Journal of Technology and Social Science (JTSS)

# Risk Impacts of Patients' Information and Extensions of $l$-Diversity

Osamu Takaki[1,a], Takayuki Asao[2,b] and Yoichi Seki[3,c]

[1]Faculty of Social and Information Studies, Gunma University, 4-2 Aramaki-machi, Maebashi, 371-8510, Japan

[2]Big Data Center for Integrative Analysis, Gunma University, 3-39-22 Showa-machi, Maebashi, 371-8511, Japan

[3]School of Science and Technology, Gunma University, 1-5-1 Tenjin-cho, Kiryu, 376-8515, Japan

[a]takaki@gunma-u.ac.jp, [b]asao@gunma-u.ac.jp, [c]seki@cs.gunma-u.ac.jp

**Keywords:** *k*-anonymity, *l*-diversity, anonymization, medical data, ontology

**Abstract.** This paper describes the construction of an ontology to clarify and quantify the severity of health information recorded in medical databases. The proposed approach extends the definition of $l$-diversity to reduce the risk of an attacker identifying a patient's health status based on medical data. The extended $l$-diversity is defined in accordance with the proposed ontology model. It is shown that the extended $l$-diversity satisfies the monotonicity property, based on which a process for anonymizing medical data is presented. Accordingly, the data satisfy the extended $l$-diversity and retain their usability to the greatest extent possible. In addition, the severity number of information regarding patients who visited or stayed in hospitals in Japan is estimated using our ontology model.

## 1. Introduction

Medical data contain sensitive patient information. Therefore, in many hospitals, when data administrators are requested to provide medical data to users for secondary applications such as data-based research, the administrators should carefully determine the scope and detail of the provided data based on the purpose and authority of the user. Moreover, data administrators must decide the anonymization level of the medical data. However, a trade-off exists between the anonymization level of medical data and their usability. Thus, it is desirable to anonymize the medical data to ensure that they satisfy hospital guidelines and remain as usable as possible.

The $k$-anonymity [1] and $l$-diversity [2] metrics, where $k$ and $l$ denote natural numbers, are useful for determining the data anonymization level. Let us consider data or a dataset to be a table $T$ in a relational database. A set of attributes in $T$ that can be linked with other tables to identify individuals is called a *quasi-identifier* in $T$. An attribute in $T$, the value of which adversaries should be prevented from discovering, is called a *sensitive attribute*. For example, Table 1 contains a quasi-identifier that consists of "age," "zip code," and "occupation," as well as a sensitive attribute, "disease." $T$ is said to satisfy $k$*-anonymity* if every record in $T$ is indistinguishable from at least $k$ records with respect to every quasi-identifier in $T$. For example, Table 1 has five records and satisfies 2-anonymity. (For simplicity, every sensitive attribute is considered to be separate from the quasi-identifier.) The $k$-anonymity of $T$ ensures that adversaries are prevented from uniquely linking an individual to a record in $T$ via a quasi-identifier.

However, for an adversary $A$ and a target individual $B$, even if $A$ cannot uniquely link $B$ to any record, $A$ might link $B$ to the value of a sensitive attribute in $T$. For example, in Table 1, even if $A$ cannot distinguish the records with ID 101 or 102, $A$ will realize that $B$ has Parkinson's disease if $A$ can link $B$ to the two above records. We call this problem a homogeneity attack [2].

To prevent homogeneity attacks, the concept of $l$-diversity has been proposed. Table $T$ is said to satisfy $l$*-diversity* if every set of records that shares the same values of attributes in a quasi-identifier

Table 1. Example table with a quasi-identifier consisting of “age,” “zip code,” and “occupation,” and a sensitive attribute “disease”

| ID (Dummy) | Age | Zip code | Occupation | Disease |
|---|---|---|---|---|
| 101 | 50 | 371-8510 | Professor | Parkinson’s disease |
| 102 | 50 | 371-8510 | Professor | Parkinson’s disease |
| 103 | 25 | 376-8515 | Nurse | Bronchitis |
| 104 | 25 | 376-8515 | Nurse | Gastric ulcer |
| 105 | 25 | 376-8515 | Nurse | Flu |

has at least $l$ different values of every sensitive attribute. For example, Table 1 will satisfy 2-diversity if the value “Parkinson’s disease” of the sensitive attribute in record 101 is replaced by “myasthenia gravis.” However, in the case of medical data, where the range of severities of the information described by a sensitive attribute is very broad, homogeneity attacks may not be completely prevented by the use of $l$-diversity alone. For example, from Table 1, replacing “Parkinson’s disease” in record 101 by “myasthenia gravis” would allow attacker $A$ to conclude that, although the disease of $B$ cannot be uniquely identified from Table 1, $B$ requires serious and specialized medical treatment before they can fully return to society.

