プライバシーを保護した垂直分割線形回帰システムの実装とDPCデータセットを用いた評価

濱永千佳　菊池浩明
明治大学総合数理学部先端メディアサイエンス学科

康永秀生　松井宏樹　橋本英樹
東京大学大学院医学系研究科

研究背景

- 2015年9月、個人情報保護法改正
  - 要配慮情報（病歴、人種、信条 など）

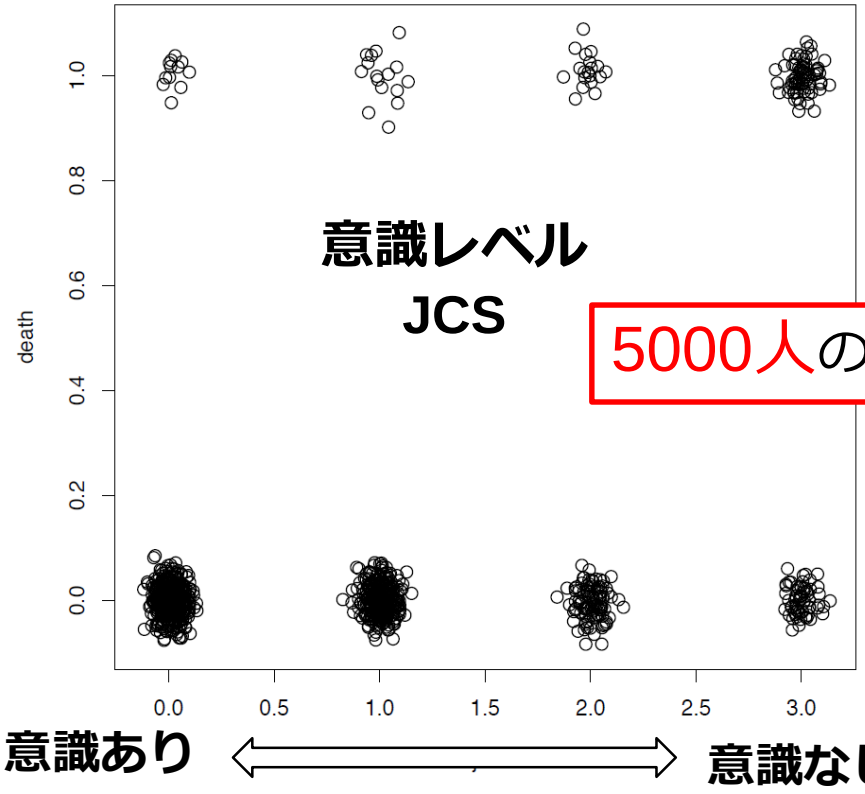
- 疾患や治療行為の表コードからなる患者の大規模データベース（DPCデータ）

意識レベル

death	age	sex	jcs	Cancer	HospitalVol	LiverDiseas	mRS	stroke_type
0	74	0	意識有り1	0	2	0	1	0
0	55	1	正常 0	0	1	0	4	0
1	71	0	意識なし3	0	0	0	5	0

