Vol.2013-CSEC-62 No.57
Vol.2013-SPT-6 No.57
2013/7/19

多目的コホート研究 (JPHC Study) データセットを用いた共通ID不在環境下におけるプライバシー保護データマイニングの事例研究

生路 茂太[1]　川村 誠[2]　魚住 高志[3]　東 貴己[4]　菊池 浩明[5]　井上 真奈美[6]

概要：対象者の疾病罹患の追跡が必須のコホート研究においては，複数の医療情報と連携することでより高い精度の研究が行えることが分かっている．しかし，医療情報は有用性の高い情報ほど機微である．そのためほとんどの場合，情報統合に必要な対象者と医療情報側の患者間の共通 ID も存在しない．本事例研究では，国立がん研究センターが保管している実際のコホート情報を用いて，共通 ID が存在しない前提でプライバシー保護データマイニングを行うための条件を考察する．

The Case Study of Privacy-Preserving Data Mining without Common Identification from the JPHC Study Data Set

Shigeta Ikuji[1]　Makoto Kawamura[2]　Takashi Uozumi[3]　Takami Azuma[4]　Hiroaki Kikuchi[5]
Manami Inoue[6]

Abstract: In the fields of medical information, an integration of the multiple data sets will lead more accurate and effective results in comparison to the research using one data set. Medical information integration has a risk of disclosure of confidential information and hence datasets don't have the common identification information in order to reduce possible risk factor to determine the identity of an individual.
In this paper, we will examine the condition of privacy-preserving data mining for the actual cohort, studied in the National Cancer Center. We study the necessary condition for data sets without common identification information to be integrated and generating more accurate and effective results.

1. はじめに

ある集団に対して，特定の要因の曝露と疾病の関係を一定期間追跡して観察することをコホート研究という．コホート研究は，コホートの観察を基に統計的手法を用いるため，その精度には観察量および観察情報の質の両方が必要であり，人的な労力を含めて実施コストが非常に大きい．

この対策として，異なる機関同士のコホートや診療情報を相互運用できることが望ましいが，プライバシーや個人情報保護との不整合により実現が難しい．国内の医療情報データベースについては，次の二つの理由により共通 ID が利用できないことが多い．1 つは，個人を特定するために必要な情報の順序や表記が，データセット毎に不統一であり，データセットの情報を隠蔽したまま hash 関数等による機械的な ID 作成が困難であること，もう 1 つは，共通 ID 情報そのものが個人を特定する機微な情報であるため流通できないという問題である．

我々は，この問題に対してプライバシー保護データマイニング技術の応用を検討している [1]．例えば，準同型暗号を用いた秘匿内積プロトコルなどにより一定の有効性が確認できているが，これはデータセット間の共通 ID の存在を前提としている．

本提案では，前者の問題について実在する大規模なコ

1 株式会社 ACCESS
2 株式会社電通国際情報サービス コミュニケーション IT 事業部
3 株式会社電通 プラットフォーム・ビジネス局
4 株式会社サイバー・コミュニケーションズ
5 明治大学 総合数理学部 先端メディアサイエンス学科
6 国立がん研究センター がん予防検診研究センター, 東京大学大学院 医学系研究科

ホートデータの内容から必要十分な個人情報の情報量を導出することにより，コホート間で必要な情報を明らかにする．後者の問題については，可換な一方向性関数を用いて互いに ID を開示せずに選択した集団同士の共通項を導出するための必要条件を考察する．さらに，データセット間の連携・データ量に増大による計算量が飛躍的に増加することが想定されるため，システムパフォーマンスについて検討を述べる．

2. 要素技術

2.1 可換な一方向性関数を用いた秘匿積集合

AES03 は Agrawal らによって提案された [9] ．2 つの集合 X と Y を，お互いに開示することなく，2 つの集合の共通集合 $X \cap Y$，または，共通集合の要素数 $|X \cap Y|$ を求めることができる．AES03 のアルゴリズムを Algorithm 1 に示す．

Algorithm 1 AES03[9](可換一方向性関数)

入力: 集合 $X = x_1, \ldots, x_{(n_A)}$ を持つ A と $Y = y_1, \ldots, y_{(n_B)}$ を持つ B.

