DPC データセットによるプライバシーを保護した治療戦略の比較

菊池 浩明 †　　橋本 英樹 ‡　　康永 秀生 ‡　　渋谷 健司 ‡

† 明治大学 総合数理学部
164-8525　東京都中野区中野 4-21-1
kikn@meiji.ac.jp

‡ 東京大学 大学院医学系研究科
113-8555 東京都文京区本郷 7-3-1
hidehashimoto-circ@umin.ac.jp,yasunagah-tky@umin.ac.jp

あらまし　日本版診断群分類 (DPC) データベースは，病名や治療行為の表コードによる患者の大規模データベースである．このデータセットにより，同一疾患に対して異なる治療戦略を施した場合の在院日数や生存率の比較検討が可能になる．同 DB は個人情報を含まないが，多くの患者情報を含むため，本人が特定されてしまう危険性がある．そこで，プライバシー保護データマイニング技術を適用して，患者が特定されるリスクを削減して，正確な治療行為の比較を実現する安全な疫学調査手法を提案する．本稿では，提案方式の試験実装を用いて，DPC 擬似データに適用した結果について報告する．

Privacy-Preserving Propensity Score Matching for evaluation of Outcomes using the DPC dataset

Hiroaki Kikuchi†　　Hideki Hashimoto‡　　Hideo Yasunaga‡　　Kenji Shibuya‡

†School of Interdisciplinary Mathematical Sciences, Meiji University
4-21-1 Nakano, Nakano Ku, Tokyo, 164-8525 Japan
‡Graduate School of Medicine, The University of Tokyo
7-3-1, Hongo, Bunkyo, Tokyo, 113-8555 Japan

Abstract　The Diagnosis Procedure Combination database is a large-scale claim-based database of Japanese hospitals, following standardized case-mix patient classification system for evaluation of clinical procedures and hospital performance. The dataset does not include confidential information, though its detailed contents may allow identification of individual patients. We developed a secure scheme to preserve patient privacy and at the same time to conduct epidemiological analysis adjusting for patient characteristics for unbiased comparison. We tested its performance with DPC database sample of 80 hospitals.

1　はじめに

疫学調査においては，ある治療法の効果を臨床試験などを通じて評価を行なうことと目標とする．[5] では，大規模な疾患，治療データセットを用いて，胃がん患者を対象とした腹腔鏡手術と開腹手術の影響を評価している．本稿では，心臓手術における狭心症の患者を対象にした治療法として，開胸手術とカテーテルを使った風船治療の比較を考える．開胸バイパス手術では

1 枝あるいは 2 本以上の血管グラフトを手術で繋ぐ．一方，カテーテル手術には，風船治療，ステント（金属メッシュを風船に載せて狭いところを内側から広げるもので，風船よりも再手術の必要性が低いと言われている）があり，手術死亡率，在院日数，医療費などの観点で両手術の差があるかどうかを明らかにしたい．

全く同じ条件の患者群が 2 つあれば比較は可能だが現実的ではない．同じ重症度の患者に対しても，病院側の施設の条件の差があったりするので，どちらの手術が選ばれるかはランダムではない．手術方式の差異よりも患者の年齢やがんの進行が影響していることも考えられる．結局のところ，これらの多くの交絡因子の影響をなくして，バイパス手術とカテーテル手術の手術手法の違いを正しく評価するためには，結局多くの病院に分散している手術データを統合するよりはない．しかしながら，単純に氏名や生年月日などの個人情報を取り除く匿名化では，疾病の種類や入院退院日，合併症，手術費用などの情報から個人が特定されてしまうリスクが残る．これらの属性は準識別子 (Quasi Identifier) と呼ばれ，後日他のデータベースの情報を突き合わせることで，個人が特定されてしまう，いわゆる，リンク攻撃の危険性が指摘されている [6]．

そこで，本研究では，加法順同型性を満たした公開鍵暗号を利用したプライバシー保護データマイニング技術 [7] [8] を適用して，各病院での患者データを秘匿したままで，様々な疫学調査を試みる．病名，手術コードなどの属性情報は共通で，それぞれの患者が異なるデータベースに分割されている分散データマイニングであるので，水平分割モデルである．