目的：DPCデータで何が知りたいのか？

- 死亡という変数には、何が深くかかわっているのか？



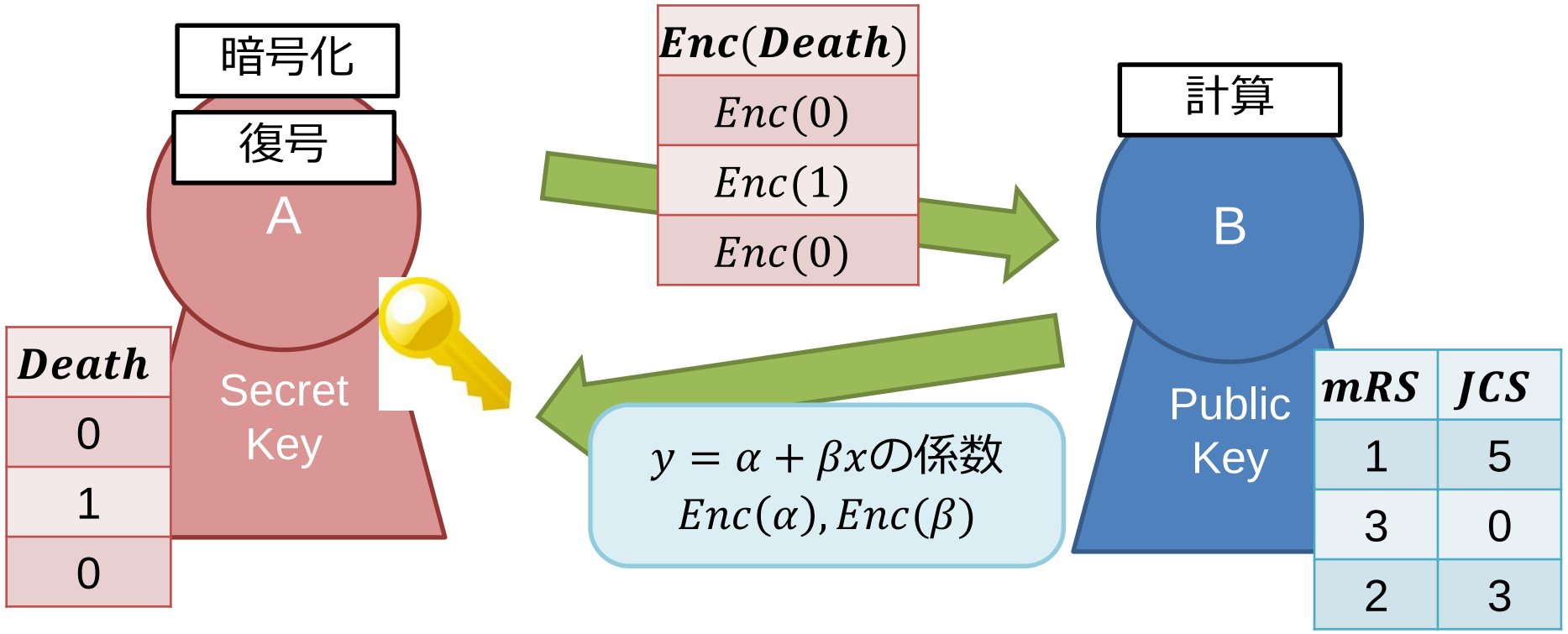
5000人の脳患者データ

- 個人情報保護の強化
  →本人の同意なく第三者提供ができない

解決手法：プライバシー保護データマイニング

目的：データを暗号化した状態のまま計算を行うことにより、
そのままのデータを公開せず、安全に活用すること。

・加法準同型性暗号
・重回帰の秘匿計算を行うこと



本研究で行ったこと

1. 3種類の秘匿線形回帰プロトコルを提案
2. システムとして実装
