What are the optimum quasi-identifiers to re-identify medical records?

Yong Ju LEE*, Kyung Ho LEE*

*School of Information Security, Korea University, Korea

sky4uni@korea.ac.kr, kevinlee@korea.ac.kr

***Abstract*—Recently, medical records are shared to online for a purpose of medical research and expert opinion. There is a problem with sharing the medical records. If someone knows the subject of the record by using various methods, it can result in an invasion of the patient's privacy. To solve the problem, it is important to carefully address the tradeoff between data sharing and privacy. For this reason, de-identification techniques are applicable to address the problem. However, de-identified data has a risk of re-identification. There are two problems with using de-identification techniques. First, de-identification techniques may damage data utility although it may decrease a risk of re-identification. Second, de-identified data can be re-identified from inference using background knowledge. The objective of this paper is to analyze the probability of re-identification according to inferable quasi-identifiers. We analyzed factors, inferable quasi-identifiers, which can be inferred from background knowledge. Then, we estimated the probability of re-identification from taking advantage of the factors. As a result, we determined the effect of the re-identification according to the type and the range of inferable quasi-identifiers. This paper contributes to a decision on de-identification target and level for protecting patient's privacy through a comparative analysis of the probability of re-identification according to the type and the range of inference.**

***Keyword*—Privacy, Re-identification, De-identification, Medical records**

I. INTRODUCTION

Recently, medical information has been shared to online for many purposes. Especially, it is needed to share patient information for the purpose of medical research and expert opinion [1]. On the other hand, sharing these data may result in an invasion of patient's privacy such as disclosure of diagnostic information via re-identification of medical records.

De-identification techniques have been used to address the problem of privacy and data sharing. In addition, once data is gathered, the conflict arises from two aspects of data use and privacy. It is expected for de-identification techniques to solve the conflict.

However, de-identified data has risk of a re-identification risk, and several studies have proven that it is possible to re-identify data which was de-identified. In addition, although de-identification strengthens the protection of privacy, it could damage the data utility. It can be found the relationship between data utility and disclosure risk like below Figure 1. In Figure 1, X-axis represents the data utility and Y-axis represents the disclosure risk. The disclosure risk of original data is the highest. When the protection techniques like de-identification are applied, the risk will become increasingly lower, however, the data utility will become increasingly lower at the same time [2]. In other words, it is important to find the level with the maximum data utility without exceeding risk threshold.

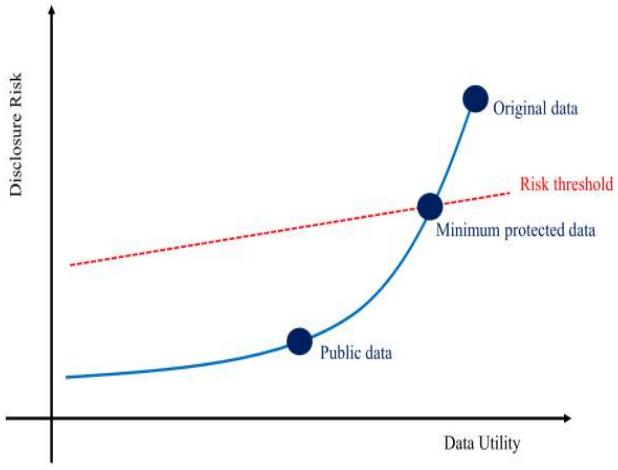

Fig. 1. Data Utility v.s Disclosure Risk

The purpose of this paper is to analyze factors affecting re-identification and to estimate probability of re-identification. From the result, we hope to decide proper de-identification level which can approve safety of personal information protection. To analyze the factors, we researched inference from background knowledge. Next, to estimate the probability of the re-identification, we used de-identified dataset provided in Statewide Planning and Research Cooperative System(SPARCS) of New York state Department of Health [3]. The result of this paper contributes to a decision on de-identification target and level for protecting patient's privacy through a comparative analysis of the probability of re-identification according to the type and the range of inference.

Manuscript received June 27, 2017. This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2017-2015-0-00403) supervised by the IITP(Institute for Information & communications Technology Promotion) .

Kyung Ho LEE is with School of Information Security, Korea University, Seoul, Korea (corresponding author to provide phone: +82-2-3290-4885; e-mail: kevinlee@korea.ac.kr).

Yong Ju LEE was with School of Information Security, Korea University Seoul, Korea (e-mail: sky4uni@korea.ac.kr).

This paper is organized as follows. In chapter 2, we describe related terms, guidelines, de-identification techniques, re-identification research, risk management. In chapter 3, we describe factors affecting re-identification, data set, and probability of a re-identification. In chapter 4, we describe the result of the re-identification simulation. Finally, in chapter 5, we describe meaning of the result, limits of this paper and future work for enhanced research.

