International Journal of Software Engineering and Its Applications
Vol. 10, No. 8 (2016), pp. 127-138
http://dx.doi.org/10.14257/ijseia.2016.10.8.12

Personal Health Information De-identified Performing Methods in Big Data Environments

Ya-Ri Lee[1], Young-Chul Chung[2], Jung-Sook Kim[3*] and Ho-Kyun Park[4]

[1]*Personal Information Protection Center of SSIS, Republic of Korea*
[2]*Big Data Research Center of KIHASA, Republic of Korea*
[3]*Division of Computer Science, Sahmyook University, Republic of Korea*
[4]*School of IT Convergence Engineering, Shinhan University, Republic of Korea*
i_lyaree@ssis.or.kr, cyc@kihasa.re.kr, kimjs@syu.ac.kr, hkpark@shinhan.ac.kr

Abstract

The field that shows the most promise among the application areas of Big Data is the medical sector. We must first deal with the problems of the invasion of privacy and misuse of personal information before we can utilize personal health information. There is a method known as the de-identification of personal health information which is one of the methods that can be used to perform the tasks of the protection and the utilization of personal health information at the same time. De-identification refers to the process of making it impossible to know the identity of an individual just by the information that is revealed. The biggest problem related to de-identification is the phenomenon of re-identification. Even with information that at first cannot be used to readily identify an individual, when enough of it has accumulated in various categories, it may become possible to identify the hidden individual behind it. What makes de-identification difficult is that one cannot completely rule out the possibility of re-identification, and that the information becomes an object of much regulatory legislation if it is re-identified as personal information. Thus, it is a priority to reach an agreement among the many parties involved since we need to organize the data into categories that are actually used. This study seeks to analyze the current state of de-identification measures which are some of the protective measures for personal information for the safe utilization of Big Data in the medical sector, both in and out of Korea, and to propose implementation plans for safe de-identification. Furthermore, this study advocates an active consideration of the establishment of a "central tower" agency which will be able to carry out the much-needed continuous monitoring for re-identification as well as the assessment of the adequacy of the de-identification methods.

Keywords*: Big Data, Personal Health Information, De-identification, Privacy Protection, Re-identification*

1. Introduction

Not only is Big Data recognized as a valuable tool in diverse fields but is also seen as a creative force behind new values for data integration across many different fields. In addition to the values of Big Data, an important issue is emerging that has to do with personal information violation and the misuse and abuse of personal information. Personal information protection and the utilization of Big Data are like two sides of a coin. This issue of personal information protection is being perceived as a limiting factor in Big Data utilization and policy implementation [1]. Thus, in the cases where Big Data makes use of information, it is crucial to have a method that can be used to protect the basic rights of the main subjects such as having the final say in the use of personal

* Corresponding Author

ISSN: 1738-9984 IJSEIA

information. Any discussion of personal information protection will invariably have at its center the methods of de-identification and anonymization. De-identification refers to the process that renders it impossible to identify a given individual only by the revealed information. If the individual cannot be identified, it will dramatically reduce the need to regulate the circulation and management of such information. There are many technical and legal problems related to de-identification but the biggest problem will be re-identification. This is due to the fact that even information that at first cannot be used to identify an individual, when it has sufficiently increased in quantity and kind, can eventually be used to identify the main subject of the information [2]. The medical field is one of the most promising fields in the utilization of Big Data. In order to utilize Big Data comprehensively in the field of medicine, we must clarify and differentiate between the types of medical information, the definition of personal medical information, and its domain [3].

This paper seeks to analyze the current state, domestic and foreign, of the steps taken in the de-identification process which is one of the measures for the protection of personal information for the safe usage of Big Data in the field of medicine. Based on that analysis, the paper proposes a 4-step plan for the implementation of safe de-identification methods. It also advocates the establishment of a "control tower" professional organization which can provide objective monitoring on a nationwide scale. Personal health information de-identification requires the continuous monitoring of the possibility of re-identification and aptitude inspections, while at the same time adhering to the laws and regulations.

2. Related Research

Each nation of the world and related organizations have their own opinion on the definition of de-identification and anonymization. The US and the European Union make a distinction between the two, while Japan and the UK have a different position. Table 1 lists the definitions of de-identification according to various nations, whereas Table 2 shows the definitions of anonymization in different nations. Recently, the term "de-identification" has been used in Korean documents [4]. The debate regarding the precise distinction between the two terms will continue, and in this paper the term "de-identification" is used throughout.

