ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20240820001932
Original scientific paper

A Risk-Based Pseudonymization Framework for Healthcare Big Data: A Korean Perspective

Donghyun KIM*, Soonseok KIM

Abstract: The utilization of healthcare big data is rapidly increasing worldwide, but privacy concerns often limit its use, particularly for sensitive information. This study proposes a new pseudonymization methodology that considers data disclosure environments to overcome these limitations. The proposed framework categorizes data and disclosure risks into three levels: low risk for secure internal environments, moderate risk for controlled but semi-open environments, and high risk for public or external environments. These levels provide pseudonymization standards that enable the safe and efficient use of sensitive information. To validate the methodology, a focus group interview and survey were conducted with 30 healthcare experts. Results showed high validity (average score 4.29) and effectiveness (average score 4.63) of the proposed framework. This approach could significantly enhance the utilization of healthcare big data while maintaining privacy protection.

Keywords: comparative analysis; data disclosure risk; health and medical bigdata; personal information; pseudonymization

1 INTRODUCTION

The amount of healthcare data generated is exploding due to the acceleration of the Fourth Industrial Revolution and AI-driven digital healthcare [1]. According to global market research firm Market and Market [2], the AI healthcare market is projected to grow rapidly from approximately USD 20.9 billion in 2024 at an annual growth rate of 48.1% to reaching about USD 148.4 billion by 2029, with the use of healthcare big data to improve healthcare services being a key factor in this growth.

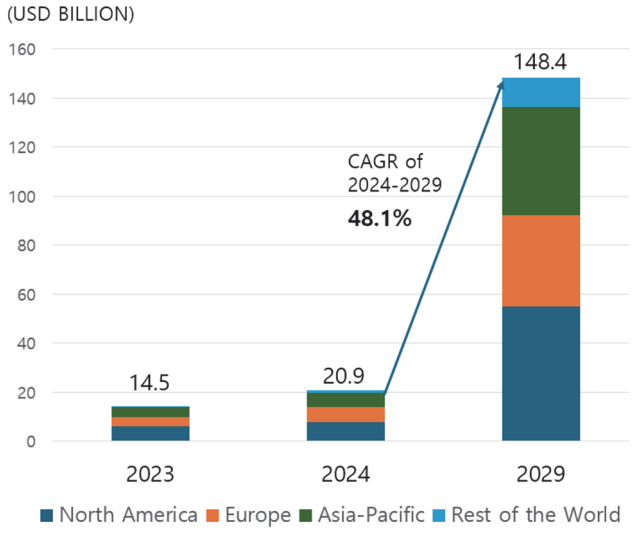

Figure 1 Status of AI healthcare big data market

Countries worldwide are keeping pace with this trend by designating the healthcare industry as a national strategic industry and promoting various policies. Representative projects include Finland's FinnGen project [3], the United States 'All of Us [4]', and Europe's' Biobank [5]'. These initiatives effectively use sensitive biological information as big data to enhance early disease detection and treatment methods, and reduce medical costs.

Korea has also established a national strategy [6] to utilize healthcare information as big data and, through legal revisions, has made it possible to use pseudonymized healthcare information for scientific research without the data subject's consent. Additionally, the Korean Ministry of Health and Welfare has issued the 'Guidelines for the Processing of Pseudonymized Information [7]', considering the sensitivity and specificity of healthcare information. However, it is the principle that mental illnesses, sexually transmitted infections, AIDS, and genomic information, which can significantly impact the privacy of the data subject if re-identified, should be used with the data subject's consent. As a result, the most valuable sensitive information in healthcare data cannot be utilized as big data, posing limitations on conducting various healthcare studies [8, 31].

The current pseudonymization guidelines in Korea have limitations, particularly in failing to quantitatively link data risk and identification risk. This disconnection leads to inefficiencies in data utilization, as non-sensitive or low-identification-risk data items may be over-pseudonymized, reducing the efficiency of big data analysis. This study addresses these limitations by proposing a new pseudonymization procedure.

This study aims to develop a method to utilize highly sensitive healthcare data, which is currently restricted, through a safer pseudonymization process that does not require the data subject's consent. It also seeks to establish criteria for uniformly pseudonymizing healthcare data based on the 'data disclosure environment'.

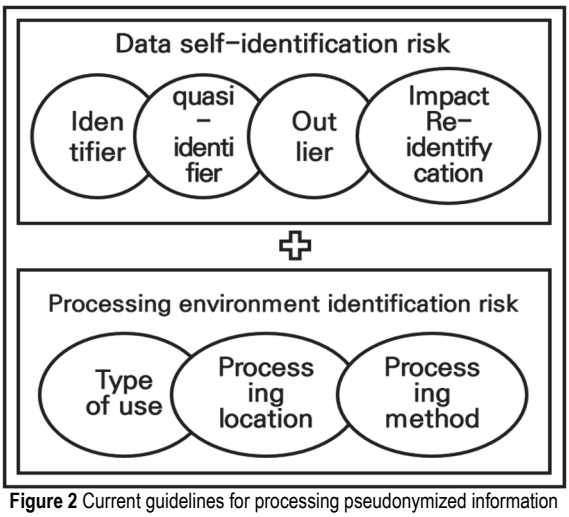

Figure 2 Current guidelines for processing pseudonymized information

The current 'Guidelines for the Processing of Pseudonymized Information [7]' in Korea suggest a risk assessment procedure, as shown in Fig. 2, specifically examining the risks of the data itself and the identification risks in the data processing environment. However, when conducting the final risk assessment, these two elements are not organically linked to produce quantitative standards, and the criteria for pseudonymization are ultimately established based on qualitative judgments by external experts. Consequently, if external experts broadly interpret the scope of personal information, even non-sensitive or low-identification-risk data items may be included in the pseudonymization process, reducing the efficiency of the pseudonymized data generated and making it difficult to achieve big data analysis goals [9].

International projects such as Finland's Finn Gen, the U.S.'s All of Us, and Europe's Biobank have successfully applied pseudonymization in big data utilization, providing relevant insights for improving Korea's guidelines. These projects demonstrate the importance of aligning pseudonymization procedures with national and international ethical and legal standards while maintaining data utility.