The main purpose of this study is to extend the concept of $l$-diversity to reduce the risk of attackers inferring a patient’s severe medical problems based on medical data. The extended versions of $l$-diversity are defined based on our proposed ontology model—the Risk-Impact Ontology for Patients’ Sensitive Information (RIOPSI)—which quantifies the severity of a patient’s health information. In this paper, it is shown that the extended $l$-diversity satisfies the monotonicity property. This enables the development of a process for anonymizing medical data, with the resulting dataset satisfying the extended $l$-diversity while maintaining its usability to the greatest possible extent. Additionally, we estimate the severity number of information about patients who visited or stayed in hospitals in Japan in accordance with our ontology model and statistical data published by the Ministry of Health, Labour and Welfare and the National Cancer Center in Japan. Based on the estimation results, we compare the risk impacts of patient information recorded in large hospitals and medical clinics. This paper is an extended version of a conference submission [3], including additional discussions of estimation of severity of patients’ information in Japan (Section 5) and related works (Section 6), and a thorough description of the concept and computation of the severity number.

The remainder of this paper is organized as follows. Section 2 presents the proposed ontology model and a method to clarify and quantify the severity of patients’ sensitive information in databases. Section 3 describes how this method is used to extend the concept of $l$-diversity based on the ontology model. Section 4 explains how to generalize medical data using the extended $l$-diversity. Section 5 estimates the severity number of information about patients in hospitals in Japan. Section 6 discusses related works, and Section 7 summarizes the results of this study.

## 2. Modeling of Patient Medical Condition Severity

This section explains the concepts in RIOPSI that allow the severity of a patient’s health condition to be defined. An approach to quantifying the severity of the information based on RIOPSI is also described.

### 2.1 Risk-Impact Ontology for Patients’ Sensitive Information

RIOPSI, denoted by $\mathbb{O}$, mainly consists of (i) an attacker’s objectives and (ii) a patient’s health conditions and severity criteria. The severity criteria are represented as sets of especially severe patient conditions that are classified by types of medical conditions. The main concepts in $\mathbb{O}$ are outlined below.

Table 2. Classification of sensitive information, related attributes, and their special values

| Large Class of Sensitive information | Middle Class of Sensitive information | Small Class of Sensitive information | Related Attributes (Underlined Part) and Sets of Special Values of the Attributes |
|---|---|---|---|
| 1. information that inflicts significant damage on patients directly only by being known by others. | 1.1. information that inflicts significant damage not only on patients but also on their offsprings | 1.1.1. information about diseases of genes | For the attribute "disease," we consider the set of values that are assigned as diseases of genes by the administration of health. |
| | 1.2. information that inflicts significant disadvantages on patients institutionally | 1.2.1. information about diseases assigned as intractable diseases | For "disease," we consider the set of values that are assigned as intractable diseases by the administration of health. |
| | 1.3. information that inflicts significant damage on patients' human rights | 1.3.1. information about significant psychiatric disorders | For "disease," we consider the set of values that specialists such as doctors of psychiatric diseases select as significant psychiatric disorders. |
| 2. Otherwise than the above | 2.1. Information about patients' life spans | 2.1.1. information about diseases that have a major impact on patients' life spans | For "disease," we consider the set of values whose survival rate (for example, 5-year survival rates) are low (for example, less than 30%). |
| | | 2.1.2. information about patients' outcomes | For the pair of "disease" and its "medical treatment" including operations, we consider the set of pairs of values whose survival rates after the medical treatments are low. |
| | 2.2. Information about patients' quality of life (QOL) | 2.2.1 and 2.2.2. information about financial burdens of patients | For "disease," we consider the set of values that the institution of expensive medical treatments can be applied to. (In the case of Japan, one can define 2-stage sets according to the criteria defined by the health ministry.) |
| | | | For "treatment" or "medicines," we consider the set of values that are assigned as expensive medical treatments by the administration of health. |
| | | 2.2.3, 2.2.4 and 2.2.5. information about degrees of disability to patients' social and/or personal life | For "disease," we consider the set of values that specialists select as diseases by which patients have difficulty to get back into society. |
| | | | For "disease," we consider the set of values that specialists select as diseases by which patients have difficulty to have their daily life. |
| | | | For attributes about durations of hospital stays, we consider the set of (tuples of) values that indicate that the patients are hospitalized for long periods. (For example, one can define 3-stage sets by more than 60 days-, more than 120 days- and more than 180 days-hospitalization.) |
| | 2.3. information that inflict damage on patients' human rights (other than the middle class 1.3) | 2.3.1. information by which others might have prejudice toward patients' life styles | For "disease," we consider the set of values that specialists select as diseases by which others might have prejudice toward patients' life styles. |
| | | 2.3.2. information by which others might have prejudice toward patients' appearances | For "disease," we consider the set of values that specialists select as diseases by which others might have prejudice toward patients' appearances. |

According to psychological studies on cyberattacks [4, 5] and insider threats [6, 7], attacker motivations can be classified as follows: (1) Emotional motivation: (1.1) pleasure, (1.2) curiosity, (1.3) revenge, (1.4) revelation, and (1.5) destruction. (2) Commercial motivation: (2.1) data sales, (2.2) business operations, and (2.3) intimidation. Let us apply the motivations above to the case of a homogeneity attack.

