Journal of Information Processing Vol.26 638–647 (Sep. 2018)

[DOI: 10.2197/ipsjjip.26.638]

Regular Paper

Privacy-Preserving Multiple Linear Regression of Vertically Partitioned Real Medical Datasets

Hiroaki Kikuchi[1,a)] Chika Hamanaga[1] Hideo Yasunaga[2] Hiroki Matsui[2] Hideki Hashimoto[2] Chun-I Fan[3,b)]

Received: November 27, 2017, Accepted: June 8, 2018

Abstract: This paper studies the feasibility of privacy-preserving data mining in epidemiological study. As for the data-mining algorithm, we focus on a linear multiple regression that can be used to identify the most significant factors among many possible variables, such as the history of many diseases. We try to identify the linear model to quantify the most significant cause of death from distributed dataset related to the patient and the disease information. In this paper, we have conducted an experiment using a real medical dataset related to a stroke and attempt to apply multiple regression with six predictors of age, sex, the medical scales, e.g., Japan Coma Scale, and the modified Rankin Scale. Our contributions of this paper include (1) to propose a practical privacy-preserving protocol for linear multiple regression with vertically partitioned datasets, (2) to show the feasibility of the proposed system using the real medical dataset distributed into two parties, the hospital who knows the technical details of diseases while patients are in the hospital, and the local government who knows the resident even after the patient has left hospital, (3) to show the accuracy and the performance of the PPDM system which allows us to estimate the expected processing time when an arbitrary number of predictors are used and (4) to study the complexity of the extended models of vertically partition.

Keywords: privacy, privacy-preserving data mining, epidemiology

1. Introduction

In a recent IT development, many kinds of electrical data are available. A smartphone keeps recording the location of its owner and a mobile device monitors our health conditions such as how many steps to walk per hour and how often we move in sleeping. These vital records allow performing epidemiological analysis easier. In hospital, every data related to patients are observed and collected to a central database for medical research. For instance, DPC dataset, which stands for Disease, Procedure and Combination, covers medical records for more than 7 million patients in more than 1,000 hospitals [1].

A privacy concern often prevents data sharing among institutes. Hospitals and medical centers have their own privacy policy that prohibits sharing data without concentration of individuals. Data protection regulation does not allow any organizations to disclose personal data in a way that any subject is identifiable. Anonymization of personal data helps mitigate the risk of identification and encourages data sharing among institutes [2]. However, there is no completely risk free anonymization method. There are data-mining and similar techniques to turn the anonymized data back into personal data.

Cryptography helps to preserve the privacy of personal data. The study, known as Privacy-Preserving Data Mining (PPDM), aims to perform a data mining algorithm with preserving confidentiality of datasets [3], [4], [6], [7], [8], [9]. Using some public-key encryption algorithms with some useful property for data mining, arbitrary algorithm is able to be performed. Tables consisting of medical records are either vertically or horizontally partitioned into partial tables owned by independent institutes, such as the hospital and the medical center. The issue of PPDM using public-key encryption [*1] is the large computational overhead in encryption. Many applications in big-data require a large scale dataset that is too large to be performed over ciphertexts.

In this paper, we study the protocol of privacy-preserving data mining in an epidemiological study. As for the data-mining algorithm, we focus a linear multiple regression because it allows to clarify what is the most significant factor related to the target death. There are some protocols for privacy-preserving linear regression from academic interests. However, there is no real application using the PPDM protocol in practical use. Hence, we aim to apply the proposed protocol to a real large scale medical dataset with more than 5,000 patients. This must be the first example used for PPDM to the real dataset of history of diseases.

Our contributions of this paper include (1) to propose a prac-

[1] Department of Frontier Media Science, School of Interdisciplinary Mathematical Sciences, Meiji University, Nakano, Tokyo 164–8525, Japan
[2] Graduate School of Medicine, The University of Tokyo, Bunkyo, Tokyo 113–8555, Japan
[3] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan
[a)] kikn@meiji.ac.jp
[b)] cifan@mail.cse.nsysu.edu.tw

The primary version of this paper was published in the IEEE 31st International Conference on Advanced Information Networking and Applications (AINA).

[*1] PPDM such as differential privacy does not require any cryptographical processing.

tical privacy-preserving protocol for linear multiple regression with vertically partitioned datasets, and (2) to show the feasibility of the proposed system using a real medical dataset distributed into two parties, the hospital who knows the technical details of diseases while the patients are in the hospital, and the local government who knows the residence even after the patients left hospital, (3) to show the accuracy and the performance of the PPDM system which allows us to estimate the expected processing time with an arbitrary number of predictors, and (4) to study the scalability of the extended models of horizontally and vertically partitions where more than two parties are involved to perform analysis.

2. DPC

2.1 DPC Datasets

The DPC dataset, Disease, Procedure and Combination, covers medical records for more than 7 million patients in more than 1,000 hospitals [1]. The 2016 DPC dataset consists of 2,553,283 records including 78,282 records of medical procedures related to a stroke.

With the international standard of disease, DPC data contains the followings; the hospital codes, the disease code, sex, age, ZIP code, the duration in hospital, the operation, the height, the weight, the degree of cancers, etc. The DPC dataset is used to study on hospital management and to provide useful statistics in hospitals. Some of the statistical data is available online and used as open data for many purposes.

2.2 Dataset 1 – Cardiac Disease

Tables 1 and **2** are the statistics of DPC datasets to be used in the later section.

The dataset 1 *Cardiac Disease* contains the records related to cardiac diseases, e.g., arrest, failure and the infarction, which were synthesized from the real DPC data with fundamental statistics. The dataset is intended to be studied from a hospital management viewpoint, i.e., the estimated expense in hospital in terms of patient attributes.

2.3 Dataset 2 – Stroke

Dataset 2 *Stroke* is a dataset related to patients who suffer from a stroke. As for significant predictors, we picked up seven variables chosen from a variety of patient information such as past history, condition of diseases. Table 2 shows the statistic of six predictor variables $x_1, \ldots, x_6$. Variable *Japan Coma Scale* gives the conscious initial state of a patient, with four scores of criteria of unconsciousness (higher is deeper). Variable *modified Rankin Scale* is a degree of disability or dependency in the daily activities of patients who have suffered a stroke. The scale runs from 0 – 6, meaning 0 – no symptoms, 1 – no significant disability, 2 – slight, 3 – moderate, 4 – moderate severe, 5 – severe disability, requiring constant nursing care. We drop scale 6 because it means that a target was dead. Variable *Stroke Type* provides the type of stroke, classified into three types: cerebral infraction, intracerebral hemorrhage, and subarachnoid hemorrhage.