This study hypothesizes that if the subjects for pseudonymization can be uniformly selected based on the control of the environment in which data is disclosed, the issue of broadly interpreting the scope of personal information can be resolved. To this end, the study will analyze the pseudonymization procedures for healthcare information suggested domestically and internationally, identify elements that can be supplemented, propose a new risk assessment procedure based on the data disclosure environment, and verify its validity.

The study aims to answer the following research questions:

1) Can a new pseudonymization procedure based on the data disclosure environment improve data utility while minimizing re-identification risk?

2) Can this procedure be effectively incorporated into Korea's pseudonymization guidelines?

Finally, if these results are reflected in Korea's guidelines, it is expected to promote the use of healthcare data in Korea in an international context.

2 LITERATURE REVIEW

2.1 Guidelines

2.1.1 Korea's De-Identification Measures for Personal Information

Korea's 2016 guidelines introduced de-identification measures for utilizing personal information as big data [10]. Unlike pseudonymized information, anonymized data can be freely used or sold, but carry a high re-identification risk when disclosed to large groups [11]. To mitigate this, privacy protection models such as k-anonymity [13], l-diversity [14], and t-closeness [15], outlined in ISO/IEC 20889 [12], are applied. These models reduce identifiability but often result in reduced data utility.

The guidelines emphasize k-anonymity, which, while protecting privacy, can overly generalize data, reducing their effectiveness in detailed analysis. This poses challenges in high-precision fields like healthcare, where data quality is crucial. Additionally, the reliance on external expert reviews for compliance may result in inconsistent application.

2.1.2 Korea's Processing of Pseudonymized Information

Korea's pseudonymization standards, based on the 'Personal Information Protection Act,' provide sector-specific guidelines for fields like finance, healthcare, and education. These guidelines classify pseudonymization criteria into 'identifiability' and 'recoverability,' and include risk assessment procedures for both the data and the processing environment. However, pseudonymized information is restricted to use in scientific research, statistics, and public records, limiting its broader application compared to international standards.

In the healthcare sector, sensitive data like rare diseases and genetic information require explicit consent from data subjects, significantly limiting the use of existing healthcare data in big data projects [16].

While Korea's pseudonymization guidelines provide a structured approach to data protection, the requirement for explicit consent in the healthcare sector restricts the full utilization of valuable datasets. This limits the potential for large-scale healthcare research and innovation, particularly when compared to more flexible international standards.

2.1.3 U.S. NIST SP800-188

The U.S. National Institute of Standards and Technology (NIST) introduced NIST IR8053 [19] in 2015 to provide guidelines for de-identifying government datasets, followed by NIST SP800-188 [22], which offers an eight-step procedure for assessing re-identification risk. This guideline emphasizes minimizing, rather than completely eliminating, the risk of re-identification, and tailors its recommendations based on the type of data shared and the data-sharing model.

Table 1 Considerations for anonymous processing in NIST800-188

Class		Content
1. Identifying Purpose of Use		Identify the purpose of using the data
2. Data Disclosure Threat Assessment		- Identity disclosures - Attribute disclosures - Inferential disclosures
3. Data Life Cycle		Review of re-identification possibilities for each step of de-identification processing
4. Data Sharing Model		-Release and Forget Model: External disclosure, unmanaged -Data Use Agreement (DUA) Model: Contracts Limit Scope of Utilization -Synthetic Data with Verification Model: Generating and Utilizing Original-Like Data
5. The Five Safes	Project	Will you use data legally and ethically to provide public benefit through this project?
	People	The data processor follows the guidelines of action and processes the data.
	Data	have the factors that can identify individuals been handled properly?
	Setting	Control of administrative, technical, and physical access to data. Is it being carried out?
	Outputs	Are the results of the data processing being disclosed?
6. Disclosure Review Boards (DRBs)		A committee that analyzes the possibility of re-identification of de-identification information and its impact on re-identification and carries out final approval
7. De-Identification Standard		The Establishment of standards for de-identification procedures should lead to reliable results
8. Education, Training, Research		Regulation establishment and procedures for de-identification and training in technology are required.

While NIST SP800-188 provides a structured and practical approach for reducing re-identification risk, its focus on merely "minimizing" risk rather than eliminating it may be insufficient for highly sensitive datasets, such as healthcare or financial information. A key characteristic of this guideline is its consideration of the type of data disclosed through the data-sharing model, which aligns with the methodology proposed in this paper. The detailed procedures suggested by the guideline are presented in Tab. 1.

2.1.4 The UK's Anonymisation Decision-Making Framework

The UK Anonymisation Network (UKAN) provides practical guidelines for utilizing anonymous information based on the 2012 'Practical Guidelines for Anonymizing Personal Data' published by the Information Commissioner's Office (ICO). The key feature of the UKAN's Decision-Making Framework on Anonymisation (ADMF) [25] is the concept of 'Data Context', which evaluates both the data itself and the environment in which it is used.

Typically, the risk of data re-identification is measured only from the perspective of the data itself, but the ADMF also considers environmental factors. For example, in highly secure environments, a lower level of de-identification may be applied to improve data utility, whereas in less secure environments, stronger measures are necessary to prevent data leaks or re-identification.

The UK's ADMF offers a flexible approach to de-identification by factoring in the data usage environment, making it more adaptive compared to static de-identification methods. However, its general applicability across various sectors, without specialization for healthcare data, limits its effectiveness in fields requiring stricter privacy controls. Nonetheless, the framework's concept of 'Functional Anonymisation' is valuable for balancing data utility and protection, which aligns well with the objectives of this paper's proposed methodology.

2.2 Law and International Standards

2.2.1 U.S. HIPAA

Enacted in 1996, the Health Insurance Portability and Accountability Act (HIPAA) protects the portability of health insurance and the privacy of U.S. citizens' health information. HIPAA provides two methods for de-identifying medical data: the Safe Harbor method, which involves deleting 18 specified attributes (e.g., name, address, SSN), and the Expert Determination method, where experts evaluate and certify that the risk of re-identification is low [17]. If data cannot be linked to specific individuals after attribute removal, it can be used for secondary purposes without HIPAA restrictions.

While the Safe Harbor method is straightforward, it can be insufficient for protecting against re-identification when additional external data is available. A well-known example is the re-identification of a governor's illness through linkage of voter records and medical data [18]. According to the US NIST 8053 report, the Expert Determination method also has limitations, as it is difficult to guarantee that re-identification risks are fully eliminated [19]. In response to these challenges, the HITECH Act of 2009 was enacted to promote the use of electronic health records (EHRs) while reinforcing penalties for non-compliance and encouraging the secure use of big data in healthcare [20].