II. RELATED WORKS

A. Research on related terms

We describe the definitions about personal information, de-identification, anonymization, and re-identification. First, we describe how to define personal information in the laws from each nation. In USA's Privacy Act, the act defines term 'record' as any item, collection, or grouping of information about an individual that is maintained by an agency, including, but not limited to, his education, financial transactions, medical history, and criminal or employment history and that contains his name, or the identifying number, symbol, or other identifying particular assigned to the individual, such as a finger or voice print or a photograph [4]. Also in Children's online privacy protection Act, the act defines term 'personal information' as individually identifiable information about an individual collected online, including (A) a first and last name, (B) a home or other physical address including street name and name of a city or town, (C) an e-mail address, (D) telephone number, (E) a Social Security number, (F) any other identifier that the Commission determines permits the physical or online contacting of a specific individual or (G) information concerning the child or the parents of that child that the website collects online from the child and combines with an identifier described in this paragraph [5].

In EU Data Protection Directive, the directive defines term 'personal data' as any information relating to an identified or identifiable natural person ('data subject'). And an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity [6]. Since then, General Data Protection Regulation(GDPR) which replaced EU Data Protection Directive appeared. According to the draft of GDPR published in 2012, it defined term 'personal data' as any information relating to a data subject [7]. In the draft, personal data was defined in a broad sense. Since then, according to final version of GDPR published in 2016, it defines term 'personal data' as any information relating to an identified or identifiable natural person ('data subject'). And an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person [8]. As seen in final version of GDPR, the notion of identification was included in defining personal data.

In Canada's Privacy Act, it defines term 'personal information' as information about an identifiable individual that is recorded in any form including, without restricting the generality of the foregoing. The Act divides the term into detailed 13 items such as information relating to the race, any identifying number, the address, the views of another individual about the individual, etc [9]. The scope of the term 'personal information' is more specifically described in Canada.

In Japan's Act on the Protection of Personal Information, it defines term 'personal information' as information about a living individual which can identify the specific individual by name, date of birth or other description contained in such information (including such information as will allow easy reference to other information and will thereby enable the identification of the specific individual) [10].

In Republic of Korea's Personal Information Protection Act, it defines term 'personal information' as information that pertains to a living person, including the full name, resident registration number, images, etc., by which the individual in question can be identified, (including information by which the individual in question cannot be identified but can be identified through simple combination with other information) [11].

So far we have discussed the definition of personal information that is described in the laws and regulations of each country. In most countries, when defining personal information, we know that their definition are based on whether it can identify the individual. On the contrary, if the criteria to identify the individual is ambiguous, there may be some confusion in defining personal information. In other words, the criteria deciding whether the individual can be identified is very important in defining personal information.

Next, we describe de-identification. In ISO/TS 25237:2008(E), it defines term 'de-identification' as general term for any process of removing the association between a set of identifying data and the data subject [12]. And, de-identification makes it hard to learn if the data in a data set is related to a specific individual, while preserving data utility [13].

In ISO/TS 25237:2008(E), it defines term 'anonymization' as process that removes the association between the identifying data set and the data subject [12].

In NISTIR 8053, it defines term 're-identification' as the process of attempting to discern the identities that have been removed from de-identified data [13]. In other words, re-identification occurs when breaking de-identification by identifying an individual who is the subject of the data [14]. Because an important goal of de-identification is to prevent re-identification, re-identification is sometimes called re-identification attack. Meanwhile, re-identification is attempted by various reasons such as testing the quality of the de-identification, gaining publicity or professional standing for performing the re-identification, etc [13]. The reasons are shown in Table 1 below.

TABLE I
THE REASONS FOR ATTEMPTING A RE-IDENTIFICATION

No	Reason
1	To test the quality of the de-identification
2	To gain publicity or professional standing for performing the re-identification
3	To embarrass or harm the organization that performed the de-identification
4	To gain direct benefit from the re-identified data
5	To cause problems such as embarrassment or harm to an individual whose sensitive information can be learned by re-identification

In ISO/TS 25237:2008(E), it explains that anonymization is another subcategory of de-identification. Because anonymization is process that removes the association between the identifying data set and the data subject, re-identification of anonymized data is not possible [12]. Therefore, this paper focuses not on anonymized data but on de-identified data.

B. De-identification guideline

In Australia's Privacy business resource 4 : De-identification of data and information, personal information is 'de-identified' if the information is no longer about an identifiable individual or an individual who is reasonably identifiable [15].

In GDPR, anonymous information is what does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. For this reason, the principles of data protection should not apply to anonymous information [8].

In UK's Anonymisation : managing data protection risk code of practice, it explains the meanings of 'not personal data' as what does not relate to and identify an individual [16].

In Republic of Korea's Personal information de-identification management guideline, de-identified information is personal information de-identified [17]. The Republic of Korea's guideline uses 'de-identification' term. In the guideline, information that is adequately de-identified is presumed to be not personal information since it can no longer identify a specific individual. In this regard, it is semantically explained as an idea of anonymization of the EU.

