J Korean Med Sci. 2021 Nov 15;36(44):e299
https://doi.org/10.3346/jkms.2021.36.e299
eISSN 1598-6357·pISSN 1011-8934

JKMS

Review Article
Medical Informatics



# Data Pseudonymization in a Range That Does Not Affect Data Quality: Correlation with the Degree of Participation of Clinicians

**Soo-Yong Shin** [1,2] **and Hun-Sung Kim** [3,4]

[1]Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology (SAIHST), Sungkyunkwan University, Seoul, Korea
[2]Center for Research Resource Standardization, Samsung Medical Center, Seoul, Korea
[3]Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Korea
[4]Division of Endocrinology and Metabolism, Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea



**Received:** May 4, 2021
**Accepted:** Oct 18, 2021

**Address for Correspondence:**
**Hun-Sung Kim, MD, PhD**
Department of Medical Informatics, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Korea.
E-mail: 01cadiz@hanmail.net



**ORCID iDs**
Soo-Yong Shin
https://orcid.org/0000-0002-2410-6120
Hun-Sung Kim
https://orcid.org/0000-0002-7002-7300

**Funding**
This study was supported by the Daewoong Pharmaceutical Company in 2020.

**Disclosure**
The authors have no potential conflicts of interest to disclose.

**Author Contributions**
Conceptualization: Kim HS. Data curation: Shin SY, Kim HS. Formal analysis: Kim HS. Writing - original draft: Kim HS. Writing - review & editing: Shin SY, Kim HS.

## ABSTRACT

Personal medical information is an essential resource for research; however, there are laws that regulate its use, and it typically has to be pseudonymized or anonymized. When data are anonymized, the quantity and quality of extractable information decrease significantly. From the perspective of a clinical researcher, a method of achieving pseudonymized data without degrading data quality while also preventing data loss is proposed herein. As the level of pseudonymization varies according to the research purpose, the pseudonymization method applied should be carefully chosen. Therefore, the active participation of clinicians is crucial to transform the data according to the research purpose. This can contribute to data security by simply transforming the data through secondary data processing. Case studies demonstrated that, compared with the initial baseline data, there was a clinically significant difference in the number of datapoints added with the participation of a clinician (from 267,979 to 280,127 points, $P < 0.001$). Thus, depending on the degree of clinician participation, data anonymization may not affect data quality and quantity, and proper data quality management along with data security are emphasized. Although the pseudonymization level and clinical use of data have a trade-off relationship, it is possible to create pseudonymized data while maintaining the data quality required for a given research purpose. Therefore, rather than relying solely on security guidelines, the active participation of clinicians is important.

**Keywords:** Cardiovascular Diseases; Data Anonymization; Data Quality; De-identification; Electronic Health Records

## INTRODUCTION

There is an increasing interest in real-world data (RWD), known as medical big data.[1-3] Clinical studies that use electronic medical record (EMR) data including a patient's history of symptoms, diagnosis, and treatment, are becoming more common.[4-7] However, the security measures required for EMR data are becoming more stringent to protect patients' privacy.[6,8] As data security increases, the utilization of data inevitably decreases. It is almost impossible

to obtain written consent when performing big data research; instead, we have to de-identify "personally identifiable information."[9,10] In terms of de-identification, a major controversy exists regarding the use of personally identifiable information for an analysis.

In Korea, revisions have been made to "three data-related bills," which collectively refer to the three laws of the Personal Information Protection Act (PIPA), Information and Communications Network Act, and Protection of Credit Information Act.[11,12] Among these data regulation bills, the amended PIPA is the most interesting for the medical field.[13] It introduces the pseudonymization of information for data use, reinforces the responsibility of honest brokers,[14] and clarifies the ambiguous "personally identifiable information" criteria. An important aspect of this act is that if data are acquired for legal use (statistics, scientific research, and record for the benefit of society), it is permissible to use them "without the consent of the data subject" if the data are pseudonymized.[11,12]

The Korean Personal Information Protection Commission (PIPC) published guidelines for pseudonymization.[15] Subsequently, PIPC and the Korean Ministry of Health and Welfare published the "Guideline for Utilization of Healthcare Data"[16] to clarify the ways in which clinical data can be used for research. However, some non-governmental organizations are still concerned about personally identifiable information being leaked[17]; therefore, researchers should be aware of these guidelines and conduct research carefully, because various unexpected problems may be encountered.

However, in practice, clinical researchers are not completely satisfied and are not able to easily comply with the government guidelines regarding the use of EMR data. In fact, the guidelines are primarily focused on processing and pseudonymizing existing medical data to suit personally identifiable information protection purposes. However, in terms of "operational definition" and "data quality management (DQM)" for actual clinical research, there are many cases in which large amounts of data must be randomly created, deleted, or are subject to secondary transformation.[18] During this process, pseudonymized data are significantly modified compared with original medical records.[8,19] The clinical data created by modifying the original data to a certain extent can protect patients' privacy. This means that the role of the researcher is important.