3. 5000人規模の実在患者データでの実験

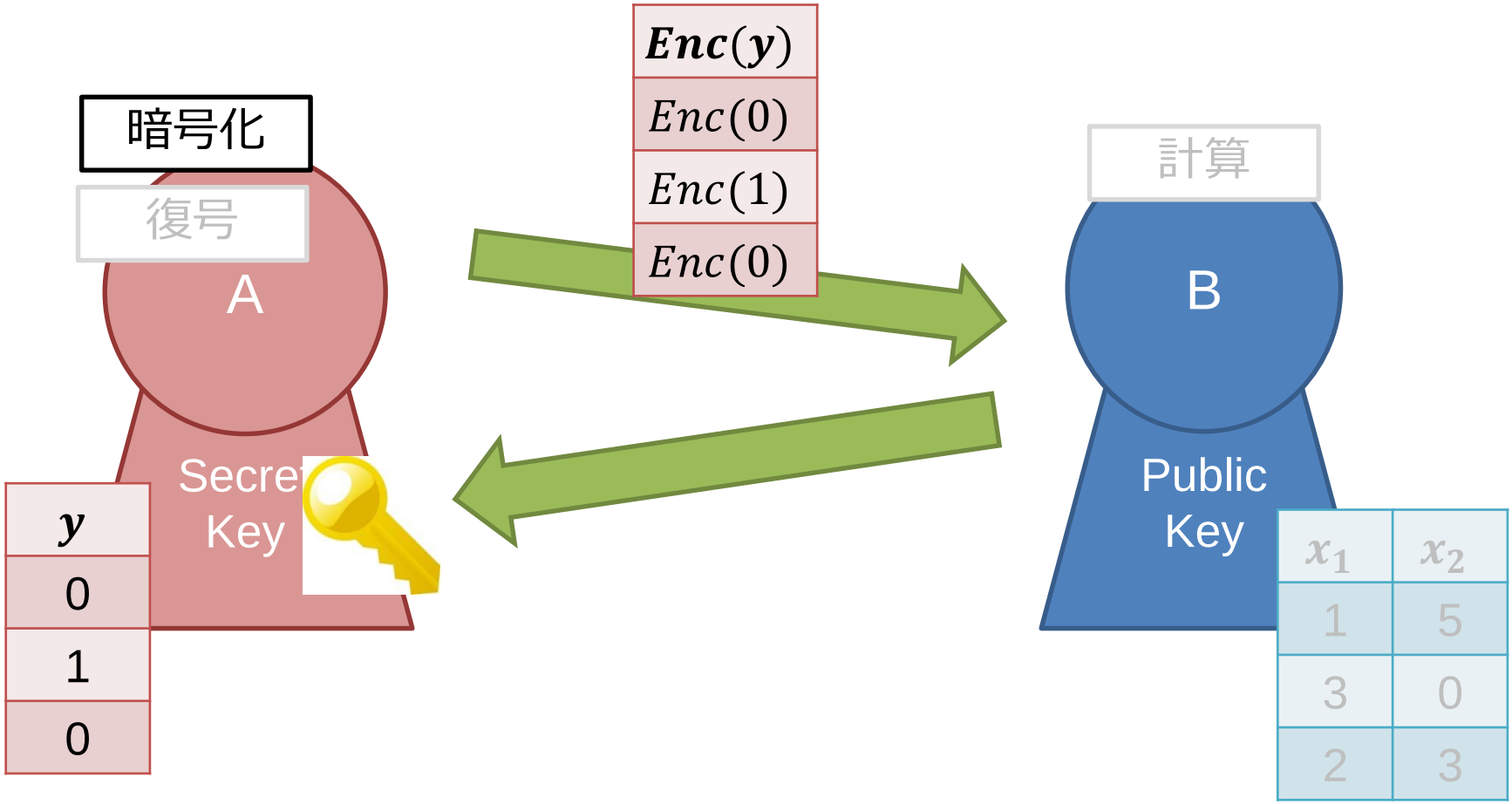
提案手法

1. 3種類の秘匿線形回帰プロトコルを提案

本発表

	（1）単回帰	（2）2変数の重回帰	（3）多変数の重回帰（n＝2）
モデル	$y = \alpha + \beta x$	$y = \alpha + \beta_1 x_1 + \beta_2 x_2$	$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$
BからAへ送るデータ暗号文数	C,D,E 3	C2,D2,E2,Σx1,Σx2 5	F2,G2 7

