International Journal of Networking and Computing – www.ijnc.org, ISSN 2185-2847
Volume 13, Number 2, pages 195–215, July 2023

Privacy preserving medical knowledge discovery by multiple “patient characteristics” formatted data

Kenta Kitamura
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8654, Japan

Mhd Irvan
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8654, Japan

Rie Shigetomi Yamaguchi
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8654, Japan

Received: February 13, 2023
Revised: April 20, 2023
Accepted: June 11, 2023
Communicated by Yasuyuki Nogami

Abstract

Statistical processing and Artificial Intelligence (AI) development utilizing big data have been actively researched recently. However, there are growing concerns about privacy violations due to the use of private data. For such concerns, the EU General Data Protection Regulation (GDPR) was introduced to regulate the handling of personal information. The GDPR makes it difficult to discover medical knowledge through big data analysis in medical studies. However, the GDPR is not concerned with handling non-personally identifiable statistical information. Statistical information is commonly published, collected, and analyzed. However, it is unknown whether collecting and analyzing such statistical information can generate medical evidence through variable-to-variable research, such as the relationship between tobacco and cancer.

In this paper, we propose to use statistical information that is not concerned by the GDPR to estimate cross-tabulation tables, which are usually generated from personal information in medical research and are widely used for analysis between medical variables. In particular, as statistical information, we use “patient characteristics” formatted data commonly published in medical research. The scope of this paper is the situation where the publisher of statistical information and the analyst of published statistical information differ. On the publisher side, we assume the publisher collects raw data from a target people group by random sampling multiple times and converts the data to patient characteristics formatted data. On the analyst side, we assume the analyst collects those published many random sampled patient characteristics formatted data and estimates the cross-tabulation table by the Law of Large Numbers (LLN). We model the publisher-analyst situation described above. In the aforementioned model, we evaluate our proposal estimation’s usefulness through both theoretical and experimental accuracy assessments. Furthermore, for quantitative Privacy Preserving Data Mining (PPDM), we evaluate the risk of anonymity when collecting multiple patient characteristics using the existing anonymity indicator, the Patient Family Detect on Overall Category (PFDOC) entropy. We theoretically and experimentally check the occurrence rate of vulnerable patient characteristics with PFDOC entropy equal to zero obtained by the analyst. In the experiment, the target people group data is 20,000 personal data which have four categorical binary values. As the publisher model, we created 10,000 patient characteristics, which are statistics for randomly sampled 50 data from the 20,000 data. As the analyst model, we estimated the cross-tabulation table by the 10,000 patient characteristics. The theoretical prediction error was 1.8% (95% CI), and the experimental error was within 1.5% (95% CI, $n = 100$), indicating a close agreement between theory and experiment. Regarding anonymity, it was theoretically expected that PFDOC entropy = 0 patient characteristics would be rare in categories with a population ratio of 25% to 75%, leading to ensured anonymity. It was confirmed in the experiment. Based on these results, we can conclude that, by using the patient characteristics formatted data release and collection model and selecting the appropriate population ratio categories, an analyst can accurately estimate cross-tabulation tables while preserving PFDOC entropy-based anonymity without legal restriction.

Keywords: random sampling, health care, patient characteristics

1 Introduction

1.1 Background

Big data has been widely used to create statistics and Artificial Intelligence (AI) but using personal data involves the privacy invasion risk [21]. Because of this concern, the EU General Data Protection Regulation (GDPR) regulates data treatment. However, the GDPR makes knowledge discovery through big data difficult especially in the medical field because medical data is sensitive.

Fig. 1 shows the traditional medical research methods' validity hierarchy that is generally considered medical evidence level [12]. In randomized control trials (RCT), cohort studies, case control studies, case series, and reports, relationships between novel medical variables, such as tobacco and lung cancer, can be uncovered. However, these traditional medical research methods are challenging to conduct publicly under the GDPR regulations because these methods handle personal information. Systematic review and meta analysis [4] require only publicly available statistical data which are not concerned by the GDPR [16]. However, systematic review and meta analysis are methods that combine the results of existing medical articles to enhance the results. Thus the analyst cannot analyze the relationship between arbitrary new variables.

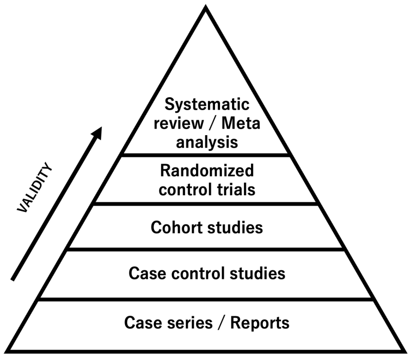


Figure 1: Medical study methods validity pyramid [12].

In contrast to these conventional medical methods, non-traditional medical research methods can be conducted by statistical information. Some of these methods can derive new variable-to-variable analyses. This type of study can be separated into medical statistical information analyzing methods [17], and non-medical statistical information analysis with medical data methods [6] [14]. However, because these methods differ from traditional medical research analysis methods, whether medical validity can be ensured is unclear. Moreover, because the input data is often not common medical data, whether traditional medical organizations can issue such data is unclear.

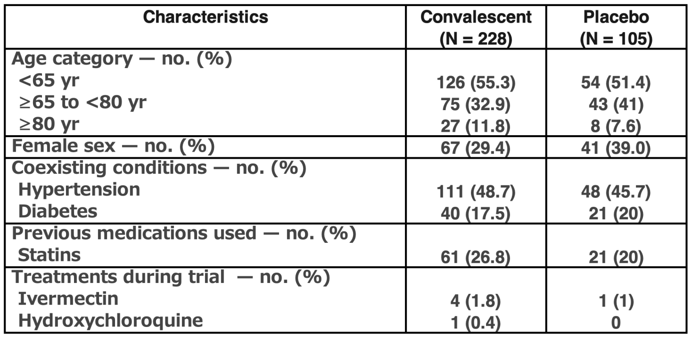
1.2 Contributions and paper outline

In this paper, we propose a new technique for estimating data required for medical evidence generation from publicly available medical statistics. Specifically, the technique generates an estimated cross-tabulation table as Table 1 by using patient characteristics as Table 2. The cross-tabulation table example is shown in Table 1. In Table 1, the ratios of "diabetes and over 65 years of age", "diabetes and under 65 years of age", "non-diabetes and over 65 years of age," and "non-diabetes and under 65 years of age" are expressed as A, B, C, and D respectively. Table 2 shows the example of the patient characteristics that summarize the patient background statistics of a drug trial. The patient characteristics are published in medical articles because the results vary depending on the patient background.

Table 1: Example of cross-tabulation table of diabetes vs. age [20].

	Diabetes	Non-Diabetes
Age ≥65 yr	A	C
Age <65 yr	B	D

Table 2: Patient characteristics from Covid-19 clinical research paper (modified as appropriate) [19].



Characteristics	Convalescent (N = 228)	Placebo (N = 105)
Age category — no. (%)		
<65 yr	126 (55.3)	54 (51.4)
≥65 to <80 yr	75 (32.9)	43 (41)
≥80 yr	27 (11.8)	8 (7.6)
Female sex — no. (%)	67 (29.4)	41 (39.0)
Coexisting conditions — no. (%)		
Hypertension	111 (48.7)	48 (45.7)
Diabetes	40 (17.5)	21 (20)
Previous medications used — no. (%)		
Statins	61 (26.8)	21 (20)
Treatments during trial — no. (%)		
Ivermectin	4 (1.8)	1 (1)
Hydroxychloroquine	1 (0.4)	0

To quantify the methodology, we propose a model where the publisher of patient characteristics statistics and the analyst estimating cross-tabulation tables are independent. We then present a method for estimating cross-tabulation tables in this model by using the Law of Large Numbers (LLN). Note that the patients included in the patient characteristics are assumed to be randomly selected from the target patient population because patients select medical institutions randomly, and the application of LLN to the estimation process becomes more appropriate as the sample size of the randomly selected data increases. In this analyst estimation model, we derive a theoretical estimation error equation and compare it to the experimental estimation error. Moreover, we quantitatively examine the relationship between usefulness and anonymity which is a famous problem in Privacy Preserving Data Mining (PPDM) field [11]. For a quantitative anonymity check on a collection of multiple patient characteristics, we utilize an existing patient characteristics vulnerability indicator called Patient Family Detect on Overall Category (PFDOC) entropy [8].

Namely, our contributions are the following four points.

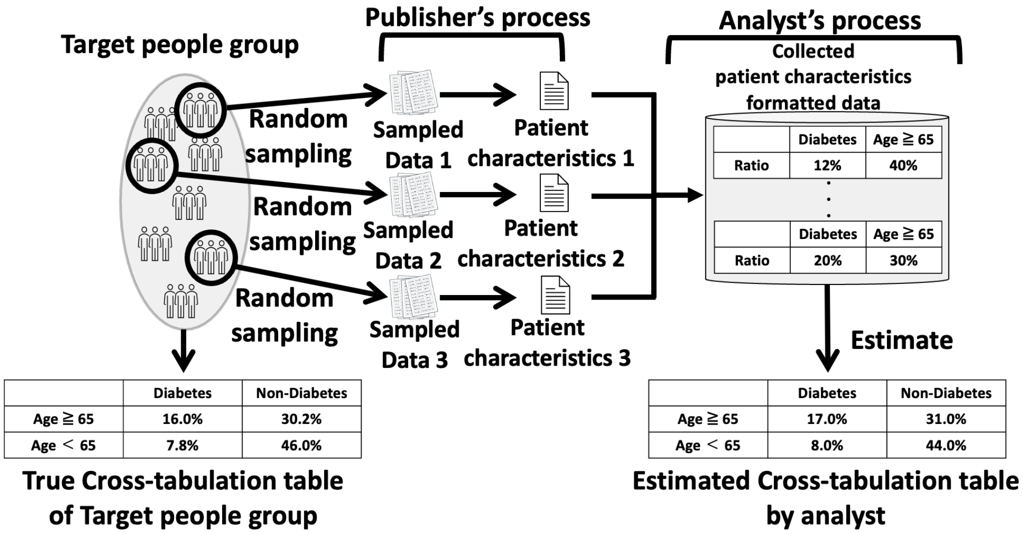
- We propose to use publicly available statistical information that is not concerned by the GDPR [16] to estimate cross-tabulation tables, which are often created from personal information in medical research and are widely used for the analysis of medical variables. More specifically, we propose to estimate cross-tabulation tables [20] from multiple "patient characteristics" [19] formatted statistics.