C. De-identification techniques

The most used method for de-identification are masking, generalization, suppression and adding random noise [18]. These methods can be described in a method of protecting statistical data. We describe the de-identification methods described above.

Masking refers to a set of direct identifier manipulation. In general, direct identifiers are removed or replaced with a random value or specific value from data set. There are redaction, randomization, and pseudonymization techniques in masking method. Redaction is a technique to remove a direct identifier from data set. Randomization is a technique to replace a direct identifier with a random value. pseudonymization is a technique to replace a direct identifier with a unique value [18].

Generalization is a set of anonymization techniques. It reduce an accuracy of data. For example, data '25 years' is generalized to '20-30 years'. In other words, generalization method is constructed from generalizing or diluting an attributes of data subjects by changing a size and scale [19]. There are hierarchy-based generalization and cluster-based generalization techniques in generalization method. Hierarchy-based generalization is based on a predefined hierarchical structure which describes that how much a reduction in an accuracy from quasi-identifier. Cluster-based generalization is based on a predefined utility policy [18].

Suppression means to delete a value of data. There are casewise deletion, quasi-identifier removal, and local cell suppression techniques in suppression method. casewise deletion is a technique to delete all records of a data set. Quasi-identifier removal is a technique to remove only quasi-identifiers of a data set. Local cell suppression, as compared to the above techniques, a more improved technique, is used to find a minimum number of quasi-identifiers required for suppression [18]. When using de-identification techniques, one important issue is data utility problem. In comparison with a casewise deletion and quasi-identifier removal techniques, local cell suppression techniques may be a better way to protect a data utility and reduce a risk of re-identification at the same time.

Adding random noise means to add noise. It may be primarily a de-identification techniques for sensitive items of personal information. This technique uses a method such as any number of addition and multiplication. Because it is added in a range of a specific mean and variance, it has special features that do not damage data utility of data set [20]. That is, it can be used as a method for solving both data utility and privacy problem.

Swapping means to replace database records with a set of predetermined variables [20]. Swapping method reduces a risk of re-identification by introducing an uncertainty of an actual data. Swapping method has an advantage, easy application, generally there is a drawback which does not hold a statistical characteristics [2].

Blank and impute means a method of filing a space portion by applying an alternative after selecting a small number of records from a micro data file, and replacing the selected filed with blank [20]. In other words, it fills blank which comes

TABLE II
THE DE-IDENTIFICATION GUIDELINES

Nation	AU	EU	UK	Republic of Korea
Guideline	Privacy business resource 4 : De-identification of data and information	General Data Protection Regulation (GDPR)	Anonymisation : managing data protection risk code of practice	Personal information de-identification management guideline
Term	de-identified information	anonymous information	not personal data	de-identified information
Definition of de-identified information	no longer about an identifiable individual or an individual who is reasonably identifiable	not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable	not relate to and identify an individual	personal information de-identified
Personal information	X	X	X	△ (presumed to be not personal information. If there is counterevidence, it is regarded as personal information) [17]

from removing original data with a calculated value by using an appropriate function(e.g., average, etc.) [2].

Blurring means a method to replace an average with a value of an item. For example, it is a typical method which replaces an average with a value of an item after classifying specific values into random groups [20].

D. Re-identification type

We confirmed the relationship between data utility and disclosure risk in Figure 1. If de-identification techniques are not sufficiently applied, data utility and disclosure risk will grow. Maybe, this is not public data but original data. On the contrary, data utility and disclosure risk may decline in public data. In other words, some de-identified data can result in a harm. When confidential information about individual such as diagnostic information is identified, disclosure happens. The types of the disclosure are identity disclosure, attribute disclosure, and inferential disclosure [21].

Identity disclosure happens when an attacker identify individual of specific data. The representative scenario which can result in identity disclosure is 're-identification by linking' [13, 21]. This is sometimes called linkage attack.

Attribute disclosure happens when confidential information about individual is identified and can be attributed to a data subject. It is similar to identity disclosure, and identity disclosure can sometimes result in attribute disclosure. However, attribute disclosure can happen without identity disclosure [13, 21].

Inferential disclosure occurs when information can be inferred with high confidence from statistical properties of the released data. For example, the data may show a high correlation between income and purchase price of a home. As the purchase price of a home is typically public information, a third party might use this information to infer the income of a data subject [22].

In this paper, we focus on inferential disclosure. Because it can result from background knowledge. This is sometimes called background knowledge attack.

E. Re-identification research

We describe studies on re-identification of medical records. According to the research by Latanya Sweeney, she collected the Group Insurance Commission(GIC) data and the voter registration list for Cambridge Massachusetts. Then, she executed the re-identification attack using linkage attack and identified the medical records about William Weld of the total 135,000 records [23]. In another research by Latanya Sweeney, she collected the Washington state de-identified medical records and the online news data. She obtained the information for a patient(e.g., gender, age, hospital, admission month, diagnostic information, address, etc.) from the collected data and identified the 35 records of the total 81 records via a linkage attack [24].