HIPAA's dual approach to de-identification provides flexibility, but the reliance on the Safe Harbor and Expert Determination methods leaves room for potential re-identification, particularly when external datasets are involved.

2.2.2 Japan's Next Generation Medical Infrastructure Law

In Japan, medical information is considered sensitive personal information, requiring explicit consent for its use in big data analysis, except in legally specified cases. This restriction has limited the effective use of healthcare data. To overcome this, Japan enacted the Next-Generation Medical Infrastructure Act in 2017, allowing "authorized business operators" to collect and process medical information anonymously. Uniquely, this law employs an Opt-Out system, meaning that patients' data can be used unless they explicitly refuse consent.

The law mandates that identifiable information, personal identification codes, and links between datasets be deleted during the anonymization process [24]. It also requires an evaluation of the purpose, scope, and nature of the data before processing, considering factors like ease of matching information and data type (static, semi-static, dynamic).

The introduction of the Opt-Out system under the Next-Generation Medical Infrastructure Act offers flexibility in utilizing healthcare data without requiring explicit consent in every instance. In order to utilize personal information as big data through pseudonymization, etc. worldwide, it is most efficient to enact and utilize such special laws, and Korea should also consider introducing revisions to special laws in the medical field.

2.2.3 ISO/IEC 20889:2018

ISO/IEC 20889 is an international standard designed to enhance privacy by defining de-identification processes and techniques. Building on ISO/IEC 29100, this standard provides a framework for identifying personal data and introduces principles and techniques for de-identification. It categorizes personal data into 'identifiers' and 'quasi-identifiers' and outlines three methods for managing re-identification risk: Single-Out, Linking, and Inference, as shown in Tab. 2.

Table 2 ISO/IEC20889 re-identification risk judgment criteria

Class	Content
Single-Out	A method of identifying some or all of a target's records by observing attributes known to identify the target uniquely
Linking	A method of combining records belonging to a specific target or target group among records belonging to a separate data set
Inference	A method of inferring the value of a specific attribute from the values of other attributes with a non-negligible probability

ISO/IEC 20889 offers a comprehensive set of tools for managing re-identification risks, making it a valuable

standard for various industries. However, its broad applicability across sectors may lead to challenges in highly specialized fields like healthcare, where stricter privacy controls are necessary. Moreover, while the standard provides a solid framework, the effectiveness of the recommended techniques often depends on the specific data environment, which may require further customization to fully protect sensitive data.

2.2.4 ISO/IEC 25237:2017

ISO/IEC 25237 is a standard that outlines the principles and requirements for pseudonymization services to protect personal medical information. Due to the sensitive nature of medical data, this standard emphasizes the importance of privacy protection through pseudonymization techniques such as data masking, generalization, and derivation. It categorizes pseudonymization into three levels based on the environment in which the data is disclosed:

Level 1: Applies pseudonymization only to direct identifiers in environments with very low re-identification risk.

Level 2: Extends pseudonymization to include indirect identifiers, considering the possibility of linkage with external data.

Level 3: Deals with situations where individuals can be identified due to outliers and considers both prior knowledge and the data disclosure environment.

ISO/IEC 25237 offers a robust framework for safeguarding medical data, addressing different levels of re-identification risk depending on the data usage environment. However, while this approach provides flexibility, the complexity of managing various levels of risk may complicate its practical application, particularly in healthcare settings where real-time data sharing is critical. Ensuring consistent application of the standard across varying environments can also be challenging, raising concerns about its effectiveness in fully mitigating privacy risks.

2.3 Summary

This study builds on existing literature concerning de-identification and pseudonymization standards from the United States, Japan, the United Kingdom, ISO standards, and Korea. These frameworks aim to balance privacy protection and data utility in the context of big data, yet they exhibit notable limitations. This research addresses these issues through the proposed methodology.

In the United States, HIPAA [17] provides two methods for de-identifying medical data (Safe Harbor and Expert Determination), but both methods fail to fully eliminate the risk of re-identification, particularly when combined with external datasets. Japan's Next-Generation Medical Infrastructure Act [23] adopts an Opt-Out consent model, allowing greater data utilization but raising concerns about patient privacy. In Korea, pseudonymization is implemented via guidelines rather than legal statutes, offering limited protection when compared to countries with more robust frameworks.

The UK's Anonymisation Decision-Making Framework (ADMF) [25] emphasizes Functional Anonymisation, adjusting the level of anonymization based on the context in which data is used. This aligns with the statistical anonymization proposed in this study, which applies pseudonymization consistently across various environments.

International standards, including ISO/IEC 25237 [26] and ISO/IEC 20889 [12], provide comprehensive frameworks for medical data pseudonymization and re-identification risk management. ISO/IEC 25237 outlines three levels of pseudonymization based on the disclosure environment, and these levels are integral to the methodology proposed in this study. The study also incorporates elements of NIST SP800-188, specifically the Release Model and Data Release and Review Boards (DRBs) [22]. These mechanisms assess the context in which data is released, ensuring consistent pseudonymization across various scenarios.

2.4 Linking the Proposed Methodology to Existing Literature

The UK's Anonymisation Decision-Making Framework (ADMF) [25] introduces Functional Anonymisation, adjusting de-identification based on data context, which aligns with the statistical anonymization proposed in this study. Complementing this, international standards like ISO/IEC 25237 [26] and ISO/IEC 20889 [12] provide comprehensive principles for pseudonymization and re-identification risk management, while NIST SP800-188 [22] further enhances these frameworks by introducing the Release Model and Data Release and Review Boards (DRBs). These boards assess the context of data disclosure, ensuring consistent pseudonymization practices across different scenarios. The proposed methodology integrates ISO/IEC 25237's three-tiered pseudonymization framework with NIST SP800-188's Release Model and DRBs, resulting in a dynamic approach that adjusts pseudonymization levels according to the security and context of data environments. This approach divides pseudonymization into three levels based on varying degrees of risk and disclosure environments. By dynamically adjusting pseudonymization and leveraging DRBs to review and control data releases, the proposed methodology ensures consistency and enhanced privacy protection. This adaptive approach addresses the limitations of static de-identification methods and provides a more practical solution for the effective use of healthcare big data.