Furthermore, cooperation between those performing pseudonymization (honest brokers) and researchers is highly essential for accurate clinical research. An honest broker should understand the purpose of a researcher given that the quality of pseudonymized data could be insufficient to fulfill the aim of targeted research. Additionally, because the degree of pseudonymization varies according to the purpose of a clinical study, researchers should carefully choose the pseudonymization methods.

Because this process is not easy, active participation of a clinician is necessary in addition to the basic role of an honest broker. Therefore, in this study, we describe a specific DQM for the protection of personally identifiable information. However, this DQM is not covered under some of the guidelines based on what should be considered when used in actual clinical research.

## PERSONALLY IDENTIFIABLE INFORMATION, PSEUDONYMIZED INFORMATION, AND ANONYMIZED INFORMATION

First, it is important to understand the concepts that define the levels of privacy protection. Based on the purpose of the investigation, researchers should clearly define in advance whether personal, pseudonymized, or anonymized information is needed. Personally identifiable information refers to information[20,21] that could help identify a specific individual. Pseudonymization is the process of processing data by deleting or replacing personally identifiable information, partially or completely, such that the individual is unrecognizable without additional information. Anonymized information refers to information that can no longer identify an individual, with or without other information combined, when time, cost, and technology are considered.

While anonymized information is the strongest in terms of data security, followed by pseudonymized and personally identifiable information, personally identifiable information is the most valuable, followed by pseudonymized and anonymized information. Although PIPA recommends using anonymized information first whenever possible, pseudonymized information is used when the purpose of the study cannot be achieved with anonymized information, and personally identifiable information is used when the purpose of the study cannot be achieved with pseudonymized information.[22-24]

It is crucial to use personally identifiable information for rare diseases because the researchers need to identify the individuals[23]; therefore, the patient's consent is required in such cases. However, anonymized information is usually sufficient for studies based on the frequency of drug use according to age.[25] On the other hand, pseudonymized information should be accurate or comprise enough personally identifiable data to determine the blood sugar control rate according to the age, sex, or prevalence of the individual. Therefore, it is necessary to consider the purpose and direction of the research before choosing the EMR data. For example, if a researcher plans and conducts a study using anonymous information during data collection, and additional information is required later, the study will have to be conducted again from scratch. It is also essential to distinguish between a situation in which it is necessary to specifically identify an individual and a situation in which a 1:1 matching (singling out) occurs. Strictly speaking, identifying an individual is “identification” and 1:1 matching is “individualization.” If the researcher attaches additional information to the individualized object and finds out who it is, it is referred to as identification. However, from a research point of view, individualization is mainly used when connecting different databases. It is not possible to attach additional information in EMR because individualization alone does not allow for attaching information. Some level of identification (e.g., the researcher knowing the patient number) is required to attach additional information. In addition, a mechanism capable of matching the same person in time-series data prepared differently with time differences may also be required. Most importantly, the quantity and quality of information that can be extracted decrease significantly as the data approaches anonymization.[8] However, considering that PIPA has strict punishment criteria, it is safer to proceed conservatively.[11,12]

Recently, anonymization was strictly required to publish highly ranked papers[26] with all identifiable data deleted and the sex, coded. Additionally, the patient's age was specified as


https://doi.org/10.3346/jkms.2021.36.e299

"in their 20s/30s/40s" to eliminate the possibility of estimating a particular patient based on the birth year.

## DIFFERENT PSEUDONYMIZATION TECHNIQUES FOR THE SAME CLINICAL DATA

Identifiable information such as the patient's name, phone number, social security number, e-mail ID, passport number, driver's license number, alien registration number, and contact information must be suppressed on pseudonymization. For epidemiological studies, the address could be included, but the specific number must be pseudonymized regardless of the purpose of the analysis (partial suppression or masking). While it is advisable to delete the date of birth, in cases where the year of birth is reflected, the month and day are deleted. Typically, the age is preferred over the date of birth in clinical studies because the age at which an event occurred is more important than the date of birth. As mentioned above, some renowned journals recommend expressing age in decades (i.e., the 20s, 30s, etc.).[26]

The guidelines stipulate those values measured by physical conditions, such as height and weight, do not require separate pseudonymization as there is no guarantee that repeated measurements would result in the same number.[11,12] However, this is the only method to protect a patient's personally identifiable information, and secondary processing must be performed separately for research purposes.

