soo-Yong Shin
https://orcid.org/0000-0002-2410-6120

Disclosure
The author has no potential conflicts of interest to disclose.

Funding
This work was supported by a grant from Kyung Hee University in 2016 (KHU-20161377).

Artificial intelligence (AI) has been highlighted as a mechanism to realize precision medicine because it contributes to analyzing healthcare big data.[1,2] Especially, among diverse AI methods, machine learning (ML) methods including deep learning algorithms are widely applied to analyze healthcare data.[2] ML requires a vast amount of data due to its nature. This means that collecting as much relevant data as possible is a critical task. The Precision Medicine Initiative[3] or Observational Health Data Sciences and Informatics (OHDSI)[4] might be representative cases for collecting healthcare big data. However, since healthcare data contain the most sensitive personal information, the concerns on protecting patients' privacy are increasing.

In Korea, the Personal Information Protection Act was passed to protect privacy. Additionally, the Bioethics and Safety Act was passed to protect the unauthorized use of patients' health information. According to the two regulations, researchers should obtain informed consent from each research participant. However, it is almost impossible to obtain written consent if researchers perform research that requires a large number of participants. An alternative method is de-identification, which is an effective method to protect privacy and comply with regulations.[5] Diverse de-identification methods have been developed for clinical texts and images,[6,7] and the Korean government published the guideline for de-identification of personal data in 2016.[8] This guideline was developed to provide clear methods on de-identification of personal data and scopes on utilizing de-identified data.

The guideline proposes four steps for de-identifying personal data: 1) The preliminary review step verifies whether specific data are personally identifiable data or not, 2) The de-identification step makes individuals unidentifiable using the necessary methods, 3) The adequacy assessment step assesses whether de-identified data can be re-identified, and 4) The follow-up management step monitors the possibility of re-identification. This guideline tries to cover all possible personal data including financial, commerce, communication, and healthcare data. Unfortunately, researchers and companies in the biomedical field face a big challenge to obey the guideline since the characteristics of biomedical data are different from other industries. Financial or commerce data consist of repetitive patterns of transactions with relatively small number of features. However, biomedical data have too diverse features, for example, extensive laboratory test results and treatments. In addition, biomedical data include structured code data as well as unstructured data such as text, images, and videos. This implies that the published guideline is not suitable for the biomedical area. Here, we criticize the current regulation on de-identification in the context of Korea and raise several issues regarding the de-identification of biomedical data.

First, the published guideline for de-identification demands "*k*-anonymity" for the mandatory privacy protection method.[8] *K*-anonymity is a well-established method to protect privacy[9] and easily provides the quantitative measure of privacy protection. Noticeably, the US Family Educational Rights and Privacy Act adopts *k*-anonymity.[10] *K*-anonymity requires there should be at least the same *k* items in the dataset. However, *k*-anonymity is difficult to achieve in healthcare datasets since raw data must be distorted. For example, if we decide to de-identify 1,000 patients' clinical data that consist of 5 different laboratory test results by keeping 5-anonymity, every combination of 5 different patients' clinical data should be the same. Therefore, all laboratory results should be generalized by replacing individual attributes with broader categories, as in **Tables 1** and **2**. As shown in **Table 2**, the modified data lost all their detailed information which is essential for analysis. Even more, all clinical images such as computed tomography (CT) or magnetic resonance imaging (MRI) images cannot be the same; therefore, images data cannot be of *k*-anonymity. Alternative rules need to be suggested in a future revised, or new, guideline.

Second, there is a controversial debate over the definition of personal information in the regulations. In the Korean Personal Information Protection Act, personal information is defined as "information that pertains to a living person, including the full name, resident registration number, images, etc., by which the individual in question can be identified, (including information by which the individual in question cannot be identified but can be identified through simple combination with other information)." However, laypersons think all private information, including height and weight, should be protected. According to the aforementioned definition, height and weight cannot identify the specific individual in most cases; therefore, they are not personal information. Most of the personal information is not personally identifiable information. To resolve this debate, the term should be clarified as "personally identifiable information" instead of "personal information."