Table 1. The Definition of De-identification According to Nation

Nation	Content
USA/EU	A state where it is very difficult to identify the individual based his/her original data due to the elimination of identifiable factors.
Korea	An action of deleting or replacing a part or the whole of a piece of personal information through data processing and making it impossible to identify a particular individual even when the personal information is combined with other pieces of information. ▶ Processing methods: deletion, use of a pseudonym, total processing, categorization, data masking, etc[6].

Table 2. The Definition of Anonymization According to Nation

Nation	Content
USA/EU	- Irretrievable de-identification - State in which the link between the individual and the identifier is irreversibly eliminated, thereby making identification of the individual impossible ▶Processing methods: falls under the category of follow-up measures after de-identification[6]
Korea	- Changing data into a form that cannot be recognized by means of deleting identifiable factors included in the data[6]. ▶Processing methods: pseudonymization, generalization, permutation, perturbation(adding an encoding mechanism of the anonymization technique[10]
Japan	- Creation of anonymous fictitious data - The process that makes it impossible to restore the concerned information. This process can be done by fabricating personal information to prevent the identification of a particular individual ▶Processing methods: Partially or completely deleting personally identifiable descriptions, symbols in the personal information[8]
U.K	- Action which turns personal information that can be used to identify the individual into a form which cannot be used to identify him/her. ▶Processing methods: Masking the face of subjects on visual data gathered by CCTV[9]

2.1. International Standards Regarding De-identification

2.1.1. ISO 25237: ISO 25237 includes requirements for personal information protection using de-identification services in a database. This defines the terms of anonymization, pseudonymization, and de-identification as shown in Table 3[11].

Table 3. The Definitions of Anonymization and De-identification According to ISO 25237

Terms	Content
Anonymization	This process blurs the relationship between data sets and data subjects that are actually used for identification
Pseudonymization	As a specific type of anonymization, this process adds a connection between a specific data subject and one or more anonymous subjects after eliminating the connection between two data subjects
De-identification	This process eliminates the connection between data subjects for one or more usage purposes with regards to an identification data set and any created data sets

Table 3 provides examples of different uses of anonymization, both reversible and irreversible. It defines a basic method of anonymization service which deals organizational and technical traits. Moreover, it offers a guideline for evaluating re-identification, and designates the minimum requirements for a policy framework and a reliable method for the operation of anonymization services.

2.1.2. ISO 27799: ISO 27799 gives definitions of health information that needs to be protected. It suggests that personal health information is not such if it cannot be used to verify the data subject when it does not include any anonymized information. Under these conditions, the ISO 27799 defines health information to be protected in eight categories as shown in Table 4 and stipulates that the secrecy, integrity, and usability of such information must be protected [11].

Table 4. Health Information Categories to be Protected According to ISO 27799

No.	Items
1	Personal health information
2	Pseudonymized data derived from personal health information
3	Statistics and research data that have been anonymized by eliminating data that can be used to identify the individual
4	Clinical/medical knowledge date including decision support data
5	Data regarding medical personnel and employees
6	Data related to public health supervision
7	Follow-up inspection data produced by the health information system
8	HIS system security data which include access control data and security-related system component data

2.1.3. HIPAA: HIPAA is a law that comprehensively deals with the privacy rights of health information subjects and stipulates the measures that must be followed by medical institutions and insurance companies in order to protect the protected health information (PHI) of the relevant subjects[13]. HIPAA privacy rule clause 164.514 stipulates that "any information that cannot be used to identify an individual or does not have any reasonable basis for identification is not personally identifiable health information." Thus, the basic rule of HIPAA says any institution that is the object of privacy regulations cannot use or reveal any personally identifiable information. But when exceptions are allowed or required by the rules, personal information can be released upon written approval by the concerned individual or a proxy. Any medical information management institution that is a covered entity of HIPAA privacy rules can de-identify personal health information by choosing one of two options. The first option is a procedural approach called the expert determination method. The second is a content-based approach where de-identification is considered to have been accomplished when certain identifiers and quasi-identifiers have been removed from the data. The content-based approach is called the safe harbor method. HIPAA's PHI (Protected Health Information) is defined as 18 main personally identifiable items as shown in Table 5[4].