出力: $|X \cap Y|$ を求める．

位数 q の巡回群 G と G を値域とするハッシュ関数 H を考える．

1. A は，乱数 $u \in Z_q$ を選び，$H_{(x_1)^u}, \ldots, H_{(x_{n_A})^u}$ を B へ送る．
2. B は，乱数 $v \in Z_q$ を選び，$H_{(y_1)^v}, \ldots, H_{(y_{n_B})^v}$ と $H_{(x_1)^{uv}}, \ldots, H_{(x_{n_A})^{uv}}$ を求めて A へシャッフルして送る．
3. A は，$H_{(y_i)^{vu}} = H_{(x_j)^{uv}}$ を満たす x_j, y_i の組の個数 $(= |X \cap Y|)$ を求める．

2.2 多目的コホート研究

コホートとは，疫学に用いるための特定集団である．また，コホートデータとは，通常追跡調査結果を含む大規模なデータセットである．多目的コホート研究は，「多目的コホート研究に基づくがん予防など健康の維持・増進に役立つエビデンスの構築に関する研究」（主任研究者 津金昌一郎 国立がん研究センターがん予防・検診研究センター長）において，全国 11 保健所と国立がん研究センター，国立循環器病研究センター，大学，研究機関，医療機関などが実施している共同研究である．

2.3 PSO モデル

安井らの定義する PSO（Privacy Search Oracle）モデルでは，国内人口 1.2 億人に対して個人を絞り込むために必要な情報量を氏名や生年月日などの個人情報に対して「絞り込み量」として定義している [2]．PSO モデルによれば，個人特定に必要な絞り込み量は，理論上 27bit であるが，安井らにより，主観的な個人情報の定量評価を表 1 に基づいて，実際のブログ等により危険度レベルと絞り込み量の分布を調査した結果，実際に個人を特定するためには，27bit 以上の絞り込み量が必要なことが分かっている．

表 2 に，多目的コホートが持つ個人情報の絞り込み量を示す．また，表中の個人情報ごとに，異体字に代表される表現の揺らぎ補正の難易度を示す．ID 重複最大母数は，多目的コホート上で実際に絞り込みを行った結果，同一同名などの一意に特定できなかった集団の最も大きな母数を指す．例えば，漢字氏名については，同姓同名が最大 24 人いることを示す．性別については，総女性数が 61,020 人であることを示す．住所についても，同一住所に 56 名の重複が見られたが，正しいデータセット上の実例である．

表 1 PSO モデルに基づく危険度レベルの定義 [2]

レベル	危険度レベルの定義	絞り込み量分布 [bit]
1	個人を特定できる情報がほとんど漏洩しておらず 個人を特定するのが非常に困難である状態	∼27
2	このままでは個人を特定できる状態とは言えないが 他の情報がかなり漏洩している状態	27∼50
3	探偵などの専門家を通すことで 個人が特定できる可能性がある状態	27∼81
4	地図や電話帳などの他のデータベースなどを 使用することで個人を特定することが可能な状態	50∼81
5	氏名・住所・電話番号など個人を 完全に特定できる情報が漏れている状態	81∼

2.4 曝露と疾病の関連評価

2.4.1 相対危険度

ある要因の曝露と疾病との関連の強さを示す最も単純な方法として，下記の式と表 3 で示されるように相対危険度（RR：Relative Risk）が知られている．

$$RR = \frac{a}{a+b} / \frac{c}{c+d}$$

表 3 曝露と疾病のデータ分布

	疾病あり	疾病なし	合計
曝露あり	a	b	$a+b$
曝露なし	c	d	$c+d$

2.4.2 交絡因子

ある要因 A の曝露と疾病 X の関係を考えた際に，A 以外に疾病 X の原因となる要因 B が存在する場合がある．このとき，要因 A を予測因子（Predictor Factor），疾病 X を結果因子（Outcome Factor），要因 B を交絡因子（Confounding Factor）と呼ぶ．