- We model how the publisher processes the personal data into the patient characteristics format and how the analyst estimates cross-tabulation tables for quantitative analysis.

- We theoretically and experimentally estimate cross-tabulation tables to confirm the proposed estimation accuracy.

- We theoretically and experimentally check the anonymity of multiple patient characteristics acquisition through PFDOC entropy = 0 [8] patient characteristics occurrence rate for PPDM.

The overview of the proposal method is shown in Fig. 2. The scope of this paper is the situation where the data publisher is different from the data analyst. For example, the data publisher is a public health official. The publisher asks local residents about their history of COVID-19 infection, age, and gender and then provides daily data formatted as patient characteristics. On the other hand, the data analyst collects much of the published patient characteristics and estimates cross-tabulation tables.


Publisher's process
Analyst's process
Target people group
Random sampling
Random sampling
Random sampling
Sampled Data 1
Sampled Data 2
Sampled Data 3
Patient characteristics 1
Patient characteristics 2
Patient characteristics 3
Collected patient characteristics formatted data
Estimate
True Cross-tabulation table of Target people group
Estimated Cross-tabulation table by analyst

	Diabetes	Age ≧ 65
Ratio	12%	40%

	Diabetes	Age ≧ 65
Ratio	20%	30%

	Diabetes	Non-Diabetes
Age ≧ 65	16.0%	30.2%
Age < 65	7.8%	46.0%

	Diabetes	Non-Diabetes
Age ≧ 65	17.0%	31.0%
Age < 65	8.0%	44.0%

Figure 2: Overview of proposal method. The publisher randomly samples from the target people group multiply and converts sampled data to patient characteristics format. The analyst collects the published data and estimates the cross-tabulation table by the Law of Large Numbers (LLN).

The challenge of the proposed method is that one patient characteristics usually shows only one ratio per category in the target group, such as "diabetes rate" or "age over 65 years old rate". Thus, the "population ratio of diabetes and over 65 years of age" is unknown, and the analyst can not get A in Table 1. However, if the analyst has many patient characteristics randomly generated from the target people group, the analyst can estimate that the average value of diabetes among patient characteristics gets close to the diabetes ratio $(A+B)/(A+B+C+D)$ in Table 1 by LLN [18]. Through such estimation, the analyst can estimate A, B, C, and D. In this paper, we present a single instance of the many possible estimations.

Our proposal method enables the discovery of new medical relationships between variables commonly used for generating medical evidence while preserving PFDOC entropy-based anonymity by selecting categories based on the population ratio. The method avoids the need for medical analysis performed primarily in healthcare institutions due to privacy concerns, by using publicly available medical statistical data that is exempt from legal privacy issues.

The remainder of this paper is organized as follows. In section 2, we review related work. In section 3, we make preliminary preparations. In section 4, we describe the publisher's model. In section 5, we describe the analyst estimation example. In section 6, we describe the analyst estimation model. In section 7, we present the theoretical evaluation of our proposal estimation model. In section 8, we present the experimental evaluation of our proposal estimation model. In section 9, we discuss the result. In section 10, we present the conclusion and future work.

2 Related Work

2.1 Common medical study methods

As shown in Fig. 1, the traditional medical studies have a validity hierarchy that is generally considered medical evidence level [12]. Case series and reports are reported by experienced physicians. Case control and some cohort studies are methods that retrospectively examine groups of patients with and without exposure. RCT and some cohort studies are methods that prospectively examine groups of patients with and without exposure. Meta analyses and systematic reviews integrate the results of several existing medical articles to increase validity [4]. In particular, meta analysis aggregates statistical results based on patient characteristics comparisons in many medical articles to ensure uniformity of patient backgrounds [19].

2.2 Medical knowledge discovery form public available statistical information

Medical statistics are published daily by various medical organizations. Based on these statistics, some studies predict the future number of COVID-19 infections in a given area [17]. In addition, medical analysis methods based on collecting non-medical statistics produced by non-medical organizations combined with medical information have also been proposed. For example, a study analyzed the relationship between national policy and COVID-19 mortality [14]. For another example, a study analyzed the relationship between COVID-19 and human flow using Google's location-based statistics [6].

2.3 PPDM

The PPDM is developed to extract utility from data without disclosing confidential information [11]. The algorithms proposed by PPDM achieve privacy based on specific privacy metrics, but it has been pointed out that there is a trade-off of reduced utility, such as decreased accuracy of the model. Therefore, it is necessary to examine how much utility can be maintained while ensuring a certain level of anonymity.

2.4 l-diversity

Releasing quasi-identifier (QI) can result in anonymity violation. A $q^\star$-block represents a tuple of individuals who possess the same non-sensitive QI combinations. When the diversity of $q^\star$-block is low, the risk of sensitive QI inference, such as homogeneity attack or background attack [10], increases. Anonymity protection through l-diversity is crucial in such cases.

Definition 1. *(l-Diversity [10]): A $q^\star$-block is l-diverse if contains at least l "well-represented" values for the sensitive QI S. A table is l-diverse if every $q^\star$-block is l-diverse.*

The value "l" of l-diversity can be expressed in different ways, such as by number, entropy, and frequency, and is well represented in each representation.

The entropy l-diversity has been proposed as a method to quantify l-diversity.

Definition 2. *(Entropy l-diversity [10]): A table is entropy l-diverse if for every $q^\star$-block*

$$-\sum_{s \in S} p_{(q^\star,s)} \log\left(p_{(q^\star,s')}\right) \geqq \log(\ell) \tag{1}$$

(where $p_{(q^\star,s)} = \frac{n_{(q^\star,s)}}{\sum_{s' \in S} n_{(q^\star,s')}}$ is the fraction of tuples in the $q^\star$-block with sensitive QI value equal to s. And $n_{(q^\star,s)}$ is the number of s in $q^\star$-block)

The concept of entropy l-diversity is that a low biased distribution of sensitive QI in a $q^\star$-block increases anonymity.

2.5 PFDOC attack and PFDOC entropy

Patient characteristics indicate what percentage of clinical trial patients belong to what category. The PFDOC attack [8] is an anonymity violation for patient characteristics.

Table 3 is vulnerable to the PFDOC attack. The patient population (333 individuals) in this clinical trial is primarily hypertensive (330 individuals). An attacker, such as a patient's relative, can infer the health condition of a specific patient with a high degree of certainty through knowledge of the patient's inclusion in the primarily hypertensive patient population (330 out of 333 individuals) in this clinical trial.

Table 3: Example of the Patient Family Detect on Overall Category (PFDOC) attack vulnerable patient characteristics [8].

	Total number (N = 333)
Hypertension	330

Table 4 displays the patient characteristics model with Na representing the total number of clinical trial participants and A representing the number of patients in a specific category. The PFDOC entropy, an indicator of PFDOC attack vulnerability, is defined in the following manner.

Table 4: Model of the patient characteristics [8].

	Total number (N = Na)
Category	A

Definition 3. *(Patient Family Detect on Overall Category Entropy (PFDOC Entropy)): In the patient characteristics, the total number of patients is Na and the number of patients belonging to the category is A, as Table 4. The PFDOC Entropy is calculated as follows.*

$$-(A/Na)\log(A/Na) - ((Na - A)/Na))\log((Na - A)/Na)) \tag{2}$$

Note that if A = 0, then the PFDOC Entropy = 0.

This anonymity violation stems from the principle of l-diversity [10], where low diversity in a population with similar characteristics leads to a violation of anonymity. When a QI has a low PFDOC entropy, patients in that group have either a high or low probability of possessing that QI, leading to a violation of anonymity. This is indicated by the PFDOC entropy l-diversity, which applies entropy l-diversity to express the anonymity violation from the PFDOC attack.

Definition 4. *(Patient Family Detect on Overall Category Entropy l-diversity (PFDOC Entropy l-diversity)): A patient characteristics is PFDOC Entropy l-diverse if for every category*

$$PFDOC\,Entropy \geqq log(l) \tag{3}$$

3 Preliminaries

3.1 The GDPR concern and anonymity

According to the GDPR Recital 26, the GDPR does not concern processing anonymous information, including for statistical or research purposes [16]. Moreover, European Data Protection Board

(EDPB) mentions that “the GDPR will no longer be applicable to these fully aggregated and anonymised datasets” [5]. However, the aggregated data should follow these rules (1) the aggregate data must not be connected directly to identifying data (2) A known systematic method of (re)identifying must not exist, and (3) the data must not be linked to a specific person [7].

3.2 Odds ratio, logistic regression, and multiple regression from cross-tabulation tables in medicine

In medical studies, the usual choice of statistical analysis method is multiple regression for continuous value responses, logistic regression for categorical responses, and Cox’s proportional hazards model for censored responses. Note that Cox’s proportional hazards model is a variant of logistic [2].

Odds ratios are used to assess the risk from exposure to some factor like tobacco [20]. In Table 1, diabetes odds ratio is calculated as $(A \times D)/(B \times C)$.

The logistic regression is expressed as follows.

$$\log(\pi/(1-\pi)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{4}$$

π is the event occurrence probability due to the exposure, x_j is a variable that reflects the exposure of j th factor by a value of 1 or 0, and β_j is the impact of the exposure to the factor on the probability of the event occurrence. The relationship $\exp(\beta_j) = OR_j$ is known, where OR_j is the odds ratio for the exposure to the j th factor. Because the odds ratio can be calculated from the cross-tabulation table, the function (4) can be determined from the cross-tabulation table [2].