The adversary's main objectives in accessing patient conditions are classified into the following problems.

I. Problems that directly inflict major damage on patients only by being known by others.
II. Problems that patients are eager to solve, even if the cost is immense.

We divide these two problem types into several categories and consider the relevant attributes in the tables stored in most hospital databases. Moreover, we consider the severity criteria of the categorized conditions by defining sets of especially severe information as values of the related attributes. The results are presented in Table 2.

Each set of special values, defined in the right-most column of Table 2, is called a *special values set*.

**Remark 1.** The survival rate, such as the five-year survival rate, is determined based on not only the disease, but also the degree of progression. However, for simplicity, we regard the attribute "disease" as a disease with its degree of progression.

**Remark 2.** In the following, we regard every pair (and every tuple) of sensitive attributes in the small class (2.1.2) (and (2.2.5), respectively) as a single sensitive attribute. Moreover, for a pair $(d, t)$ of a disease $d$ and a medical treatment $t$, which is defined in the small class (2.1.2), if $d$ is contained in a special values set $S$, then we regard $(d, t)$ as being contained in $S$.

### 2.2 Quantification of Patient Condition Severity Based on $\mathbb{O}$

In this section, we quantify the severity of the values and records in medical database tables from the viewpoint of potential adversaries. In Definition 1, we define the "severity number" $\mathrm{s}(v)$ of a value $v$ to satisfy the following principles.

i. For values $v$ and $u$, if $v$ is contained in a special values set in the first large class of sensitive information, but $u$ is not, then $\mathrm{s}(u) < \mathrm{s}(v)$.
ii. For values $v$ and $u$ in the first large class, if $\mathbb{G}_u \subsetneq \mathbb{G}_v$, then $\mathrm{s}(u) < \mathrm{s}(v)$, where $\mathbb{G}_u$ and $\mathbb{G}_v$ are the special values sets in the first large class that contains $u$ and $v$, respectively.
iii. For values $v$ and $u$ that are not in the first large class, but are in the second large class, if $v$ and $u$ are contained in special values sets in different middle classes, then the severities of $v$ and $u$ are not comparable.
iv. However, if the severities of $v$ and $u$ are compared from the specified viewpoint represented by the middle class (2.1), (2.2), or (2.3) in Table 2, and if $\mathbb{G}_u \subsetneq \mathbb{G}_v$, then $\mathrm{s}(u) < \mathrm{s}(v)$, where $\mathbb{G}_u$ and $\mathbb{G}_v$ are the special values sets in the same middle class (2.1), (2.2), or (2.3) and $\mathbb{G}_u$ and $\mathbb{G}_v$ contain $u$ and $v$, respectively.

We believe that principles (i) and (ii) are reasonable. Note that the severity of a patient's information from an adversary's viewpoint differs from the patient's severity itself. We also consider principles (iii) and (iv), because it is not easy to compare the severities of patients' information if they are considered from different viewpoints that are represented by the middle classes (2.1), (2.2), and (2.3) in Table 2. Therefore, we define the severity number of $v$ according to not only the large class type, but also the middle class type.

**Definition 1.** (1) Let $m$ be the number of the middle classes with item number $(2.m)$ in Table 2, which we call the *middle class type*. Then, we define the *severity number* $s_m(v)$ of value $v$ with respect to $m$ as follows.

1. If $v$ is contained in a special values set in the first large class, then $s_m(v)$ is defined independently of $m$ to be $n+N$. Here, $n$ denotes the number of special values sets in the first large class that contains $v$, and $N$ denotes the maximum number in the set of all severity numbers satisfying condition (2) below.
2. If $v$ is not contained in any special values sets in the first large class, then $s_m(v)=n$. Here, $n$ denotes the number of special values sets in the middle class with type $m$ that contain $v$.

(2) For record λ in a table and middle class type $m$, $s_m(\lambda)$ is the maximum value of $s_m(v)$ for all values $v$ in λ.

## 3. $\mathbb{O}$-Based $\boldsymbol{l}$-Diversity

In this section, we extend the concept of $l$-diversity based on $\mathbb{O}$ and the severity numbers defined in Section 2. Let $T$ be a table, λ be a record in $T$, π be a quasi-identifier in $T$, and $q$ be a tuple of values of π in λ. Then, the set of records in $T$ that share $q$ as the tuple of the values of π is called the π-*block* of $q$.

**Definition 2.** Let $l$ be a natural number and $m$ be a middle class type. Then, a π-block Λ is said to satisfy $\mathbb{O}(m)$-*l-diversity* if, for every sensitive attribute σ, there exist at least *l* values $v_1, \dots, v_l$ of $\sigma$ in Λ with severity numbers $s_m(v_1), \dots, s_m(v_l)$ that differ from each other. Moreover, Λ is said to satisfy $\mathbb{O}$-*l-diversity* if Λ satisfies $\mathbb{O}(m)$-*l*-diversity for all types $m$ of middle classes. Furthermore, $T$ is said to satisfy $\mathbb{O}(m)$-*l-diversity* (or $\mathbb{O}$-*l-diversity*) if all blocks in $T$ satisfy $\mathbb{O}(m)$-*l*-diversity (or $\mathbb{O}$-*l*-diversity, respectively).