例えば，身体活動量とがんの関連を評価する際の肥満度を考える．身体活動量の導出には，予め計測された運動と運動強度の対応表を用いる [3]．この際，運動強度の単位

表 2 多目的コホート保有個人情報毎の絞り込み量と絞り込み結果について

	PSO モデル 絞り込み量 [bit][2]	揺らぎ 補正	ID 重複 最大母数	備考
氏名漢字	27	困難	24	データ入力者によって利用する漢字体表記に揺らぎがある． 異体字の解決は困難のため「ひらがな」や「カタカナ」の利用が望ましい．
氏名カナ	未定義	可能	30	データ入力者によって拗音や促音の表記に揺らぎがある．
性別	1	不要	61,020	揺らぎはほぼ存在しない．
生年月日	15	不要	86	揺らぎはほぼ存在しない．
住所	26	可能	56	市区町村以降の住所を示す．番地まで保証されるが号は任意．方書は含まない． 全角数字，半角数字，漢数字などの表示の揺らぎが存在する．
市区町村	14	要検討	12,131	異体字の揺らぎに加え，地域や時期また入力者の判断による 群や字などの扱いポリシーの揺らぎが存在する．住所の完全従属情報となる．
都道府県	6	不要	22,336	揺らぎはほぼ存在しない．市区町村の完全従属情報となる．
方書	2	困難	未計測	集合住宅のマンション名や号室等の補足的な住所情報を示す． 入力自体が任意のため，データ入力者によって入力有無を含めた揺らぎが存在する．

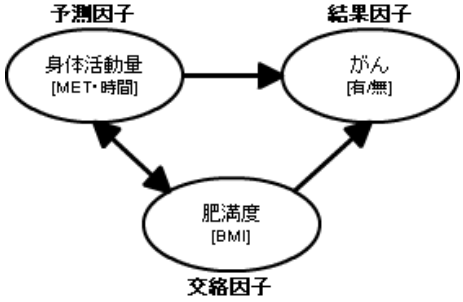

図 1 予測因子と結果因子と交絡因子の関係

は，運動強度指数 MET（Metabolic Equivalent）によって示され，身体活動量は「MET・時間」で示される．肥満度は BMI（Body Mass Index）で示される．図 1 に示すように，MET・時間と BMI は相互に関連し，かつ両方共にがんの原因と疑われる．このとき，身体活動量は予測因子，がんは結果因子，肥満度は交絡因子となるため，予測因子と結果因子を正しく評価する際は，交絡因子の影響を補正する必要がある．

2.4.3 カイ二乗検定

導出された相対危険度の妥当性は，統計学的有意水準 α に基づく統計学的有意差 p 値で評価される．$N > 8$ である場合 [7]，を 5%とした際の p 値は，χ^2 検定によって次のように求められる．

$$\chi^2_{(\alpha)} = \chi^2_{(0.05)} = 3.84$$

$$N = a + b + c + d$$

$$\chi^2_{(p)} = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

上記より，$\chi^2_{(p)} < \chi^2_{(\alpha)} = 3.84$ の際に $p \leq 0.05$ と同義となるため，統計学的に有意である．

3. 提案方式

3.1 アイデア

互いに秘匿されたデータセット間であっても，名前や生年月日などの個人情報から個人を特定できる．よって，十分な個人情報があれば，AES03[9] を用いることで，互いに公開された共通 ID を用いずにデータの共通集合から疫学に有益な特定の要因の曝露と疾病の関係を示すことができる．

本提案では，個人情報から個人を特定する条件を明らかにし，AES03 によって機微な情報を公開せずに相対危険度が導出できることを示す．

3.2 個人特定方法

3.2.1 氏名の利用に関する問題

PSO モデルで氏名の絞り込み量を 27bit としていることからも分かるように，氏名は個人特定に有効な情報である．しかし，実際にシステム実装を行うにあたっては二つの問題がある．

一つは，異体字やシステム独自拡張外字などの文字コード問題であるが，「住基ネット統一文字と戸籍統一文字を拠りどころに国内で運用中のデータセットの外字を整理し UTF-16 と IVS/IVD で対応しても．市町村に残存するそれ以外の外字 (約 37,000 字) に対応できない」[6]．そのため，より広い範囲でデータセット間で個人情報より個人特定を行うためには，外字が介在する氏名漢字の利用を避け，氏名カナを用いる必要がある．

もう一つは，同姓同名問題である．2001 年発行 NTT ハローページ登録者人中の 2 人以上の漢字氏名の同姓同名者総数は，母数 30,552,849 人中 19,281,386 人となり過半数が一意に特定できない．そのため，多目的コホートのデータセットより氏名カナの同姓同名集団の分布を導出した上，