The multiple regression is expressed as follows.

$$Y = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_\mathrm{k} \mathbf{x}_\mathrm{k} + \mathrm{e} \tag{5}$$

Y is the target response variable, x_j is a variable that reflects the exposure of j th factor, β_j is the impact of the exposure to the factor, and e is a residual term. For $k = 1$, the least-squares method determines (5) for as follows.

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \beta_0 = \bar{y} - \beta_1 \bar{x} \tag{6}$$

x_i, y_i are the values determined by the exposure and the occurrence of the event in the i th data, and the number of $(x_i, y_i) = (1,1), (1,0), (0,1), (0,0)$ can be obtained from the cross-tabulation table. In addition, $\bar{x}$ and $\bar{y}$ are the means of x_i and y_i, which can also be obtained from the cross-tabulation table. In the case where k is general, the function (5) is determined similarly from the cross-tabulation table [1].

3.3 LLN and normal distribution

LLN is the law stating as follows [18]. “The average of the results from a large number of trials tends to get close to the expected value, and the larger the number of trials, the closer to the expected value.”

In the normal distribution, as LLN, the error in the 95% confidence interval (CI) between the sample proportion and the population proportion gets smaller in many samples as the following relationship [3].

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} \tag{7}$$

where n is number of samples, $\hat{p}$ is the sample proportion, $\hat{q} = 1 - \hat{p}$, and p is the population proportion.

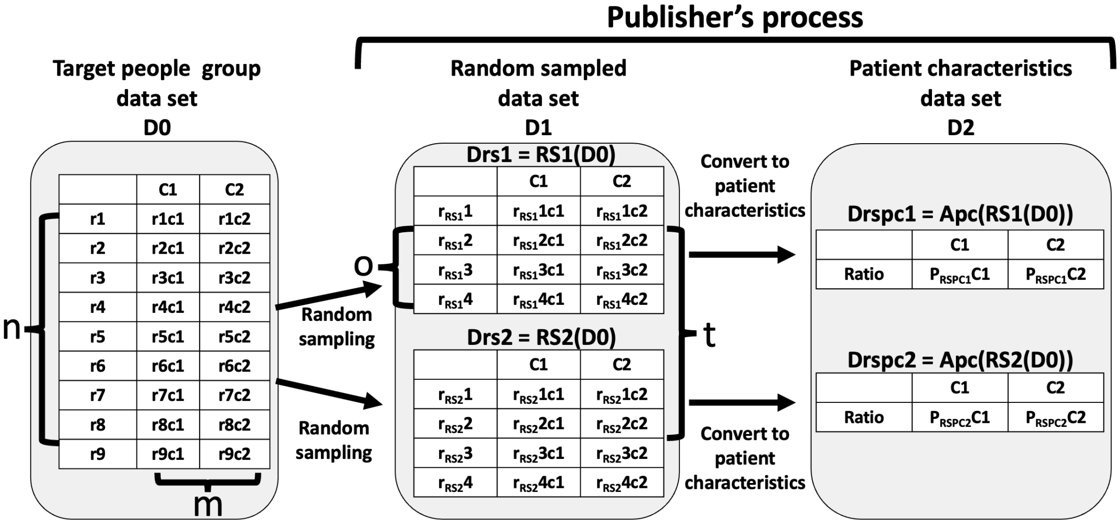
4 Publisher’s process model

4.1 Target people group data set

As shown in Fig. 3 left side, the publisher selects the target people group. The target people group data set $D0$ contains n people’s data with m binary categorical values. The n people’s data are denoted as $r1, \ldots, rn$, and the m categories are denoted as $C1, \ldots, Cm$. And i th person’s data ri contains m binary categorical values, so ri is expressed by a row vector containing m matrix elements as $ri = (ric1, \ldots, ricm)$. If the person has category Cj attribute, we express it as $ricj = 1$, and if not, $ricj = 0$.

Definition 5. *(Target people group data set): Let $D0$ be a target people group data set consisting of n rows and m columns matrix with 0 or 1 value on each matrix component. Let ri be i th row of $D0$. $D0$ and ri are expressed as follows.*

$$\begin{aligned} D0 &= (r1, \ldots, rn) \\ &= (d0_{ij})_{[n \times m]} \, (d0_{ij} \in \{0, 1\}) \\ ri &= (ric1, \ldots, ricm) \end{aligned} \tag{8}$$


Publisher’s process
Target people group data set D0
Random sampled data set D1
Patient characteristics data set D2
n
m
o
t
Random sampling
Random sampling
Convert to patient characteristics
Convert to patient characteristics
Drs1 = RS1(D0)
Drs2 = RS2(D0)
Drspc1 = Apc(RS1(D0))
Drspc2 = Apc(RS2(D0))

	C1	C2
r1	r1c1	r1c2
r2	r2c1	r2c2
r3	r3c1	r3c2
r4	r4c1	r4c2
r5	r5c1	r5c2
r6	r6c1	r6c2
r7	r7c1	r7c2
r8	r8c1	r8c2
r9	r9c1	r9c2

Drs1 = RS1(D0)

	C1	C2
$r_{RS1}1$	$r_{RS1}1c1$	$r_{RS1}1c2$
$r_{RS1}2$	$r_{RS1}2c1$	$r_{RS1}2c2$
$r_{RS1}3$	$r_{RS1}3c1$	$r_{RS1}3c2$
$r_{RS1}4$	$r_{RS1}4c1$	$r_{RS1}4c2$

Drs2 = RS2(D0)

	C1	C2
$r_{RS2}1$	$r_{RS2}1c1$	$r_{RS2}1c2$
$r_{RS2}2$	$r_{RS2}2c1$	$r_{RS2}2c2$
$r_{RS2}3$	$r_{RS2}3c1$	$r_{RS2}3c2$
$r_{RS2}4$	$r_{RS2}4c1$	$r_{RS2}4c2$

Drspc1 = Apc(RS1(D0))

	C1	C2
Ratio	$P_{RSPC1}C1$	$P_{RSPC1}C2$

Drspc2 = Apc(RS2(D0))

	C1	C2
Ratio	$P_{RSPC2}C1$	$P_{RSPC2}C2$

Figure 3: Model of the publisher. $D0$ ($n \times m$ matrix) is the target people group data set, which contains n people’s m categorical binary values. $n = 9$ and $m = 2$ in this figure. $D1 = \{Drs1, \ldots, Drst\}$ is the data set created by sampling t times from $D0$ with the o row random sampling operator RS. $t = 2$ and $o = 4$ in this figure. The dataset $D2 = \{Drspc1, \ldots, Drspct\}$ is the result of applying the patient characteristics converter Apc to each element of the dataset $D1$.

4.2 Random sampling

As shown in Fig. 3 middle, the publisher does random sampling with a sample size of o individuals t times from n people containing $D0$ and gets random sampled data set $D1$.

Definition 6. *(Random sampling): A random sampling operator RS converts $\mathbb{R}^n \times \mathbb{R}^m$ matrix to $\mathbb{R}^o \times \mathbb{R}^m$ matrix by choosing o rows from $\mathbb{R}^n \times \mathbb{R}^m$ matrix with uniformly random.*

Definition 7. *(Random sampling on target population data set): Let $D1$ be t times random sampled data set from target population group $D0$. Let RSi be i th random sampling operation. Let $r_{RSi}j$ be j th person with m categorical values in $RSi(D0)$. $D1$, $RSi(D0)$, and $r_{RSi}j$ are expressed as follows.*

$$
\begin{aligned}
\mathit{D1} &= \{Drs1, \ldots, Drst\} \\
&= \{RS1(D0), \ldots, RSt(D0)\} \\
RSi(D0) &= \{r_{RSi}1, \ldots, r_{RSi}o\} \\
r_{RSi}j &= \{r_{RSi}jc1, \ldots, r_{RSi}jcm\}
\end{aligned}
\tag{9}
$$

4.3 Convert to patient characteristics format

As shown in Fig. 3 right side, the publisher makes patient characteristics data set $D2$ from $D1$. Patient characteristics are the statistics of categories ratio, so the publisher makes statistics of categories ratio on each $RSi(D0)$.

Definition 8. *(Convert to patient characteristics from data set): Let D be the data set with o rows and m columns.*

$$
D = (d_{ij})_{[o \times m]}\,(d_{ij} \in \{0, 1\}) \tag{10}
$$

A patient characteristics operator Apc converts $\mathbb{R}^m \times \mathbb{R}^o$ *matrix to* $\mathbb{R}^1 \times \mathbb{R}^o$ *matrix by column-wise averaging. Let* $P_{PC}Ci$ *be i th Apc(D) column. Apc(D) and* $P_{PC}Cj$ *are expressed as follows.*

$$
\begin{aligned}
Apc(D) &= (P_{PC}C1, \ldots, P_{PC}Cm) \\
P_{PC}Cj &= (d_{1j} + \ldots + d_{oj})\,/o
\end{aligned}
\tag{11}
$$

Definition 9. *(patient characteristics conversion on randomly sampled data set): Let* $D2$ *be the patient characteristics data set with t elements of* $1 \times m$ *matrix, which are converted by Apc from randomly sampled data set* $D1$*'s components. Let* D_{rspc} *i be i th element of* $D2$*. Let* $P_{RSPCi}Cj$ *be the j th element of* $D_{rspc}i$*.* $D2$ *and* D_{rspc} *i are expressed as follows.*

$$
\begin{aligned}
D2 &= \{D_{rspc}1, \ldots, D_{rspc}t\} \\
&= \{Apc(RS1(D0)), \ldots, Apc(RSt(D0))\} \\
D_{rspc}i &= \{P_{RSPCi}C1, \ldots, P_{RSPCi}Cm\}
\end{aligned}
\tag{12}
$$