Generally, when a patient visits the hospital more than once, the different values of the measured height do not match (**Fig. 1**). While this may not be problematic for pseudonymization-based studies, it is difficult for researchers because it is their concern whether to use the average height value or the actual height. Using the average height alters some or all personally identifiable information is changed, thus reducing the amount of missing data and ensuring pseudonymization. If the average height value is not used, the data are missing if the height was not checked during the visit. Therefore, for the actual height, entering the average value is beneficial when there are many missing data. However, care should be taken in the case of elderly patients (as in osteoporosis studies) because height may decrease over time.[27] Therefore, it is not always possible to recommend using only the average value. That is, to process statistics, the opinions of clinical experts associated with

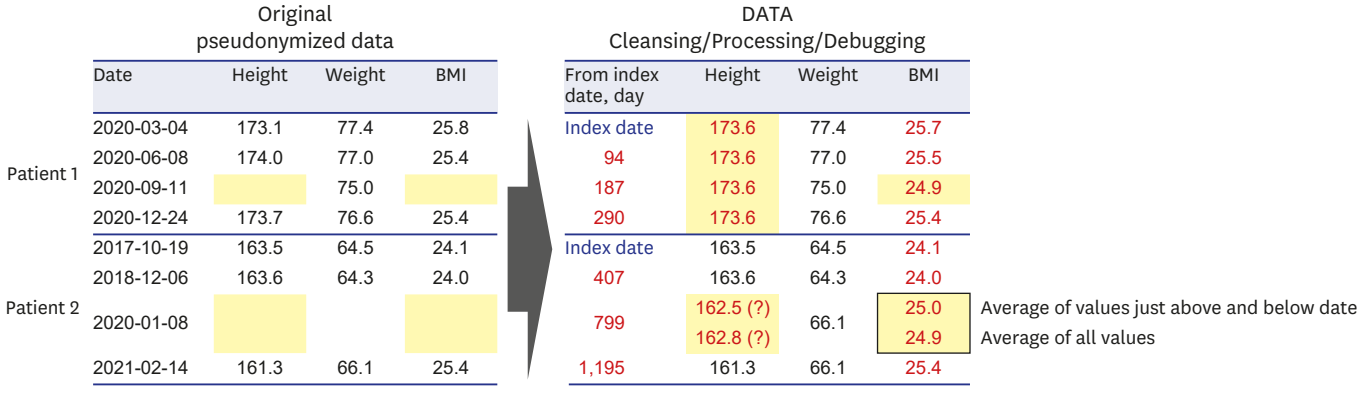



**Fig. 1.** Example of changing the height of the patient. Based on the BMI value of 25.0 kg/m$^2$, which is the criterion for obesity, it is important for researchers to consider that the treatment method may be completely different because of this simple conversion of values.
BMI = body mass index.


https://doi.org/10.3346/jkms.2021.36.e299

the research purpose are required. For example, if height was measured four times and one value is different from the other three, we can delete it and replace it with the average of the other three values. However, in actual clinical practice, many cases demonstrated two values differing from the other two values. Although the judgment of the clinician is essential, it is necessary to consider it deeply because it might be viable to simply delete ambiguous values.

As another example, systolic blood pressure and diastolic blood pressure require DQM. In actual EMR data, many of these values are recorded in reversed order. While medical staff may recognize a mistake in the values, a researcher cannot change the data arbitrarily. When values are out of range, a data table specification is created through operational definition prior to the study to change or delete that information.[5] Additionally, modifying these data also prevents the re-identification of personally identifiable information.

The most significant problem is that it modifies data quality, and we need to apply different techniques for the same clinical data based on the context.[28] This has already been carefully emphasized in various studies.[28,29] We cannot overlook the fact that the clinical quality of data suitable for a research purpose reduces with the increase in the level of pseudonymization and anonymization.[8] If the researcher changes the value according to the purpose of the study, the data are also altered. This limits the ability to identify a particular individual. For example, if a researcher changes a person's height from 180 to 170 cm, it could identify a completely different person. Moreover, clinically sensitive areas may be affected. If the body mass index (BMI) changes owing to a change in height, it may result in a change in the clinical criteria. When BMI changes, the boundaries between normal and obesity also change, and the diagnosis and treatment may change accordingly. Utmost care should be taken when changing a value from the patient's original data.