Table 5. 18 Main Personally Identifiable Items

No.	Items	No.	Items	No.	Items
1	Name	8	Medical record number	15	Internet address, IP address
2	Address	9	Health insurance number	16	Bank account number
3	Dates(birth, hospitalization, discharge, death, etc.)	10	Biological information including fingerprint and voice record	17	Identification photo or other personally identifiable photos
4	Phone number	11	Certification number	18	Others(excluding cases that are permitted by the re-identification related clauses in the laws). Other unique identifiable variables, characteristics, and symbols
5	Fax number	12	Vehicle identification number, serial number, registration number		
6	Email address	13	Equipment identification number, serial number		
7	Social security number	14	URL information		

2.2. The Present State of De-identification in Korea

In Korea, *Personal Information Protection* Act enacted in September 2011 contains a clause that stipulates that "one who manages personal information shall anonymize the personal information when such a process is possible."

The *Personal Information Protection Act* speaks of anonymization only in a broad sense and does not give a definition of anonymization or discuss the scope and use of anonymized data. For this reason, additional guidelines regarding personal information protection as it relates to the recent controversy involving Big Data are shown in Table 6.

Table 6. Guidelines for the Processing of Personally Identifiable Data in Korea

Title	Date of Announcement	Institution	Usage Area
Protective guidelines for personal information as it relates to the opening up and sharing of public information[14]	Sep. 2013	Safety Administration	Public
Casebook of personal information de-identification for the utilization of Big Data[6]	May 2014	Ministry of Science, ICT and Future Planning/ National Information Society Agency	Public/Private
Big Data personal information protection guidelines[15]	Dec. 2014	Korea Communications Commission	Public/Private
Handbook for a free assessment of the adequacy of personal information de-identification[16]	Dec. 2014	Ministry of Government Administration and Home Affairs/ National Information Society Agency	Public/Private
Handbook of techniques for personal information de-identification for the utilization of Big Data[17]	May 2015	Ministry of Science, ICT and Future Planning/ National Information Society Agency	Public/Private

Following the recent changes in the IT environments, studies using data that contain personal information (Pintech, SNS, Cloud Computing, Big Data, personal health information statistics, *etc*.) are increasing in number. Also, amidst the rising interests and concerns regarding a more efficient use of data through the de-identification of personal information, discussions for countermeasures are being continually held.

In Korea, a de-identification technique is defined as a series of measures that makes it impossible to identify a particular individual by deleting or replacing a part or the whole of a piece of information and preventing it from readily combining with other data. As shown in Table 7, de-identification techniques include, pseudonymization, total processing, data deletion, categorization, and data masking [6].

Table 7. De-identification Techniques

Terms	Content
Pseudonymization	Replacing main identifiable elements with different figures and making it difficult to identify the individual
Total processing	Showing the total sum of the data and hiding the values for individual data
Data deletion	Deleting unnecessary values or values that are important to identification among the component values of the data set depending on the purpose of data release or sharing
Categorization	Placing the values of the data in broad categories and hiding the precise values
Data masking	Making it impossible to identify an individual by hiding main personal identifiers that have a high probability of contributing to the identification of the individual by combining with revealed data

3. Personal Health Information De-identification Methods

This study proposes a four-step methodology for de-identification as shown in Figure 1 for the safe de-identification in the utilization of Big Data in personal health information. The four steps are: 1) defining standard personal health information, 2) offering de-identification guidelines, 3) suggesting limiting factors, 4) performing constant monitoring for the prevention of re-identification.