3 PROPOSED METHODOLOGY

The pseudonymization and anonymization procedures for safely utilizing personal information were examined. In this section, the limitations of using Korea's health care information as big data will be discussed, solutions to these limitations will be suggested, and a supplemented pseudonymization procedure will be proposed.

3.1 Problems with Current Pseudonymization Guidelines

The current guidelines consist of four stages, as shown in Fig. 3: pre-review, risk assessment, pseudonymization, appropriateness review, and post-management. Risk

assessment establishes pseudonymization criteria by considering the data itself and the environment in which it is used. The problems with the above procedure are as follows:

- First, the data itself and the environment in which it is used are considered. Still, the risks between the two are not organically linked, and they are not quantitatively calculated.
- Second, pseudonymized information is judged by external experts through the appropriateness review procedure. Split the sentence to Still, suppose the external experts lack expertise and expand the scope of personal information. In that case, it is impossible to use pseudonymization because they do not receive approval, or low-quality pseudonymized information that cannot achieve the purpose of the analysis is created.
- Third, a uniform pseudonymization standard cannot be applied to various personal information attributes, so establishing pseudonymization criteria is expensive and time-consuming.

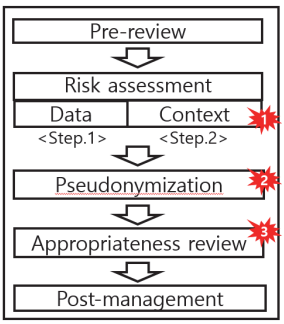

Figure 3 Problems with the current pseudonymization process

3.2 Solutions for the Use of Sensitive Medical Information

It is very difficult to quantitatively measure the data and data environment presented as the first problem by organically examining them. This is because the level of risk varies depending on the situation in which the data is used/utilized.

For example, suppose personal and pseudonymized information is provided to an institution with high protection and analyzed in a closed environment. In that case, the possibility of re-identification is very low because linkage with other information is impossible. In such an environment, pseudonymization standards that lower the level of protection of pseudonymized information and increase efficiency are necessary according to the purpose of analysis. On the other hand, if the level of protection of the environment in which it is used/analyzed is very low, processing standards that strengthen the level of processing of pseudonymized information, increase the level of protection, and decrease efficiency are necessary. Therefore, even if a methodology is presented to solve the first problem, a technology that can mathematically prove the long review time and quantitative calculation in practice must also be prepared.

This study cannot present a perfect methodology for this problem. Still, prior research determined that if certain criteria are met through a review of the data usage environment, the scope of data pseudonymization can be uniformly applied by distinguishing between 'identifiers' and 'quasi-identifiers'.

The method of examining this data usage environment is based on the 'statistical anonymization' type of the UKAN anonymization decision framework [25] examined in Section 2.4 and the level of pseudonymization in ISO/IEC 25237[26] examined in Section 2.5, which are presented based on the environment in which data is disclosed. Here, statistical anonymization is to control or limit the risk of exposure situations through a technology called Statistical Disclosure Control (SDC) and to control the scope of exposure by limiting the scope of use, similar to the legal restriction of the scope of use of pseudonymized information in Korea.

The definition of personal information is generally divided into 1) information that can identify a specific individual with only a single information and 2) information that can identify a specific individual through combination/linkage with other information. ISO/IEC20889 [12] presented in Section 3.1 divides the scope of personal information into 'direct identifiers' and 'indirect identifiers', and NIST standard [23] also divides personal information into 'identifiers' and 'quasi-identifiers' with different terms but the same meaning, and Korean guidelines [7] define them as 'personally identifiable information' and 'personally identifiable information'. Since the interpretation scope of identifiers and quasi-identifiers is similar worldwide, if only risk criteria according to the data disclosure environment are established, uniform pseudonymization can be performed for identifiers and quasi-identifiers.

Even if personal information is classified above, questions may arise regarding distinguishing the 'quasi-identifier' elements in various healthcare fields. In healthcare, the government often defines data in advance (Ministry of Health and Welfare, Korea Disease Control and Prevention Agency, Korea Centers for Disease Control and Prevention, etc.) for rare diseases or the scope of genetic information. Therefore, when utilizing such various information, it means that 'quasi-identifiers' that can identify specific individuals by linking with other information among sensitive information have already been distinguished.

Based on these analysis results, a solution can be derived if standards are established by considering the data disclosure environment, and uniform pseudonymization standards for sensitive information can be established.

Finally, establishing uniform pseudonymization standards based on the data disclosure environment can also solve the second and third problems examined in 3.1. Uniformly applying pseudonymization standards can mean that there are fewer things to review from the perspective of external experts, and consistent processing standards can reduce the time and cost required for pseudonymization.

I proposed a new pseudonym information processing procedure that complements the above solution through 3.3.

3.3 Proposed Pseudonymization Procedure

The proposed procedure is based on the 'functional anonymization' of HITRUST [21] and UKAN [25], examined in previous studies. The guidelines define that the 'data environment (Context)', which considers both the

data itself and the environment in which it is used, should be reviewed. Kim D.H [29] defines the data situation by classifying it into data utilization methods (three perspectives), and utilization environments (three perspectives) as shown in Fig. 4.

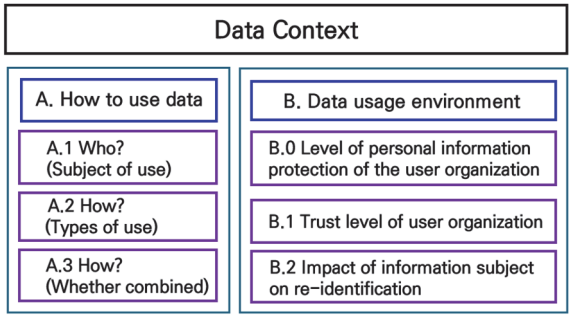

Figure 4 Considerations for pseudonymization based on data environment

This classification is an excellent method for reviewing the risk of the environment in which data is utilized when anonymizing. However, the classification focuses only on the data utilization environment, so a procedure that can measure the data's risk and environment should be supplemented. In other words, the risk of the data itself can be organically measured through a review of the data environment.