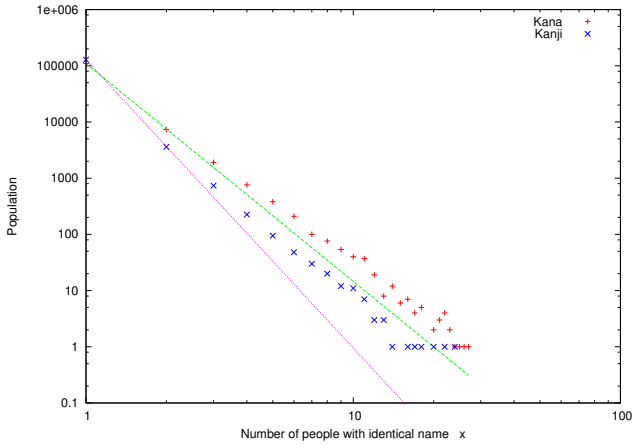

図 2 氏名漢字/氏名カナ 同姓同名集団母数分布（多目的コホート）

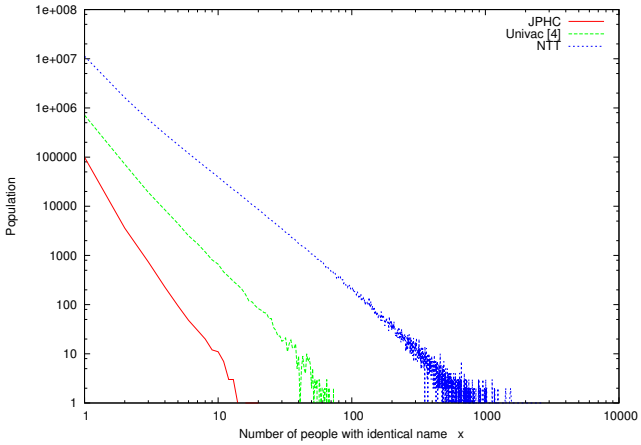

図 3 母集団別同姓同名集団母数分布

氏名カナの絞り込み量を再定義する必要がある．

3.2.1.1 同姓同名集団の分布

多目的コホートやハローページの同姓同名集団のランクとサイズの分布を図 2, 3 に観察したところ，千田，間瀬らの先行研究 [5] 同様に以下の式に示す Zipf の法則が観察された．すなわち，同姓同名人数 x 人となる氏名の数 $f(x)$ は，出現頻度についての第 x 位の数について全体の割合が $1/x$ に比例するとする Zipf の経験則でモデル化できる．これを，$1/x^s$ と一般化して，最小二乗法で多目的コホートにあてはめを行い，次の式を得る．

$$f(x) = \frac{a}{x^s} = \frac{110000}{x^{3.87}}$$

この近似式を用いて，全人口 $D = 1.2$ 億人とした際の氏名カナの 2 人以上の同姓同名数 a を導出する．

$$D = a \sum_{k=1} \frac{1}{k^{3.87}}$$

$$a = \frac{D}{(\sum_{k=1} \frac{1}{k^{3.87}})} \simeq \frac{D}{1.1} \simeq 109\mathrm{e}^6$$

導出の結果，氏名カナについて全人口 1.2 億人中の 1.09 億人には同姓同名の存在が予想されるため，他の個人情報との組み合わせの検討が必須である．

3.2.1.2 氏名カナの絞り込み量

氏名漢字と氏名カナの名前を情報源とみなすと，それぞれのエントロピーは以下の式で与えられる．

$$H(S) = \sum_k P(k) \log(P(k)) \ \ [\mathrm{bit/symbol}]$$

上記の式に基づいて，多目的コホート 140,420 件のうち個人情報が完全に記入された 111,458 件のデータに対して導出を行った．その結果，氏名漢字と氏名カナのエントロピーはそれぞれ 14.63 bit と 13.71 bit となった．