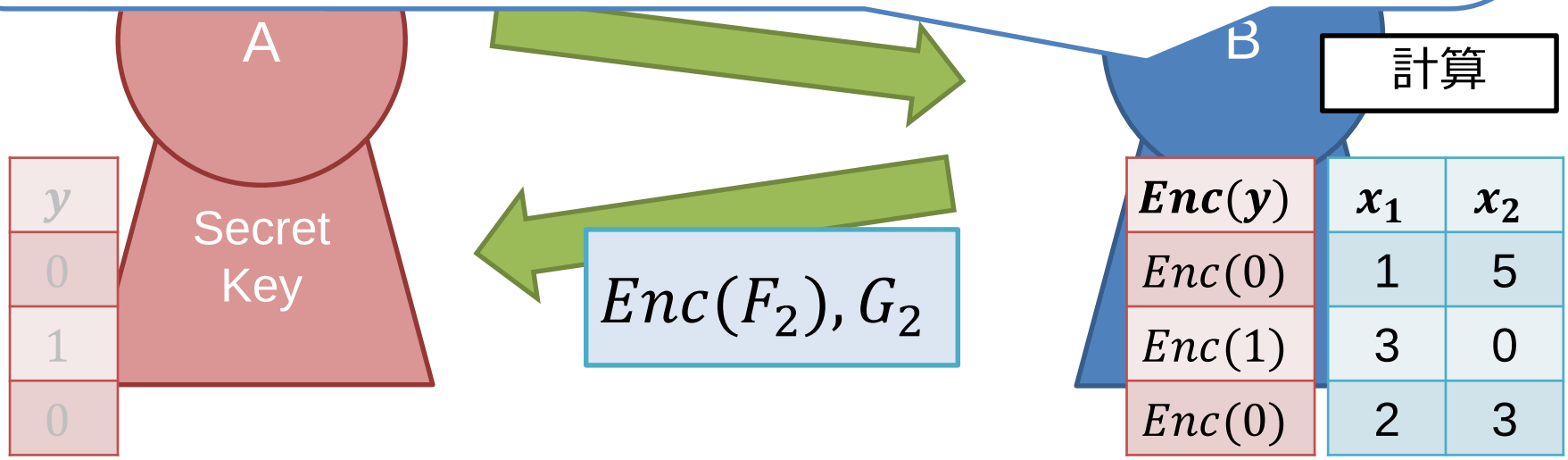
多変数に対応した重回帰　計算モデル　1/3



多変数に対応した重回帰　計算モデル　2/3

$$F_2 = \begin{pmatrix} \Sigma {x_1}^2 & \Sigma x_1 x_2 & \Sigma x_1 \\ \Sigma x_1 x_2 & \Sigma {x_2}^2 & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_2 & \Sigma 1 \end{pmatrix} = \begin{pmatrix} 14 & 11 & 6 \\ 11 & 34 & 8 \\ 6 & 8 & 3 \end{pmatrix}$$

$$G_2 = \begin{pmatrix} \Sigma x_1 y \\ \Sigma x_2 y \\ \Sigma y \end{pmatrix} = \begin{pmatrix} Enc(0)^1 \times Enc(1)^3 \times Enc(0)^2 \\ Enc(0)^5 \times Enc(1)^0 \times Enc(0)^3 \\ \Sigma y \end{pmatrix}$$