5 Analyst's estimation example

5.1 Analyst's situation and goal

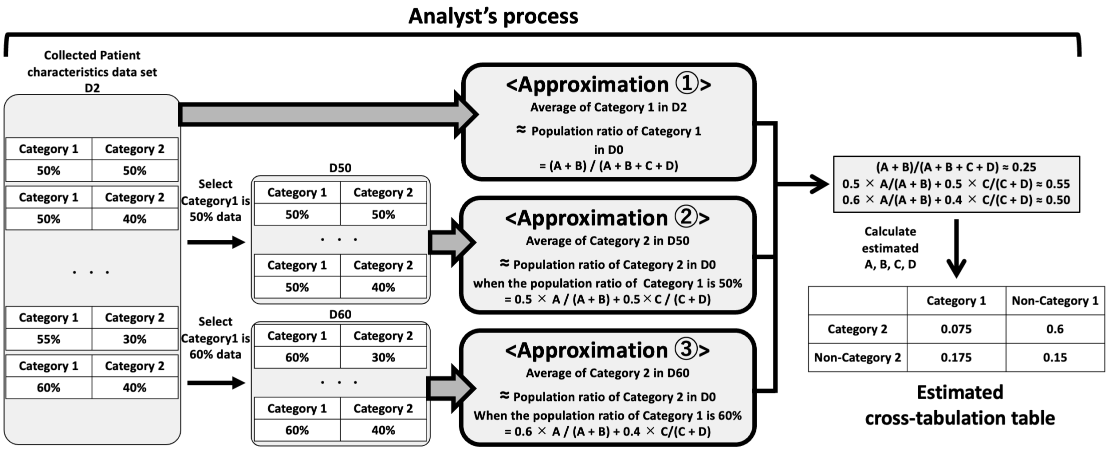
As shown in Fig. 4 left side, the analyst has $D2$ containing many patient characteristics with two categories, category 1 and category 2. Table. 5 is the category 1 vs. category 2 cross-tabulation table of $D0$. On Table. 5, A, B, C, D are the ratio of "category 1 and category 2," "category 1 and non-category 2," "non-category 1 and category 2," "non-category 1 and non-category 2" respectively. The analyst's goal is to estimate A, B, C, D. A, B, C, D can be expressed as follows.

$$
\begin{aligned}
&A = (A/(A+B)) \times (A+B)/(A+B+C+D) \\
&B = (1 - A/(A+B)) \times (A+B)/(A+B+C+D) \\
&C = C/(C+D) \times (1 - (A+B)/(A+B+C+D)) \\
&D = (1 - C/(C+D)) \times (1 - (A+B)/(A+B+C+D))
\end{aligned}
\tag{13}
$$

5.2 Estimate category 1 population ratio

As shown in Fig. 4 approximation ①, the average value of category 1 in $D2$ gets close to the population ratio of category 1 in $D0$ $(= (A+B)/(A+B+C+D))$ by LLN [18], as the patient characteristics are increased through random sampling. Then, if the average value of category 1 in $D2$ is 25%, the analyst can estimate as follows.

$$
(A+B)/(A+B+C+D) \approx 0.25 \tag{14}
$$


Analyst's process
Collected Patient characteristics data set D2
Category 1
Category 2
50%
50%
Category 1
Category 2
50%
40%
. . .
Category 1
Category 2
55%
30%
Category 1
Category 2
60%
40%
Select Category1 is 50% data
D50
Category 1
Category 2
50%
50%
. . .
Category 1
Category 2
50%
40%
Select Category1 is 60% data
D60
Category 1
Category 2
60%
30%
. . .
Category 1
Category 2
60%
40%
<Approximation ①>
Average of Category 1 in D2
≈ Population ratio of Category 1 in D0
= (A + B) / (A + B + C + D)
<Approximation ②>
Average of Category 2 in D50
≈ Population ratio of Category 2 in D0 when the population ratio of Category 1 is 50%
= 0.5 × A / (A + B) + 0.5 × C / (C + D)
<Approximation ③>
Average of Category 2 in D60
≈ Population ratio of Category 2 in D0 When the population ratio of Category 1 is 60%
= 0.6 × A / (A + B) + 0.4 × C/(C + D)
(A + B)/(A + B + C + D) ≈ 0.25
0.5 × A/(A + B) + 0.5 × C/(C + D) ≈ 0.55
0.6 × A/(A + B) + 0.4 × C/(C + D) ≈ 0.50
Calculate estimated A, B, C, D
Estimated cross-tabulation table

	Category 1	Non-Category 1
Category 2	0.075	0.6
Non-Category 2	0.175	0.15

Figure 4: Example of the analyst's estimation. The analyst has patient characteristics data set $D2$ containing two categories of statistics. The analyst has selected 50% and 60% of category 1's patient characteristics from data set $D2$ and has denoted them as $D50$ and $D60$, respectively. The analyst makes three approximations by the Law of Large Numbers (LLN) [18]. First is "average value of category 1 in $D2$ gets close to the population ratio of category 1 in $D0$ $(= (A+B)/(A+B+C+D))$." The second and third approximations are "average value of category2 in $D50$ gets close to the population ratio of category 2 in $D0$ when category1 is 50%," and "average value of category 2 in $D60$ gets close to the population ratio of category 2 in $D0$ when category 1 is 60% ." Then by three approximation equations, the analyst estimates the cross-tabulation table contains A, B, C, D category ratios.

Table 5: Cross-tabulation table of category 1 vs category 2 created from $D0$. A, B, C, D are each category's ratio.

	Category 1	Non-Category 1
Category 2	A	C
Non-Category 2	B	D

5.3 Estimate category 2 population ratio on two percentage cases of category 1

As shown in Fig. 4 middle, the analyst selects category 1 value is 50% patient characteristics from $D2$ and denotes this data set as $D50$. And the analyst also selects category 1 value is 60% patient characteristics from $D2$ and denotes this data set as $D60$.

As shown in Fig. 4 approximation ② and approximation ③, if the people in $D0$ are chosen as the category 1 population is 50%, the population rate of category 2 becomes $0.5 \times A/(A+B) + 0.5 \times C/(C+D)$. $D50$ gets close to the value by LLN [18] as more patient characteristics are collected through random sampling. Similarly, $D60$ get close to $0.6 \times A/(A+B) + 0.4 \times C/(C+D)$ by LLN [18]. If $D50$ is 0.55 and $D60$ is 0.5, the approximations are as follows.

$$\begin{aligned} 0.5 \times A/(A+B) + 0.5 \times C/(C+D) &\approx 0.55 \\ 0.6 \times A/(A+B) + 0.4 \times C/(C+D) &\approx 0.50 \end{aligned} \tag{15}$$

By solving this simultaneous equation,

$$A/(A+B) \approx 0.3, \quad C/(C+D) \approx 0.8 \tag{16}$$

5.4 Estimate cross-tabulation table

As Fig. 4 right side, at last, the analyst calculates estimated A, B, C, D by (13), (14), (16).

$$(A, B, C, D) \approx (0.075, 0.175, 0.6, 0.15) \tag{17}$$

6 Analyst's estimation model

Fig. 5 is the analyst's estimation model. The analyst has $D2$ containing t patient characteristics with two categories, C1 and C2. Table 6 is the cross-tabulation table of C1 vs. C2 created from $D0$ containing A, B, C, D. Let AA, BB, CC, DD be the replaced values of A, B, C, D in the approximate formulas described below by LLN. Let EE, AA', CC' as follows.

$$EE = (AA + BB)/(AA + BB + CC + DD), AA' = AA/(AA + BB), CC' = CC/(CC + DD) \tag{18}$$

By (18), AA, BB, CC, DD can be expressed as follows.

$$\begin{gathered} AA = AA' \times EE, BB = (1 - AA') \times EE \\ CC = CC' \times (1 - EE), DD = (1 - CC') \times (1 - EE) \end{gathered} \tag{19}$$

From (19), the values of AA, BB, CC, and DD are obtained from EE, AA', and CC'.

Let $Pr[C1\ in\ D2]$ be the average ratio of C1 in D2. Let $Pr[C1\ in\ D0]$ be the population ratio of C1 in D0. As Fig. 5 approximation ①, $Pr[C1\ in\ D2]$ gets close to $Pr[C1\ in\ D0]$.

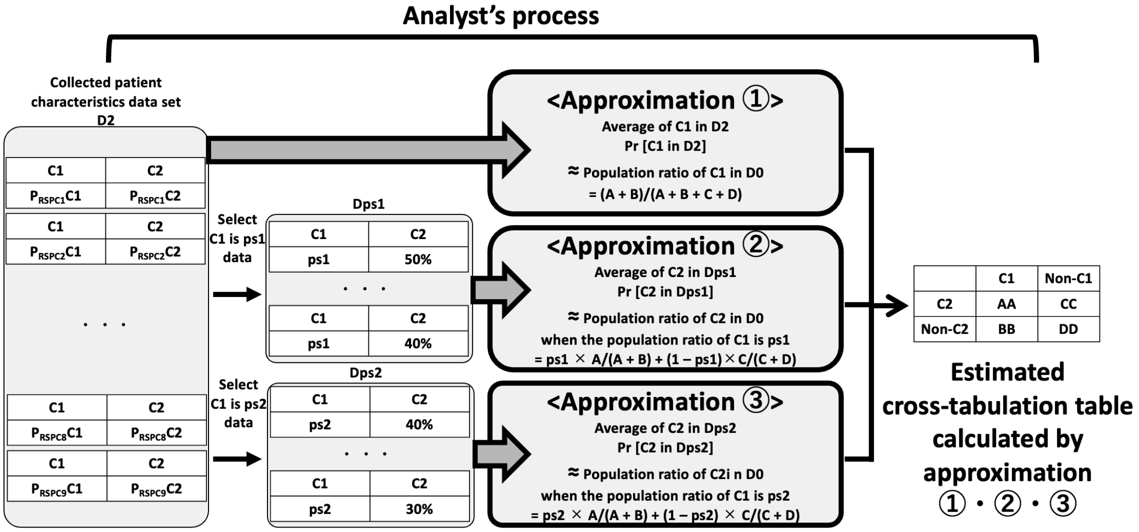


Figure 5: Analyst's estimation model using three approximations by the Law of Large Numbers (LLN) [18]. $D2$ is the collected patient characteristics data set. $Dps1$ is selected from $D2$ with C1 ratio is $ps1$. $Dps2$ is selected from $D2$ with C1 ratio is $ps2$. First approximation is "average value of category 1 in $D2$ ($= Pr[C1\ in\ D2]$) gets close to the population ratio of C1 in $D0$ ($= (A + B)/(A + B + C + D)$)." Second and third approximations are "average value of C2 in $Dps1$ ($= Pr[C2\ in\ Dps1]$) gets close to the population ratio of C2 in $D0$ when the population ratio of C1 is $ps1$ ($= ps1 \times A/(A + B) + (1 - ps1) \times C/(C + D)$)" and "average value of C2 in $Dps2$ ($= Pr[C2\ in\ Dps2]$) gets close to the population ratio of C2 in $D0$ when the population ratio of C1 is $ps2$ ($= ps2 \times A/(A + B) + (1 - ps2) \times C/(C + D)$)." By three approximation, the analyst gets AA, BB, CC, DD.