Table 1 Dataset 1 *Cardiac Disease*.

predictor		# records	min–max	average	owner
duration [days]	x_1	655	0–101	14.56	*B*
age	x_2	655	0–95	52.25	
expense [JPY]	y	655	597–1,291,937	77,930	*A*

Table 2 Dataset 2 *Stroke*.

predictor		min, max	average	owner
Death	y	0–1	0.12	*A*
Age	x_1	40–106	72.03	*A*
Sex	x_2	1–2	1.431	*B*
Japan Coma Scale	x_3	0–3	0.957	*B*
modified Rankin Scale	x_4	0–5	3.556	*B*
Stroke Type	x_5	1–3	1.432	*B*
Liver Disease	x_6	0–1	0.022	*B*

2.4 Effect of Predictors to Death

There are several candidates of variable to predict that a target is dead, which is a target variable y. To make the significance visible, we plot the status of death in terms of predictors Japan Coma Scale (jcs) and modified Rankin Scale (mRS) in **Figs. 1** and **2**, respectively. Since both x and y are the discrete values, we add small noise of normal random numbers (mean 0 and standard

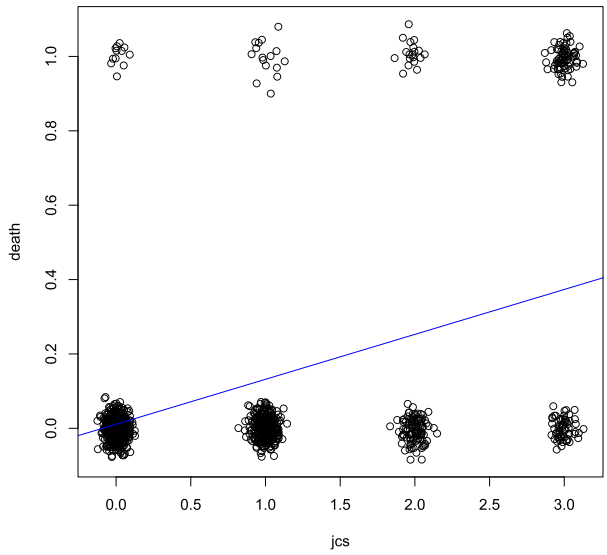

Fig. 1 Effect of Japan Coma Scale (jcs) on death.

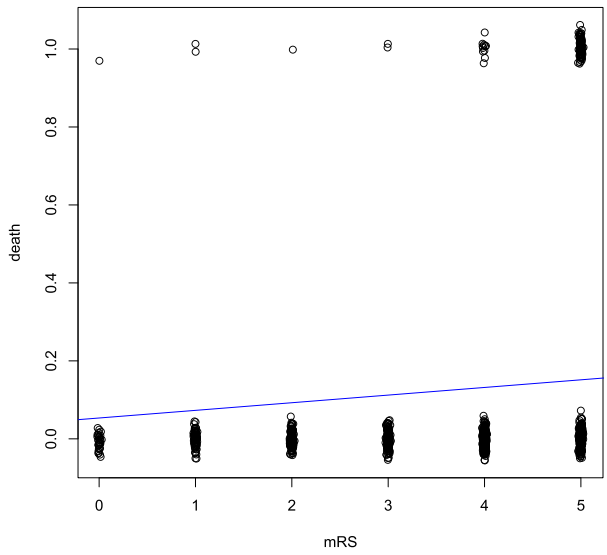

Fig. 2 Effect of modified Rankin Scale (mRS) on death.

deviation 0.05 (0.03) to x (jcs) (y), and mean 0 and standard deviation 0.01 (0.02) to x (mRS) (y)) in plotting. At a glance, both predictors have positive effect to the target variable (death), that is, a patient is likely to be dead as either jcs or mRS is higher. Our purpose is to quantify the degree of significant of predictor using secure multiple regression.

2.5 Distributed Datasets

The history of diseases is classified as one of the most sensitive attributes in personal data. Hospitals should take the greatest considerations of security of the dataset as much as possible. However, in order to reduce a risk to be compromised, the personal identity must be detached from the history dataset and some dataset should be distributed into several institutes.

Hence, we propose a vertically distributed datasets model in which multiple datasets are stored by multiple institutes with common identity which must be maintained by some other identity authority. Even if one of the stores was compromised, the security of the whole dataset is preserved in the model.

In our study, we assume the following two institutes owning datasets associated with common identities.

- Local government A
  is an authority that maintains residence data including name, address, household information, marriage status, and whether dead or alive. It also aims to help resident in living healthily but sometime does not allow to have access to the history of diseases stored in some hospitals. In Tables 1 and 2, it knows target variable y (death) and one predictor x (age).
- Hospital B
  has all history of diseases and the corresponding procedures, operations, and medicine taken. It also knows the detailed status of patients while the patients are in the hospital but does not allow to be trucked after they have left the hospital. Hence, in our model in Tables 1 and 2, they have all variables $x_2, \ldots, x_6$ except the target one y (death).

3. Building Blocks

3.1 Secure Scalar Product Protocol [10]

The scalar product of two vectors is performed in a secure manner using an additive homomorphic public key algorithm as shown in Algorithm 1. In this protocol, two parties who own the private X_A and X_B collaborate to obtain the size of the intersection $|X_A \cap X_B|$ without revealing X. The results are shared by the parties using the random numbers s_A and s_B such that $s_A + s_B = |X_A \cap X_B|$.

Algorithm 1 Secure Scalar Product

Input: Alice has the n-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_n)$. Bob has the n-dimensional vector $\boldsymbol{y} = (y_1, \ldots, y_n)$.

Output: Alice has s_A and Bob has s_B such that $s_A + s_B = \boldsymbol{x} \cdot \boldsymbol{y}$.

(1) Alice generates a homomorphic public-key pair and sends the public key to Bob.

(2) Alice sends Bob n ciphertexts $E(x_1), \ldots, E(x_n)$.

(3) Bob chooses s_B at random and computes

$$c = E(x_1)^{y_1} \cdots E(x_n)^{y_n} / E(s_B),$$

and sends c to Alice.

(4) Alice decrypts c to obtain $s_A = D(c) = x_1 y_1 + \cdots + x_n y_n - s_B$.

3.2 Linear Regression

Given n tuples of m input values $x_{i,1}, x_{i,2}, \ldots, x_{i,m}$ and output value y_i for $i = 1, \ldots, n$, a linear regression determines a linear model $f(X)$ of the form

$$y = f(X) = \alpha + \beta_1 x_1 + \cdots + \beta_m x_m. \tag{1}$$

By differentiating the sum of squared differences between $f(x)$ and y, and making it to be zero, we have the following $(m+1)$ simultaneous equations

$$\begin{aligned}
&\frac{\partial}{\partial \alpha} \sum_i^n (y_i - f(X_i))^2 \\
&= \sum_i^n 2(y_i - \alpha - \beta_1 x_{i,1} - \cdots - \beta_m x_{i,m})(-1) = 0 \\
&\frac{\partial}{\partial \beta_1} \sum_i^n (y_i - f(X_i))^2 \\
&= \sum_i^n 2(y_i - \alpha - \beta_1 x_{i,1} - \cdots - \beta_m x_{i,m})(-x_{i,1}) = 0 \\
&\cdots
\end{aligned}$$

To have $\beta_1, \ldots, \beta_m$, we solve X such that

$$AX = B \tag{2}$$

where

$$A = \begin{pmatrix} \sum x_{i,1}^2 & \sum x_{i,1}x_{i,2} & \cdots & \sum x_{i,1} \\ \sum x_{i,1}x_{i,2} & \sum x_{i,2}^2 & \cdots & \sum x_{i,2} \\ \vdots & & & \vdots \\ \sum x_{i,1} & \sum x_{i,2} & \cdots & \sum 1 \end{pmatrix},$$

$$X = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \\ \alpha \end{pmatrix}, \qquad B = \begin{pmatrix} \sum x_{i,1} y_i \\ \vdots \\ \sum x_{i,m} y_i \\ \sum y_i \end{pmatrix}$$

By taking the inverse matrix of A from the left side, we have X.