## SECONDARY DATA PROCESSING BY CLINICIANS, NOT BY THE HONEST BROKER

In some cases, new data are added by the researcher. For example, if one value out of the four values for total cholesterol (TC)/triglyceride (TG)/high-density lipoprotein cholesterol (HDL-C)/low-density lipoprotein cholesterol (LDL-C) is missing, the null data can be filled (official example; LDL-C = TC − TG/5 − HDL-C) (**Fig. 2**).[30] Values outside the range that can be measured by medical devices in hospitals need to be treated according to the purpose of the study. The most typical case is when it is expressed as a string or outlier, not a number, such as "TG < 5 mg/dL" or "TG > 1,995 mg/dL."[18] This requires a deep understanding of RWD. RWD are not intended for clinical research purposes. Therefore, to conduct clinical research on RWD, it is necessary to understand their limitations.[2] In fact, looking at changes in LDL-C levels in retrospective cohort studies is not a good idea. This is because, in RWD, there are many limitations in viewing the average change in laboratory findings owing to differences in drug compliance and visit intervals. Rather, in RWD, it is more advantageous to check the target achievement rate rather than the average change (**Fig. 2**). For example, if we want to check the target rate of less than 200 mg/dL, we can change it to "0" for less than 200, and "1" for more than 200. According to this method of operation, no data omission occurs, and data security is improved. If we want to know the overall distribution of a certain value rather than the target rate, a more subdivided group can be created (less than 200 is "1," 200–300 is "2," 300–400 is "3," 400–500 is "4," and 300 or more is "5," etc.). If we want


https://doi.org/10.3346/jkms.2021.36.e299

**A** Original pseudonymized data

| | TC | TG | HDL-C | LDL-C |
|---|---|---|---|---|
| 1 | 248 | 188 | 35 | |
| 2 | 185 | 225 | 54 | |
| 3 | 175 | 188 | | 94 |
| 4 | 249 | 198 | | 154 |
| 5 | 278 | | 44 | 191 |
| 6 | 150 | | 63 | 69 |
| 7 | | 298 | 44 | 207 |
| 8 | | 288 | 45 | 196 |
| 9 | 322 | > 1,995 | 45 | 177 |
| 10 | 289 | > 1,995 | 33 | 156 |
| 11 | 145 | < 5 | 22 | 122 |
| 12 | 125 | < 5 | 31 | 93 |
| Total slot, n | | | 392,688 | |
| Existing value, n | | | 267,979 | |
| Existing value, % | | | **68.20%** | |

**B** In case of observation a change in value

| | TC | TG | HDL-C | LDL-C |
|---|---|---|---|---|
| 1 | 248 | 188 | 35 | 175 |
| 2 | 185 | 225 | 54 | 86 |
| 3 | 175 | 188 | 43 | 94 |
| 4 | 249 | 198 | 55 | 154 |
| 5 | 278 | 215 | 44 | 191 |
| 6 | 150 | 88 | 63 | 69 |
| 7 | 311 | 298 | 44 | 207 |
| 8 | 299 | 288 | 45 | 196 |
| 9 | 322 | | 45 | 177 |
| 10 | 289 | | 33 | 156 |
| 11 | 145 | | 22 | 122 |
| 12 | 125 | | 31 | 93 |
| Total slot, n | | | | 392,688 |
| Existing value, n | | | + 12,133 | 280,112 |
| Existing value, % | | | + 3.1% | **71.30%** |

**C** When checking the target achievement rate

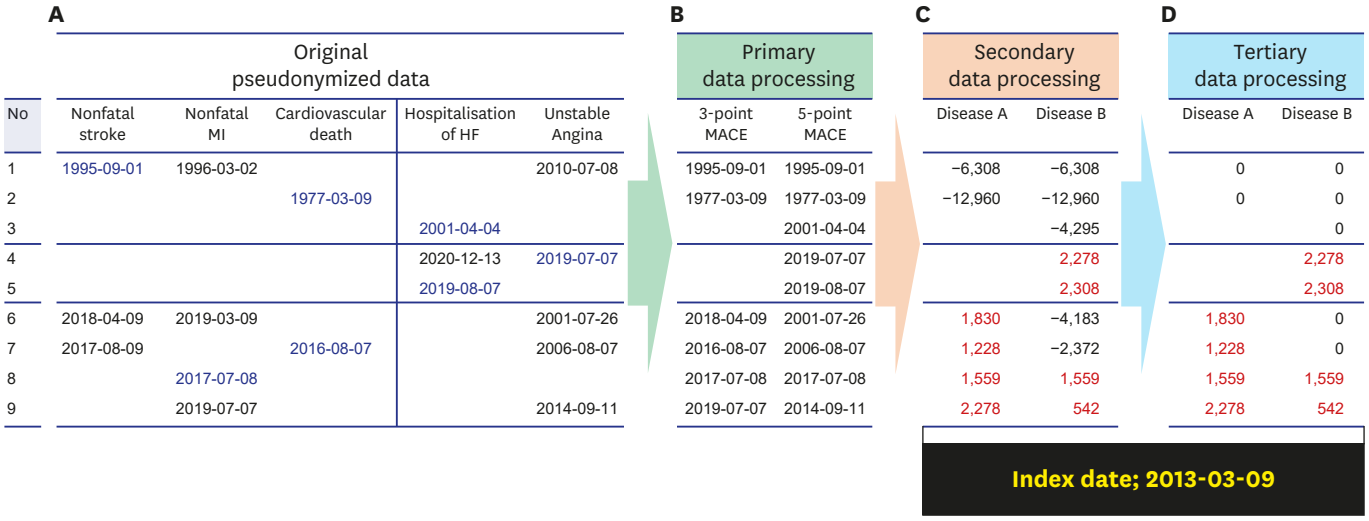
| | TC | TG<br>0; < 200 mg/dL<br>1; 200–500 mg/dL<br>2; ≥ 500 mg/dL | HDL-C<br>0; normal<br>1; abnormal | LDL-C<br>0; < 700 mg/dL<br>1; < 100 mg/dL<br>2; ≥ 100 mg/dL |
|---|---|---|---|---|
| 1 | 248 | 0 | 1 | 2 |
| 2 | 213 | 0 | 0 | 1 |
| 3 | 195 | 0 | 1 | 1 |
| 4 | 249 | 0 | 0 | 2 |
| 5 | 278 | 1 | 1 | 2 |
| 6 | 200 | 0 | 0 | 0 |
| 7 | 311 | 1 | 0 | 2 |
| 8 | 299 | 1 | 0 | 2 |
| 9 | 322 | 2 | 0 | 1 |
| 10 | 289 | 2 | 0 | 1 |
| 11 | 145 | 1 | 0 | 1 |
| 12 | 125 | 1 | 0 | 1 |
| Total slot, n | | | | 392,688 |
| Existing value, n | | | + 15 | 280,127 |
| Existing value, % | | | | **71.30%** |