If table $T$ satisfies $\mathbb{O}(m)$-*l*-diversity for some type $m$, then it clearly satisfies *l*-diversity. Moreover, one can easily extend other *l*-diversity types defined in [2] by replacing the values with the corresponding severity numbers.

For most values, the severity number will be very low. Thus, the conditions of $\mathbb{O}(m)$-*l*-diversity and $\mathbb{O}$-*l*-diversity in Definition 2 may be too strict. If it is not necessary to focus on records with a low severity number, Definition 3 would be more useful.

**Definition 3.** Let $l$ be a natural number and $m$ be a middle class type. Then, π-block Λ is said to satisfy *downward*-$\mathbb{O}(m)$-*l-diversity* if there exists a record λ in Λ that satisfies $s_m(\lambda) \leq N - l$, where $N$ is the maximum severity number. Moreover, Λ is said to satisfy *downward*-$\mathbb{O}$-*l-diversity* if Λ satisfies downward-$\mathbb{O}(m)$-*l*-diversity for all types $m$ of middle classes. Furthermore, $T$ is said to satisfy *downward*-$\mathbb{O}(m)$-*l-diversity* (resp. *downward*-$\mathbb{O}$-*l-diversity*) if all blocks Λ in $T$ satisfy downward-$\mathbb{O}(m)$-*l*-diversity (resp. downward-$\mathbb{O}$-*l*-diversity).

Note that $N$ in Definition 3 is determined independently of table $T$ and that the downward version of $\mathbb{O}(m)$-*l*-diversity has no logical strength relationship with *l*-diversity. Actually, if $T$ has no value with a severity number greater than $N-l$, then $T$ automatically satisfies *downward*-$\mathbb{O}$-*l-diversity*.

## 4. Generalization of Medical Data Based on $\mathbb{O}$-Based $\boldsymbol{l}$-Diversity

In this section, we outline a process to anonymize medical data based on the extended $l$-diversities defined in Definitions 2 and 3, which we call $\mathbb{O}$-*based* $l$-*diversities*. To this end, we first show the monotonicity of these $\mathbb{O}$-based $l$-diversities. For simplicity, we consider that all quasi-identifiers in table $T$ are integrated and that $T$ has only one quasi-identifier.

Let $T$ be a table that consists of the (integrated) quasi-identifier π and a sensitive attribute σ. In addition, λ denotes a record in $T$, and $q$ represents a tuple of values of π in λ. Moreover, let $Q$

be the domain of $\pi$, and let $Q^*$ be a set $\{Q_1, \dots, Q_n\}$ of subsets of $Q$ that satisfies $\mathrm{Q} = Q_1 \cup \dots \cup Q_n$ and $Q_i \cap Q_j = \emptyset$ for each $i, j \leq n$ with $i \neq j$. Furthermore, $T^*$ denotes a table that consists of the quasi-identifier $\pi^*$ and the sensitive attribute $\sigma$, where $\pi^*$ has $Q^*$ as its domain. Then, $T^*$ is called a *generalization* of $T$ if $T$ and $T^*$ share the same number $I$ of records and if, for every $i \leq I$, $q_i \in q_i^*$ and $s_i = s_i^*$, where $q_i, q_i^*$, $s_i$, and $s_i^*$ denote the values of $\pi$, $\pi^*$, $\sigma$, and $\sigma^*$ in the $i$-th records of $T$ and $T^*$, respectively.

For tables $T$ and $T^*$, let $i$ be a number with $i \leq I$, $\Lambda$ be the $\pi$-block in $T$ of $q_i$, and $\Lambda^*$ be the $\pi^*$-block in $T^*$ of $q_i^*$. Then, for every $j \leq I$, if the $j$-th record $\lambda_j$ of $T$ is contained in $\Lambda$, then the $j$-th record $\lambda_j^*$ of $T^*$ is contained in $\Lambda^*$ and the value $s_j$ of $\sigma$ in $\lambda_j$ is the same as the value $s_j^*$ of $\sigma^*$ in $\lambda_j^*$. Thus, using Definitions 2 and 3, one can easily demonstrate Proposition 1.

**Proposition 1.** All $\mathbb{O}$-based $l$-diversities satisfy the monotonicity property. For example, if table $T$ satisfies $\mathbb{O}(m)$-$l$-diversity for a type $m$, then every generalization $T^*$ also satisfies $\mathbb{O}(m)$-$l$-diversity. The same holds for other $\mathbb{O}$-based $l$-diversities.