Hypothesis testing can be made by verifying whether a test statistic defined as $Z = \hat{\beta}_j / S.E.(\hat{\beta}_i)$ follows a normal distribution $N(0, 1)$.

3.3 Paillier Cryptosystem

Additively homomorphic public-key schemes – Paillier [13] or the modified ElGamal cryptosystems are both widely used. Both allow for key generation and decryption to be distributed among partially trusted authorities sharing private key. A cryptosystem E is said to satisfy the additively homomorphic property if: taking messages M_1 and M_2,

$$\begin{aligned}
&D[E[M_1] \oplus E[M_2]] = M_1 + M_2, \\
&D[E[M_1]^{M_2}] = M_1 M_2,
\end{aligned}$$

where $\oplus$ is a binary operator over ciphertext space and is a multiplication in $Z_{n^2}^*$ for Paillier cyptosystem. For indistinguishably,

we write the property as above using decryption D rather than writing $E[M_1] \oplus E[M_2] = E[M_1 + M_2]$

The Paillier cryptosystem consists of three stages: key generation, encryption, and decryption.

- *Key generation:* Let n be pq, a product of two large prime numbers p and q, and $g \in Z^*_{n^2}$ be a generator whose order divides n. Compute $\lambda = \mathrm{LCM}(p-1, q-1)$ and $\mu = (L(g^\lambda \pmod{n^2}))^{-1} \pmod{n}$, where L is defined by $L(u) = (u-1)/n$. The public key is (n, g) and the private key is (λ, μ).
- *Encryption:* A ciphertext c of M is defined with randomly chosen $r \in Z^*_{n^2}$ as:
$$c = E(M) = g^M r^n \pmod{n^2}.$$
- *Decryption:* Given ciphertext c, plaintext M is computed as $M = L(c^\lambda \pmod{n^2}) \cdot \mu$.

4. Proposed Scheme

The purpose of our proposal is to allow parties A and B owning private datasets to perform secure linear regression without revealing their datasets. In a *simple linear regression* model, a single response measurement Y is related to a single predictor X such that

$$E(Y|X) = \alpha + \beta x \tag{3}$$

where α is called the intercept and β is called coefficient. A party A has X and B has Y, but doesn't know what the other party has, and vice versa.

In most cases, more than one predictor variable will be available. This leads to the *two-variable regression* model of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 \tag{4}$$

More generally, with more than two predictors, we have a *multiple (linear) regression* model as

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{5}$$

where m is a number of predictors.

4.1 Threats

Our players are assumed as *honest-but-curious* in which they follow the protocol as defined (honest) but would try to reveal any information from any intermediate data (curious). Since our model involves two parties, the honest-but-curious assumption is reasonable.

We should consider if the intermediate data could reveal any private information to the others.

4.2 Simple Linear Regression

In order to estimate β, we take a least square approach, i.e., minimizing $S = \sum_{i=1}^n (y_i - \alpha - \beta x)^2$ over all possible values of coefficients. By differentiating it, we have the simultaneous equation

$$\begin{cases} \dfrac{\partial S}{\partial \alpha} = -2 \displaystyle\sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0 \\ \dfrac{\partial S}{\partial \beta} = -2 \displaystyle\sum_{i=1}^n x_i (y_i - \alpha - \beta x_i) = 0 \end{cases}$$

Table 3 Shared information owned by parties.

	A	B
task	encryption/decryption	regression
own data	$y_1, y_2, \ldots, y_n$	$x_{1,1}, x_{2,1}, \ldots, x_{n,1}$
keys	private key $(p, q), n$	public key (n, g)

Algorithm 2 scLinear (simple regression)

	A	generate key
	$A \to B$	public key
1.	A	encrypts $y_1, y_2, \ldots, y_n$
	$A \to B$	sends $Enc(y_1), \ldots, Enc(y_n)$
2.	B	Using $x_{1,1}, x_{2,1}, \ldots, x_{n,1}$,
		computes $Enc(C) = Enc(n \sum x_i^2 - \sum x_i \sum x_i)$
		computes $Enc(D) =$
		$(\prod Enc(y_i)^{x_i n})((\prod Enc(y_i))^{\sum x_i})^{-1}$
		$Enc(E) =$
		$(\prod Enc(y_i))^{\sum x_i^2}((\prod Enc(y_i)^{x_i})^{\sum x_i})^{-1}$
3.	$B \to A$	sends $Enc(C)$, $Enc(D)$, $Enc(E)$
4.	A	decrypts ciphertexts and obtains
		$\beta = D/C$ and $\alpha = E/C$.

which leads to estimate coefficient

$$\beta = \frac{n \sum_{i=0}^n x_i y_i - \sum_{i=0}^n x_i \sum_{i=0}^n y_i}{n \sum_{i=0}^n x_i^2 - (\sum_{i=0}^n x_i)^2} = \frac{D}{C} \tag{6}$$

and intercept

$$\alpha = \frac{\sum_{i=0}^n x_i^2 \sum_{i=0}^n y_i - \sum_{i=0}^n x_i \sum_{i=0}^n x_i y_i}{n \sum_{i=0}^n x_i^2 - (\sum_{i=0}^n x_i)^2} = \frac{E}{C}. \tag{7}$$

By employing the secure scalar product protocol in Algorithm 1, two parties can compute the numerator and the denominator without revealing x_i and y_i at all. However, the division of additive homomorphic ciphertexts is feasible only when the numerator is a multiple of the denominator. It is hard to assume. Instead, we allow them to decrypt the ciphertexts of the numerator and the denominator and then divide in plaintext to obtain the coefficient.

Table 3 summarizes the primary tasks and the information owned by two parties. We show Algorithm 2 for the simple regression.

4.3 Two-variable Linear Regression

Before we generalize multiple regression, we study the two-variable regression of the form $y = \alpha + \beta_1 x_1 + \beta_2 x_2$. Similar to the simple regression, we minimize $S = \sum_{i=1}^n (y_i - \alpha - \beta_1 x_{1i} - \beta_2 x_{2i})^2$ by solving the simultaneous equations

$$\begin{cases} \dfrac{\partial S}{\partial \beta_1} = -2 \displaystyle\sum_{i=1}^n x_{1i}(y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \alpha) = 0 \\ \dfrac{\partial S}{\partial \beta_2} = -2 \displaystyle\sum_{i=1}^n x_{2i}(y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \alpha) = 0 \\ \dfrac{\partial S}{\partial \alpha} = -2 \displaystyle\sum_{i=1}^n (y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \alpha) = 0 \end{cases}$$

with the estimated coefficients

$$\beta_1 = \frac{\sum x_{1i} y_i \sum x_{2i}^2 - \sum y_i x_{2i} \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \sum x_{1i} x_{2i}{}^2} = \frac{D_2}{C_2}$$

$$\beta_2 = \frac{\sum x_{2i} y_i \sum x_{1i}^2 - \sum y_i x_{1i} \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \sum x_{1i} x_{2i}{}^2} = \frac{E_2}{C_2}$$

$$\alpha = n \sum y_i - \beta_1 \sum x_{1i} - \beta_2 \sum x_{2i}.$$

The whole steps are given in Algorithm 3.