**Fig. 2.** Real clinical example of transformation of data according to research purpose. There was a clinically significant difference in the number of datapoints added from the number of initial baseline data (from 267,979 to 280,127 points, $P < 0.001$).
HDL-C = high-density lipoprotein cholesterol, LDL-C = low-density lipoprotein cholesterol, TC = total cholesterol, TG = triglyceride.

to compare the average values in various situations, it is recommended to delete the string; however, reliability will decrease.

In addition to subdividing the groups mentioned above, group consolidation is also frequently used (**Supplementary Table 1**). As the date of visit is also identifiable, it is deleted or processed, but it cannot be done mechanically or automatically. For various reasons, even when de-identifying, the clinician must be careful to do it in accordance with the purpose of the study. For example, suppose we want to check the effects of a specific drug for three months and later in RWD. From the RWD, it can be observed that some patients visited after a month instead of three months, some patients visited at four months, and some patients discontinued visits. There is no definite three-month time point, as in the randomized control study. Even if a patient visits the hospital on a regular basis within three months, there is no evidence that the patient has taken the specific drug as prescribed. It is difficult to properly comprehend the effects of the drug after three months. RWD cannot determine the inherent effects of the drugs. In this case, to protect personally identifiable information and minimize data loss, the patient's visit date is segmented to lower the possibility of identification. After the visit is segmented, all patients visit dates are deleted. It is possible to reduce de-identification of data by grouping similar diseases based on patient's "disease information (International Classification of Diseases, 10th Revision [ICD-10] classification)," as well as the patient's visit date (**Fig. 3**). New disease definitions can be created and written by combining ICD-10 codes with various treatment codes. This secondary processing of data not only minimizes data loss but also contributes to data security. However, at the moment, this work is not done by clinicians, but by honest brokers, which can lead to various unpredictable problems.


https://doi.org/10.3346/jkms.2021.36.e299



A

| No | Original pseudonymized data: Nonfatal stroke | Nonfatal MI | Cardiovascular death | Hospitalisation of HF | Unstable Angina |
|---|---|---|---|---|---|
| 1 | 1995-09-01 | 1996-03-02 | | | 2010-07-08 |
| 2 | | | 1977-03-09 | | |
| 3 | | | | 2001-04-04 | |
| 4 | | | | 2020-12-13 | 2019-07-07 |
| 5 | | | | 2019-08-07 | |
| 6 | 2018-04-09 | 2019-03-09 | | | 2001-07-26 |
| 7 | 2017-08-09 | | 2016-08-07 | | 2006-08-07 |
| 8 | | 2017-07-08 | | | |
| 9 | | 2019-07-07 | | | 2014-09-11 |

B

| Primary data processing: 3-point MACE | 5-point MACE |
|---|---|
| 1995-09-01 | 1995-09-01 |
| 1977-03-09 | 1977-03-09 |
| | 2001-04-04 |
| | 2019-07-07 |
| | 2019-08-07 |
| 2018-04-09 | 2001-07-26 |
| 2016-08-07 | 2006-08-07 |
| 2017-07-08 | 2017-07-08 |
| 2019-07-07 | 2014-09-11 |