According to the research by Khaled El Emam and Patgricia Kosseim, they collected Pharmacy data from the Children's Hospital of Eastern Ontario. Then, they executed the re-identification attack using background knowledge and identified the 1 record of the total 3,510 records [25]. The research demonstrated that if they had a sufficient background knowledge about a particular individual, it is possible to re-identify medical records. In other words, strong background knowledge for a particular individual can increase the probability of re-identification.

According to the research by Grigorios Loukides, Joshua C Denny and Bradley Malin, they confirmed that it was possible to re-identify the de-identified medical records from the linkage attack based on a diagnosis code. They found that more than 96% of the total 2,762 records were uniquely identified from a diagnosis code and that de-identified medical records may be combined with DNA information via a re-identification attack [26].

Sean Hooley and Latanya Sweeney surveyed every state and the District of Columbia to find what state released about medical records and how much identifiable information were released. They found that 33 states released hospital discharge data. Also it differed from the level that protected the hospital discharge data for each state, and they found that most of 33 states did not meet the HIPAA criteria [27]. So if de-identification techniques are not sufficiently applicable, it may be vulnerable to re-identification attacks.

F. Re-identification risk management

Khaled El Emam and Bradley Malin developed de-identification process. The process consists of 11 steps [28]. The steps of the process are like below Table 3.

Step 1 : determine that which data fields are direct identifiers in data set.

Step 2 : masking methods are applied to the direct identifiers which have been determined in Step 1.

Step 3 : there is two activities. First, we can identify adversaries and what information they may be able to access. Second, we can determine quasi-identifiers in data set.

Step 4 : determine the minimal acceptable data utility.

Step 5 : determine acceptable re-identification risk. This is called re-identification risk threshold.

Step 6 : import data from the origin database.

Step 7 : evaluate the risk of re-identification.

Step 8 : compare the actual re-identification risk with the threshold determined in Step 5.

Step 9 : if the actual re-identification risk is higher than the threshold, it is necessary to apply additional de-identification techniques to data set.

Step 10 : if the actual re-identification risk is lower than the threshold, it is necessary to perform diagnostics on the solution.

Step 11 : export de-identified data to external data set. This is final data.

TABLE III
THE DE-IDENTIFICATION PROCESS

Step	Action
Step 1	Determine direct identifiers in the data set
Step 2	Mask (transform) direct identifiers
Step 3	Perform threat modeling
Step 4	Determine minimal acceptable data utility
Step 5	Determine the re-identification risk threshold
Step 6	Import (sample) data from the source database
Step 7	Evaluate the actual re-identification risk
Step 8	Compare the actual risk with the threshold
Step 9	Set parameters and apply data transformations
Step 10	Perform diagnostics on the solution
Step 11	Export transformed data to external data set

III. METHODS

A. Factors

Since the direct identifier is specific to an individual, the direct identifier is usually removed and quasi-identifier is usually processed through Generalization. It is difficult to re-identify the de-identified data by itself. The risk of re-identification will increase when additional data which can be linked is available [13]. Data sets should be linked for re-identification. To link the data sets, they should include common quasi-identifiers. Meanwhile, the term 'quasi-identifier', first introduced by Tore Dalenius, unlike direct identifiers, can not be identified by itself. However, it allows be linked to an individual who is the subject of the data [29]. That is, quasi-identifier is a variable that allows re-identification via a connection to an individual like below Figure 2.

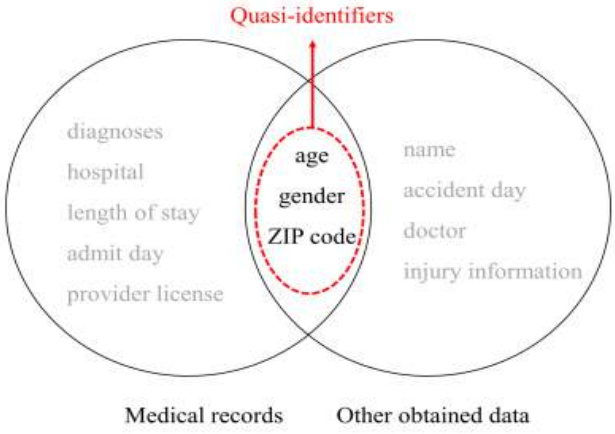

Fig. 2. Linkage attack based on Quasi-identifier

On the other hand, it is possible to infer quasi-identifiers from background knowledge. For example, if someone knew that the diagnosis of a particular patient was a prostate cancer, it can be inferred that the gender of the patient is male. Through the previous research and this study, we drew six factors which affected the probability of re-identification and could be inferred from background knowledge like below Table 4.