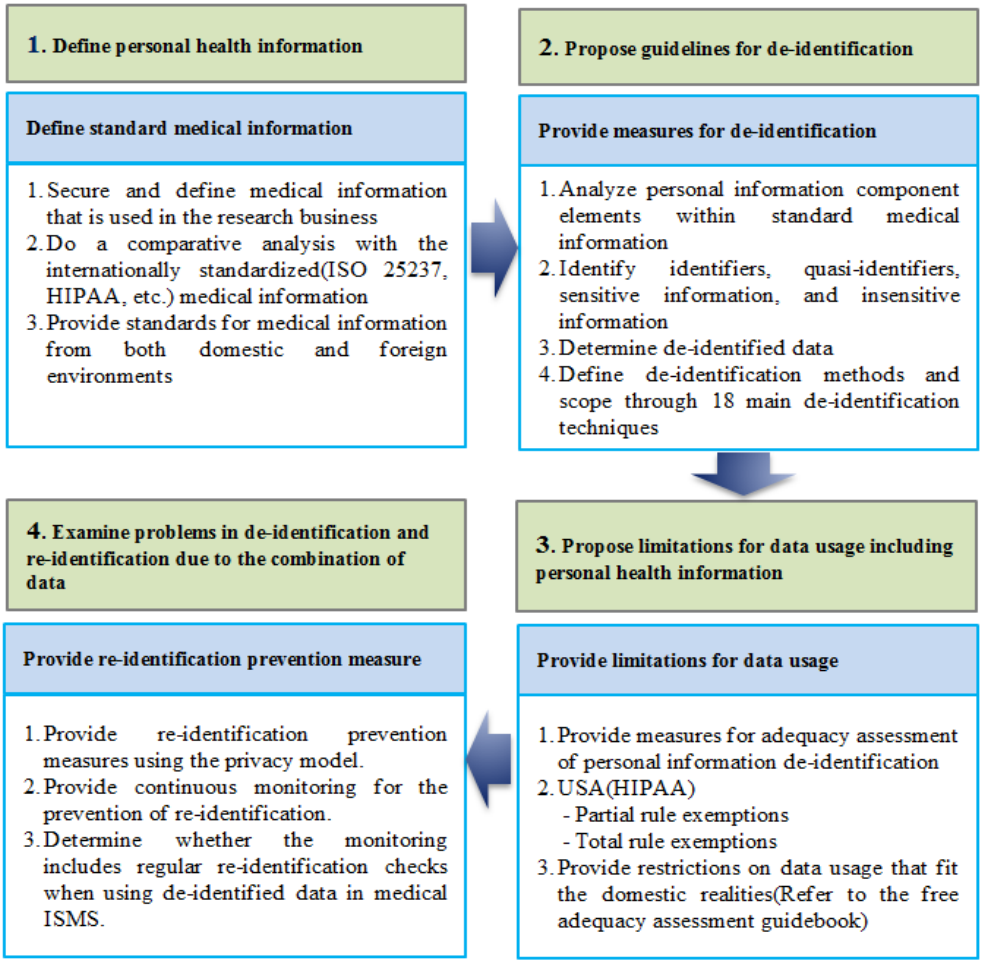

Figure 1. A Four-Step De-identification Process

3.1. Definition of Standard Health Information

For the sharing of health information in Big Data and the exchange of information among systems, we must first establish a standardized system of terminology regarding personal health information. For this purpose, we must secure and define health information that is used throughout domestic research and related businesses. We need to have comprehensive health information standards for both the domestic and international environments through the comparative analysis of health information that is used under international standards.

3.2. De-identification Guidelines

Personal information is composed of four elements which are identifiers, quasi-identifiers, sensitive information, and insensitive information.

Table 8. Definitions of the Component Elements of Personal Information

Terms	Content
Identifier	-That which shows the clear identity of the individual. Examples: Resident Registration Number, Passport Number, Driver's License Number, Alien Registration Number.(in USA: Social Security Number)
Quasi-Identifiers	-Also known as Key Attributes. -An attribute that has a certain degree of ambiguity and sometimes can be used to identify an individual when combined with other quasi-identifiers. Example: name, address, gender, age, telephone number, *etc*.
Sensitive Information	-A characteristic that includes sensitive information about an individual -Also known as secret information deduction attributes Example: monthly salary, region, political affiliation, health, *etc*.
Insensitive Information	-Any information that is not included in any of the above three categories. -Also known as non-secret information deduction attributes

In this step, we confirm the data that need to be de-identified by analyzing the personal information component elements in the medical information that was defined in the first step-defining standard medical information-and categorizing them as identifiers, quasi-identifiers, sensitive information, or insensitive information. Lastly, we define the method and scope of de-identification by applying de-identification techniques to the 18 personally identifiable items defined under HIPAA.

3.3. Limitations on Data Usage

In this step, we must regulate the limitations on data use and determine whether certain data can be used as Big Data or not. In this process, we provide limitations on data usage that fit domestic conditions after referring to and analyzing regulations from both domestic and foreign examples to assess the adequacy of personal information de-identification.

In the HIPAA rules and regulations the partial waiver and exemption rules across the board, as shown in the following Table 9.

Table 9. Limitations on Data Usage in HIPAA

Partial rule Exemptions	Total rule Exemptions
Name	Name
Address	Address
-	Dates (birth, hospitalization, discharge, death, etc.)
Phone number	Phone number
Fax number	Fax number
Email address	Email address
Social security number	Social security number
Medical record number	Medical record number

Health insurance number	Health insurance number
Biological information including fingerprint and voice record	Biological information including fingerprint and voice record
Certification number	Certification number
Vehicle identification number, serial number, registration number	Vehicle identification number, serial number, registration number
Equipment identification number, serial number	Equipment identification number, serial number
URL information	URL information
Internet address, IP address	Internet address, IP address
Bank account number	Bank account number
Identification photo or other personally identifiable photos	Identification photo or other personally identifiable photos
-	Others (excluding cases that are permitted by the re-identification-related clauses in the laws). Other unique identifiable variables, characteristics, and symbols