交絡因子のバランスを保ち，臨床試験における治療方法の因子を正しく評価するために，ロジステック回帰を用いる．要請条件と研究目的を次のように整理する．

1. 病院間で統計値以外のデータは交換しない．
2. 部分的な統計情報の暗号文は交換する．復号鍵は，信頼出来る第三者，あるいは，閾値以上のメンバーの協力による復号できる分散復号スキームを取る．
3. 全病院で，治療法に関する年齢分布，平均治療費などの統計値を得る（クロス集計）．
4. 全病院で，平均死亡率における手術方法の調整済みオッズ比とその検定結果を得る（ロジスティック回帰）．

なお，垂直分割ロジステック回帰については，[9] にて PPDM プロトコルが提案されている．ロジステック回帰での初めての水平分割 PPDM プロトコルである．試験実装で分析を行った DPC データの運用実験も，本方式の実用可能性を評価するために有益である．

2 疫学調査と秘匿計算

2.1 DPC データセット

DPC データセットは，病院の経営や診療の質を公平に評価する目的で，疾患 (Diseases)，治療 (Procedure) の組み合わせ (Combination) のデータからなる [1]．1015 施設から収集された，700 万人分の患者データが構築されている．緊急入院患者のほぼ 50%がカバーされている．データセットには，病院コード，データ識別番号，性別，年齢，郵便番号，在院日数，手術名，疾病名，身長，体重，喫煙歴，がんステージ分類，重症度などの診療情報を含む．

2.2 ロジステック回帰

ロジステック分析は，疫学コホート研究において，m 個の要因 $x_1, x_2, \cdots, x_m$ からある期間内に疾病の発症する確率

$$p(y = 1|x_1, \ldots, x_m) = \frac{1}{1 + \exp(-z(x_1, \ldots, x_m)}$$

で推定して，発生要因を特定する．ここで，

$$z(x_1, .., x_m) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

と定義される多項式である．ある要因 x_j の疾病に対する効果は，他の要因の影響の調整済みのオッズ比，$p/(1-p) = \exp(\hat{\beta}_j)$ で与えられる．仮説検定は，統計量 $Z = \hat{\beta}_j / S.E.(\hat{\beta}_i)$ が正規分布 $N(0,1)$ に従うことで確かめられる．

係数 β を推定するには，β を変数とする尤度関数の対数

$$\begin{aligned} L(\beta_0,\ldots,\beta_m) &= \log\prod_{i=1}^{n} p(z(\boldsymbol{x}_i))(1-p(z(\boldsymbol{x}_i)) \\ &= \sum_{i=1}^{n} y_i\beta + \log(1-p(\boldsymbol{x}_i)) \end{aligned}$$

を偏微分して，連立方程式を解く．代数的には解けないので，Newton 法などで逐次的に最尤推定する [2][3].

3 プライバシー保護疫学調査

加法准同型性を満たす公開鍵暗号 $Enc()$ と復号関数 $Dec()$ を用いて，分散したプレイヤー間で互いの入力を秘匿したままで，共通の目的関数の値を計算する．プレイヤーは，プロトコルに従ってデータの送受信を行うが，受動的にデータを見ようとする（セミオネストモデル）.

3.1 概要

N を患者集合 $\{1,\ldots,n\}$ とする．患者 i の m 個の属性を，m 次元ベクトル $\boldsymbol{x}_i = (x_{i,1},\ldots,x_{i,m})$ で表す. 疾病データセットは $D = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ とする．

D は h 個の病院に分散管理されており，病院 k は，患者 $N_k \subset N$ の患者に関する属性情報を持つ（患者はどこかの病院に属し，多重に属することは考えない）．すなわち，$N = N_1 \cup N_2 \cup \cdots \cup N_h$ と直和分割されている．