Third, the definition of re-identification is also unclear. Usually, re-identification implies that the de-identified data are matched to the specific individual. However, there is no

Table 1. Example of *k*-anonymity: original dataset

Patients	WBC, × 10^3/μL	Hb, g/dL	AST (SGOT), IU/L	ALT (SGPT), IU/L	Cholesterol, mg/dL
Patient 1	5.6	17.0	39	64	199
Patient 2	5.4	17.5	44	67	173
Patient 3	4.2	16.4	28	58	179
Patient 4	4.7	16.1	36	64	180
Patient 5	6.1	18.4	101	151	231
Patient 6	7.5	15.6	33	42	195
Patient 7	8.2	17.1	35	54	175

WBC = white blood cell, Hb = hemoglobin, AST = aspartate aminotransferase, SGOT = serum glutamic oxaloacetic transaminase, ALT = alanine aminotransferase, SGPT = serum glutamic pyruvic transaminase.

Table 2. Example of *k*-anonymity: modified dataset for 5-anonymity

Patients	WBC, × 10^3/μL	Hb, g/dL	AST (SGOT), IU/L	ALT (SGPT), IU/L	Cholesterol, mg/dL
Patient 1	Normal	Normal	Normal	> 40	Normal
Patient 2	Normal	> 17.0	> 40	> 40	Normal
Patient 3	Normal	Normal	Normal	> 40	Normal
Patient 4	Normal	Normal	Normal	> 40	Normal
Patient 5	Normal	> 17.0	> 40	> 40	> 200
Patient 6	Normal	Normal	Normal	> 40	Normal
Patient 7	Normal	Normal	Normal	> 40	Normal

Patients 1, 3, 4, 6, and 7 are the same to obey 5-anonymity. As a result, all data were distorted.
WBC = white blood cell, Hb = hemoglobin, AST = aspartate aminotransferase, SGOT = serum glutamic oxaloacetic transaminase, ALT = alanine aminotransferase, SGPT = serum glutamic pyruvic transaminase.

clear definition in the Korean regulations. Some argue that re-identification is finding the individual's identity, for example, resident registration number, name, and phone numbers. While others argue that re-identification includes finding the same entity among different databases even though the identity is not confirmed. Technically, the second opinion should not be considered as re-identification for healthcare research. For big data research, the researcher should be able to combine the patient data from electronic medical records of hospital A with those of hospital B. In this case, there should exist the key (the information to find the same patient) to link the different databases. If linking two different databases is re-identification, big data research cannot be performed without written consents. In this regard, the regulations should be revised to include the precise definition of re-identification, for example, pointing out the exact identity of an individual by discovering the individuals' name or phone numbers. Additionally, the concept of "motivated intruder" in the Information Commissioner's Office (ICO) of the United Kingdom should be introduced.[11] A motivated intruder is the layperson who can access resources such as the internet, libraries, and all public documents, and is not assumed to have any specialist knowledge such as computer hacking skills or domain knowledge. This concept is important in clinical research since physicians can easily recognize their patients despite de-identification.

Fourth, there is no list of personal health identifiers. Unfortunately, the definition of personal information includes the following sentence "(including information by which the individual in question cannot be identified but can be identified through simple combination with other information)." This definition implies all candidate identifiers need to be protected. Technically, personal identifiers can be categorized into direct identifier and indirect identifier. The direct identifier can uniquely identify the individual, for example, name, address, etc. Indirect identifier or quasi-identifier is the above-quoted information. It cannot immediately identify individuals, but has a chance to identify individuals when linked with other identifiers. For example, we may guess the individual by combining diverse indirect identifiers including one's occupation, race, place of birth, and education information. The problem is all indirect identifiers cannot identify the individual. To distinguish the individual, the combination of indirect identifiers heavily depends on background knowledge for the target individual. For example, the famous re-identification of Governor William Weld's medical information in 1997 was possible since he was a public figure with a highly publicized hospitalization.[12] Therefore, a full list of personal health identifiers is indispensable for the practical de-identification. Current Korean de-identification guideline suggests each organization should review and decide the identifiers at their own risk. However, the Health Insurance Portability and Accountability Act (HIPAA) in the US defines the 18 protected health information (PHI).[5] This approach can reduce the burdens of the de-identification processes.