TABLE IV
ELEMENTS FOR INFERRING DATA FIELD

Data field	Background knowledge affecting in inferring data field	Study
Zip Code	accident information	[14]
	SNS information (profile, etc.)	[30]
	Geographic information	
Length of Stay	accident information	This study
	admit and discharge day information	
Admit day (of Week)	accident information	[14]
Discharge day (of Week)	discharge information	This study
Diagnosis*	injury information	[14]
	prescription information	
Provider license†	physical information	This study

* We replace name of the data field 'APR MDC Code' with 'Diagnosis' when defining factors.

† We replace name of the data field 'Attending Provider License Number' with 'Provider license' when defining factors.

In the previous research, it inferred admit information and diagnostic information by analyzing the information on an accident [24]. This is possible because online news can contain the information such as accident date and injury contents. In addition to them, in this research, we found that it was possible to infer length of stay, discharge day and provider license from background knowledge. Length of stay can be inferred by using the information about an admission of a patient from accident information. For example, if someone knew the admit day of a particular patient and the publication date of the online news which contained the information, the patient has been hospitalized, it is possible to infer length of stay(strictly, range of length of stay). Also, if someone knew the admit day and the discharge day of a particular patient, it is possible to infer a length of stay by calculating the difference between them. It is possible to infer a discharge day of a particular patient from a publication date of an online news or from a discharge day contained in online news. It is possible to infer a provider license of the physician in charge of treatment of a particular patient. This is possible because the name of the physician in charge of treatment of the patient could be contained in online news, and the provider license could be searched in online site such as profession official database. On the other hand, it is possible to infer Zip Code of a particular patient from accident information. Also, if someone knows SNS account of a particular patient, it is possible to infer Zip Code from account information of the patient. In addition, there is a research which could infer address information from map information [30], it is possible to confirm Zip Code from the address information.

B. Data set and Subject of simulation

To calculate the probability of re-identification, we used the Hospital Inpatient Discharges 2014 (Public Use data) provided in Statewide Planning and Research Cooperative System(SPARCS) of New York state Department of Health. This data set contained de-identified data and could not include protected health information(PHI) according to HIPAA. It included total 2,365,208 records of the hospitalized patients and total 39 fields like below Table 5.

Prior to calculating the probability of re-identification, it was necessary to select the subjects of re-identification simulation. Extraction procedure of the subjects is as follows. Step 1 : Select the most frequently existed 'Facility Name' for each 'APR MDC Code' in the data set. Step 2 : Select the most frequently existed 'Attending Provider License Number' for each 'APR MDC Code' and 'Facility Name' selected Step 1 in the data set. Step 3 : From result of Step 2, total 9,160 records were generated(Table 6).

The reasons for having the extraction procedure of the subjects are as follows. Reason 1 : To include the subjects with a variety of diagnosis. Reason 2 : If the frequency of provider license for each diagnosis was low, it is difficult to analyze how much a provider license impacts on the re-identification.

TABLE V
DATA SET FIELD NAME AND EXAMPLE OF MEDICAL RECORD

No	Field name	Example
1	Health Service Area	New York City
2	accident information	Manhattan
3	Operating Certificate Number	1234567
4	Facility Id	1234
5	Facility Name	New York Hospital
6	Age Group	30 to 49
7	Zip Code - 3 digits	112
8	Gender	F
9	Race	White
10	Ethnicity	Spanish/Hispanic
11	Length of Stay	34
12	Admit Day of Week	WED
13	Type of Admission	Emergency
14	Patient Disposition	Home or Self Care
15	Discharge Year	2014
16	Discharge Day of Week	TUE
…	…	…
23	APR MDC Code	22
24	APR MDC Description	Burns
…	…	…
32	Attending Provider License Number	123456
33	Operating Provider License Number	123456
…	…	…
39	Total Costs	$4,000

TABLE VI
THE SUBJECT OF RE-IDENTIFICATION SIMULATION

No	Facility Name	Age Group	Zip Code	Gender	Length of Stay	Admit Day of Week	Discharge Day of Week	APR MDC Code	Attending Provider License Number
1	New York Hospital	30 to 49	103	F	34	WED	TUE	22	123456
…	…	…	…	…	…	…	…	…	…
9160	…	…	…	…	…	…	…	…	…

C. *Probability of re-identification*

We introduce the formula for measuring the probability of re-identification [31]-[32]. Next, we interpreted the meaning of the probability of re-identification.

$$\theta_j = \frac{1}{f_j}$$

θ refers to the probability of re-identification. f refers to the size of equivalence class. j refers to the number of equivalence class in data set. When f, the size of equivalence class, is minimum value, θ, the probability of re-identification, will be maximum value. Here it is important to find the value of j which makes f be minimized.

Next, we generated the total 64 of the combination of the six factors that affect the re-identification. In other words, j was 1, 2, …, 64. Then, the value of f could be calculated according to the value of each of j. Finally, we calculated the value of θ, the probability of re-identification, which is the inverse of f.