To implement this, this paper proposes a new risk measurement method that reflects the solution derived in 3.2 and expands the 'risk review' stage suggested in the current guideline into four stages, adding a processing standard for personal information attributes based on the data utilization environment and an environmental review to improve data efficiency, as shown in Fig. 5. The specific processing methods for each stage are as follows.

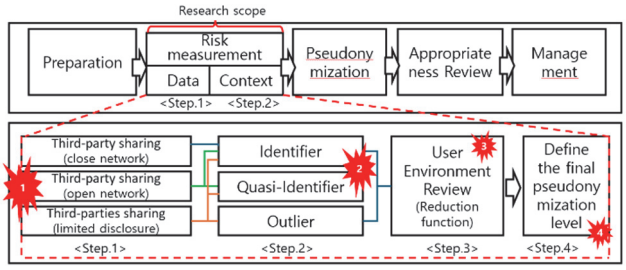

Figure 5 Proposed pseudonymization procedure

• Step 1: Reflecting the guarantee level according to the level of data disclosure presented by ISO/IEC25237 [26], the data disclosure environment is categorized into three levels to balance data utility and privacy protection.

First, the closed network environment refers to a scenario where data is utilized within a restricted network by a specific third party. This environment offers the highest level of security since data is confined to an internal network with minimal exposure, making re-identification risks extremely low.

Second, the public network environment allows for a broader use of data where software provided to a specific third party can be connected with other external information. This environment poses a higher risk of re-identification due to the potential for data to be combined with other external datasets, necessitating stricter pseudonymization measures to ensure privacy protection.

Lastly, the restricted (generalized) environment is defined as an environment where no specific third party utilizing the data is pre-determined. Data may still reside within a closed framework, but access is more generalized, such as in cases where an unspecified number of users may interact with the data, as seen in the recently announced "Personal Information Safe Zone." In this zone, pseudonymized information is stored in a closed network, yet it may be accessed by multiple users under controlled conditions.

In Korea, when using pseudonymized information, the specific third party receiving the information must be clearly designated by law, and the scope of its use is legally restricted. As such, pseudonymized data can currently only be provided in two environments: closed networks and public networks. However, the recently introduced "Personal Information Safe Zone" system provides a new approach where pseudonymized information can be utilized even without designating a specific third party. In this system, pseudonymized information is stored in a closed network that is disconnected from the internet, and this data can be made accessible to an unspecified number of users within the Safe Zone. This allows for more flexible and secure data utilization. For this reason, the Personal Information Safe Zone has been incorporated as the third data disclosure environment, along with the existing closed network and public network, reflecting a broader range of data usage scenarios.

• Step 2: This is the step of selecting subjects for pseudonymization according to the 'data disclosure model' set in Step.1. In the case of 'specific third parties sharing (closed networks)', re-identification during the analysis process (an act of attempting to identify a specific individual) Even if this occurs, it is an environment in which privacy infringement due to external leakage, etc. cannot occur, and since there is no other data or separate S/W to attempt re-identification, only the 'identifier' is defined as the subject of the processing.

In the case of a 'specific third party sharing (open network)', unlike in a closed network environment, it is a model that takes into account accessibility to other information and attackers using external data, which may lead to a situation where an individual can be identified by combining it with other information. Although domestic laws and regulations strictly prohibit re-identification, even quasi-identifiers are defined as processing targets to reduce risks that may occur when identifying a specific individual in combination with other information.

In the case of the last 'third party sharing (limited disclosure)', although it is a closed network, the possibility of re-identification can be very high depending on what prior knowledge an unspecified number of people (to be analyzed in the future) have and what data they have processed. Therefore, in addition to identifiers and quasi-identifiers, 'outliers' were defined as subjects of pseudonymization.

• Step 3: As a procedure to enhance the efficiency of the generated pseudonymized information, the possibility of achieving the analysis purpose is reviewed through interviews with data processors, etc., and the level of pseudonymization is reduced by reviewing administrative / technical protection measures for processors.

• Step 4: After conducting the risk review as above, the final level of pseudonymization is defined through review by the department holding the original and the person in charge of personal information protection.

The following advantages exist when applying the proposed methodology to review the risks of processing pseudonymized information. First, by uniformly defining the scope of pseudonym processing according to the 'data disclosure environment', ambiguous judgment standards that arise when reviewing personal information risks can be clarified. Second, sensitive health and medical information such as genomic information can be safely used through strict control of the data use environment. Lastly, the time required to review the overall level of pseudonym processing can be dramatically reduced, and the efficiency of the pseudonym information generated can be increased by providing a procedure for analysts to achieve the purpose of big data analysis.

4 VALIDATION OF THE PROPOSED PSEUDONYMIZATION PROCEDURE

To verify the newly proposed pseudonymization procedure, researchers conducted two stages of validation. First, 13 experts and employees from medical institutions, each with more than 10 years of experience in the medical field, participated in a Focus Group Interview (FGI). Based on the FGI results, they then conducted a feasibility study from February 10 to February 22, 2024, targeting 230 participants, including pseudonymization experts and consulting experts in pseudonymization processes. This section will present the survey design and the results in detail.

4.1 FGI Review and Survey Design

The FGI involved 13 experts from medical institutions with a minimum of 10 years of experience in healthcare and data management as shown in Fig. 6.

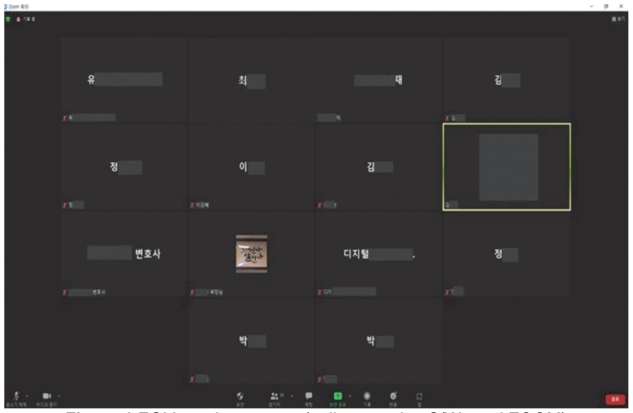

Figure 6 FGI interview screen (online meeting S/W used ZOOM)

The main topics of discussion were the current challenges in pseudonymization processes and how the proposed methodology could improve efficiency and security. Key questions included:

• What are the limitations of the current pseudonymization guidelines?