属性には，年齢などのカテゴリーデータ，症状の有無などを表すブール値データ，医療費などの連続値がある．例えば，変数 x_1: 年齢, x_2: 医療費，x_3: バイパス手術，y: 退院時死亡を表すこととすると，x_1 がカテゴリーデータ，x_2 が連続値，x_3, y がブール値である．i 番目のレコードは，$\boldsymbol{x}_i = (x_{i,1},\ldots,x_{i,3}) = (30, 37\text{万円}, 1,)$ とする．

問題 (1) 連続値の変数 (x_3) の平均値 $b_3 = 1/n\sum_i^n x_{3,i}$, カテゴリー変数 (x_1) の各値のカウント $a_{1,20} = |\{i \in N | x_{i,1} = 20\}$, $a_{1,30}$, ... を求める（クロス集計）.

問題 (2) D について, $y_i - p(y|\boldsymbol{x})$ を最小化するロジステック回帰モデル $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ を求めて，オッズ比（要因のリスク）を算出する（ロジステック分析）.

3.2 提案方式

病院 $k = 1,\ldots,h$ が互いに協力して，問題を解く．全病院は公開鍵を有し，病院からのデータを集計する集計者と秘密鍵を持つ信頼できる復号者（閾値以上の不正がない信頼出来るグループを考えれば良い）.

プロトコル 1 は問題 (1) のクロス集計，プロトコル 2 は問題 (2) のロジステック分析を行なう．説明を簡単にするために，カテゴリー変数 x_1 の値域 $\{20, 30, 40\}$, 連続変数 x_3 の値域 Z として説明する.

3.2.1 プロトコル 1:（水平分割クロス集計）

1. 病院 k は，$\boldsymbol{x}_i \in N_k$ について，各手術の患者カウント $b_{1,k}$

   $b_{1,20,k} = |\{i \in N_k | x_{i,1} = 20\}|, b_{1,30,k}, b_{1,40,k}$

   と医療費の総和 $a_{3,k}$

   $$a_{3,k} = \sum_{i \in N_k} x_{i,3}.$$

   を求める.

2. 病院 k は，カウント $b_{1,20,k}, b_{1,30,k}, b_{1,40,k}$ と総和 $a_{3,k}$ を公開鍵で暗号化して，$Enc(b_{1,20,k})$, $Enc(b_{1,30,k})$, $Enc(b_{1,40,k})$ と $Enc(a_{3,k})$ を集計者に送る.

3. 集計者は $c_{1,20} = \prod_k^h Enc(b_{1,20,k})$, $c_{1,30}$, $c_{1,40}$ と，$d_3 = \prod_k^h Enc[a_{3,k}]$ をそれぞれ求めて，復号者に送信する.

4. 復号者は $c_{1,20}, d_3$ 復号し，年齢分布 $Dec[c_1] = \sum_k b_{1,k}$ と平均治療費 $\hat{a}_3 = Dec[d_{20}]/n = 1/n\sum_k a_{3,k}$ を得る.

3.2.2 プロトコル 2: (水平分割ロジスティック回帰)

1. $t = 0$ とする．パラメータ $(\beta_0, \beta_1, \ldots, \beta_m) = (0, \cdots, 0)$ を初期化し，各病院 $k = 1, \ldots, h$ で共有する．

2. 病院 k では，N_k の患者のそれぞれ i について，説明変数 $\boldsymbol{x}_i$ について，

$$\begin{aligned} \nabla \boldsymbol{\beta}_k^{(t)} &= \left(\frac{\partial L}{\partial \beta_0}, \ldots, \frac{\partial L}{\partial \beta_m} \right) \\ &= \sum_{i=1}^{N_k} (p(\boldsymbol{x}_i) - y_i) x_i \end{aligned}$$