According to the research by Khaled El Emam and Bradley Malin, they introduced 'minimum cell size' concept in determining the threshold of re-identification like below Table 7 [28]. Cell size means the number of response corresponding to a particular condition in data set [20]. Therefore, cell size is seen to have the same concept as the equivalence class. On the other hand, k-anonymity, as one of the privacy protection models, is used to determine whether the propriety of de-identification measures is appropriate in Republic of Korea's Personal information de-identification management guideline. For example, the propriety of de-identification measures is presumed appropriate if the value of k is five for k-anonymity [17].

TABLE VII
MEANING OF IDENTIFIABLE RECORD EACH CELL SIZE

Cell size (Probability)	< 3 (> 0.33)	3 (0.33)	5 (0.2)	11 (0.09)	20 (0.05)
Meaning	Identifiable data	Highly trusted data disclosure	-	-	Highly untrusted data disclosure

In this paper, we had two assumptions for the simulation. First, it was assumed that when the size of equivalence class was 3 or less, the data was identifiable data. Second, it was assumed that when estimating the probability of re-identification, patient information about 'Facility Name', 'Age Group', and 'Gender' was known. They can be sufficiently collected from information such as online news [24], they were excluded from inferable quasi-identifier group we extracted.

IV. RESULTS

We estimated the probability of re-identification by using both prepared data set and previously extracted subject of re-identification simulation. The result of the simulation is shown in Table 8 below. The table shows the probability of re-identification according to the combinations of inferable quasi-identifiers.

Based on the results, if the number of inferable quasi-identifiers was 1, the quasi-identifier which was the most effective factor for re-identification was 'length of stay'. If the number was 2, we knew that the combination of 'length of stay and provider license' was the highest. If the number was 3, the most effective combination was 'length of stay and discharge day and provider license'. If the number was 4 or 5, we knew that the most effective combination included patient's 'zip code'. The most effective combination according to the number of the inferable quasi-identifiers is shown in Table 9 below.

This allows us to know which combination of quasi-identifiers is the most affecting re-identification of medical records. In other words, it helps us decide which quasi-identifier we must de-identify to decrease the probability of re-identification using inference attack through background knowledge.

TABLE VIII
THE PROBABILITY OF RE-IDENTIFICATION AS QUASI-IDENTIFIER COMBINATIONS

No	Combination	Probability (Number of re-identification)	No	Combination	Probability (Number of re-identification)
1	-	0% (0/9160)	33	Zip Code	0.41% (38/9160)
2	Provider license	0.28% (26/9160)	34	Zip Code & Provider license	7.89% (723/9160)
3	Diagnosis	0.01% (1/9160)	35	Zip Code & Diagnosis	2.67% (245/9160)
4	Diagnosis & Provider license	0.67% (61/9160)	36	Zip Code & Diagnosis & Provider license	9.81% (899/9160)
5	Discharge day	0% (0/9160)	37	Zip Code & Discharge day	2.22% (203/9160)
6	Discharge day & Provider license	4.44% (407/9160)	38	Zip Code & Discharge day & Provider license	23.36% (2140/9160)
7	Discharge day & Diagnosis	0.57% (52/9160)	39	Zip Code & Discharge day & Diagnosis	12.47% (1142/9160)
8	Discharge day & Diagnosis & Provider license	7.18% (658/9160)	40	Zip Code & Discharge day & Diagnosis & Provider license	25.67% (2351/9160)
9	Admit day	0% (0/9160)	41	Zip Code & Admit day	2.15% (197/9160)
10	Admit day & Provider license	4.04% (370/9160)	42	Zip Code & Admit day & Provider license	21.53% (1972/9160)
11	Admit day & Diagnosis	0.43% (39/9160)	43	Zip Code & Admit day & Diagnosis	12.05% (1104/9160)
12	Admit day & Diagnosis & Provider license	7.13% (653/9160)	44	Zip Code & Admit day & Diagnosis & Provider license	24.08% (2206/9160)
13	Admit day & Discharge day	0.04% (4/9160)	45	Zip Code & Admit day & Discharge day	9.9% (907/9160)
14	Admit day & Discharge day & Provider license	21.74% (1991/9160)	46	Zip Code & Admit day & Discharge day & Provider license	42.87% (3927/9160)
15	Admit day & Discharge day & Diagnosis	6.84% (627/9160)	47	Zip Code & Admit day & Discharge day & Diagnosis	30.72% (2814/9160)
16	Admit day & Discharge day & Diagnosis & Provider license	25.85% (2368/9160)	48	Zip Code & Admit day & Discharge day & Diagnosis & Provider license	45.94% (4208/9160)
17	Length of Stay	0.71% (65/9160)	49	Zip Code & Length of Stay	5.28% (484/9160)
18	Length of Stay & Provider license	13.1% (1200/9160)	50	Zip Code & Length of Stay & Provider license	28.48% (2609/9160)
19	Length of Stay & Diagnosis	4.67% (428/9160)	51	Zip Code & Length of Stay & Diagnosis	16.89% (1547/9160)
20	Length of Stay & Diagnosis & Provider license	16.1% (1475/9160)	52	Zip Code & Length of Stay & Diagnosis & Provider license	31.12% (2851/9160)
21	Length of Stay & Discharge day	3.46% (317/9160)	53	Zip Code & Length of Stay & Discharge day	16.98% (1555/9160)
22	Length of Stay & Discharge day & Provider license	30.72% (2814/9160)	54	Zip Code & Length of Stay & Discharge day & Provider license	49.16% (4503/9160)
23	Length of Stay & Discharge day & Diagnosis	16.46% (1508/9160)	55	Zip Code & Length of Stay & Discharge day & Diagnosis	39.08% (3580/9160)
24	Length of Stay & Discharge day & Diagnosis & Provider license	33.56% (3074/9160)	56	Zip Code & Length of Stay & Discharge day & Diagnosis & Provider license	51.31% (4700/9160)
25	Length of Stay & Admit day	3.48% (319/9160)	57	Zip Code & Length of Stay & Admit day	17.03% (1560/9160)
26	Length of Stay & Admit day & Provider license	30.53% (2797/9160)	58	Zip Code & Length of Stay & Admit day & Provider license	49.04% (4492/9160)
27	Length of Stay & Admit day & Diagnosis	16.44% (1506/9160)	59	Zip Code & Length of Stay & Admit day & Diagnosis	38.89% (3562/9160)
28	Length of Stay & Admit day & Diagnosis & Provider license	33.55% (3073/9160)	60	Zip Code & Length of Stay & Admit day & Diagnosis & Provider license	51.33% (4702/9160)
29	Length of Stay & Admit day & Discharge day	3.59% (329/9160)	61	Zip Code & Length of Stay & Admit day & Discharge day	17.64% (1616/9160)
30	Length of Stay & Admit day & Discharge day & Provider license	31.36% (2873/9160)	62	Zip Code & Length of Stay & Admit day & Discharge day & Provider license	49.84% (4565/9160)
31	Length of Stay & Admit day & Discharge day & Diagnosis	17.15% (1571/9160)	63	Zip Code & Length of Stay & Admit day & Discharge day & Diagnosis	39.93% (3658/9160)
32	Length of Stay & Admit day & Discharge day & Diagnosis & Provider license	34.24% (3136/9160)	64	Zip Code & Length of Stay & Admit day & Discharge day & Diagnosis & Provider license	52.06% (4769/9160)