• How could the proposed framework improve data privacy without compromising usability?

• Is the proposed method applicable to various healthcare data environments?

I supplemented some proposed methodologies by reflecting the FGI results, and I divided the survey into two categories: the validity of the pseudonymization procedure and the effectiveness of the proposed methodology when applied in practice.

For the validity verification, the survey design included detailed factors such as the proposed procedure, processing criteria, mitigation methods, and additional risk review. In the effectiveness part, detailed factors like assistance with pseudonymization, safety enhancement, quality improvement, and rapid processing were derived to structure the survey.

Following the FGI, I designed a comprehensive survey to assess both the validity and effectiveness of the proposed pseudonymization procedure. I administered the survey to 230 participants, categorizing them into two groups: pseudonymization experts (150 participants) and pseudonymization consulting experts (80 participants). The survey consisted of two main sections.

• Validity: This section assessed the appropriateness of the risk measurement methods and the criteria for selecting pseudonymization targets.

• Effectiveness: This section evaluated whether the new procedure enhanced the safety and usability of pseudonymized data, and whether it could lead to improved efficiency in data processing.

Table 3 Key elements of survey design with FGI

Validity	Effectiveness
Pseudonym processing procedures	Help to process pseudonyms
Selection of processing criteria	Strengthen safety
Mitigation method	Quality enhancement
Extend risk review procedures	Expedited processing

I used a 5-point Likert scale in the survey, where respondents rated their agreement with various statements. I chose the 5-point Likert scale for its simplicity and wide acceptance in the field of social science research. The scale allows for a clear differentiation of responses and provides a balance between granularity and respondent ease. In addition, I used IBM SPSS 26.0 to measure the Cronbach α value to evaluate the internal consistency of the survey.

4.2 Results of Survey

As the first type, the newly proposed feasibility study for expanding the risk review procedure for processing pseudonym information showed a high level of reliability, with an average of 4.29 for all responses, an average of standard deviation of 0.673, and a Cronbach α value of 0.736.

As a detailed question, in the case of V1, I found that the overall validity of measuring the risk according to the type of data disclosure was high. However, some expressed the opinion that 'a review of various public models that may occur in the future will also be necessary.' Therefore, I decided that it would be necessary to supplement this paper by adding a new public model if the law is systematically revised in the future.

In the case of V2, the results of the investigation on the validity of the criteria for selecting pseudonymization targets according to V1 (data disclosure type) were lower

than other questions, with an average of 4.03 and a standard deviation of 0.85. As for the opinion of the survey respondents who gave low scores, 'I don't know what criteria should be used to distinguish between identifiers and quasi-identifiers', to supplement this, a checklist that can determine the risk of the data itself was prepared as in Tab. 5.

Table 4 Results of the validity survey

No.	Question	M	SD
V1	Is the risk measurement method appropriate to the data disclosure level in Step 1?	4.53	0.57
V2	Are the criteria for selecting subjects for pseudonymization based on the level of data disclosure in Step 2 appropriate?	4.03	0.85
V3	Is it appropriate to reduce the level of pseudonym processing through user environment analysis in Step 3?	4.23	0.77
V4	Is it appropriate to expand the 2nd stage of existing risk measurement to 4th stage?	4.4	0.50

Table 5 Checklist for assessing the risk to the data itself

No.	Question	Yes	No
C1	Does that data contain items to identify a particular individual with only one item?		
C2	Does the data contain legally limited sensitive and stigmatizing information?		
C3	Does the data include the consent priority information in the Guidelines [7]?		
C4	Does it include items that could infringe on the privacy of individuals if the data is re-identified?		
C5	Does it include items likely to identify individuals through published medical information?		
C6	Does it include items likely to identify individuals through published medical information?		
C7	Is there any item in continuous numerical data withsignificantly lower information at both ends of the distribution?		

Table 6 Results of effectiveness survey

No.	Question	M	SD
E1	Do you think the new proposed procedures will help in the pseudonymization of health care data?	4.67	0.55
E2	Do you think the safety of pseudonymization will be enhanced by using the new proposed procedures?	4.63	0.56
E3	Do you think the quality of pseudonymized information generated using the newly proposed procedure will be higher than before?	4.53	0.51
E4	Do you think the new proposed process will enable selection of subjects for pseudonymization more quickly?	4.67	0.48

In the case of V3, I confirmed a satisfactory level of validity with an average of 4.23 and a standard deviation of 0.77 in reducing pseudonymization based on the environmental analysis of users who utilize pseudonymized information. However, one of the survey responses in Step 4 raised the question, 'How can we guarantee the method of reducing the level of pseudonymization?' This concern prompted me to consider the possibility of re-identification through prior knowledge or other information unknown to the general public. To address this issue, I benchmarked the Data Disclosure Review Boards (DRBs) [22]. In addition, I referenced the Caldicott Guardian system [30], suggested by the UK, and the US NIST 800-188 framework, both of which introduced a procedure to have data analysis experts in the healthcare field evaluate the disclosure risks. These systems established a method to review whether additional pseudonymization is necessary, providing a structured approach to safeguard against re-identification risks.

In the second type of survey, I found that the average of the total responses for the effectiveness of the proposed pseudonymous information processing procedure was 4.62, with a standard deviation of 0.522. Additionally, I calculated the Cronbach α value as 0.757, indicating a high level of reliability. There was no separate opinion on the survey results, and it is judged that the survey respondents as a whole reflected the complexity of the guideline [7] procedure currently presented by the government, excessive pseudonymous processing standards, and the demand to improve problems such as a decrease in the efficiency of the generated pseudonymous information. Fig. 7 shows the final procedure for analyzing the survey results and reflecting expert opinions.

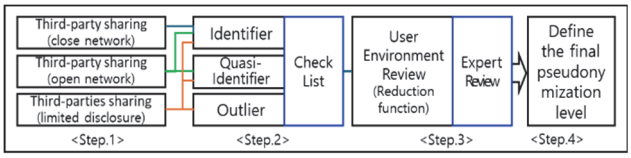

Figure 7 Final proposal reflecting survey opinions pseudonymization procedure

The final proposed procedure reflects the opinions of the survey respondents in V2, adds a checklist to [step 2] to assess the risk of the data itself, and supplements it so that it can be used to assess whether it is a quasi-identifier. In addition, in order to prevent the possibility or concern of re-identification in the future when the level of data protection is reduced through the review of the user environment presented in V3, the procedure for expert review is supplemented, and a procedure is presented to recommend additional pseudonymization in case there is a concern of re-identification.