3.4. Re-identification Prevention Measures

Privacy protection model [16] (k-anonymity, l-diversity, t-access) applies the model and determines the values that are needed for evaluation by thoroughly examining the possibility of re-identification and results and data attributes from an analysis of the effects of a possible personal information leakage. The causes for re-identification of de-identified personal information that reflect the limiting factors can include a third-party re-identification agent besides the personal information manager, such as a criminal, stalker, or hacker. For example, a re-identification agent may find out the actual user of a Twitter account and figure out the behavior patterns of the user, and obtain the email address or phone number and send him/her spam messages or figure out one of his/her behavior patterns. An event like this could happen at any time. Especially, an ill-meaning re-identification agent like a stalker can acquire re-identification techniques and seriously infringe on someone's personal privacy. Threats of re-identification of personal data can be grouped into 3 broad categories as shown in Table 10 [16].

Table 10. Elements of a Personal Information Re-identification Threat

Category	Content
Increase in data release	• The largest background for the claim that the danger of re-identification of de-identified personal information is very high • An increase in the possibility of personal identification from unidentifiable de-identified information by means of a new, third securing of data by a researcher or an ill-meaning hacker who possesses de-identified data • Can occur between data from certain areas like medicine and Internet service and revealed data from different fields. • Re-identified data leads to a serious breach of privacy for the individual

Increase in custom-made advertisement data and intensification of data concentration	• Custom-made advertisements are taking up more and more space among advertisements, and there is a rising trend of cases where various industries are adopting and utilizing custom-made advertisements • Possibility of de-identified personal information that is used for online custom-made advertisements containing personal data such as the behavior, tendencies, location, etc. of the individual • Advertising businesses can indiscriminately collect quasi-identifier-related data such as personally identifiable IP addresses, session information, online search records, etc. through information-gathering applications • It is anticipated that there will be an increasing possibility that industries that provide custom-made advertisements will possess data attributes that can be used to identify an individual as they gather more and more data • The possibility of re-identification from concentrated data and of the violation of personal information is increasing
Machine To Machine environment and increase in de-identified data	• Techniques in Machine To Machine(IoT7), mobile environment, etc. create new forms of data such as the location of an individual • Extensive data such as location and biological information that can be gathered through sensors, IP addresses, instruments, etc. have emerged as personally identifiable elements • Increase in the possibility of identification of specific devices and their location through IP addresses in a super-connected society where many devices have an IP address • Information linked with Machine To Machine and location information can be used to re-identify an individual more than existing identifiable data as it is sensitive de-identified information that reveals an individual's behavior patterns and life patterns

From the perspective of the industry's personal information processors, the consumers' various behavior patterns can be used as very important factors in product sales and marketing strategies. As a result, one can reach the conclusion that re-identification prevention measures must be able to prevent re-identification through continuous monitoring of de-identified personal health information. Furthermore, we must carry out verification work regarding de-identified personal health information through authentification (ISMS) of medical institutions.

4. Conclusion

While the US, European Union, and Japan make a distinction between anonymization and de-identification in actual usage, there has not yet been a clear definition of de-identification in Korea.

Therefore, it should be anonymized, together with a clear definition of the area by establishing a standard medical term for de-identifying information about the Tuesday prior to the Personal Health Information Protection.

According to the findings of related research, the following have the HIPAA privacy rules as their basis in domestic and foreign personal information de-identification regulations: de-identification processing rules, step-by-step measures for de-identification, de-identification processing techniques and detailed application methods, cases of de-identification usage in Big Data processing according to field, evaluation procedures and detailed evaluation methods of the adequacy of de-identification, and re-identification threat management measures, *etc*. Therefore, it appears that applying Big Data to the

medical sector and health information de-identification will not create any major problems.

On the other hand, guidelines as well and based on HIPAA regulations to be prescribed the restrictions on big data use requires a plan prepared for the domestic situation, and constant monitoring for the re-identification prevention to conduct periodic verification is important. This paper proposed 4 step processes safe de-identification methods as they apply to the utilization of Big Data in personal health information based upon an analysis of de-identification trends both in and outside Korea.

In addition, adequacy verification of de-identification methods and continuous monitoring of any possible re-identification in accordance with the laws and guidelines are essential to the de-identification of personal health information. Thus, we propose that the establishment of a third-party professional organization that can play the role of a "control tower" that provides reliable and objective monitoring on a nationwide scale be actively considered.