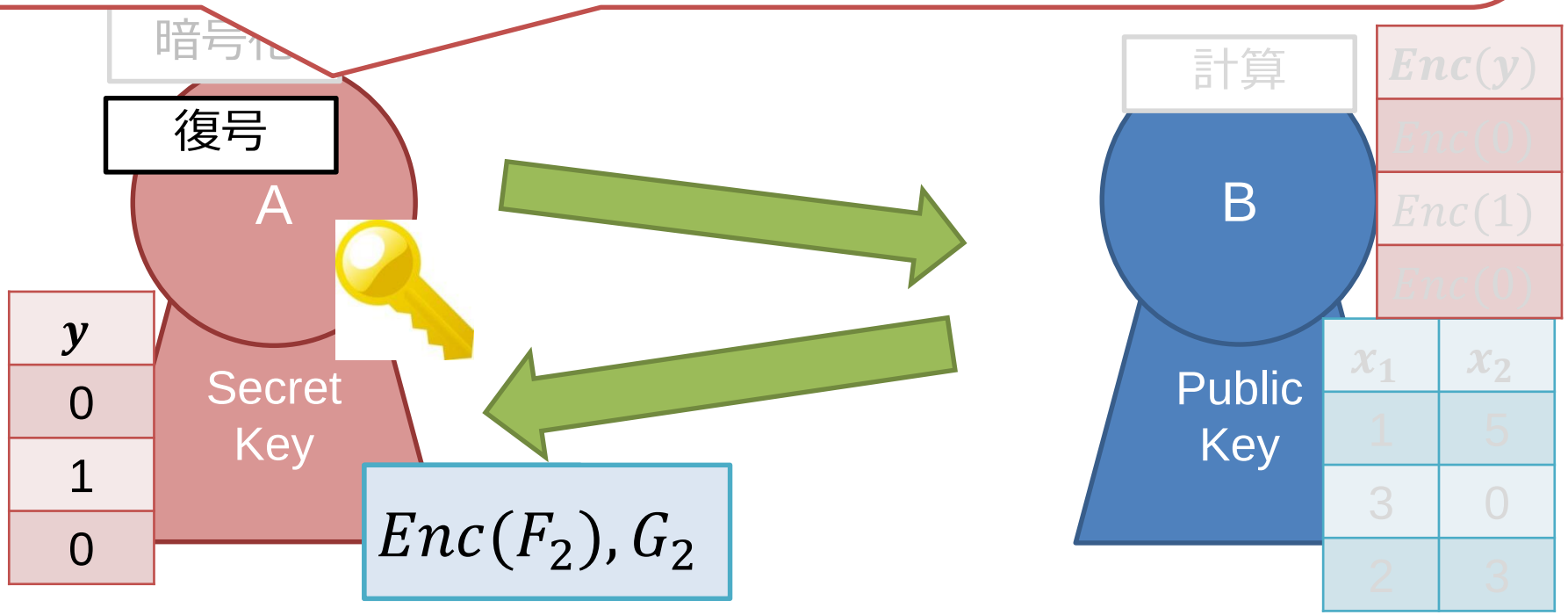


多変数に対応した重回帰　計算モデル　3/3

$$F_2 = \begin{pmatrix} 14 & 11 & 6 \\ 11 & 34 & 8 \\ 6 & 8 & 3 \end{pmatrix} \quad G_2 = \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}$$

$F_2 \begin{pmatrix} \beta_1 \\ \beta_2 \\ \alpha \end{pmatrix} = G_2$ を解き、

係数を求める



2．実装

- システム“scLinear”として実装

OS	Windows 7
メモリ	4 GB
CPU	Intel® Core™ i5-3337U
クロック	1.8 GHz
使用言語	Java(1.8.0_91-b14) R(3.1.0)
鍵長	2048[bit]