Table 4 Comparison of the proposed regression schemes.

	(3) simple	(4) two-variable	(5) multiple ($m = 2$)
linear model	$y = \alpha + \beta x$	$y = \alpha + \beta_1 x_1 + \beta_2 x_2$	$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$
data sent from B to A	C, D, E	$C_2, D_2, E_2, \sum x_1, \sum x_2$	F_2, G_2
total number of ciphertexts	3	5	7

Algorithm 3 scLinear (two-variable regression)

1.		same to Algorithm 2
2.	B	Using x_1, x_2, $Enc(y_1), \ldots, Enc(y_n)$, computes $\sum x_1$, $\sum x_2$ $Enc(C_2) = Enc(\sum x_1^2 \sum x_2^2 - \sum x_1 {x_2}^2)$, $Enc(D_2) = (\prod Enc(y_i)^{x_{1i}})^{\sum {x_{2i}}^2}$ $((\prod Enc(y_i)^{x_{2i}})^{\sum x_{1i}x_{2i}})^{-1}$ $Enc(E_2) = (\prod Enc(y_i)^{x_{2i}})^{\sum {x_{1i}}^2}$ $((\prod Enc(y_i)^{x_{1i}})^{\sum x_{1i}x_{2i}})^{-1}$
3.	$B \to A$	sends $Enc(C_2)$, $Enc(D_2)$, $Enc(E_2)$
4.	A	decrypts and have C_2,D_2,E_2 obtains β_1, β_2, α.

Algorithm 4 scLinear (multiple regression)

1.		same to Algorithm 2
2.	B	computes matrix F, $Enc(G) = \begin{pmatrix} \prod_{i=0}^{n} Enc(y_i)^{x_{i,1}} \\ \vdots \\ \prod_{i=0}^{n} Enc(y_i)^{x_{i,1}} \\ \prod_{i=0}^{n} Enc(y_i) \end{pmatrix}$,
3.	$B \to A$	sends F and $Enc(G)$.
4.	A	decrypts $Enc(G)$ and gets G solves $FX = G$ to obtain X.

4.4 Multiple Regression

In multiple regression model of the form $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$, by solving $m + 1$ simultaneous equations,

$$\begin{cases} \dfrac{\partial S}{\partial \alpha} = -2 \displaystyle\sum_i^n (y_i - \alpha - \beta_1 x_{i,1} - \cdots - \beta_m x_{i,m}) = 0 \\ \dfrac{\partial S}{\partial \beta_1} = -2 \displaystyle\sum_i^n x_{i,1}(y_i - \alpha - \beta_1 x_{i,1} - \cdots - \beta_m x_{i,m}) = 0 \\ \cdots \end{cases}$$

which can be represents by matrix-vector product

$$FX = G \tag{8}$$

where

$$F = \begin{pmatrix} \sum x_{i,1}^2 & \sum x_{i,1}x_{i,2} & \cdots & \sum x_{i,1} \\ \sum x_{i,1}x_{i,2} & \sum x_{i,2}^2 & \cdots & \sum x_{i,2} \\ \vdots & & & \vdots \\ \sum x_{i,1} & \sum x_{i,2} & \cdots & \sum 1 \end{pmatrix},$$

$$X = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \\ \alpha \end{pmatrix}, G = \begin{pmatrix} \sum x_{i,1}y_i \\ \vdots \\ \sum x_{i,m}y_i \\ \sum y_i \end{pmatrix}.$$

By multiplying the inverse of F, we estimate coefficients β in plain text.

4.5 Security Evaluation

We show the comparison of three proposed regression schemes in **Table 4** in terms of the bandwidth consumption and the degree of security.

Algorithms 2 (two-variable) and 3 (multiple regressions) looks similar but the former improves the degree of privacy in terms of private information leaked to the other party. This is why the Algorithm 2 needs more ciphertexts to be sent to the other party.

In Algorithm 2, we minimize the information to be revealed to the other party. To make the difference clear, let us consider the simple case of $m = 2$. Algorithm 3 reveals all elements of the matrix

$$F_2 = \begin{pmatrix} \sum x_1^2 & \sum x_1x_2 & \sum x_1 \\ \sum x_1x_2 & \sum x_2^2 & \sum x_2 \\ \sum x_1 & \sum x_2 & \sum 1 \end{pmatrix},$$

$$G_2 = \begin{pmatrix} \sum x_1y_i \\ \sum x_2y_i \\ \sum y_i \end{pmatrix}$$

By eliminating the duplications and $\sum y_i$, $\sum 1$ that can be computed without help of B, there are a total of seven statistical values available to the other party. On the other hand, Algorithm 2 needs to send only five date C_2, D_2, E_2, $\sum x_1$, $\sum x_2$ in encrypted way. Therefore, Algorithm 2 is more secure than Algorithm 3 in terms of quantity of information leaked in executing the protocol.

When the size of dataset is limited, there is concern that the intermediate data such as C and F may reveal partial information of the other party. The revealed confidential information decreases as either size n and dimension m of dataset increases.

In Ref. [17], Shirakawa et al. studies the risk from the intermediate data of linear regression from statistical disclosure control (SDC) perspective. They show that standard deviation (variance), skewness and kurtosis confirm that cells with frequency of 10 or higher are unsafe based on synthetic data.

5. Experiment

5.1 Purpose of the Experiment

We have implemented the proposed protocol and developed the privacy-preserving regression system called "scLinear". Our system aims to clarify the following

(1) accuracy of the estimation,

(2) performance of regression.

5.2 Experimental Method

Table 5 shows the experimental environments. We use a proprietary protocol for exchange data between parties.

5.2.1 Experiment 1 (performance)

We measure the processing time of the proposed simple and two-variable regressions for the synthesized DPC data, with $n = 100, 300, 500$ and 655. Repeating 10 times for each we have the average.

5.2.2 Experiment 2 (real medical data)

We use the scLinear to analyze the real medical DPC datasets with the number of predictors, $m = 3, 4, 5$, and 6, and the num-

ber of patients, $n = 1{,}000, 2{,}000$, and 5,000. The purpose of the experiment is to make sure if the proposed system estimates the outcome (death) given multiple predictors (medical records) and also to find the most significant predictor out of multiple variables.

Note that we use integer variables in the DPC dataset. As shown in Table 2, all variables are integer in our experiment. Our system encodes the values as plaintext and does not use any floating nor fixed point real values. After decrypting, the coefficients are represented as real values.

5.3 Results of Experiment 1

5.3.1 Accuracy

Tables 6 and **7** show the estimated coefficients in the simple regression, and the two-variable regressions, respectively.