Acknowledgments

This article is a revised and expanded version of a paper entitled “A Study on Personal Health Information De-identification Status for Big Data” presented at The 5th International Conference on Information Technology and Computer Science (ITCS 2016) held on July 5-8, 2016 at The Mercure Bali Harvestland Kuta Hotel, Bali, Indonesia.

References

[1] K. Gwan-Hyung, L. Joon-Yun and O. Am-Suk, “Fusion of Medical IT and Big Data”, Korea Computer and Information Society, vol. 21, no. 2, **(2013)**, pp. 17-26.
[2] K. Sang-Chun, “Personal information protection and usage, the issue of "de-identification" must be resolved first.”, Security News(www.boannews.com), **(2016)**.
[3] Y.-C. Chung, “De-identification Policy of Personal Information and Tasks on Healthcare Big Data”, Health and Welfare Forum, vol. 227, **(2015)**, pp. 50-60.
[4] Privacy Commission, “A study on the effects of personal information de-identification on personal information protection.”, **(2015)**.
[5] HHS OCR, “Guidance Regarding Methods for De-identifacation of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule”, **(2012)**.
[6] National Information Society Agency, “Casebook of personal information de-identification for the utilization of Big Data”, Ministry of Science, ICT and Future Planning, **(2014)**.
[7] Article 29 Data Protection Working Party, [Opinion 05/2014 on Anonymisation Techniques], http://ec.europa.eu/justice/data-protection/index_en.htm, **(2014)** April.
[8] H. Eun-Young, “Japan's revised personal information protection laws”, KISDI, **(2015)**.
[9] U.K ICO: Managing data protection risk code of practice.
[10] Korea Information and Communications Technology Association, Dictionary of IT Terminology.
[11] ISO/TS 25237:2008(en) Health informatics - Pseudonymization, http://www.iso.org
[12] ISO 27799:2008 Health informatics - Information security management in health using ISO/IEC 27002, http://www.iso27001security.com/html/27799.html
[13] HIPAA Privacy Rule, Code of Federal Regulations, 45CFR164.514.
[14] Ministry of Government Administration and Home Affairs, “Guidelines on personal information protection according to the opening up and sharing of public data”, **(2013)**.
[15] Korea Communications Commission, “Guidelines on Big Data personal information protection”, **(2014)**.
[16] National Information Society Agency, “Handbook for free study of the adequacy of personal information de-identification”, Ministry of Government Administration and Home Affairs, **(2014)**.
[17] National Information Society Agency, “Handbook of techniques on personal information de-identification for the utilization of Big Data”, Ministry of Science, ICT and Future Planning, **(2015)**.

Authors

Ya-Ri Lee, she received B.S. degree in Electronics & Computer Engineering from Korea University and M.S. and Ph.D. degrees in Computer Engineering from Dongguk University, Seoul Korea, in 1990, 1999 and 2002 respectively. She now works as a principle research engineer in the Personal Information Protection Center commissioned by Ministry of Health and Welfare, Korea, since June 2012. Her major research interests include privacy protection in health and welfare, security technology, data protection and security, and big data.

Young-Chul Chung, she received B.S. degree in Human Ecology from Yonsei University and M.S. degree in Public Health from Yonsei University, Seoul Korea, in 1984. She completed PhDs in Public Health from Catholic University and also in Management Information from KAIST College of Business, in 1996 and 2003 respectively. She now works as a research fellow in Big Data Research Center of KIHASA,, Korea, since June 2007. Her major research interests include privacy protection in health and welfare, health and welfare information system and evaluation, health information service, welfare information service and big data.

Jung-Sook Kim, she received B.S degrees in computer science from Kwangwoon University, Seoul Korea, in 1984. And She obtained M.S. and Ph.D. degrees in Computer Engineering from Dongguk University, Seoul Korea, in 1988 and 1999 respectively. She is currently a professor with the Division of Computer Engineering & Science at the Sahmyook University, Korea, since March 2001. Her major research interests include network security, web programming, embedded system and ubiquitous computing system.

Ho-Kyun Park, he received his B.S. and M.S. degrees in computer science from Kwangwoon University, Seoul Korea, in 1987 and 1989, respectively. He obtained his Ph.D. in multimedia network from same university, 1998. He is currently a professor with the School of IT Convergence Engineering at the Shinhan University, Korea, since March 1992. His major research interests include network security, smart home-network system, and ubiquitous computing system.