を計算し，暗号化して

$$\boldsymbol{e}_k = Enc[\nabla \boldsymbol{\beta}_k^{(t)} / B]$$

を集計者へ送る．ここで，B は整数化の為の定数であり，例えば小数点以下 3 桁の精度の時は，$B = 1000$ とする．

3. 集計者は $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_h$ の暗号文から $\prod_{k=1}^{h} \boldsymbol{e}_k$ を求めて，復号者に送り，復号した結果から，$\boldsymbol{d} = (d_1, \ldots, d_m) = Dec(\prod \boldsymbol{e}_k)/B$ を全病院へ同報する．

4. 各病院 k は，

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \eta \boldsymbol{d}$$

により各自のパラメータを更新する（全病院でパラメータを同期している）．ここで，η は収束の為の定数であり，例えば，$\eta = 0.001$ とする．

5. パラメータが収束条件（$\sum_{j=1}^{m} |d_j| < \theta$）を満たすまで，goto 2.

3.3 評価

プロトコル 1 は，分布を求める属性値の数に比例するコストがかかる．連続値の属性については，h 回の暗号化，カテゴリー値の属性は，値域の数について h 回の暗号化がかかる．プロトコル 2 は，各ラウンドについて，mh 個の暗号文を送信と m 回の復号処理がかかる．収束するまでのラウンド数に比例したコストがかかる．従って，1 回の暗号化と復号化に T_{enc}, T_{dec} かかり，収束までのラウンド数 t^* とすると，総処理時間は $t^* m(hT_{enc} + T_{dec})$ と表される．

患者データそのものは病院内に閉じており，要請条件 1 を満たす．集計者は秘密情報を持たないため，正直にプロトコルを守る限りでは，集計者と復号者が結託しない限り，患者情報は漏洩しない．プロトコル 2 では，偏微分した値が同報されてしまうので，各病院の患者の統計量に関する部分情報が漏れる．それについて，患者が特定されるかどうかは自明ではない．

4 試験実装と実験結果

4.1 擬似データセットによる解析結果

擬似 DPC データは，DPC データセットから，心臓手術に関する属性に制限し，個人識別に係わる情報を削除し，ランダム置換を行ったものである．データセットの属性数 $m = 45$, 総レコード数 $n = 5892$, 病院数 $h = 5$ 病院から成る．

4.2 心臓病術式評価における術式の比較

図 1 と図 2 に，術式大分類についての平均入院日数と手術費用の分布を表す．バイパス手術 (1) はカテーテル手術 (2) に比べて，長い間入院しなくてはならず，手術費用も高く付いていることがわかる．

表 1 に術式の違い (バイパス開胸手術とカテーテル手術，および細分類) による患者数と平均死亡率の比較を示す．

どちらの術式を選ぶかは，患者の状態や病院の設備などの様々な交絡因子があり，単純に平均死亡率だけでどちらのリスクが高いかと評価するのは早急である．例えば，バイパス手術を受けるのは体力のある若年が多いかも知れず，死亡率は年齢に大きく依存することを考慮すると，その評価は自明ではない．

そこで，プロトコル 1 を適用して，5 病院間でデータを人くしたままで年齢分布を算出する．結果を，図 3 に示す．両手術で患者の年齢分布

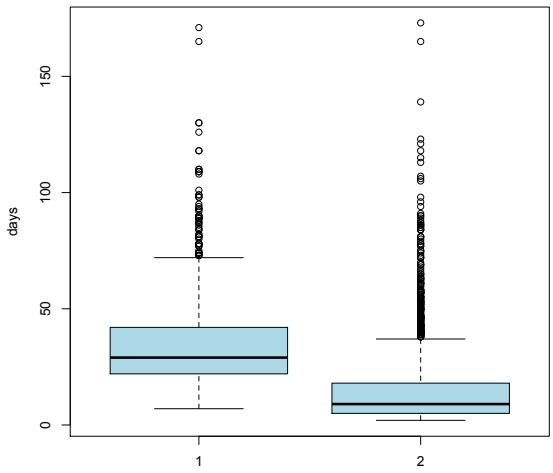

図 1: 入院日数の比較 (1=バイパス手術，2=カテーテル手術)