The estimated result of the scLinear is compared with that of the computed value in R (function `lm`). Table shows that the developed system estimates the coefficient with very high precision of the third decimal place. When $n = 100, 300, 500$, the estimated coefficients are exactly the same to that of R. The possible reason of the small error might be introduced by estimating α using the estimated β, which results in cumulating errors.

5.3.2 Performance

Figure 3 shows the processing time of the scLinear with respect to the size of the dataset, n. Most time spends in performing encryption for each of n records. To see the internal overhead, we show the processing time spent only for Step 2 in Algorithm 2 in **Fig. 4**.

The average processing time per record is 1.29 [ms] and 16.7 [ms] for Algorithms 2 (simple regression) and 3 (two-variable regression), respectively. The source of the overhead of two-variable regression is of the modular exponentiation in the protocol. **Table 8** shows the portfolio of the average processing time. No significant difference in encryption and decryption can be found from the table.

Table 5 Experimental environment.

	Experiment 1	Experiment 2
OS	Windows 7	Windows 7
memory	4 GB	11.7 GB
CPU	Intel(R) Core(TM) i5-3337U	Intel Xeon X5460
clock	1.8 GHz	3.16 GHz
lang.	Java(1.8.0_91-b14) R(3.1.0)	Java(1.8.0_45-b15) R(3.1.2)
key length	2,048 [bit]	

Table 6 Experiment result (simple regression).

	β		α	
n	scLinear	R	scLinear	R
655	5,099.358	5,099.358	5,002.521	5,002.521
500	67,751.810	67,751.810	1,021.751	1,021.751
300	72,369.082	72,369.082	322.378	322.378
100	89,508.076	89,508.076	786.790	786.790

5.4 Result of Experiment 2

5.4.1 Accuracy

Table 9 shows the result of multiple regression with six predictors. We compare the resulting coefficients of the scLinear with that of R (non partitioned dataset) and verify that the scLinear estimate the exactly same results for any of $n = 1{,}000, 2{,}000$.

5.4.2 Performance

Figures 5 and **6** shows the processing time for multiple regression of the scLinear. The former shows the total time in terms of number of records, n, while the latter shows the time difference in terms of number of variables, m. Note that we exclude the time spent in encryption in the Fig. 6 for making the inference visible.

From the observation of the results, the total processing time is linear to n and increases with square of m. As we will study the scalability of the proposed protocol in Section 6.2, the algorithm runs in time of $O(nm^2)$.

Above all, we have the expected time given n as

$$\text{time} = 224.576n + 9{,}023.692 \text{ [ms]}.$$

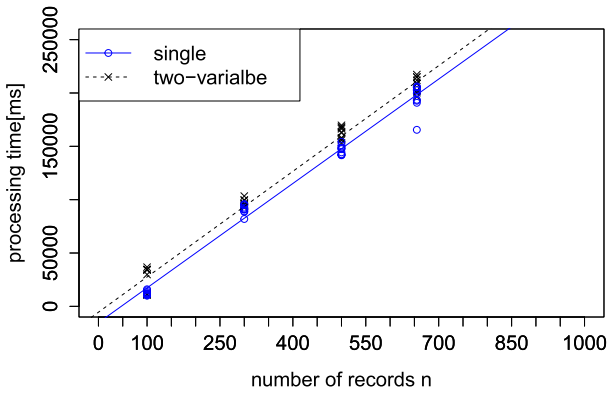

Fig. 3 Experiment 1, system processing time.

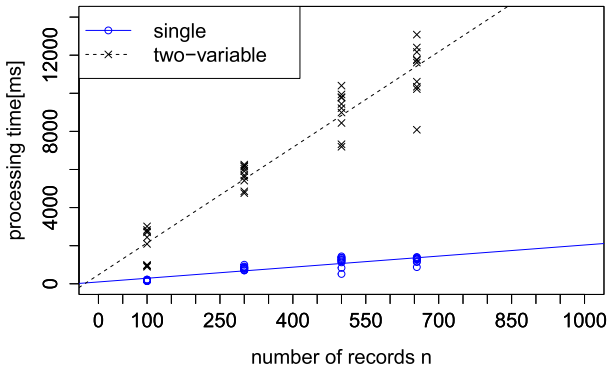

Fig. 4 Experiment 1, processing time for Step 2 in Algorithm 2.

Table 8 Experiment 1: average processing time per record [ms].

	encryption	homomorphic op.	decryption
model	A	B	A
simple	320	1.9	1.0
two-variable	320	16.7	0.6

Table 7 Experimental result (two-variable regression).

	β_1		β_2		α	
n	scLinear	R	scLinear	R	scLinear	R
655	4,995.554	4,995.555	41.304	41.304	3,042.752	3,042.759
500	998.417	998.417	245.842	245.842	54,332.010	54,332.010
300	302.882	302.882	128.136	128.136	65,339.083	65,339.084
100	4,730.629	4,730.629	273.939	273.939	1,744.523	1,744.523

Table 9 Coefficients for each of predictor in multiple regression (n = 5,000).

variables	Proposed	R				
	scLinear	coefficient	Std. Error	t value	$Pr(>\|t\|)$	
α	−0.1731982	−0.1731982	0.0290099	−5.970	$2.53e^{-09}$	***
Age	0.0015410	0.0015410	0.0003576	4.310	$1.67e^{-05}$	***
Sex	−0.0217865	−0.0217865	0.0083993	−2.594	0.009519	**
Japan Coma Scale	0.1283596	0.1283596	0.0049296	26.039	$< 2e^{-16}$	***
modified Rankin Scale	0.0121227	0.0121227	0.0034845	3.479	0.000507	***
Stroke Type	0.0292522	0.0292522	0.0073582	3.975	$7.12e^{-05}$	***
Liver Disease	0.0095770	0.0095770	0.0324591	0.295	0.767970	

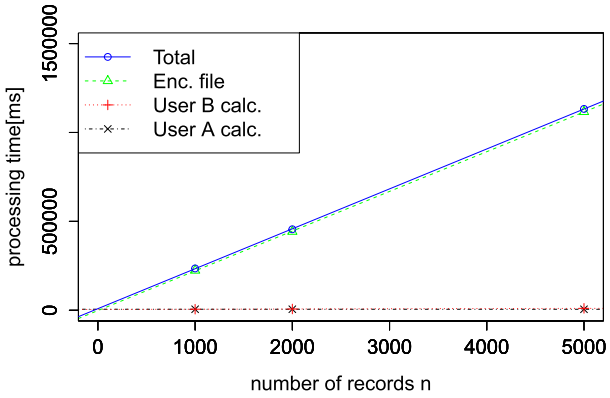

Fig. 5 Experiment 2: Processing time with respect to the number of records n.