C

| Secondary data processing: Disease A | Disease B |
|---|---|
| −6,308 | −6,308 |
| −12,960 | −12,960 |
| | −4,295 |
| | 2,278 |
| | 2,308 |
| 1,830 | −4,183 |
| 1,228 | −2,372 |
| 1,559 | 1,559 |
| 2,278 | 542 |

D

| Tertiary data processing: Disease A | Disease B |
|---|---|
| 0 | 0 |
| 0 | 0 |
| | 0 |
| | 2,278 |
| | 2,308 |
| 1,830 | 0 |
| 1,228 | 0 |
| 1,559 | 1,559 |
| 2,278 | 542 |

Index date; 2013-03-09

**Fig. 3.** Example of emphasizing anonymization by consolidating diagnosis names. Anonymization is emphasized, but the quality of the data is not affected. All data has been completely changed or added differently from the original data.
HF = heart failure, MACE = major adverse cardiac events, MI = myocardial infarction.

## FIRST CASE EXAMPLE: CHANGING PATIENT'S DATA TO AN AVERAGE VALUE

### Patient 1

After determining the first visit as an index date, all subsequent visit dates were numerically indicated by the interval from the index date. This allows the data to be anonymized. As a result of changing the height, a new height value that was not included in the original data was created. As a result, the measured BMI values also changed from 25.8/25.4/null/25.4 to 25.7/25.5/24.9/25.4 kg/m$^2$). Numerical values that are different from the original data clearly help to protect personally identifiable information. In addition, the method of changing null data to an average value is the most advantageous for minimizing data loss (**Fig. 1**).

### Patient 2

We cannot always recommend an average value for older people, because there is a gradual decrease in their height. However, when one value is missing, it is worth considering the average of the values immediately before and after it. It is also necessary to properly view and calculate the back-and-forth interval. Before the study, it is also necessary to remove the abnormal values that contradict common sense.[18] It should be noted that owing to methodological differences, one BMI value is 25.0 kg/m$^2$ or more, and the other BMI value is less than 25.0 kg/m$^2$. Based on the BMI value of 25.0 kg/m$^2$,[31] above which indicates obesity, this simple conversion of values can completely change the treatment method. This is a serious consideration for clinicians. Finally, the most important point is that clinical researchers need to scan the data directly (**Fig. 1**).


https://doi.org/10.3346/jkms.2021.36.e299

## SECOND CASE EXAMPLE: REAL CLINICAL EXAMPLE OF TRANSFORMATION OF DATA ACCORDING TO RESEARCH PURPOSE

Let us consider the example of a statin data mart built for research purposes at St. Mary's Hospital in Seoul (**Fig. 2**).[5] We extended the extraction time in the previous statin data mart to obtain a larger number of samples. Because there are four lipid profiles (TC/TG/HDL-C/LDL-C) and 98,172 patients, there should be a total of 392,688 data points (4 × 98,172), but only 68.2% of the data points were present (61,497 in LDL-C + 62,209 in HDL-C + 71,324 in TG + 72,949 in TC = 267,977). As shown in **Fig. 2B**, a total of 12,133 data points could be newly generated using the expression, LDL-C = TC − TG/5 − HDL-C.[30] Certainly, these values did not exist in the original dataset. In the case of a study on the target achievement rate rather than the change in the average value (**Fig. 2C**), 15 new data points were applied without being deleted. An additional 3.1% {(15 + 12,133)/392,688} of data were generated. There was a clinically significant difference in the number of data points added from the number of initial baseline data points (from 267,979 to 280,127 points, $P < 0.001$). This is the result of preventing data loss and ensuring proper anonymization. In conclusion, this is a way to increase the number of data samples while enhancing anonymization.

## THIRD CASE EXAMPLE: EMPHASIZING ANONYMIZATION BY CONSOLIDATING DIAGNOSIS NAMES

This method is often used to determine the occurrence of a specific disease (**Fig. 3**). Several diseases were combined into one category. For example, nonfatal stroke, nonfatal myocardial infarction, and cardiovascular death were integrated into "3-point major adverse cardiac events (MACE)," or hospitalization due to heart failure and unstable angina are added to the "3-point MACE" and integrated into a "5-point MACE."[32,33] If there is an index date for a specific research purpose, the index and disease onset dates can be calculated and expressed. In this case, the data were anonymized. A negative value denotes the patient's history because it is the disease before the index date, and a positive value denotes the interval from the index date to the occurrence. If there is no history before the index date, it can be replaced with "null (or other value)," and if there is a past history, it can be replaced with "0." Thus, data anonymization was strengthened, but the quality of the data remained unchanged.