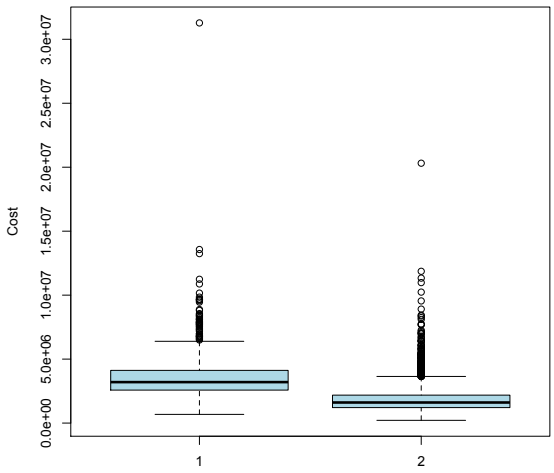

図 2: 入院費用の比較 (1=バイパス手術，2=カテーテル手術)

表 1: 術式の違いによる死亡率

術式	手術	個体数	死亡数	割合
バイパス	CABG			%
1 枝	K5881	114	1	0.88
2 枝	K5882	861	11	1.28
計		975	17	1.74
カテーテル	PCI			
風船治療	K614	2418	7	0.29
ステント	K615	2354	14	0.59
計		4772	21	0.44

には，総量に差があるだけで同様な分布をしていることがわかった．

4.3 水平分割ロジステック分析

表 2 に，平均死亡率に関してロジステック回帰を行い，要因を算出した結果（全データ）を示す．死亡率を上げている主要因は，緊急入院 (emergency = 1 緊急入院，0 計画入院）と腎不全 (jinfu = 1 あり，0 なし）であり，それぞれオッズ比で 4.7 倍，8.4 倍である．本分析の主要な興味である術式については，カテーテル手術 (PCI=2) がバイパス手術よりもオッズ比で 0.377 倍に死亡率を下げている．すなわち，バイパス手術は 1/0.377 = 2.65 倍カテーテル手術よりもリスクが高いことを表している．確率検定による P 値は，99.9%の確からしさで緊急と腎不全が有意であり，術式の差は 99%の水準で有意であった．

4.4 プライバシー保護ロジステック回帰実験

DPC データは属性数が多く大規模であるために，$m = 2$ 属性（ブール値と連続値）の $n = 21$

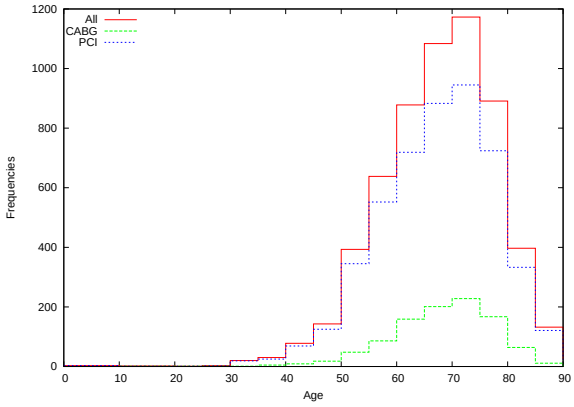

図 3: プロトコル 1 の結果，「治療法毎の年齢分布」

表 2: 平均死亡率に関する因子のロジステック分析結果

要因		予測 β	標準誤差 S.E.	$z(x)$	P		
定数	(Intercept)	-8.67926	1.39315	-6.23	4.67E-10	***	0.00017
術式大分類	(CABG=1,PCI=2)	-0.97433	0.3693	-2.638	0.00833	**	0.3774
緊急入院	emergency	1.56326	0.37422	4.177	2.95E-05	***	4.7743
性別	sex	0.56554	0.35589	1.589	0.11204		1.760398141
年齢	age	0.02808	0.01774	1.582	0.11356		1.028477959
心不全	sinfu	0.58581	0.3946	1.485	0.13766		1.796445517
脳血管	kekkan	0.21075	0.7394	0.285	0.77562		1.234603666
腎不全	jinfu	2.13025	0.39709	5.365	8.11E-08	***	8.416970791