The monotonicity property of $\mathbb{O}$-based $l$-diversities implies that one can easily apply the "Incognito" algorithm [8] to $\mathbb{O}$-based $l$-diversities. This algorithm has been employed for data generalization to satisfy $k$-anonymity and $l$-diversity (see also [2]). As an application, the following outlines an Incognito-based process to generalize medical data (tables) to satisfy an $\mathbb{O}$-based $l$-diversity while maintaining data usability to the greatest possible extent.

i. Prior to generalization of the medical data, experts in medical informatics, administrators of medical databases, and hospital research ethics committees collaborate in formulating guidelines for the appropriate use of medical data for research purposes in the hospital.
ii. Administrators of medical databases and users (researchers) of medical data collaborate in deciding the scope of data $T$ that the users desire, as well as the conditions and priority for generalization of the (integrated) quasi-identifier $q$ in $T$ based on the guidelines. For domain $Q$ of $q$, a sequence $\mathbb{Q} = \{Q, Q^*, Q^{**}, \dots, Q^{*\cdots*}\}$ of iterative generalizations of $Q$ is specified, which we call a *strategy of generalization* of $T$.
iii. Based on the guidelines determined in (i) and the discussion in (ii), one of the $\mathbb{O}$-based $l$-diversities $\mathbb{D}$, including parameters, is determined and Incognito is adjusted based on $\mathbb{D}$.
iv. By applying strategy $\mathbb{Q}$ to $\mathbb{D}$-adjusted Incognito, the user can obtain generalized data $T^*$ that satisfy $\mathbb{D}$ through the fewest iterations of generalizations according to strategy $\mathbb{Q}$.

## 5. Estimation of Severity of Patients' Information in Japan

In this section, we estimate the severity number of information about patients who visited or stayed in hospitals in Japan. This is helpful for understanding the distribution of severity numbers of patient information across a whole society or hospital, allowing criteria of extended i-diversities to be determined based on the severity of inpatients' information for a given dataset. Thus, we estimate the severity numbers of information about patients who visited or stayed in hospitals and medical clinics with more than 20 beds during 2014. These estimations are based on the following data.

1. Estimated number of patients by classification of diseases and by type of healthcare facility in Japan during 2014, as published by Ministry of Health, Labour and Welfare [9].
2. Five-year relative survival rate of cancer patients by classification of diseases who were diagnosed during 2006 and 2008, as published by National Cancer Center Japan [10].

The estimated number of patients in dataset 1 above is determined by the classification of diseases and by the type of healthcare facility. The international classification of diseases (ICD-10) is used, and this is the same as used for dataset 2 and Table 3. Therefore, we could easily associate patients in dataset 1 with those in dataset 2 by use of ICD-10.

Strictly speaking, the severity number should be determined by a set of medical data. However, based on datasets 1 and 2 alone, it is not possible to calculate the severity number of patient information for each set of medical data. Therefore, we focus on the distribution of patients according to the classification of diseases and the type of healthcare facility in Japan, and calculate the severity numbers of patient information for patients in dataset 1.

The severity number of patient information is calculated as follows.

i. Based on the five-year relative survival rate of cancer patients by classification of diseases (dataset 2), cancer patients with survival rates of greater than 90% are assigned severity number 1, cancer patients with survival rates of 60–90% are assigned severity number 2, and cancer patients with survival rates of less than 60% are assigned severity number 3. This assignment is conducted in accordance with class 2.1.1 (information about diseases that have a major impact on patients' life spans) in RIOPSI (Table 2).

ii. Alzheimer's patients with "psychiatric disorders" (see Table 3) are assigned a severity number of 4, based on class 1.3.1 (information about significant psychiatric disorders) in Table 2. Patients with "other" psychiatric disorders (see Table 3) are assigned severity number 1, based on class 2.3.2 (information by which others might have prejudice toward patients' appearances) in Table 2.

iii. Patients who have "vascular brain disease" or "chronic renal failure" (see Table 3) are assigned severity number 1, based on class 2.2.2 (information about financial burdens of patients), 2.2.3, or 2.2.4 (information about degrees of disability to patients' social and/or personal life) in Table 2.

iv. We consider patients with "congenital abnormality, chromosome abnormality" to be assigned severity number 4 or 1, based on class 1.1.1 (information about diseases of genes), 1.2.1 (information about diseases assigned as intractable diseases), or 2.3 (information that inflict damage on patients' human rights). However, it is impossible to classify the patients into these three cases based on datasets 1 and 2 alone. Therefore, we assigned these patients with a severity number of 2.5 (the mean of 1 and 4).

Based on steps (i)–(iv) above, we aggregated the severity number of patient information by patient according to the classification of diseases and type of healthcare facility. The results are presented in Table 3.

Moreover, to compare the severity numbers of patients in hospitals and medical clinics, we aggregated the severity numbers according to type of healthcare facility and calculated the severity number per patient in hospitals and medical clinics. The results are given in Table 4.