Table 6: Cross-tabulation table created from $D0$.

	C1	Non-C1
C2	A	C
Non-C2	B	D

$$Pr[C1\ in\ D2] \approx Pr[C1\ in\ D0] = (A+B)/(A+B+C+D) \tag{20}$$

LLN converts A, B, C, D to AA, BB, CC, DD because the patient characteristics are random sampled data.

$$Pr[C1\ in\ D2] = (AA+BB)/(AA+BB+CC+DD) = EE \tag{21}$$

From (21), the analyst gets the value of EE because the analyst can know $Pr[C1\ in\ D2]$.

Let Dps1 be the data set of selected data from D2 with a ratio of C1 is ps1. Let Dps2 be the data set of selected data from D2 with a ratio of C1 is ps2. Let $Pr[C2\ in\ Dps1]$ be the average ratio of C2 in Dps1. Let $Pr[C2\ in\ D0\ when\ C1\ is\ ps1]$ be the population ratio of C2 in D0 when the population ratio of C1 is ps1. Let $Pr[C2\ in\ Dps2]$ be the average ratio of C2 in Dps2. Let $Pr[C2\ in\ D0\ when\ C1\ is\ ps2]$ be the population ratio of C2 in D0 when the population ratio of C1 is ps2. As shown in Fig. 5 approximation ② and approximation ③, $Pr[C2\ in\ Dps1]$ gets close to $Pr[C2\ in\ D0\ when\ C1\ is\ ps1]$, and $Pr[C2\ in\ Dps2]$ gets close to $Pr[C2\ in\ D0\ when\ C1\ is\ ps2]$.

$$\begin{aligned} Pr[C2\ in\ Dps1] &\approx Pr[C2\ in\ D0\ when\ C1\ is\ ps1] = ps1 \times A/(A+B) + (1-ps1) \times C/(C+D) \\ Pr[C2\ in\ Dps2] &\approx Pr[C2\ in\ D0\ when\ C1\ is\ ps2] = ps2 \times A/(A+B) + (1-ps2) \times C/(C+D) \end{aligned} \tag{22}$$

The conversion of A, B, C, D into AA, BB, CC, DD is achieved through the application of LLN on the randomly sampled patient characteristics data.

$$\begin{aligned} Pr[C2\ in\ Dps1] &= ps1 \times AA/(AA+BB) + (1-ps1) \times CC/(CC+DD) \\ &= ps1 \times AA' + (1-ps1) \times CC', \\ Pr[C2\ in\ Dps2] &= ps2 \times AA/(AA+BB) + (1-ps2) \times CC/(CC+DD) \\ &= ps2 \times AA' + (1-ps2) \times CC' \end{aligned} \tag{23}$$

By (23), AA' and CC' can be expressed as follows.

$$\begin{aligned} AA' &= Pr[C2\ in\ Dps1] \times (ps2-1)/(ps2-ps1) - Pr[C2\ in\ Dps2] \times (ps1-1)/(ps2-ps1), \\ CC' &= Pr[C2\ inDps1] \times ps2/(ps2-ps1) - Pr[C2\ in\ Dps2]) \times ps1/(ps2-ps1) \end{aligned} \tag{24}$$

From (24), the analyst gets the value of AA' and CC' because the analyst can know $Pr[C2\ in\ Dps1]$, $Pr[C2\ in\ Dps2]$, $ps1$ and $ps2$ values.

In conclusion, by (21) and (24), the analyst gets EE, AA', CC' values. Then, the analyst gets AA, BB, CC and DD values by (19) as Fig. 5 right side.

7 Theoretical estimation error and anonymity evaluation

7.1 Theoretical estimation error equation

For numerical analysis, we derive the theoretical formula for estimating error. As Fig. 5 approximation ①, EE is estimated from $D2$, while AA' and CC' are estimated from $Dps1$ and $Dps2$ extracted from $D2$. Therefore, the error in EE is negligible compared to the error in AA' and CC'. As Fig. 5 approximation ② and approximation ③, the analyst estimates $Pr[C2\ in\ D0\ when\ C1\ is\ ps1]$

and $Pr[C2\ in\ D0\ when\ C1\ is\ ps2]$. Let $\hat{p}1 = Pr[C2\ in\ Dps1]$, $\hat{p}2 = Pr[C2\ in\ Dps2]$, $p1 = Pr[C2\ in\ D0\ when\ C1\ is\ ps1]$, $p2 = Pr[C2\ in\ D0\ when\ C1\ is\ ps2]$. Let the estimation errors of $p1$ and $p2$ be $\varepsilon1$ and $\varepsilon2$. Then, $\hat{p}1 = p1 + \varepsilon1$, $\hat{p}2 = p2 + \varepsilon2$. By (24) and (19), AA is as follows.

$$\begin{aligned} AA &= EE \times (-p1 + p2 + ps2 \times p1 - ps1 \times p2)/(ps2 - ps1) \\ &+ EE \times (-\varepsilon1 + \varepsilon2 + ps2 \times \varepsilon1 - ps1 \times \varepsilon2)/(ps2 - ps1) \end{aligned} \tag{25}$$

The terms including $\varepsilon1$ and $\varepsilon2$ in (25) are errors as follows.

$$EE \times (-\varepsilon1 + \varepsilon2 + ps2 \times \varepsilon1 - ps1 \times \varepsilon2)/(ps2 - ps1) \fallingdotseq EE \times (-\varepsilon1 + \varepsilon2)/(ps2 - ps1) \tag{26}$$

Let $p3 = p2 - p1$, and let $\hat{p}3$ be the observed value of an event with probability $p3$ when observed with the same number of observations as $\hat{p}1$ or $\hat{p}2$. Let $\varepsilon3$ is the error of $p3$. Then, we can denote as $\hat{p}3 = p3 + \varepsilon3$, and we can get the following equation.

$$\varepsilon3 = \hat{p}3 - p3 \tag{27}$$

From $\hat{p}1 = p1 + \varepsilon1$ and $\hat{p}2 = p2 + \varepsilon2$, we can denote $\varepsilon2 - \varepsilon1$ as follows.

$$\varepsilon2 - \varepsilon1 = \hat{p}2 - p2 - (\hat{p}1 - p1) = \hat{p}2 - \hat{p}1 - (p2 - p1) \tag{28}$$

By (27), (28) and $p3 = p2 - p1$, we can calculate as follows.

$$\varepsilon3 - (\varepsilon2 - \varepsilon1) = \hat{p}3 - p3 - (\hat{p}2 - \hat{p}1) + (p2 - p1) = \hat{p}3 - (\hat{p}2 - \hat{p}1) \tag{29}$$

In many samples, $\hat{p}3$ gets close to $(\hat{p}2 - \hat{p}1)$ and (29) gets close to 0. This results in the following.

$$\varepsilon3 \approx (\varepsilon2 - \varepsilon1) \tag{30}$$

In conclusion, (26) can be approximated as follows.

$$EE \times (-\varepsilon1 + \varepsilon2)/(ps2 - ps1) \fallingdotseq EE \times (\varepsilon3)/(ps2 - ps1) \tag{31}$$

7.2 Theoretical estimation error evaluation

To know the theoretical range of estimation error, we examine the range of values that (31) could potentially take. From the (31), EE, $\varepsilon3$ and $(ps2 - ps1)$ determine the error of AA.

Regarding EE, it is reasonable to assume that $EE = 0.1 \sim 0.9$. That is because EE is a population ratio of a particular category of patient characteristics and the value should not be extremely low or high for estimation in a reasonable sample size to adapt LLN. Regarding $(ps2-ps1)$, it can be arbitrarily chosen for each EE, thus it can be fixed at $ps2 - ps1 = 0.08$.

Regarding $\varepsilon3$, we can consider the $p3$ 95% CI as the theoretical range of $\varepsilon3$ values, and the 95% CI is determined by (7). From (7), we can calculate the $\varepsilon3$ as follows.

$$\varepsilon3 = 1.96\sqrt{\frac{p3\ (1 - p3)}{n3}} \tag{32}$$

Where $n3$ is $p3$ sample size in (32).

Regarding $p3(= p2 - p1)$ in (32), we can assume that the value is small. This is because $p1$ and $p2$ are the ratios of C2 corresponding to C1 at ps1 and ps2, respectively. With $ps2 - ps1 = 0.08$, it is expected that $p2 - p1$ would be a small value. We can consider the case when $p3 = p2 - p1 = 0.01 \sim 0.03$. Regarding $n3$ in (32), if $p2 - p1(= p3)$ is $0.01 \sim 0.03$ and $10,000 \sim 100,000$ patient characteristics with 50 people are collected by the analyst, the size of $p2$ or $p1$ is about $1,000 \sim 10,00$ based on a normal distribution. Then, we can consider $n3 = 1,000 \sim 10,000$.