4.3 Results Summary

In the second survey, we focused on further evaluating the items V2 and V3, which had received comparatively lower scores in the first survey. The aim was to assess the impact of the modifications made to the pseudonymization procedure based on the initial feedback.

The V2 and V3 items were included to evaluate specific aspects of the pseudonymization process. V2 assesses the adequacy of the risk evaluation criteria under various data disclosure environments, while V3 evaluates the efficiency and practicality of the pseudonymization procedure in real-world healthcare settings. These items were selected for further analysis due to their relatively lower scores in the initial survey.

According to the results of the second survey, the overall average score of V2 and V3 improved to 4.4, which is higher than the score observed in the first survey. The Cronbach α value was 0.732, indicating a satisfactory level of internal consistency and reliability improvement compared to the first survey, and the results are as shown in Fig. 8.

Furthermore, a t-test was conducted to compare the results from the first and second surveys. The t-value for V2 was –7.779, and for V3 it was –4.583, with both items showing highly significant p-values of 0.000. These negative t-values indicate that the average scores for V2

and V3 in the second survey were significantly higher thanin the first survey, reflecting a notable improvement in the participants' perceptions of these aspects of the pseudonymization procedure. The highly significant p-values confirm that these improvements were not due to random variation, but represent meaningful enhancements in the evaluated criteria.

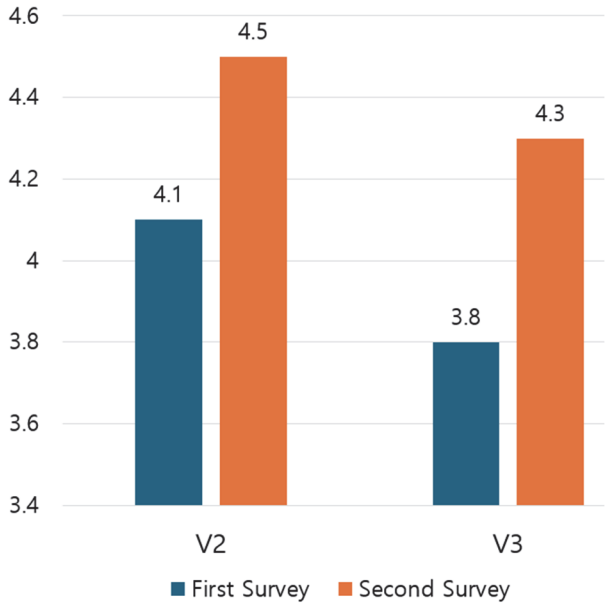

Figure 8 Comparison between first and second surveys for V2 and V3

The improvements observed in the second survey suggest that the adjustments made to the pseudonymization procedure led to substantial progress in both the risk assessment criteria (V2) and the efficiency of pseudonymization (V3). These findings support the conclusion that the proposed modifications are not only theoretically sound but also practically applicable in healthcare data environments, enhancing both the safety and usability of pseudonymized data.

5 CONCLUSIVE REMARKS

5.1 Limitations of the Study

This study presents a method for safely utilizing Korean healthcare information by pseudonymizing it, and proposes a new pseudonymization information processing procedure after identifying problems in the current guidelines. However, I mainly evaluated the proposed pseudonymization procedure in a controlled environment, and I have not yet verified its applicability in a wider range of real-world scenarios. Furthermore, since I based this study on the Korean pseudonymization guidelines, comparative studies that reflect the frameworks of other countries are necessary.

5.2 Directions for Future Research

Future research should verify the proposed procedure in various real data environments and conduct comparative analysis with pseudonymization methods in other countries to increase generalizability. In many cases overseas, there are no government restrictions or regulations on using pseudonymized information. This includes fostering global companies such as Google and Netflix through the free use of pseudonymized information to strengthen the country's data competitiveness.

Therefore, to reflect the results of this study in the Korean government's guidelines, efforts are needed to suggest various best practices through empirical research targeting actual medical institutions. Empirical research will verify the effectiveness of the proposed pseudonymization procedure and reflect the best practices obtained through this in government guidelines, thereby maximizing the efficiency of data utilization.

5.3 Conclusion

This study proposed a novel pseudonymization procedure that addresses the limitations of current big data utilization in healthcare information, specifically by incorporating Korea's pseudonymization guidelines.

The proposed procedure enhances the efficiency of pseudonymized data creation by introducing a smoother and more secure pseudonymization process. Despite the relatively small sample size, a survey of healthcare experts demonstrated that the proposed pseudonymization procedure is both practical and beneficial in real-world applications. Through this approach, organizations that utilize healthcare information can safely leverage sensitive data for research and analysis. The key contributions of this study are as follows:

1. Uniform criteria were presented for selecting pseudonymization targets based on the data disclosure environment, addressing the gap in existing methodologies that fail to fully account for different disclosure risks.
2. The efficiency and safety of pseudonymized data were enhanced through the implementation of mitigation measures tailored to the data use environment, further strengthening data protection through expert evaluation.
3. The practical applicability of the proposed procedure was demonstrated by collecting and incorporating opinions from healthcare experts, ensuring that the method is feasible and effective in real-world settings.

However, one limitation of this study is its focus on the Korean pseudonymization guidelines, which may limit the broader applicability of the findings in other regions or regulatory frameworks. To address this, future research should conduct comparative studies incorporating pseudonymization frameworks from other countries. By doing so, the proposed procedure can be refined and adapted to meet international standards, thus enhancing its global applicability and usefulness.

Moreover, this study makes a significant contribution to the field by offering a pseudonymization framework that balances data utility and privacy protection. By introducing a risk-based approach, we overcome the limitations of existing guidelines and provide a more adaptable and secure method for healthcare data utilization.

Policymakers and practitioners should consider incorporating the proposed pseudonymization procedures into data protection regulations to enhance both the utility and security of healthcare data. Healthcare organizations are encouraged to adopt these procedures to improve data governance practices, enabling secure data sharing and analysis. By implementing these recommendations, stakeholders will contribute to the development of a more

robust, privacy-conscious data utilization ecosystem in the healthcare sector.