Table 4 indicates that the sum of the severity numbers of patients who visited or stayed in hospitals was approximately three times that in medical clinics, and that the severity numbers per patient in hospitals was more than four times that in medical clinics. These facts are probably natural, as patients who have difficult diseases to treat or need long-term hospitalization generally prefer hospitals to medical clinics. Thus, from the viewpoint of privacy preservation, some might claim that medical data should be treated with more prudence in hospitals than in medical clinics. However, Table 4 also suggests that even medical clinics have 1 in 10 patients for whom information has a severity number greater than 1. This is also reasonable, as there are many patients who are not admitted to hospital but require support from medical clinics. Hence, even in the case of medical clinics, researchers should give their full attention to the appropriate treatment of medical data.

In this section, we have calculated the severity number of patient information based on just two datasets [9, 10]. We calculated the severity numbers with respect to diseases, but not with respect to medical treatments, medicines, and so forth. Therefore, the severity numbers we calculated would have underestimated the true values. To calculate the true values, it is necessary to assess the monetary costs of medical treatments and medicines. Moreover, to calculate the severity numbers more precisely, we need expert opinions regarding the severity and costs of diseases and healthcare.

Table 3. Severity numbers of patients' information according to classification of diseases and type of healthcare facility

| Disease | Subclass | 5 year survival rate | Sevirity number (S.N.) | Hospitals | | Medical clinics | |
|---|---|---|---|---|---|---|---|
| | | | | Patients (x 1000) | Sum of S.N. (x 1000) | Patients (x 1000) | Sum of S.N. (x 1000) |
| All disease | | | | 2914.9 | | 4278.8 | |
| Cancer | Stomach | 64.6 | 2 | 28.3 | 56.6 | 4.5 | 9 |
| | Colon, Rectosigmoid junction and rectum | 71.1 | 2 | 41.6 | 83.2 | 5.4 | 10.8 |
| | Liver and intrahepatic bile ducts | 32.6 | 3 | 11.3 | 33.9 | 1.1 | 3.3 |
| | Trachea, bronchus and lung | 31.9 (lung) | 3 | 33 | 99 | 1.9 | 5.7 |
| | Breast | 91.1 | 1 | 26.8 | 26.8 | 2.9 | 2.9 |
| Psychiatric disorders | Alzheimer | | 4 | 59.9 | 239.6 | 32.1 | 128.4 |
| | Others | | 1 | 375.4 | 375.4 | 148.2 | 148.2 |
| Vascular brain disease | | | 1 | 199.6 | 199.6 | 53.8 | 53.8 |
| Chronic renal failure | | | 1 | 68.9 | 68.9 | 62.5 | 62.5 |
| Congenital abnormality, chromosome abnormality | | | 1 or 4 | 14.7 | 36.75 | 5.3 | 13.25 |

Table 4 Severity numbers of patients' information according to type of healthcare facility

| | Hospitals | Medical clinics |
|---|---|---|
| Number of patients with severity numbers ≧1 (x 1000) | 859.5 | 317.7 |
| Sum of severity number (x 1000) | 1219.75 | 437.85 |
| Sum of severity number / patient | 0.42 | 0.1 |

## 6. Related Work

For clarity, this section refers to the tables in relational databases as "datasets" rather than "data."

Both $k$-anonymity [1] and $l$-diversity [2] are well-known indicators of the anonymization of a dataset, especially in relational databases, which are the most popular database systems. Whereas the $k$-anonymity is employed to measure the risk of identification of target individuals in a given dataset, the $l$-diversity is employed to measure the risk of identification of sensitive information about the target individuals.

Ensuring the $k$-anonymity of the given data often requires significant anonymization of the dataset, which may destroy its value for researchers. Complete $k$-anonymization provides a strong assurance of the information security of a given dataset, but sometimes makes it very ambiguous for both adversaries and researchers. Therefore, $k^m$-anonymity [11] has been introduced as an indicator of $k$-anonymity in the limiting condition for the knowledge that adversaries have about the dataset. A dataset $D$ is said to satisfy $k^m$-anonymity if every sub-dataset $E$ of $D$ given by restricting the length of the quasi-identifier to not more than $m$ satisfies $k$-anonymity. Whereas $k$-anonymity assumes that an adversary knows everything about the quasi-identifier of $D$, $k^m$-anonymity assumes that an adversary knows at most $m$ attributes that constitute the quasi-identifier of $D$. Compared with $k$-anonymization, $k^m$-anonymization may weaken the anonymity of the data, but it prevents the situation whereby the given dataset $D$ is of significantly diminished value for researchers.

Though an Incognito-based approach for $k^m$-anonymization was presented in [11], a different approach for $k^m$-anonymization has also been developed [12]. This latter approach uses a process called "disassociation." Disassociation anonymizes a dataset so that it satisfies $k^m$-anonymity by masking the relationships between quasi-identifiers or sensitive attributes, rather than masking the quasi-identifiers or sensitive attributes themselves. In many cases of statistical analysis, masking elements in a dataset $D$ has a significantly negative effect on the quality of analysis, whereas masking the relationships between elements in $D$ has an insignificant effect. Thus, disassociation is expected to reduce the risks of identification of a dataset $D$ by an adversary while retaining the quality of analysis of $D$.