## REQUEST DATA ANALYSIS FROM OTHER ORGANIZATIONS

Another aspect to be considered is whether the researcher will conduct the study with the data extracted from the hospital, or whether the data were requested from an analysis institution other than the hospital. This is because it is necessary to determine the pseudonymization/anonymization method and the application environment suitable for research purposes and methods. When data analysis is entrusted to an institution other than the hospital, not only a pledge not to provide the data to a third-party is specified in the guidelines but also a contract with additional details is required. To prevent the risk of leakage, the following statements are required: a statement that the data will be stored in a secure in-house network that cannot be accessed from outside, a statement that the data will

be managed in such a manner that only authorized personnel among internal employees can access them, and that the person who accesses the data as well as the access history are also recorded. This includes a statement that the company should write an accurate list (specify the number of people) of the personnel who can access the data, and a statement that if anyone other than people in this list accesses the data, the company will be responsible. From the story of ownership of the results after analysis, the statement that the provided pseudonymized data will be completely destroyed after the end of the study, and the liability for damages in case of data leakage should be mentioned. As many of these measures must be prepared in advance, it is recommended to continue to move conservatively until dismissal, precedent, and interpretation of the rules.

## CONCLUSION

Anonymization of medical data is essential for protecting personally identifiable information in the field of medicine, and for promoting the development of new medical products and techniques. Excessive or weak pseudonymization can result in the loss of data or the inability to provide an adequate level of protection for the data. Information that was absent in the original dataset could be added or modified when pseudonymizing data, which can make a significant difference compared to the actual data because modified data may span most of the data (**Figs. 2** and **3**). Therefore, when considering such factors, the most suitable approach is to solve them internally instead of using an analysis agency. However, data analysis institutions are excellent in terms of equipment and facilities (e.g., artificial intelligence), and it is inevitable to engage them. Considering the laws that have been in force, it is better for each hospital to immediately establish an internal management plan considering it is important to be observant of the interpretation and precedents of data. Moreover, it is better to emphasize more on proper DQM along with data security when considering the fundamental purpose of clinical research.[18]

Although the level of pseudonymization and clinical use of data have a trade-off relationship, it is possible to create pseudonymized data while maintaining the quality of data for the given research purpose. However, while pseudonymization can be performed by the honest broker, researchers should carefully guide the process. Because both privacy and practical use of clinical data, and a harmonious balance are required, the active participation of clinicians is essential as it can induce data generation instead of the loss of data while maintaining the level of data security based on the research purpose. Furthermore, depending on the degree of active participation of the clinician, data anonymization can be strengthened without affecting the data quality.

## SUPPLEMENTARY MATERIAL

### Supplementary Table 1

The technique used to set date data in an arbitrary standard range is expressed as a range or interval of the corresponding value

**Click here to view**


https://doi.org/10.3346/jkms.2021.36.e299

## REFERENCES


1. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? *Clin Pharmacol Ther* 2017;102(6):924-33.
PUBMED | CROSSREF
2. Kim HS, Kim JH. Proceed with caution when using real world data and real world evidence. *J Korean Med Sci* 2019;34(4):e28.
PUBMED | CROSSREF
3. Kim HS, Lee S, Kim JH. Real-world evidence versus randomized controlled trial: clinical research based on electronic medical records. *J Korean Med Sci* 2018;33(34):e213.
PUBMED | CROSSREF
4. Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *BMJ* 2020;369:m958.
PUBMED | CROSSREF
5. Kim HS, Kim H, Jeong YJ, Kim TM, Yang SJ, Baik SJ, et al. Development of clinical data mart of HMG-CoA reductase inhibitor for varied clinical research. *Endocrinol Metab* 2017;32(1):90-8.
PUBMED | CROSSREF
6. Lee J, Kim TM, Kim H, Lee SH, Cho JH, Lee H, et al. Differences in clinical outcomes between patients with and without hypoglycemia during hospitalization: a retrospective study using real-world evidence. *Diabetes Metab J* 2020;44(4):555-65.
PUBMED | CROSSREF
7. Choi J, Bove LA, Tarte V, Choi WJ. Impact of simulated electronic health records on informatics competency of students in informatics course. *Healthc Inform Res* 2021;27(1):67-72.
PUBMED | CROSSREF
8. Shin SY. Privacy protection and data utilization. *Healthc Inform Res* 2021;27(1):1-2.
PUBMED | CROSSREF
9. 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
10. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. *J Med Internet Res* 2019;21(5):e13484.
PUBMED | CROSSREF
11. Korea Legislation Research Institute. Personal Information Protection Act. https://elaw.klri.re.kr/eng_service/lawView.do?hseq=53044&lang=ENG. Updated 2020. Accessed Mar 1, 2021.
12. Ministry of Culture, Sports and Tourism (KR). Three Data Bills. http://www.korea.kr/special/policyCurationView.do?newsId=148867915. Updated 2020. Accessed Mar 1, 2021.
13. Lee D, Park M, Chang S, Ko H. Protecting and utilizing health and medical big data: policy perspectives from Korea. *Healthc Inform Res* 2019;25(4):239-47.
PUBMED | CROSSREF
14. Choi HJ, Lee MJ, Choi CM, Lee J, Shin SY, Lyu Y, et al. Establishing the role of honest broker: bridging the gap between protecting personal health data and clinical research efficiency. *PeerJ* 2015;3:e1506.
PUBMED | CROSSREF
15. Personal Information Protection Commission (KR). Pseudonymization, Combination of Pseudonymized Information. https://www.pipc.go.kr/eng/user/lgp/bnp/pseudonymization.do. Updated 2020. Accessed Mar 1, 2021.
16. Ministry of Health and Welfare (KR). Establish guidelines for the use of health care data for safe use of pseudonym information in the field of health care. http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&page=2&CONT_SEQ=360056. Updated 2020. Accessed Mar 1, 2021.
17. The Hankyoreh. Civil society "The government encourages commercial use of medical information". http://www.hani.co.kr/arti/economy/it/963693.html. Updated 2020. Accessed Mar 1, 2021.
18. Kim HS, Kim DJ, Yoon KH. Medical big data is not yet available: Why we need realism rather than exaggeration. *Endocrinol Metab* 2019;34(4):349-54.
PUBMED | CROSSREF
19. Shin SY. Issues and solutions of healthcare data de-identification: the case of South Korea. *J Korean Med Sci* 2018;33(5):e41.
PUBMED | CROSSREF