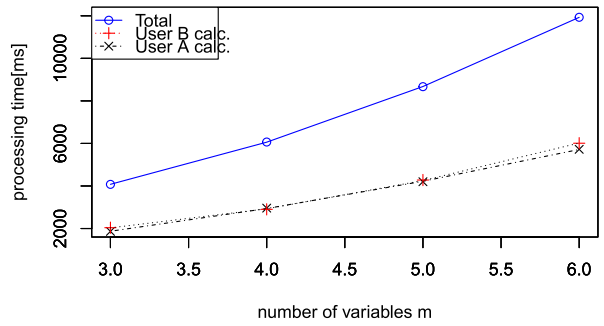

Fig. 6 Experiment 2: Processing time with respect to variable m (n = 1,000, without encryption).

and for instance of n = 10,000,000, the estimated time is $225 \times 10^7 + 9{,}024 = 2.25 \times 10^9$ [ms] = 26 days. The 99.6% of processing time is spent for performing encryptions.

5.5 Consideration from the Results

We have some remarks on the result of multiple regression in Table 9.

The table shows that variables *age*, *Japan Coma Scale*, *modified Rakin Scale* and *Stroke Type* are statistically significant with confidence level more than 95% (indicated with $***$). The most significant predictor is Japan Coma Scale with probability less than 2^{-16}. Its coefficient is 0.128 that can be seen as the slope of the linear function of x_3 (jcs) in Fig. 1, which is higher than that of function of x_4 (mRS) in Fig. 2.

Note the negative coefficient of Sex, −0.02. Assigned 1 and 2 as male and female, respectively, the coefficient of predictor x_2 implies that the risk of women to be dead is slightly smaller than that of men.

We also note that the variable x_6 (Liver Disease) does not have the significant confidence. It makes sense because the patients used to be examined in the experiment are restricted within who suffered from stroke. The history if the patient had liver disease has nothing to do with the risk of death by stroke.

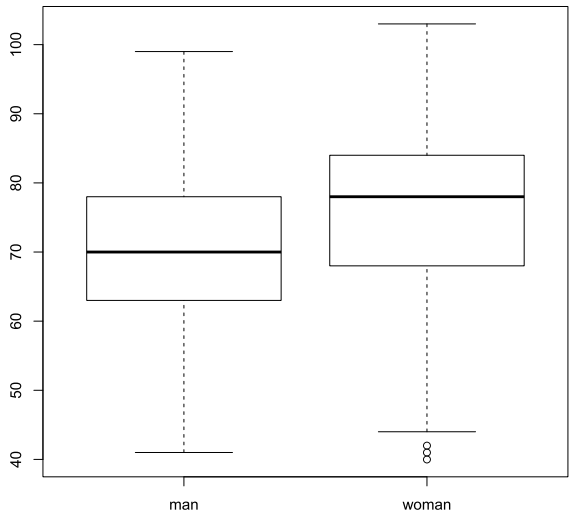

Fig. 7 Experiment 2: Difference in sex (n = 1,000).

6. Scalability of Privacy-Preserving Analysis

6.1 Two Styles of Partition

In this section, we extend our study from the viewpoint of scalability with more than two parties. In privacy preservation, we distribute the whole dataset S of n records in either of the following partitions.

(1) (Horizontal Partition) A set of n-records (row) is partitioned into N disjoint subset such that $S = S_1 \cup \cdots \cup S_N$, which are distributed N parties (hospitals). Let n_i be the number of records in subset S_i.

(2) (Vertical Partition) A set of input variables (column) is partitioned into M disjoint subsets $X_1, \ldots, X_M$ such that $X_1 \cup \cdots \cup X_M = \{x_1, \ldots, x_m\}$. Let m_i be the number of variables in X_i.

Both partitions have models in epidemiological study.

The horizontal partition is a model of N hospitals that maintain the common format of medical records such as disease and operation codes (DPC dataset) for different sets of patients. In the example of DPC database, the number of hospitals is N = 1,000 and a number of patients n_i varies from 100 to 6,000. In this model, all hospitals jointly estimate some quantity in terms of common medical records more accurately than local estimate.

The vertical partition is a model of heterogeneous institutes and hospitals. Multiple parties maintaining distinct quantities for common set of patients. For instance, a medical center that keeps a medical examination collaborates with a governmental cancer registry in order for a model of risk of cancer.

In horizontally partitioned privacy-preserving linear regres-

Algorithm 5 Vertically Partitioned Linear Regression

Input: M parties own records for n patents. A set of variables $\{x_1, \ldots, x_m\}$ and y is partitioned and distributed into M parties with subset of variables $A_1, \ldots, A_M$.

(1) ℓ-th party (institute) ($\ell = 1, \ldots, M$), independently, computes the sum of variable $j \in A_\ell$ in U, $\sum_{i \in U} x_{i,j}$, and the sum of products of two different variables $i \neq k \in A_\ell$, $\sum_{i \in U} x_{i,j} x_{i,k}$. The ℓ-th party publishes the two sums.

(2) For each pair of two institutes $\ell_1 \neq \ell_2 \in \{1, \ldots, M\}$, jointly computes the scalar products of n-dimension vectors $\boldsymbol{x}j = (x_{1,j}, \ldots, x_{n,j})$ and $\boldsymbol{x}k = (x_{1,k}, \ldots, x_{n,k})$ for all pairs of variable $j \in A_{\ell_1}, k \in A_{\ell_2}$ and publishes the results $\boldsymbol{x}j \cdot \boldsymbol{x}k$.

(3) Arbitrary party computes coefficients $\alpha, \beta_1, \ldots, \beta_m$ according to Eq. (2).

sion, all parities (hospitals) share the common set of attributes but own records for different partitions. We omit the construction of protocol for the horizontally partitioned data since it is out of scope of the work.

Note that the horizontal partition performs well in terms of parallel processing since an addition is commutative and hence the order of operation does not affect the outcome. With making N parties structured in binary tree, the summation of N runs in $\log N$ time.

The protocol for a vertically partitioned linear regression is performed in Algorithm 5.

6.2 Complexity

6.2.1 Horizontal Partition

The largest overhead happens at public-key encryption at the Step (1). Given m variables, the number of ciphertexts necessary to perform the protocol is

$$\frac{m^2}{2} + \frac{3}{2}m + 1 = O(m^2).$$

Note that it does not depends on n, the number of records in the table and the fact implies that the horizontal protocol scales well in terms of number of patients in the epidemic applications. Also note the Step (1) can be performed in parallel even if N parties are involved. The multiple ciphertexts are generated independently. Hence, the total processing time is as same as that of a single parity. Therefore, the scalability of the step is achieved with n and N.

The Step (2) requires a multiplication of N ciphertexts, some decryptions for each variable and pair of variables. The cost of multiplication is negligible in comparison to modular exponentiations. The number of decryptions is m for variables plus $\binom{m}{2} = (m^2 - m)/2$ for pairs of two variables.