TABLE IX
THE MOST EFFECTIVE COMBINATION ACCORDING TO THE NUMBER OF THE QUASI-IDENTIFIERS

Number of the inferable quasi-identifiers	Combination
1	Length of Stay
2	Length of Stay & Provider license
3	Length of Stay & Discharge day & Provider license
4	Zip Code & Length of Stay & Discharge day & Provider license
5	Zip Code & Length of Stay & Admit day & Diagnosis & Provider license

V. CONCLUSION

In this paper, before analyzing solutions to problems related to de-identification, which were data utility and re-identification, we derived optimum quasi-identifiers which have the greatest impact on re-identification of medical records. We analyzed the factors affecting re-identification and estimated the probability of re-identification based on extracted factors by using a de-identified data set. The factors were 'Zip Code', 'Length of Stay', 'Admit day', 'Discharge day', 'Diagnosis', and 'Provider license'. Especially, compared with the previous paper, we added 'Zip Code' factor affecting re-identification. As a result, we found 'Zip Code' factor had a greater impact on re-identification than the other factors when the number of inferable quasi-identifiers was more than four.

We simulated the re-identification of medical records by using the Hospital Inpatient Discharges 2014 (Public Use data) provided in SPARCS of New York state Department of Health. From the results of the simulation, we found that the probability of re-identification was depending on the type of inferable quasi-identifier. In other words, the probability of re-identification would be either higher or lower according to the type of quasi-identifier inferred. This allows us to find the optimum quasi-identifiers for re-identification of medical records. But at the same time, this shows what we prevent from being inferred to decrease the probability of re-identification. Although it is hard to completely block the inference of information related to patient, it will be possible to decrease the probability of re-identification by means such as increasing de-identification level.

On the other hand, we describe two limitations of this paper. First, the number of inferable quasi-identifiers for re-identification of medical records may be more than six presented in this paper. Second, although this paper shows that the probability of inference is either 0 or 1, the probability of inference may actually be various. For example, the probability of inference may have various value because of variables such as the amount of collected background knowledge, the characteristics of the quasi-identifier, etc.

In order to overcome the limitations presented above, the research on extending the range of inferable quasi-identifiers and estimating the probability of inference should be done in future works.