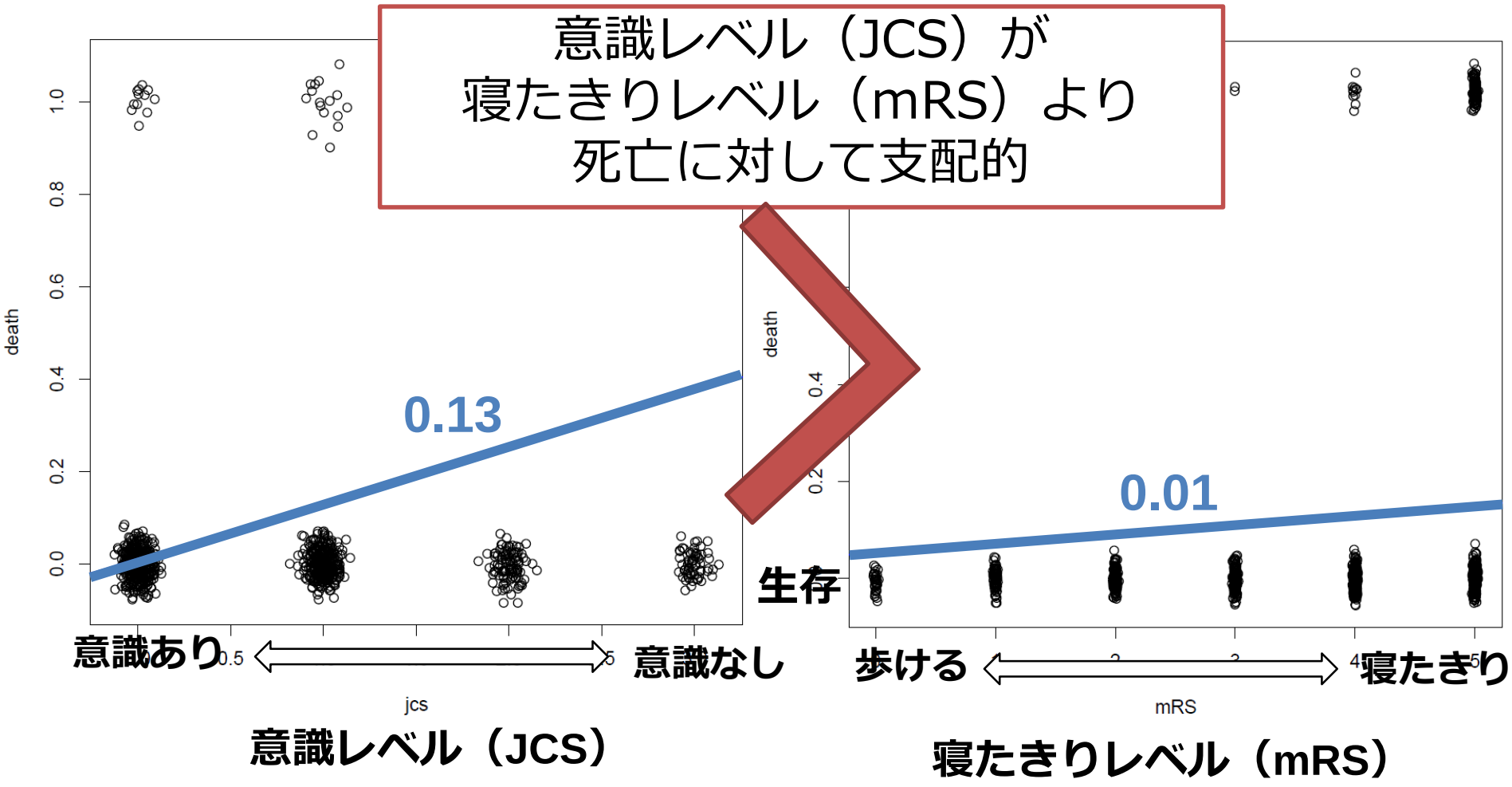
３．実験

- 互いに異なるデータセットを持つユーザ間での秘匿回帰分析を実施
  1. 単回帰と2変数の重回帰を、擬似データについてそれぞれ10回ずつ実施する。
  2. 5000人の脳患者DPCデータセットについて、m=3,4,5,6の重回帰を実施する。