6.2.2 Vertical Partition

At Step (2), it requires n-dimension scalar products, which is the bottleneck in the vertical partition protocol. The number of ciphertexts depends on the size of dataset, n. In addition, the scalar products need to be performed as many as $m_1 \times m_2$ times where m_1 and m_2 are numbers of variables owned by two parties. Moreover, if set of parties are divided into M groups, the above processing happen for every pair of two out of M. Consequently, the total number of ciphertexts (computations) is

$$n(m^2 + m)\frac{M(M-1)}{2} = O(nm^2M^2).$$

Table 10 Computation Costs in Privacy-Preserving Linear Regression.

partition	horizontal $U = U_1 \cup \cdots \cup U_N$	vertical $A = A_1 \cup \cdots \cup A_M$
cost	$\frac{m^2}{2} + \frac{3}{2}m + 1$	$n(m^2 + m)\frac{M(M-1)}{2}$
complexity	$O(m^2)$	$O(nm^2M^2)$

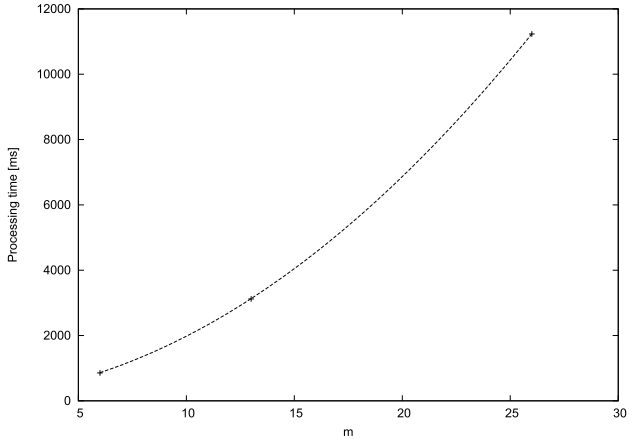

Fig. 8 Processing Time for Horizontal Partition Linear Regression ($N = 1$, $n = 257{,}997$).

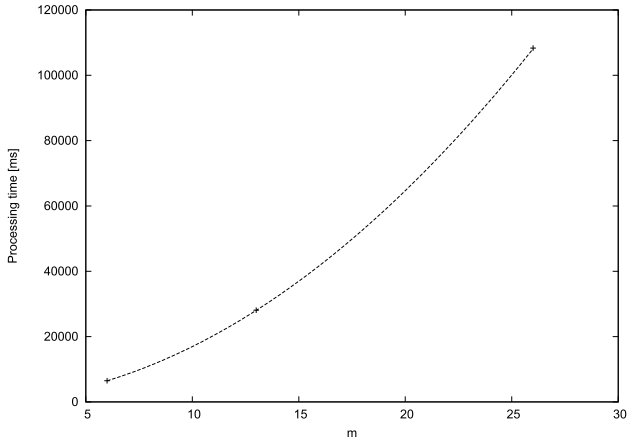

Fig. 9 Processing Time for Vertical Partition Linear Regression ($N = 1$, $n = 10{,}000$, $M = 1$).

This is a large burden for epidemic study. The size of dataset and the number of partitions should be carefully chosen for practical level.

We show the summary of scalability in **Table 10**.

6.3 Performance Estimation

Figures 8 and **9** show the processing time of the proposed protocol in terms of the number of variables m in horizontal and vertical partitions, respectively. The increase of processing time looks similar in both partitions. However, the complexity of vertical partition is much higher than that of horizontal model. For instance, the regression with $n = 250{,}000$ and $m = 26$ takes about 11 second in horizontal partition and remains the same even if the size of database n increases. While, the regression with $n = 10{,}000$ takes about 108 seconds or 1.83 minutes in vertical partition but it is estimated about 1 hour for $n = 250{,}000$.

6.4 Related Works

In Ref. [5], Hall et al. proposed a protocol for linear regression as well as certain goodness of fit statistics using homomor-

Table 11 Comparison with related works.

scheme	Nikolaenko [15]	Gascon [14]	Aono [16]	Ours
Partition	holizontal	virtical	virtical	virtical
building	Paillier enc. + garbled circuit	Paillier enc. + garbled circuit	LWE-based enc.	Paillier enc.
scheme	multi-party w. CSP	multi-party w. CSP	multi-party	two-party
security	semi-honest servers	semi-honest servers	semi-honest servers, output privacy	semi-honest parties
dataset	UCI DS. (forestFires)	UCI DS. (9 datasets)	UCI DS.	real DPC data

phic encryption. Their protocol provides only the final result of the linear regression and hide the intermediate values. With Current Population Survey with 51,016 cases and 22 variables, they demonstrate how practical their proposed protocol is. However, their protocol needs multiple-round interaction over institutes to get convergence.

Karr et al. [9] proposed a secure regression based on the similarity on secure matrix product. Hall claims that the protocol compromises with respect to privacy.

Nikolaenko et al. [15] studies a privacy-preserving linear regression of horizontally-partitioned data. Their solution combines an additive homomorphic encryption with the garbled circuit. Their setting introduces two semi-trusted third parties, Crypto Service Provider (CSP) and Evaluator and construction does not reveal any private data into the third parties.

Gascon et al. proposed the version of vertically-partitioned dataset in Ref. [14]. They also used homomorphic encryption and the gabled circuit. They proposed a new Conjugate Gradient Descent (CDG) algorithm that scales well and works with privacy-preserving building schemes.

Aono et al. [16] proposed a privacy-preserving protocol for linear regression that satisfies input and output privacy. They combined the LWE-based homomorphic public key encryption with differential privacy. They implemented their proposed scheme and reported with open data of 10^4 records of 100 Kbytes.

Table 11 summaries the comparison with some existing schemes. The most schemes assumes trusted party and allow multiple parties, while our scheme assumes two party in order to simplify the transaction. Our scheme shows the experimental result with real medical data.

7. Conclusions

We have proposed some secure protocol for some independent institutes to collaborate to perform multiple linear regression to show the significant variables to predict a given target variable. Our experiment used the real medical data records of $n = 5{,}000$ patients and revealed that the Japan Coma Scale (unconsciousness state) was the most significant predictor for death by stroke and the modified Rakin Scale (degree of disability) followed. Our developed system running on the consumer PC specification allows hospitals to perform multiple regression with arbitrary number of variables without revealing any confidential history of diseases to the other party such as a local government who maintains personal database of the residences.

Based on the experimental data, we estimate the processing time for $n = 100{,}000{,}000$ patients is 2.6 days. The large overhead comes form the performing public homomorphic encryption. By replacing it with latest technologies such as lattice encryptions, we expect that the proposed multiple regression protocol is feasible to do arbitrary epidemically analysis over distributed datasets connected in secure network.

We have studied the scalability of privacy-preserving linear regression protocols in horizontal and vertical partitions and showed that the horizontal partition protocol scales well in terms of size of databases but the vertical partition suffers the complexity in terms of numbers of variables and records.

Acknowledgments This work was supported by JSPS KAKENHI Grand-in-Aid Research (C), Grant Number 15K00194 and the Japan Science and Technology Agency as part of the Japan-Taiwan Collaborative Research Program.

References

[1] Yasunaga, H., Horiguchi, H., Kuwabara, K., Matsuda, S., Fushimi, K. and Hashimoto, H.: Outcomes After Laparoscopic or Open Distal Gastrectomy for Early-Stage Gastric Cancer: A Propensity-Matched Analysis, *Annals of Surgery*, Vol.257, No.4, pp.640–646 (2012).

[2] Sweeney, L.: k-anonymity: A model for protecting privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol.10, No.5, pp.557–570 (2002).