Last, due to the advance of technologies, more types of possible direct identifiers are introduced and used, but have not been considered yet, i.e., the artificially reconstructed facial images using skull CT images[13] and genetic/genomic data.[14] There is no clear answer whether we should treat the artificially reconstructed facial images as normal full-face images or not. Furthermore, there is a debate over genetic/genomic data if they are personally identifiable information or not. The re-identification of the personal genome project data was possible,[15] but the genetic genealogy database was used for identification. This implies DNA itself cannot identify an individual's identity. To identify the individual, there should be a reference database such as a criminal DNA database. We need to discuss these emerging candidate identifiers related to the above issues.

De-identification is indispensable for big data research and open data, which can strengthen academic research and commercial solution development. To accelerate the development of healthcare AI, the aforementioned equivocal issues should be clearly resolved by revising guideline and regulations continuously. And then, the research on the de-identification method to obey regulations should be followed. Unfortunately, healthcare data de-identification research is not popular in Korea since most IT engineers or researchers cannot access the clinical data in hospitals. However, privacy-preserving data mining or data de-identification methods have been widely applied to other areas. Promising methods for healthcare data de-identification include differential privacy[16] and homomorphic encryption.[17] To advance the research, as well as to comply with the regulation, both the on-going efforts to revise the governmental regulations and to develop methods for privacy protection should be made together by multidisciplinary collaboration with jurists, bioethicists, doctors, basic science researchers, and IT engineers.

REFERENCES

1. Chen JH, Asch SM. Machine learning and prediction in medicine - beyond the peak of inflated expectations. *N Engl J Med* 2017;376(26):2507-9.
PUBMED | CROSSREF
2. Deo RC. Machine learning in medicine. *Circulation* 2015;132(20):1920-30.
PUBMED | CROSSREF
3. Ashley EA. The precision medicine initiative: a new national effort. *JAMA* 2015;313(21):2119-20.
PUBMED | CROSSREF
4. Park RW. Sharing clinical big data while protecting confidentiality and security: observational health data sciences and informatics. *Healthc Inform Res* 2017;23(1):1-3.
PUBMED | CROSSREF
5. McGraw D. Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data. *J Am Med Inform Assoc* 2013;20(1):29-34.
PUBMED | CROSSREF
6. Shin SY, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A De-identification method for bilingual clinical texts of various note types. *J Korean Med Sci* 2015;30(1):7-15.
PUBMED | CROSSREF
7. Monteiro E, Costa C, Oliveira JL. A De-identification pipeline for ultrasound medical images in DICOM format. *J Med Syst* 2017;41(5):89.
PUBMED | CROSSREF
8. Guidelines for De-identification of personal data. https://www.privacy.go.kr/cmm/fms/FileDown.do?atchFileId=FILE_000000000827161&fileSn=0. Updated 2016. Accessed August 1, 2017.
9. El Emam K, Dankar FK. Protecting privacy using k-anonymity. *J Am Med Inform Assoc* 2008;15(5):627-37.
PUBMED | CROSSREF
10. Family Educational Rights and Privacy Act (FERPA). https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html. Updated 2015. Accessed October 13, 2017.
11. Anonymisation: managing data protection risk code of practice. https://ico.org.uk/media/1061/anonymisation-code.pdf. Updated 2012. Accessed July 21, 2017.
12. The 'Re-identification' of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. http://www.ssrn.com/abstract=2076397. Updated 2015. Accessed July 24, 2017.
13. Wilkinson C. Facial reconstruction--anatomical art or artistic anatomy? *J Anat* 2010;216(2):235-50.
PUBMED | CROSSREF
14. Erlich Y, Williams JB, Glazer D, Yocum K, Farahany N, Olson M, et al. Redefining genomic privacy: trust and empowerment. *PLoS Biol* 2014;12(11):e1001983.
PUBMED | CROSSREF
15. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science* 2013;339(6117):321-4.
PUBMED | CROSSREF

16. Dwork C. Differential privacy: a survey of results. In: Agrawal M, Du DZ, Duan Z, Li A, editors. *Theory and Applications of Models of Computation*. Berlin: Springer-Verlag Berlin Heidelberg; 2008, 1-19.

17. Fontaine C, Galand F. A survey of homomorphic encryption for nonspecialists. *EURASIP J Inf Secur* 2007;2007:013801.