Fig. 6 shows the theoretical range of $\varepsilon3$ values on $p3 = p2 - p1 = 0.01 \sim 0.03$ and $n3 = 1,000 \sim 10,000$ for (32). From Fig. 6, the reasonable theoretical range of $\varepsilon3$ values is about $0.002 \sim 0.010$. Fig. 7 shows the theoretical error on $EE = 0.1 \sim 0.9$, $\varepsilon3 = 0.002 \sim 0.010$, and $ps2 - ps1 = 0.08$ for

(31). As shown in Fig. 6, by taking $n3$ to be around 10,000, $0.002 \leqq \varepsilon 3 \leqq 0.004$ can be obtained. This indicates that with $0.1 \leqq EE \leqq 0.9$, the theoretical error can be kept below 0.05 (5%) as shown in Fig. 7. Additionally, to keep the theoretical error below 0.02 (2%), EE must be $EE \leqq 0.4$.

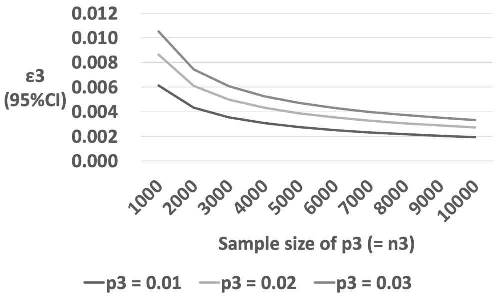


Figure 6: Theoretical effect of sample size of $p3$ on $\varepsilon 3$ for $p3$ levels of 0.01, 0.02, and 0.03.

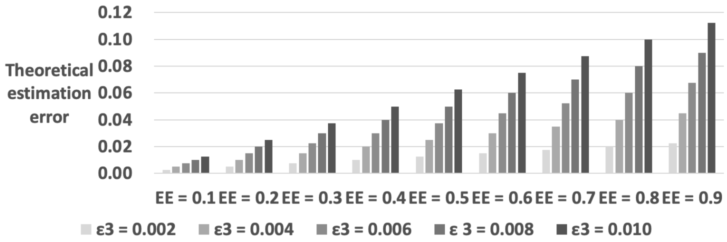


Figure 7: Theoretical estimation error for EE ranging from 0.1 to 0.9 and $\varepsilon 3$ levels of 0.002, 0.004, 0.006, 0.008, and 0.010.

7.3 Theoretical anonymity equation

In the field of PPDM, it is important to quantify and evaluate anonymity. The risk of anonymity violations from the PFDOC attack can arise from the analyst collecting a large number of patient characteristics because there could be low PFDOC entropy patient characteristics. Particularly, when a patient characteristic with PFDOC entropy = 0 is included in the multiple patient characteristics collected by the analyst, the risk of anonymity violation is maximized. Therefore, the aim is to investigate how many patient characteristics with PFDOC entropy = 0 are contained in the multiple patient characteristics collected by the analyst. The anonymity of patient characteristics is affected by the ratio of categories in the target population, the number of patients in the patient characteristics, and the number of patient characteristics collected by the analyst. Let p be the ratio of individuals with a certain category, q be the number of patients in the patient characteristics, and R be the number of patient characteristics collected by the analyst. The probability that all q patients in a patient characteristics belong to a specific category is p^q, while the probability that no individual belongs to that category is $(1-p)^q$. In this case, when all the patient characteristics given belong or do not belong to a specific category, the PFDOC entropy becomes zero. Therefore, when the analyst collects R patient characteristics, the number of PFDOC entropy = 0 patient characteristics is approximated as follows.

$$R \times p^q \text{ or } R \times (1-p)^q \tag{33}$$

7.4 Theoretical anonymity evaluation

We aim to consider the impact of the criteria set by the analyst for collecting patient characteristics on the level of anonymity. Therefore, first, we aim to identify the variables that the analyst can control and those that cannot be controlled. In (33), the analyst has control over the number of patient characteristics R, as well as the number of individuals q included in each patient characteristic. However, the population ratio of categories p included in the patient characteristics is determined by the publisher's selection of categories contained in the patient characteristics and cannot be controlled by the analyst. Based on the discussion above, our objective can be rephrased to evaluate the level of anonymity that is preserved when the analyst chooses the criteria by R and q for collecting the patient characteristics. Therefore, we calculate the theoretical number of PFDOC entropy $= 0$ patient characteristics assuming $R = 10,000$ and $q = 50$, which are within the control of the analyst. In other words, we examine the values of $10000p^{50}$ or $10000(1-p)^{50}$ for each p.

Fig. 8 shows the PFDOC entropy $= 0$ patient characteristics in 10,000 patient characteristics with 50 individuals by each p. Fig. 9 a, b, c, and d are expanded Fig. 8 for $0.00 \leqq p \leqq 0.10$, $0.10 \leqq p \leqq 0.25$, $0.75 \leqq p \leqq 0.90$, and $0.90 \leqq p \leqq 1.00$ respectively. Note that, in $0.25 \leqq p \leqq 0.75$, the analyst acquires almost zero PFDOC entropy $= 0$ patient characteristics. From Fig. 8 observation, if the publisher uses $0.25 \leqq p \leqq 0.75$ categories, the anonymity is safe for the PFDOC attack, and if the publisher uses $0.20 \leqq p \leqq 0.25$ and $0.70 \leqq p \leqq 0.75$ categories, the anonymity is approximately safe for the PFDOC attack. But if the publisher uses $0 \leqq p \leqq 0.20$ and $0.80 \leqq p \leqq 1.00$ categories, the anonymity is not safe because the PFDOC attack could succeed.

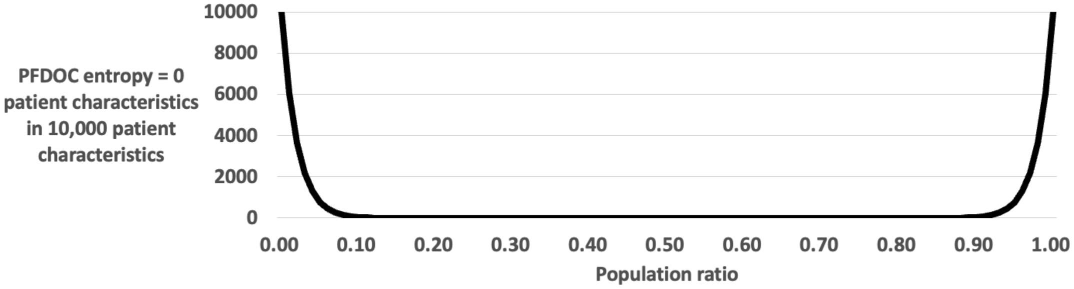


Figure 8: Theoretical Patient Family Detect on Overall Category (PFDOC) entropy $= 0$ patient characteristics occurrence number in 10,000 patient characteristics for category's population ratio p. Note that each patient characteristics contains 50 patients' data.

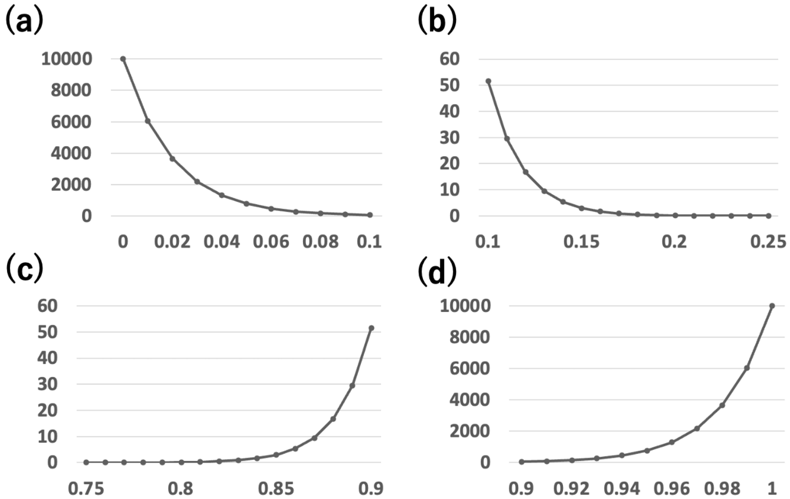
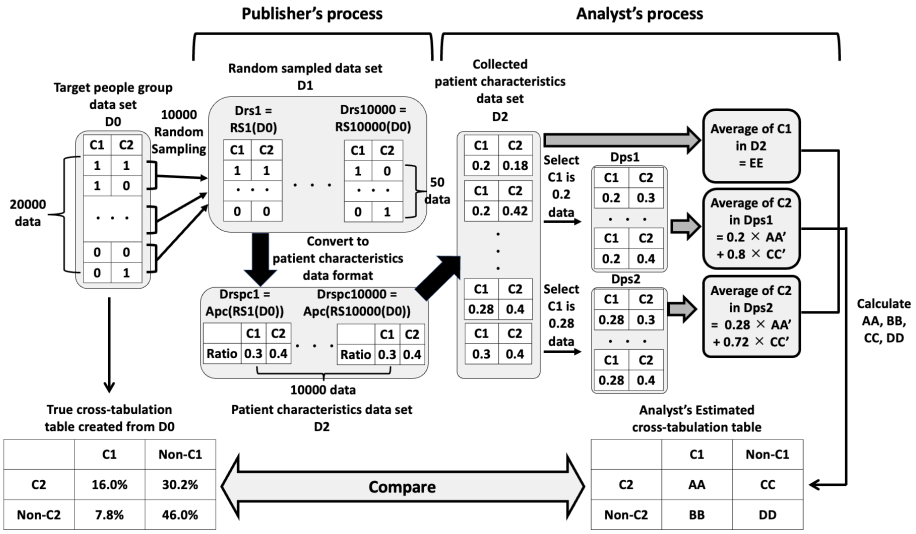


Figure 9: Magnified Fig. 8. a, population ratio 0.00 to 0.10. b, population ratio 0.10 to 0.25. c, population ratio 0.75 to 0.90. d, population ratio 0.90 to 1.00.