[3] Vaidya, J. and Clifton, C.: Privacy preserving association rule mining in vertically partitioned data, *The 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp.639–644, ACM Press (2002).

[4] Wu, S., Teruya, T., et al.: Privacy-preservation for Stochastic Gradient Descent, *The 27th Annual Conference of the Japanese Society for Artificial Intelligence*, 3L1-OS-06a-3 (2013), available from ⟨https://kaigi.org/jsai/webprogram/2013/paper-596.html⟩.

[5] Hall, R., Fienberg, S.E. and Nardi, Y.: Secure Multiple Linear Regression Based on Homomorphic Encryption, *Journal of Official Statistics*, Vol.27, No.4, pp.669–691 (2011).

[6] Kikuchi, H. and Sakuma, J.: Bloom Filter Bootstrap: Privacy-Preserving Estimation of the Size of an Intersection, *Journal of Information Processing*, Vol.22, No.2, pp.388–400 (2014).

[7] Chaudhuri, K. and Monteleoni, C.: Privacy-preserving logistic regression, *Proc. 21st International Conference on Neural Information Processing Systems* (*NIPS 2008*), pp.289–296 (2008).

[8] Du, W., Han, Y.-S. and Chen, S.: Privacy-preserving multivariate statistical analysis: Linear regression and classification, *2004 SIAM International Conference on Data Mining* (2004).

[9] Karr, A.F., Lin, X., Sanil, A.P. and Reiter, J.P.: Privacy-preserving analysis of vertically partitioned data using secure matrix products, Journal of Official Statistics, Vol.25, No.1, pp.125–138 (2009).

[10] Goethals, B., Laur, S., Lipmaa, H. and Mielikainen, T.: On Private Scalar Product Computation for Privacy-Preserving Data Mining, *The 7th Annual International Conference in Information Security and Cryptology* (*ICISC 2004*), LNCS, Vol.3506, pp.104–120 (2004).

[11] Freedman, M.J., Nissim, K. and Pinkas, B.: Efficient private matching and set intersection, *EUROCRYPT 2004*, LNCS, Vol.3027, pp.1–19, Springer-Verlag (2004).

[12] Kantarcioglu, M., Nix, R. and Vaidya, J.: An Efficient Approximate Protocol for Privacy-Preserving Association Rule Mining, *13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining* (*PAKDD 2009*), LNCS, Vol.5476, pp.515–524, Springer (2009).

[13] Paillier, P.: Public-Key Cryptosystems Based on Composite Degree Residuosity Classes, *EUROCRYPT*, pp.223–238, Springer (1999).

[14] Gascon, A.G.A., Schoppmann, P., Balle, B., Raykova, M., Doerner, J., Zahur, S. and Evans, D.: *Privacy-Preserving Distributed Linear Regression on High-Dimensional Data*, PoPETS, Vol.2017, No.4, pp.345–364 (2017).

[15] Nikolaenko, V., Weinsberg, U., Ioannidis, S., Joye, M., Boneh, D. and Taft, N.: Privacy-preserving ridge regression on hundreds of millions of records, *IEEE Symposium on S&P 2013*, pp.334–348 (2013).

[16] Aono, Y., Hayashi, T., Phong, L.T. and Wang, L.: Input and Out-

put Privacy-Preserving Linear Regression, *IEICE Trans. Inf. & Syst.*, Vol.100-D, No.10, pp.2339–2347 (2017).
[17] Shirakawa, K., Abe, Y. and Ito, S.: Empirical Analysis of Sensitivity Rules: Cells with Frequency Exceeding 10 that Should Be Suppressed Based on Descriptive Statistics, *Privacy in Statistical Databases* (*PSD 2016*), Lecture Notes in Computer Science, Vol.9867, pp.28–40, Springer (2016).

Hiroaki Kikuchi received his B.E., M.E. and Ph.D. degrees from Meiji University in 1988, 1990 and 1994. After he working in Fujitsu Laboratories Ltd. in 1990, he had worked in Tokai university from 1994 through 2013. He is currently a professor in at Department of Frontier Media Science, School of Interdisciplinary Mathematical Sciences, Meiji University. He was a visiting researcher of the school of computer science, Carnegie Mellon University in 1997. His main research interests are network security, cyrptographical protocol, privacy-preserving data mining, and fuzzy logic. He received the Best Paper Award for Young Researcher of Japan Society for Fuzzy Theory and Intelligent Informatics in 1990, the Best Paper Award for Young Researcher of IPSJ National Convention in 1993, the Best Paper Award of Symposium on Cryptography and Information Security in 1996, the IPSJ Research and Development Award in 2003, the JIP Outstanding paper Award in 2010 and 2017 and the IEEE AINA Best Paper Award in 2013. He is a member of IEICE, SOFT, IEEE and ACM. He receives IPSJ Fellow.

Chika Hamanaga received her B.S. degree from Meiji University in 2017. She is engaged in a system development.

Hideo Yasunaga graduated from Faculty of Medicine, University of Tokyo in 1994. After he worked at Tokyo University hospital and some community hospitals as a surgeon from 1994 to 2000, he studied as a graduate student in Department of Public Health, Graduate School of Medicine, University of Tokyo, from 2000 to 2003, and he worked at Department of Planning, Information and Management, The University of Tokyo Hospital, as an assistant professor from 2003 to 2008. He then worked at Department of Health Management and Policy, The University of Tokyo, as an associate professor from 2008–2013. He is currently the professor of Department of Clinical Epidemiology and Health Economics, The University of Tokyo. His main research interests are clinical epidemiology and health economics. He is an associate editor of Journal Epidemiology. He is a director of the Society for Clinical Epidemiology.

Hiroki Matsui received his M.P.H. at the University of Tokyo School of Public Health. He is currently the Research Associate in Department of Clinical Epidemiology and Health Economics at the University of Tokyo School of Public Health.

Hideki Hashimoto received his M.D. at the University of Tokyo School of Medicine, and DPH at Harvard School of Public Health. He is currently the Professor in Health and Social Behavior at the University of Tokyo School of Public Health.

Chun-I Fan received his M.S. degree in computer science and information engineering from the National Chiao Tung University, Hsinchu, Taiwan, in 1993, and the Ph.D. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1998. From 1999 to 2003, he was an Associate Researcher and a Project Leader with Telecommunication Laboratories, Chunghwa Telecom Company, Ltd., Taoyuan, Taiwan. In 2003, he joined the faculty of the Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, where has been a Full Professor since 2010. His current research interests include applied cryptology, cryptographic protocols, and information and communication security. Prof. Fan is the Chairman of the Chinese Cryptology and Information Security Association, Taiwan, and was the Chief Executive Officer (CEO) of "Aim for the Top University Plan" Office at National Sun Yat-sen University. He was the recipient of the Best Student Paper Awards from the National Conference on Information Security in 1998, the Dragon Ph.D. Thesis Award from Acer Foundation, the Best Ph.D. Thesis Award from the Institute of Information and Computing Machinery in 1999, and the Engineering Professors Award from Chinese Institute of Engineers - Kaohsiung Chapter in 2016. He is also an Outstanding Faculty in Academic Research in National Sun Yat-sen University.