The concept of $l$-diversity is basically premised on $k$-anonymity, and compensates for the lack of effectiveness of $k$-anonymity. However, requesting additional conditions for $l$-diversity often reduces the research value of a dataset $D$. The possibility of $l$-diversity having a negative effect on a dataset has been considered [2], and derivatives of $k$-anonymity such as entropy $l$-diversity and recursive $l$-diversity have been introduced. However, Li et al. [13] mentioned the impracticality of anonymizing a given dataset $D$ to satisfy $l$-diversity or its derivative versions, and found that $l$-diversity and its derivatives were sometimes unnecessary for protecting $D$ from adversaries. To ensure the anonymity of a dataset $D$, Li et al. [13] considered it important to prevent an adversary of $D$ from having additional useful information about $D$. In this context, useful information was considered to be that which gave them some special knowledge about the statistical distribution of sensitive attributes in $D$. Therefore, for a given dataset $D$, Li et al. [13] first considered a completely anonymized dataset $D_0$ of $D$, which gave adversaries no special information about the statistical distribution of sensitive attributes in $D$. This was the most desirable anonymized version of $D$ in terms of privacy preservation, but might be useless for researchers. Li et al. then considered the actual anonymized data $D_1$ of $D$ based on $D_0$. They defined $D_1$ as satisfying the t-closeness property if the difference between the statistical distributions of sensitive attributes in $D_0$ and $D_1$ was below a given threshold t. The issues with this approach are what actually is $D_0$ and how should the "difference" between the statistical distributions be defined. $D_0$ is the dataset in which, for sensitive attributes, the only statistical distributions provide no additional information to the adversary. Such a dataset might have no value to researchers. This may be unavoidable, as there is generally a trade-off between the effect of anonymization and the research value. As an indicator of the difference between the statistical distributions of attributes, Li et al. [13] employed the "Earth Mover's Distance" (EMD) [14]. The EMD between statistical distributions $u_1$ and $u_2$ is the least cost that is necessary to remove the differences between $u_1$ and $u_2$. By preventing adversaries from having useful information about the statistical distributions of sensitive attributes in a given dataset $D$, one can actually preserve privacy without unnecessarily reducing the value of $D$ for researchers.

Our extension of $l$-diversity is also intended to anonymize a dataset $D$ by clarifying and quantifying the damage that patients or informants suffer from an attack to obtain information about their sensitive attributes based on $\mathbb{O}$. Anonymization based on the degree of informants' damage has been attempted by others, such as in the p-sensitive $k$-anonymity model [15]. However, the quantification of informants' damage in [15] is no more than a simple classification of diseases, and this approach requires further research to conceptualize or quantify the severity of patients' risks. Our approach for clarifying and quantifying the severity of patients' risks constructs an ontology of risks by conceptualizing the viewpoints and values of adversaries and clarifying the situations in which the adversaries enhance their profits, which corresponds to the situation whereby patients suffer from breaches of their privacy. We first conceptualize elements of a given dataset that enhance the severity of damage to individuals and quantify the severity from the viewpoints of diseases or medical services based on the ontology. This approach quantifies the severity of patient damage more precisely than in the method of [15]. Although our approach employs the ontology RIOPSI $\mathbb{O}$ and quantifies the severity of damage in extending the $l$-diversity, our approach can also be employed to extend the concept of t-closeness. The extension of t-closeness based on this approach and the development of an algorithm for the extended t-closeness will be considered in future studies.

The severity of patient damage from the viewpoint of breaches of privacy in a dataset also has been investigated outside of the anonymization of medical data. For example, the Privacy Impact Assessment (PIA) evaluates the risks of damage through data breaches in target information systems [16, 17, 18]. The purpose of PIA is to ameliorate the issues arising from privacy breaches in which data are recorded in the target information system. This is done by making a preliminary assessment of the risks and damages of privacy breaches in the target system prior to its development. Therefore, the assessment covers not only data in the target system, but also the system itself and work processes within the system. Hence, PIA assesses not only the severity of damage following privacy breaches, but also the possibility of emerging problems and the vulnerability of the system against attacks by adversaries. Furthermore, as PIA forms the basis of guidelines for introducing and managing information systems, it is a highly public matter. Therefore, the severity of privacy breaches in PIA cannot be assessed very thoroughly. In fact, the sensitivity of patient data is classified into at most three levels [16, 17]: basic information that identifies each patient (level-1), information that indicates the medical characteristics of patients (level-2), and information that indicates the situations of patients during medical care (level-3).

## 7. Conclusion

In this paper, we have proposed several extensions of $l$-diversity to reduce the risk of attackers identifying severe patient medical conditions from medical data. To this end, we defined the RIOPSI ontology model and quantified the severity of patient health information. Moreover, we showed that our extended versions of $l$-diversity satisfy the monotonicity property. This allowed us to outline a process to generalize medical data (tables) so as to satisfy the extended $l$-diversities while maintaining data usability to the greatest extent possible using the adjusted Incognito algorithm.