8 Experimental estimation error and anonymity evaluation

8.1 Experimental method

Fig. 10 shows the experiment overview. The notation of variables follows sections 4 and 6. We use 20,000 US Census [9] data as target people group data set $D0$. We use the data set's categories of fin_flag as C1 and age, education or marital-status as C2 and convert to binary data as Table 7. As the publisher's process, we do random sampling with sample size $o = 50$ and do $t = 10,000$ times. Random sampled data set is $D1 = \{RS1(D0), \ldots, RS10000(D0)\}$ and $RSi(D0) = \{r_{RSi}1, \ldots, r_{RSi}50\}$. $D1$ is converted to patient characteristics as $D2 = \{Apc(RS1(D0)), \ldots, Apc(RS10000(D0))\}$. As the analyst's process, we make three approximated equations by $ps1 = 0.2$, $ps2 = 0.28$, and estimate the cross-tabulation table. We compare the estimated cross-tabulation table from $D2$ with the true cross-tabulation table created from $D0$. To evaluate the standard deviation of the estimated error and the PFDOC entropy-based anonymity, we perform $n = 100$ experiments.


Publisher's process
Analyst's process
Target people group data set D0
20000 data
C1 C2
1 1
1 0
0 0
0 1
10000 Random Sampling
Random sampled data set D1
Drs1 = RS1(D0)
Drs10000 = RS10000(D0)
C1 C2
1 1
0 0
1 0
0 1
50 data
Convert to patient characteristics data format
Drspc1 = Apc(RS1(D0))
Drspc10000 = Apc(RS10000(D0))
Ratio 0.3 0.4
10000 data
Patient characteristics data set D2
Collected patient characteristics data set D2
0.2 0.18
0.2 0.42
0.28 0.4
0.3 0.4
Select C1 is 0.2 data
Select C1 is 0.28 data
Dps1
0.2 0.3
0.2 0.4
Dps2
0.28 0.3
0.28 0.4
Average of C1 in D2 = EE
Average of C2 in Dps1 = 0.2 × AA' + 0.8 × CC'
Average of C2 in Dps2 = 0.28 × AA' + 0.72 × CC'
Calculate AA, BB, CC, DD
True cross-tabulation table created from D0
Compare
Analyst's Estimated cross-tabulation table

	C1	Non-C1
C2	16.0%	30.2%
Non-C2	7.8%	46.0%

	C1	Non-C1
C2	AA	CC
Non-C2	BB	DD

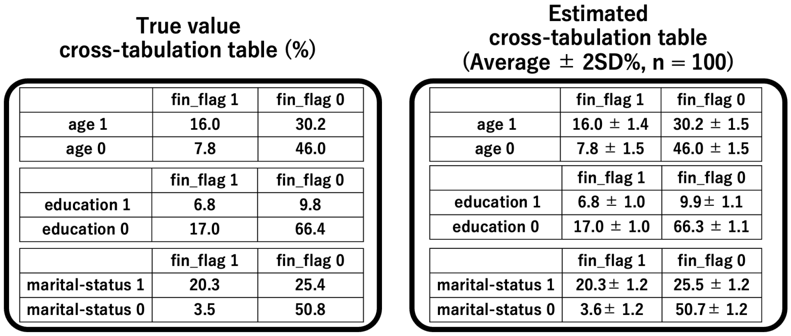
Figure 10: Experiment overview. The target people group data set D0 is US Census [9] 20,000 data, where C1 is fin_flag, and C2 is age, education, or marital-status. The publisher's processes are 10,000 random samplings with a size of 50 and conversion to patient characteristics formatted data. The analyst's process is making three approximated equations from $D2$ and calculating the equations to get the estimated cross-tabulation table. The estimated cross-tabulation table is compared with the true cross-tabulation table created from $D0$. The notation of variables follows sections 4 and 6.

Table 7: Binary allocation for variables of US Census [9].

	1	0
fin_flag	>50K	<= 50K
age	>38	<= 38
education	Bachelors, Masters, Some-college, Assoc-acdm, Assoc-voc, Doctorate, Prof-schoo	11th, HS-grad, 9th, 7th-8th, 12th,1st-4th, 10th,5th-6th, Preschoo
marital-status	Married-civ-spouse, Married-spouse-absent, Married-AF-spouse	Divorced, Never-married, Separated, Widowed

8.2 Experimental estimation error evaluation

Fig. 11 shows the true value and the experimental estimation result. We can compare this experimental result with theoretical value by applying experimental value on (31). EE is the sample ratio of fin_flag and $EE = 0.24$. $ps2 - ps1$ is $ps2 - ps1 = 0.28 - 0.20 = 0.08$. Furthermore, in fin_flag vs. age case, $p2 - p1 \fallingdotseq 0.01$ and the sample size of $ps1$ or $ps2$ $(= n3)$ is about 1,000 patient characteristics, leading to $\varepsilon 3 \fallingdotseq 0.006(95\%CI)$ according to (32). Then, the theoretical estimation error is $EE \times (\varepsilon 3)/(ps2 - ps1) = 1.8\%(95\%CI)$. The experimental values were within 1.5% ($2SD$, $n = 100$) and were found to fall within the $95\%CI$ of the theoretical value. This result indicates consistency between the experimental and theoretical results.



True value cross-tabulation table (%)

	fin_flag 1	fin_flag 0
age 1	16.0	30.2
age 0	7.8	46.0

	fin_flag 1	fin_flag 0
education 1	6.8	9.8
education 0	17.0	66.4

	fin_flag 1	fin_flag 0
marital-status 1	20.3	25.4
marital-status 0	3.5	50.8

Estimated cross-tabulation table (Average ± 2SD%, n = 100)

	fin_flag 1	fin_flag 0
age 1	16.0 ± 1.4	30.2 ± 1.5
age 0	7.8 ± 1.5	46.0 ± 1.5

	fin_flag 1	fin_flag 0
education 1	6.8 ± 1.0	9.9± 1.1
education 0	17.0 ± 1.0	66.3 ± 1.1

	fin_flag 1	fin_flag 0
marital-status 1	20.3± 1.2	25.5 ± 1.2
marital-status 0	3.6± 1.2	50.7± 1.2

Figure 11: True cross-tabulation tables created from raw data and estimated cross-tabulation tables (average ± 2 standard deviations (SD) ($n = 100$)).

Fig. 12 shows the absolute error distributions for fin_flag vs. age, fin_flag vs. education, and fin_flag vs. marital-status estimations in $n = 100$ experiments. The results of fin_flag vs. age and fin_flag vs. education do not contain over 2.0 % absolute error result, and the results of fin_flag vs. marital-status contain one over 2.0 % absolute error result ($n = 100$). No significant deviation from the theoretical value of 1.8%(95%CI) was observed in the collected data.

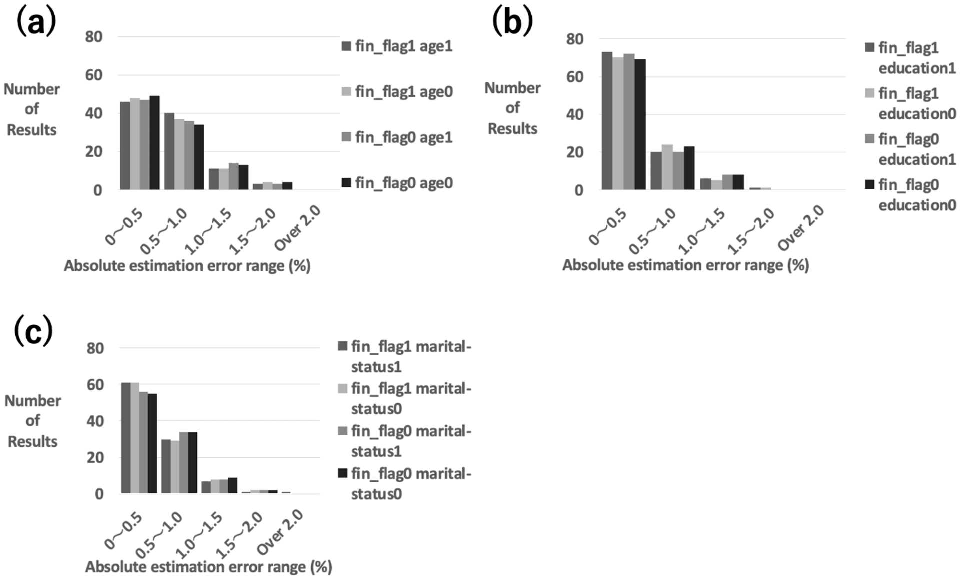


Figure 12: Number of results within the absolute estimation error range for fin_flag vs. age, fin_flag vs. education, fin_flag vs. marital-status (n = 100).

8.3 Experimental anonymity evaluation

Fig. 13 shows the experimental values of how many PFDOC entropy = 0 patient characteristics occur if the analyst gets 10,000 patient characteristics that contain 50 people in 100 experiments. Age has a 46.2% population rate and the proportion of having zero patient characteristics with PFDOC entropy = 0 is 100%. Education has a 16.6% population ratio and the results of 100 experiments showed that the analyst obtained 0 PFDOC entropy = 0 patient characteristics in 28% of cases, 1 in 39% of cases, 2 in 21% of cases, 3 in 9% of cases, 4 in 3% of cases. No cases were recorded with 5 or more PFDOC entropy = 0 patient characteristics. The average number of PFDOC entropy = 0 patient characteristics obtained by the analyst was 1.2. Marital-status has a 45.7% population rate and the proportion of having zero patient characteristics with PFDOC entropy = 0 is 100%. fin_flag has a 23.8% population rate and the proportion of having zero patient characteristics with PFDOC entropy = 0 is 100%.