レコードの擬似データ ([4]) を用意し，それを 2 分割 ($h = 2$) して，提案方式の実行可能性を検証した．暗号化には，1024 ビットの Paillier 暗号を用いて，非標準の TCP ソケットで 2 者間での同期処理を行い，プロトコルを実行した．

図 4 に，最小勾配法 (BGD) による誤差の収束の様子を示す．学習データに対する誤差は，最初は振動しているが，繰返し回数 $t > 1000$ から緩やかに減少していく．係数の予測値が収束するまでの t は要求する精度に依存するが，単純な逐次解法である BGD では 10 万回は繰返しが必要であった．

提案方式の正しさは，暗号化をするために，実数を定数倍して整数化する際に生じる丸め誤差により影響を受ける．本実験ででは，小数点 3 位まで（定数 $B = 1000$) の固定長実数で行った．図 5 と 6 に，統計ソフト R を用いたロジステック回帰の結果と提案方式で算出した結果の差を示す．小さな誤差が確認できるが，これらは前述した丸め誤差よりも，収束に至る終了条件の違いの方が大きい．図 5 を見ても，誤差は R の結果でも生じており，誤差は変わらない．

試験実装においては，$T_{enc} = 10$ ms 程度であり，一ラウンド 0.25[s] であった．$m = 3$ で収束までに $t^* = 30,000$ かかる試験データで，ロジスティック回帰を実施するのに約 2 時間かかった．

5 おわりに

病院間で患者情報を交換することなく，治療法の比較を実現可能な方式を提案した．より効率の良い収束アルゴリズム，複数の病院間での故障に強い非同期通信，パラメータから漏れる情報量の評価を今後の課題とする．

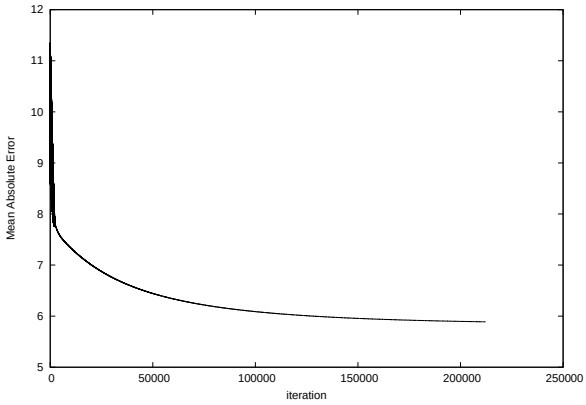

図 4: 最小勾配法 BGD による絶対平均誤差 MAE の収束

謝辞

本研究に際して，有益な議論を頂いた ACCESS の生路茂太氏，電通国際情報サービスの川村誠氏，電通の魚住高志氏，サイバー・コミュニケーションズの東貴己氏，小柳肇氏に感謝いたします．

参考文献

[1] 松田，伏見，診療情報による医療評価，DPC データから見る医療の質，東京大学出版会．

表 3: 術式細分類についてのロジステック分析

K5881		予測値	SE	Z 値	P 値 (両側検定)	有意
	(Intercept)	-4.542196	0.632116	-7.186	6.69E-13	***
sex	sex	0.258322	0.217781	1.186	0.236	
age	age	0.001392	0.009548	0.146	0.884	
心不全	sinfu	1.034634	0.229206	4.514	6.36E-06	***
脳血管	kekkan	0.552804	0.355418	1.555	0.12	
AIC	1105.9					
K5882		予測値	SE	Z 値	P 値 (両側検定)	有意
	(Intercept)	-2.183243	0.246997	-8.839	¡ 2e-16	***
sex	sex	-0.02899	0.089992	-0.322	0.747	
age	age	0.004989	0.003704	1.347	0.178	
心不全	sinfu	0.845791	0.103179	8.197	2.46E-16	***
脳血管	kekkan	0.812289	0.139003	5.844	5.11E-09	***
AIC	4883.6					
K614		予測値	SE	Z 値	P 値 (両側検定)	有意
	(Intercept)	-0.550259	0.184217	-2.987	0.00282	**
sex	sex	0.013174	0.068931	0.191	0.84844	
age	age	-0.002626	0.002777	-0.946	0.34438	
心不全	sinfu	-0.642185	0.108275	-5.931	3.01E-09	***
脳血管	kekkan	-0.378749	0.140872	-2.689	0.00718	**
AIC	7292.3					
K615		予測値	SE	Z 値	P 値 (両側検定)	有意
	(Intercept)	0.1359374	0.1715033	0.793	0.428	
sex	sex	-0.0176208	0.0637667	-0.276	0.7823	
age	age	-0.0003421	0.0025832	-0.132	0.8946	
心不全	sinfu	-0.1515618	0.0883675	-1.715	0.0863	.
脳血管	kekkan	-0.2658165	0.1214471	-2.189	0.0286	*
AIC	8151.6					