In Section 5, we estimated the severity number of information about patients who had visited or stayed in hospitals in Japan using our ontology model and statistical data published by the Ministry of Health, Labour and Welfare and the National Cancer Center in Japan [9, 10]. The results show that the sum of the severity numbers of patients in hospitals was approximately three times that in medical clinics, and that the severity number per patient in hospitals was more than four times that in medical clinics. However, our analysis also shows that, even in the case of medical clinics, 1 in 10 patients have information with a severity greater than 1.

There are several limitations to this study. For instance, in Table 2, it would be more effective to have more refined subsets of the special values sets. For example, it would be better to define subsets in the small class (2.1.1) using a one-year survival rate and/or a three-year survival rate. An approach to creating more refined subsets will be investigated in future work.

## Acknowledgements

This work was supported by JSPS KAKENHI (Grant Number 15K00297) and the JUTEN-SHIEN project of Gunma University.

## References

[1] L. Sweeney, “k-anonymity: a model for protecting privacy”, ***Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*** Vol. 10, No. 5, pp. 557-570, 2002.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “L-diversity: Privacy beyond k-anonymity”, ***TKDD*** Vol. 1, No. 1(3), pp. 1-52, 2007.

[3] O. Takaki, T. Asao, and Y. Seki, “Extensions of l-diversity to reduce the risk of revealing patient severe health conditions”, in ***Proc. International Conference on Mechanical, Electrical and Medical Intelligent System 2017 (ICMEMIS2017)*** (Kiryu, Japan) November 2017.

[4] C. A. Meyers, S. S. Powers, and D. M. Faissol, “Taxonomies of cyber adversaries and attacks: a survey of incidents and approaches”, ***Technical Report in Lawrence Livermore National Laboratory (LLNL), Livermore***, CA: LLNL-TR-419041, DOI: 10.2172/967712, 2009.

[5] S. Atkinson, “Psychology and the hacker: Psychological incident handling”, White Paper, SANS Institute, 2015.

[6] D. Cappelli, A. G. Desai, A. P. Moore, T. J. Shimeall, E. A. Weaver, and B. J. Willke, “Management and education of the risk of insider threat (MERIT): system dynamics modeling of a computer system”, White Paper, Carnegie Mellon University, 2008.

[7] The Nikkoso Research Foundation for Safe Society, “Report on measure against human threats to information security”, White Paper, 2010 (in Japanese).

[8] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Incognito: efficient full-domain K-anonymity”, in ***Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data (SIGMOD '05)***. ACM, New York, NY, USA, 2005, pp. 49-60.

[9] Ministry of Health, Labour and Welfare, “Heisei 26 nen (2014) kanja-chousa no gaikyou” (in Japanese), https://www.mhlw.go.jp/toukei/saikin/hw/kanja/14/ (Accessed 19 March 2019).

[10] National Cancer Center Japan, “Gan touroku · toukei” (in Japanese), https://ganjoho.jp/reg_stat/statistics/dl/index.html (Accessed 19 March 2019).

[11] M. Terrovitis, N. Mamoulis, and P. Kalnis, “Privacy-preserving anonymization of set-valued data”, ***In Proceedings of the VLDB Endowment***, 2008, pp. 115-125. DOI:10.14778/1453856.1453874.

[12] M. Terrovitis, J. Liagouris, N. Mamoulis, and S. Skiadopoulos, “Privacy Preservation by Disassociation”, in ***Proceedings of the VLDB Endowment***, Vol. 5, No. 10, pp. 944-955, 2012.

[13] N. Li, T. Li and S. Venkatasubramanian, “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity”, in ***Proceedings of 2007 IEEE 23rd International Conference on Data Engineering***, Istanbul, 2007, pp. 106-115.

[14] Y. Rubner, C. Tomasi, and L. J. Guibas, “The Earth Mover’s Distance as a metric for image retrieval”, ***Int. J. Comput. Vis.***, Vol. 40, No. 2, pp.99-121, 2000. DOI: 10.1023/A:1026543900054.

[15] X. Sun, H. Wang, J. Li, and T. M. Truta, “Enhanced P-Sensitive K-Anonymity Models for Privacy Preserving Data Publishing”, ***Trans. Data Privacy*** Vol. 1, No. 2, pp. 53-66, 2008.

[16] Y. Seto, “Jissenteki puraibashii risuku hoyukagihou”, ***Kindaikagakusha***, 2014 (in Japanese).

[17] Y. Seto, “Puraibashii eikyou hyouka gaidorain jissen tekisuto”, ***Impress R&D***, 2016 (in Japanese).

[18] R. Clark, “Privacy impact assessment: Its origins and development”, ***Comput. Law Secur. Rev.*** Vol. 35, pp.123-135, 2009.