Acknowledgments

This research was supported by the Korea Health Technology R&D Project grant through the Korea Health Industry Development Institute(KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI23C0733).

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT)(IITP-2025-RS-2024-00438056, 50%) and the Korea Health Industry Development Institute(KHIDI) grant funded by the Korea government(MOHW)(HI23C0733, 50%).

6 REFERENCES

[1] Liu, F. & Panagiotakos, D. (2022). Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology*, *22*(1), 287. https://doi.org/10.1186/s12874-022-01768-6

[2] https://www.marketsandmarkets.com/MarketReports/artificial-intelligence-healthcare-market-54679303.html

[3] Jukarainen, S., Kiiskinen, T., & Kuitunen, S. (2022). Genetic risk factors have a substantial impact on healthy life years. *Nat Med*, *28*, 1893-1901. https://doi.org/10.1038/s41591-022-01957-2

[4] Tartof, S., Slezak, J., Fischer, H., & Hong, V. (2021). Effectiveness of mRNA BNT162b2 COVID-19 vaccine up to 6 months in a large integrated health system in the USA: a retrospective cohort study. *The Lancet*, *398*(10309), 1407-1416. https://doi.org/10.1016/S0140-6736(21)02183-8

[5] Coppola, L., Cianflone, A., & Grimaldi, A. M. (2019). Biobanking in health care: evolution and future directions. *Journal of Translational Medicine*, *17*, 172. https://doi.org/10.1186/s12967-019-1922-3

[6] Ha, T. J., Kang, S. G., Yeo, N. Y., & Park, S. W. (2024). Status of My Health Way and Suggestions for Widespread Implementation, Emphasizing the Utilization and Practical Use of Personal Medical Data. *Healthcare Informatics Research*, *30*(2), 103-112. https://doi.org/10.4258/hir.2024.30.2.103

[7] Ministry of Health and Welfare (2024). *Guidelines for Use of Health and Medical Data.*

[8] Kim, G. R. & Lee, D. H. (2018). Review on Healthcare Big Data Analysis. *Hannam Journal of Law & Technology*, *24*(3), 57-90. https://doi.org/10.32430/ilst.2018.24.3.57

[9] Jung, Y. J. (2022). A Review on the Relationship between Health Medical Data and Privacy. *Kookmin Law Review*, *34*(3), 191-225. https://doi.org/10.17251/legal.2022.34.3.191

[10] Joint Government Departments in Korea. (2016). Guidelines for De-Identification of Personal Information.

[11] Cecaj, A., Mamei, M., & Bicocchi, N. (2014). Re-identification of anonymized CDR datasets using social network data. *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, 237-242. http://doi.org/10.1109/PerComW.2014.6815210

[12] ISO/IEC 20889 (2018). *Privacy enhancing data de-identification terminology and classification of techniques*.

[13] Sweeney, L. (2002). k-ANONYMITY: A Model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, *10*(5), 557-570. https://doi.org/10.1142/S0218488502001648

[14] Machanavajjhala, A., Gehrke, J., & Kifer, D. (2006). ℓ-Diversity: Privacy Beyond k-Anonymity. *IEEE International Conference on Data Engineering*, *1*(1), 24. https://doi.org/10.1145/1217299.1217302

[15] Li, N. & Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *2007 IEEE 23rd International Conference on Data Engineering*, 106-115. https://doi.org/10.1109/ICDE.2007.367856

[16] Jeon, S. R. (2024). The Burden of Clostridioides difficile Infection in Korea. *Journal of Korean Academy of Medical Sciences*, *39*(12), e122. https://doi.org/10.3346/jkms.2024.39.e122

[17] U.S. Department of Health and Human Service (2012). *HIPAA Privacy Rule*.

[18] Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. *Data Privacy Working Paper*, 3. https://doi.org/10.1184/R1/6625769.V1

[19] Garfinkel, S. L. (2015). NISTIR 8053 De-Identification of Personal Information. *NIST*. https://doi.org/10.6028/NIST.IR.8053

[20] U.S. Department of Health and Human Service (2009). *HITECH Act*.

[21] Eman, K. E. (2013). *Guide to the De-identification of Personal Health Information*. CRC Press.

[22] Garfinkel, L. (2022). NIST800-188 De-Identifying Government Data Sets. *NIST*. https://doi.org/10.6028/NIST.SP.800-188.3pd

[23] Secretariat for the Promotion of Health and Medical Strategy (2022). *Next Generation Health Care Act*.

[24] Kim, S. T. (2022). A Study on the Legal System Related to the Use of Medical Information in Japan. *Korea National Institute for Bioethics Policy*, *6*(1), 1-31. https://doi.org/10.23183/konibp.2022.6.1.001

[25] UK Anonymisation Network (2016). *The anonymisation decision making framework*.

[26] ISO/IEC 25237 (2017). Health informatics-Pseudonymization.

[27] Sweeney, L. (2017). Re-identification Risks in HIPAA Safe Harbor Data. *Technology science*, *2017*.

[28] Hiramatsu, K., Barrett, A., & Miyata, Y. (2021). Current Status, Challenges, and Future Perspectives of Real-World Data and Real-World Evidence in Japan. *Drugs - Real World Outcomes*, *8*, 459-480. https://doi.org/10.1007/s40801-021-00266-3

[29] Kim, D. H. (2022). A study on Data Context-Based Risk Measurement Method for Pseudonymized Information Processing. *Journal of The Korea Society of Computer and Information*, *27*(6), 53-63. https://doi.org/10.9708/jksci.2022.27.06.053

[30] Roch-Berry, C. (2003). What is a Caldicott guardian? *Postgraduate Medical Journal*, *79*, 516-518. https://doi.org/10.1136/pmj.79.935.516

[31] Gulati, R. (2023). Digital health and technological interventions. *International Journal of Advanced Nursing Education and Research*, *8*(2), 37-44. https://doi.org/10.21742/IJANER.2023.8.2.04

Contact information:

Donghyun KIM
(Corresponding author)
Halla University,
(26404) 28, Halla Dae-gil, Wonju-si, Gangwon-do, Republic of Korea
E-mail: dh.kim@halla.ac.kr

Soonseok KIM
Halla University,
(26404) 28, Halla Dae-gil, Wonju-si, Gangwon-do, Republic of Korea
E-mail: sskim@halla.ac.kr