[2] 丹後，山岡，高木,「ロジステック回帰分析，SAS を利用した統計解析の実際」朝倉書店.

[3] 椿，岩崎,「R による健康科学データの統計分析」，朝倉書店.

[4] 高橋 信, 井上 いろは, トレンドプロ, マンガでわかる統計学 回帰分析編, オーム社, 2005.

[5] H. Yasunaga, H. Horiguchi, K. Kuwabara, S. Matsuda, K. Fushimi, H. Hashimoto, “Outcomes After Laparoscopic or Open Distal Gastrectomy for Early-Stage Gastric Cancer: A Propensity-Matched Analysis”, Annals of Surgery, Volume 257, Issue 4, pp. 640 □ 646, 2012.

[6] L. Sweeney, “k-anonymity: A model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10.05, pp. 557-570, 2002.

[7] 菊池，佐久間，三上，“プライバシーを保護したピロリ菌疫学調査”, 第 26 回人工知

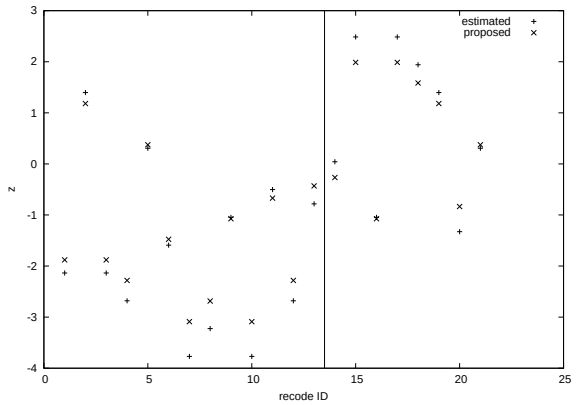

図 5: 提案プロトコルの回帰結果と学習データの関係 ($ID \geq 14$ が $y = 1$, 他が $y = 0$ である. 中央の線より右が 1, 左が 0 で, 誤りが 4 点 (2,5, 16, 20) 生じている

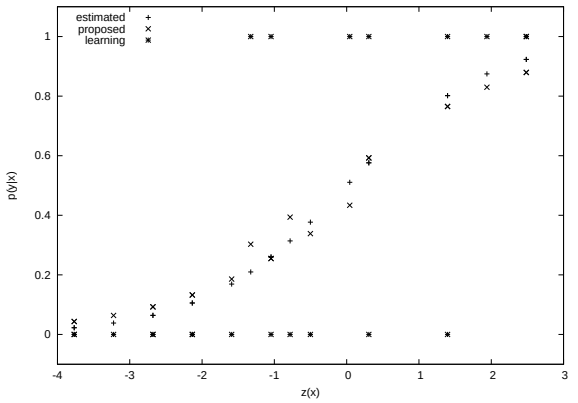

図 6: 提案プロトコルの z とメンバーシップ関数で算出した確率

能学会, 3I2-OS-20-9, pp. 1-4, 2012.

[8] Vaidya, J. and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data”, The Eighth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, SIGKDD, ACM Press, Edmonton, Canada, pp. 639-644, 2002.

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