ACKNOWLEDGMENT

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2017-2015-0-00403) supervised by the IITP(Institute for Information & communications Technology Promotion). This paper is an extension of the Yong Ju LEE's Master's thesis and ICACT2017 conference proceeding. Thanks to the Master's thesis reviewers, who were Hun Yeong KWON, In Seok KIM and Kyung Ho LEE from School of Information Security, Korea University and the ICACT2017 reviewers.

REFERENCES

[1] H Taneja, AK Singh. "Preserving Privacy of Patients Based on Re-identification Risk," *Procedia Computer Science* 2015; 70: 448-454.
[2] V Ciriani, SDC Di Vimercati, S Foresti, P Samarati. "Microdata protection," In: *Secure data management in decentralized systems* 2007; 33: 291-321.
[3] New York State Department of Health. Hospital Inpatient Discharges (SPARCS De-Identified): 2014. [Online]. Available: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/rmwa-zns4.
[4] 5 U.S.C. §552a.
[5] 15 U.S.C. §6501.
[6] Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
[7] European Commission. "Proposal for a Regulation of the european parliament and of the council, on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012PC0011&from=EN.
[8] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
[9] R.S.C., 1985, c. P-21 Privacy Act.
[10] Act on the Protection of Personal Information.
[11] Personal Information Protection Act.
[12] International Organization for Standardization. ISO/TS 25237:2008(E) Health Informatics — Pseudonymization.
[13] National Institute of Standards and Technology. "NISTIR 8053 De-Identification of Personal Information," [Online]. Available: http://dx.doi.org/10.6028/NIST.IR.8053.
[14] L Sweeney. "Only You, Your Doctor, and Many Others May Know," *Technology Science* 2015; 2015092903.
[15] Office of the Australian Information Commissioner. Privacy business resource 4: De-identification of data and information.
[16] Information Commissioner's Office. Anonymisation : managing data protection risk code of practice.
[17] Office for Government Policy Coordination. Personal information de-identification management guideline.
[18] K El Emam. "Methods for the de-identification of electronic health records for genomic research," *Genome medicine* 2011; 3.4: 1.
[19] ARTICLE 29 DATA PROTECTION WORKING PARTY. "Opinion 05/2014 on Anonymisation Techniques," [Online]. Available: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.
[20] Park WH, HWANG JY. "Disclosure Limitation Techniques for Statistical Tables and Microdata," *Journal of The Korean Official Statistics* 2004; 9.2: 146-172.
[21] L. Xiong, J. Gardner, P. Jurczyk, J. J. Lu. "Privacy-Preserving Information Discovery on EHRs," in *Information Discovery on Electronic Health Records*. CRC Press 2009.
[22] Glossary of statistical terms, OECD. [Online]. Available : https://stats.oecd.org/glossary/detail.asp?ID=6932 .
[23] L Sweeney. "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2002; 10.05: 557-570.
[24] L Sweeney. "Matching known patients to health records in Washington State data," *Available at SSRN 2289850* 2013.

[25] K El Emam, P Kosseim. "Privacy interests in prescription data, part 2: patient privacy," *IEEE Security & Privacy* 2009; 7.2: 75-78.
[26] G Loukides, JC Denny, B Malin. "The disclosure of diagnosis codes can breach research participants' privacy," *Journal of the American Medical Informatics Association* 2010; 17.3: 322-327.
[27] S Hooley, L Sweeney. "Survey of Publicly Available State Health Databases," *Available at SSRN 2277688* 2013.
[28] K El Emam, B Malin. "CONCEPTS AND METHODS FOR DE-IDENTIFYING CLINICAL TRIAL DATA," *Paper commissioned by the Committee on Strategies for Responsible Sharing of Clinical Trial Data* 2014.
[29] T Daleniusl. "Finding a needle in a haystack," *Journal of official statistics* 1986; 2.3: 329-336.
[30] Brownstein, John S., Christopher A. Cassa, and Kenneth D. Mandl. "No place to hide—reverse identification of patients from published maps," *New England Journal of Medicine* 2006; 355.16: 1741-1742.
[31] K El Emam, FK Dankar, R Vaillancourt, T Roffey, M Lysyk. "Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records," *The Canadian journal of hospital pharmacy* 2009; 62.4: 307-319.
[32] K El Emam. "Guide to the de-identification of personal health information," *CRC Press* 2013.

Yong Ju LEE , was born in Republic of Korea, October 7, 1989. Yong Ju Lee earned Master's degree from School of Information Security at Korea University. His main research interests include risk management, privacy policy, de-identification and re-identification of personal information.

Kyung Ho LEE , was born in Republic of Korea, September 9, 1967. Kyung Ho Lee earned his Ph.D. degree from Korea University. He is now a Professor in School of Information Security at Korea University, and leading the Risk management Laboratory in Korea University since 2011. He was the former CISO in Naver corporation and CEO of Secubase corporation. His main research interests include information security management system(ISMS), risk management, information security consulting, privacy policy, and privacy impact assessment(PIA).