We can compare this experimental value with the theoretical value shown in Fig. 8. Theoretically, the occurrence of patient characteristics with PFDOC entropy = 0 is rare when $0.25 \leqq p \leqq 0.75$. $0.20 \leqq p \leqq 0.25$ and $0.70 \leqq p \leqq 0.75$ are considered to be approximately safe. Furthermore, $0 \leqq p \leqq 0.20$ and $0.80 \leqq p \leqq 1.00$ are vulnerable to a successful PFDOC attack. This theoretical conclusion was also confirmed through the experiment.

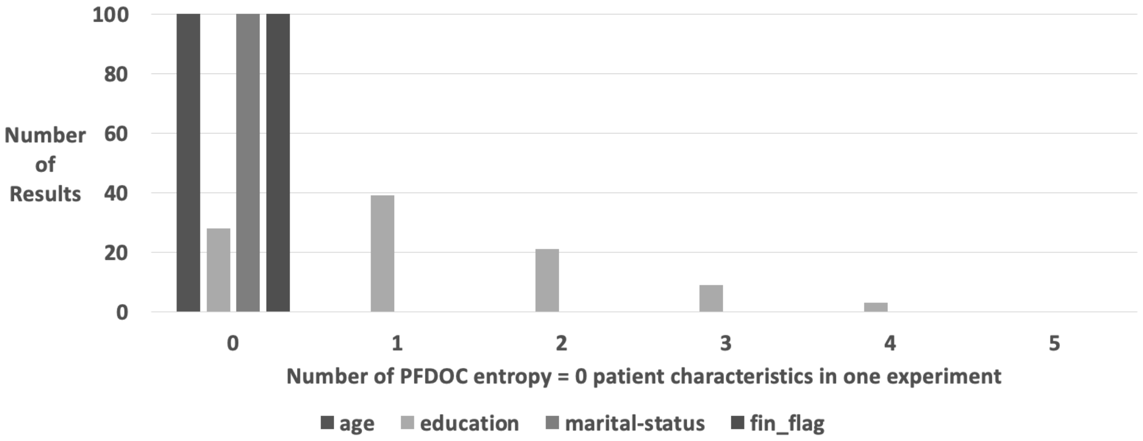


Figure 13: Experimental number of Patient Family Detect on Overall Category (PFDOC) entropy = 0 patient characteristics out of 10,000 patient characteristics with 50 patients each for age, education, marital-status, and fin_flag category. Age (46.2% population ratio), marital-status (45.7% population ratio), and fin_flag (23.8% population ratio) have 100% results of 0 PFDOC entropy = 0 patient characteristics out of 10,000 patient characteristics in 100 experiments. Education (16.6% population ratio) has an average of 1.2 PFDOC entropy = 0 patient characteristics out of 10,000 patient characteristics in 100 experiments.

9 Discussion

First, we confirm that our estimation method can reduce the $95\%CI$ estimation error by any desired amount. As seen from (31) and (32), the estimation error decreases as the size of $n3$ increases. This indicates that obtaining additional patient characteristics will enable the analyst to minimize the estimation error to an arbitrarily small value.

Second, we evaluate experimental estimation error as a medical study. The experimental estimation error in section 8 is within 1.5% ($2SD$). In traditional medical research, an error of 1.5% ($2SD$ = 95% CI [22]) in a normal distribution is an error of about 1,000 cases. Since there are four categories in the cross-tabulation table, the 1.5% error ($2SD$) is equivalent to an error of 4,000 cases in traditional medical research. Medical analysis cases using cross-tabulation tables range from less than 100 cases to tens of thousands of cases [13] [15]. The estimation error of our method is equivalent to that of medical research methods, which guarantees the usefulness of the proposed estimation.

Third, we discuss anonymity when adopting our proposal estimation method. From the anonymity result of Fig. 8 and Fig. 9, the category's population ratio p which has $0 \leqq p \leqq 0.25$ or $0.75 \leqq p \leqq 1.00$ is vulnerable to the PFDOC attack. However, Table 2 also shows that some of the patient characteristics categories can fall within these ranges. This means that some categories in Table 2 have the risk of anonymity violation. Therefore, it is recommended to choose categories within $0.25 \leqq p \leqq 0.75$ when applying the proposed estimation method to ensure anonymity protection.

Fourth, we consider the impact of extreme probability categories that could be included in the patient characteristics. In the experiment, we could observe the effect of errors on the true value of 3.5 to 66.4%. However, extreme probability cases can also be included in the patient characteristics, such as when the true value is less than 3.5%. In such instances, additional random samples of patient characteristics may be required in order to reach the desired statistical probability before applying LLN in the proposal estimation method.

10 Conclusion and future work

We proposed to use patient characteristics, medical statistics that would not be concerned by the GDPR, to estimate the cross-tabulation table, which is usually generated from personal information in medical research and widely used for the analysis of medical variables. To quantify the proposed method, we modeled a publisher of randomly sampled patient characteristics and the analyst estimating cross-tabulation tables. In this model, we theoretically evaluated the effectiveness of estimating multiple patient characteristics. For quantitative PPDM, we also theoretically evaluated anonymity as a vulnerability to the PFDOC attack by PFDOC entropy = 0 patient characteristics occurrence rate in multiple patient characteristics. Furthermore, we confirmed that the effectiveness and anonymity of the estimation method are consistent with the theoretical evaluation. For effectiveness, we confirmed in the experiment that the estimation can be made with an error of 1.8% (95% CI) using 10,000 patient characteristics with 50 patients each. For anonymity, though the analyst can get patient characteristics as our model without legal problems, we showed that using categories within the range of 25% to 75% population ratio in our proposal estimation method ensures safety from the risk of the PFDOC entropy = 0 patient characteristics occurrence.

As future work, we need to create a more efficient model, examine whether the method can be extended to general privacy-preserving data mining fields, and find methods to mitigate the impact of large amounts of statistical information on anonymity.

References

[1] Ralf Bender. Introduction to the use of regression models in epidemiology. *Cancer Epidemiology*, 1:179–195, 2009.

[2] Ralf Bender and Ulrich Grouven. Ordinal logistic regression in medical research. *Journal of the Royal College of physicians of London*, 31(5):546, 1997.

[3] Ewan Cameron. On the estimation of confidence intervals for binomial population proportions in astronomy: the simplicity and superiority of the bayesian approach. *Publications of the Astronomical Society of Australia*, 28(2):128–139, 2011.

[4] Rebecca DerSimonian and Raghu Kacker. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary clinical trials*, 28(2):105–114, 2007.

[5] European Data Protection Supervisor (EDPS). Edps opinion on safeguards and derogations under article 89 gdpr in the context of a proposal for a regulation on integrated farm statistics, 2017.

[6] Avi J Hakim, Kerton R Victory, Jennifer R Chevinsky, Marisa A Hast, D Weikum, L Kazazian, S Mirza, R Bhatkoti, MM Schmitz, M Lynch, et al. Mitigation policies, community mobility, and covid-19 case counts in australia, japan, hong kong, and singapore. *Public Health*, 194:238–244, 2021.

[7] Mike Hintze. Viewing the gdpr through a de-identification lens: a tool for compliance, clarification, and consistency. *International Data Privacy Law*, 8(1):86–101, 2018.

[8] Kenta Kitamura, Mhd Irvan, and Rie Shigetomi Yamaguchi. Anonymity test attacks and vulnerability indicators for the “patient characteristics” disclosure in medical articles. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 186–193. IEEE, 2022.

[9] Ronny Kohavi and Barry Becker. Uci machine learning repository: adult data set. Retrieved may, 10, 2022 from https://archive.ics.uci.edu/ml/datasets/adult, 1996.

[10] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[11] Ricardo Mendes and João P Vilela. Privacy-preserving data mining: methods, metrics, and applications. *IEEE Access*, 5:10562–10582, 2017.

[12] M Hassan Murad, Noor Asi, Mouaz Alsawas, and Fares Alahdab. New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127, 2016.

[13] Akiko Nanri, Tohru Nakagawa, Keisuke Kuwahara, Shuichiro Yamamoto, Toru Honda, Hiroko Okazaki, Akihiko Uehara, Makoto Yamamoto, Toshiaki Miyamoto, Takeshi Kochi, et al. Development of risk score for predicting 3-year incidence of type 2 diabetes: Japan epidemiology collaboration on occupational health study. *PLoS One*, 10(11):e0142779, 2015.

[14] Yusuke Narita and Ayumi Sudo. Curse of democracy: Evidence from 2020. *Available at SSRN 3827327*, 2021.

[15] Bum J Park, Yoon J Kim, Dong H Kim, Won Kim, Yong J Jung, Jung H Yoon, Chung Y Kim, Young M Cho, Se H Kim, Kyoung B Lee, et al. Visceral adipose tissue area is an independent risk factor for hepatic steatosis. *Journal of gastroenterology and hepatology*, 23(6):900–907, 2008.

[16] Emanuela Podda. Shedding light on the legal approach to aggregate data under the GDPR & the FFDR. *Conference of European statisticians Expert Meeting on Statistical Data Confidentiality*, 2021.

[17] Iman Rahimi, Fang Chen, and Amir H Gandomi. A review on covid-19 forecasting models. *Neural Computing and Applications*, pages 1–11, 2021.

[18] Eugene Seneta. A tricentenary history of the law of large numbers. *Bernoulli*, 19(4):1088–1121, 2013.

[19] Ventura A Simonovich, Leandro D Burgos Pratx, Paula Scibona, María V Beruto, Marcelo G Vallone, Carolina Vázquez, Nadia Savoy, Diego H Giunta, Lucía G Pérez, Marisa del L Sánchez, et al. A randomized trial of convalescent plasma in covid-19 severe pneumonia. *New England Journal of Medicine*, 384(7):619–629, 2021.

[20] Magdalena Szumilas. Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227, 2010.

[21] Rohit Valecha, Shambhu Upadhyaya, and H Raghav Rao. An activity theory approach to leak detection and mitigation in patient health information (PHI). *Journal of the Association for Information Systems*, 22(4):6, 2021.

[22] Elise Whitley and Jonathan Ball. Statistics review 2: Samples and populations. *Critical Care*, 6(2):1–6, 2002.