- 実装したシステム“scLinear”の２項目を評価
  - 計算結果の正確さ
  - パフォーマンス（処理時間）

実験2　結果：線形回帰直線

意識レベル（JCS）が
寝たきりレベル（mRS）より
死亡に対して支配的



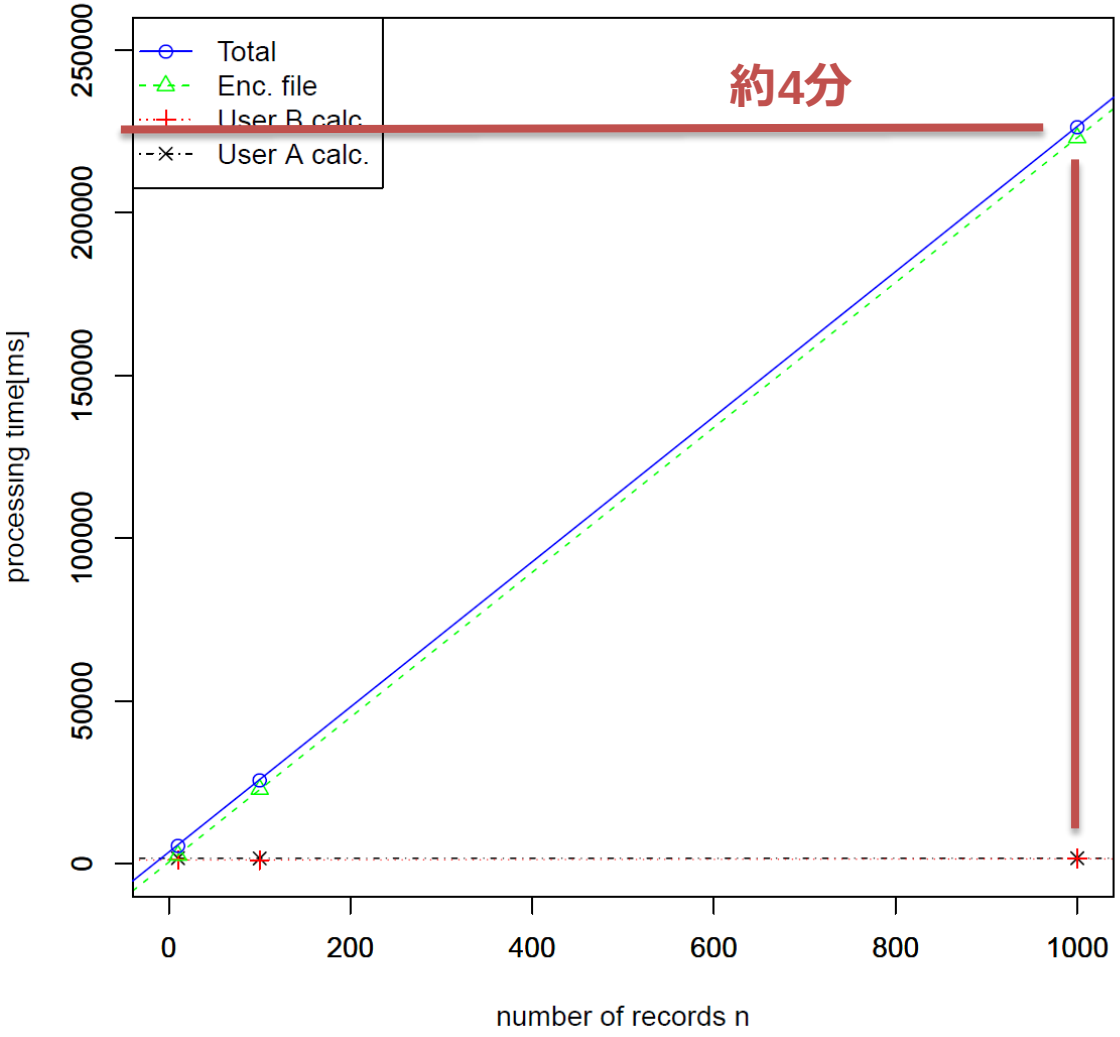
実験2　結果：計算結果の正確性

表 9　線形回帰モデルの係数と提案方式の比較 ($n = 5000$)

	提案方式	R				
variables	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	**
JapanComaScale	0.1283596	0.1283596	0.0049296	26.039	$< 2e-16$	***
modifiedRankinScale	0.0121227	0.0121227	0.0034845	3.479	0.000507	***
StrokeType	0.0292522	0.0292522	0.0073582	3.975	$7.12e-05$	***
LiverDisease	0.0095770	0.0095770	0.0324591	0.295	0.767970	

- 提案方式と統計ソフトRの結果に、差は見られなかった
  （*n=1000, 2000*の場合においても差がなかった）
  →提案システムは正確に計算できている

実験2　結果：パフォーマンス



- レコード数n対して線形に、システムの実行時間が増加している
- n=1000のとき、
  システム全体：226s = 約4分の処理時間を要している
- このシステムにおいて
  n=100万件のとき、
  　システム全体：2.60日
  であり、暗号化を除くシステム実行時間は約20分かかる

おわりに

- 本研究の結果
  - 2組織間におけるプライバシーを保護した線形回帰を求める3つのプロトコルを提案した
  - 5000人の実在脳患者データを用いた実験により、意識レベル（JCS）が寝たきりレベル（mRS）より死亡に対して10倍支配的であった
  - 実験より、小数第8位までの有効精度であることを示した
- 今後の課題
  - 相手に漏れてしまう情報の安全性評価