20. Jones W, Bruce H, Bates MJ, Belkin N, Bergman O, Marshall C. Personal information management in the present and future perfect: reports from a special NSF-sponsored workshop. *Proc Am Soc Info Sci Tech* 2005;42(1).
CROSSREF
21. Waling L, Sell A. A new vision on personal information managing and sharing using instant messaging. https://www.researchgate.net/publication/31597236_A_New_Vision_on_Personal_Information_Managing_and_Sharing_Using_Instant_Messaging. Updated 2004. Accessed Mar 1, 2021.
22. Mandl KD, Perakslis ED. HIPAA and the leak of "deidentified" EHR data. *N Engl J Med* 2021;384(23):2171-3.
PUBMED | CROSSREF
23. Kim H, Baik SY, Yang SJ, Kim TM, Lee SH, Cho JH, et al. Clinical experiences and case review of angiotensin II receptor blocker-related angioedema in Korea. *Basic Clin Pharmacol Toxicol* 2019;124(1):115-22.
PUBMED | CROSSREF
24. Mehra MR, Ruschitzka F, Patel AN. Retraction-Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet* 2020;395(10240):1820.
PUBMED | CROSSREF
25. Kim HS, Kim H, Lee H, Park B, Park S, Lee SH, et al. Analysis and comparison of statin prescription patterns and outcomes according to clinical department. *J Clin Pharm Ther* 2016;41(1):70-7.
PUBMED | CROSSREF
26. Huh S. Protection of personal information in medical journal publications. *Neurointervention* 2019;14(1):1-8.
PUBMED | CROSSREF
27. Galloway A. Estimating actual height in the older individual. *J Forensic Sci* 1988;33(1):126-36.
PUBMED | CROSSREF
28. Hartman T, Howell MD, Dean J, Hoory S, Slyper R, Laish I, et al. Customization scenarios for de-identification of clinical notes. *BMC Med Inform Decis Mak* 2020;20(1):14.
PUBMED | CROSSREF
29. Purdam K, Elliot M. A case study of the impact of statistical disclosure control on data quality in the individual UK samples of anonymised records. *Environ Plan A Econ Space* 2007;39(5):1101-18.
CROSSREF
30. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem* 1972;18(6):499-502.
PUBMED | CROSSREF
31. Nam GE, Park HS. Perspective on diagnostic criteria for obesity and abdominal obesity in Korean adults. *J Obes Metab Syndr* 2018;27(3):134-42.
PUBMED | CROSSREF
32. El Sanadi CE, Ji X, Kattan MW. 3-point major cardiovascular event outcome for patients with T2D treated with dipeptidyl peptidase-4 inhibitor or glucagon-like peptide-1 receptor agonist in addition to metformin monotherapy. *Ann Transl Med* 2020;8(21):1345.
PUBMED | CROSSREF
33. Hermans WR, Foley DP, Rensing BJ, Rutsch W, Heyndrickx GR, Danchin N, et al. Usefulness of quantitative and qualitative angiographic lesion morphology, and clinical characteristics in predicting major adverse cardiac events during and after native coronary balloon angioplasty. *Am J Cardiol* 1993;72(1):14-20.
PUBMED | CROSSREF