3.2.2 共通 ID 生成に必要な個人属性組み合わせ

多目的コホートの個人情報が記入済み 111,458 件のデータに対して個人情報を用いて ID 化を行った結果を表 4 に示す．

表中の絞り込み量は，安井らの PSO モデルから算出した理論値だが，氏名カナについては前項で再定義した 13bit を用いた．ID 重複最大母数は，同姓同名などの ID 重複集団の最大母数を示す．未解決レコードは，一意に ID が特定できなかった数であり，組み合わせ D と E が個人特定に成功した．このことから，組み合わせ D と E が氏名カナを用いた個人絞り込みの必要条件であることが示された．

3.2.3 AES03 による相対危険度の導出

コホートを持つ研究機関は，ライフログやヘルスケアなど外部のデータセットが持つ多用な要因と疾病の関連を調べたい．しかし，機微な疾病に関わる情報は，ID 情報を含めて隠蔽したい．そのため，機微な疾病情報を管理する医療機関 Alice と身体活動量や生活習慣など比較的機微でない情報を持つ事業者 Bob のやり取りを設定する．

WHO の発表 [10] によれば，科学的根拠の基づき，結腸がんに相関が「確実」なリスク要因として，リスク低下については身体活動量，リスク増加については肥満がそれぞれ挙げられている．そこで，本実験では，多目的コホート 140,420 件を元に，結果因子 of=がん情報（結腸がん有=1/無=0），交絡因子 cf=肥満度（BMI 値 27 以上=1/未満=0），予測因子 $pf = 4$ 段階の身体活動量（Lowest, Second, Third, Highest）とする．このうち，Alice は属性 of と cf を，Bob は属性 pf を持つように垂直分割している．

実験の結果，運動をほとんど行わない集団に対して，運動を行う各集団の結腸がんのリスクが低い傾向が観察された．導出過程で明らかになった共通集合は，表 6, 7 と図 4 に示す．

表 4 データセット要素組み合わせの情報エントロピー

組み合わせ	データ要素	PSO モデル絞り込み量合計 [bit]	ID 重複最大母数	未解決レコード数
A	氏名カナ＋性	14	30	30,180
B	氏名カナ＋性＋生年月日	30	2	16
C	氏名カナ＋性＋生年月日＋都道府県	36	2	12
D	氏名カナ＋性＋生年月日＋住所	56	重複 ID なし	0
E	氏名カナ＋生年月日＋住所	55	重複 ID なし	0
F	氏名カナ＋住所	40	2	16
G	性＋生年月日＋住所	42	2	10

表 5 Algorithm 2 の分割表

	$\lvert X \cap Y_p \rvert$	$\lvert Y_p - (X \cap Y_p) \rvert$	Y_p	RR	$N_{(p)}$	$\chi^2_{(p)}$
Y_1	c	d	$c+d$	1.0	-	Reference
Y_2	a_2	b_2	a_2+b_2	$\frac{a_2}{a_2+b_2} / \frac{c}{c+d}$	a_2+b_2+c+d	$\frac{N_{(2)}(a_2d-b_2c)^2}{(a_2+b_2)(c+d)(a_2+c)(b_2+d)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_q	a_q	b_q	a_q+b_q	$\frac{a_q}{a_q+b_q} / \frac{c}{c+d}$	a_q+b_q+c+d	$\frac{N_{(q)}(a_qd-b_qc)^2}{(a_q+b_q)(c+d)(a_q+c)(b_q+d)}$

Algorithm 2 AES03 を用いた $q \times 2$ 分割表の相対危険度の導出

入力:

i. 全体集合 U と結果因子 $of = 1$ と交絡因子 $cf = 0$ となるような U の部分集合 $X = x_1, x_2, \ldots, x_n$ を保有する Alice.

ii. 予測因子 $pf \in 1, 2, \ldots, q$ を持つ集合 Y_p によって U を分割した $Y_1, Y_2, \ldots, Y_q$ を保有する Bob. すなわち, $Y_1 \cup Y_2 \cup \ldots \cup Y_q = U$, $Y_i \cap Y_j = \varphi$ for all $\neq j$

iii. 有意水準 α.

出力:

iv. Y_1 から Y_q の表 5 に示す分割表.

v. Y_p の Y_1 に対する相対危険度 $RR for p = 2, \ldots, q$.

vi. 有意水準 α に基づく Y_p の Y_1 に対する統計量 $\chi^2_{(p)}$.

Step1. Alice は，Bob と協力し，AES03 を用いて表 5 の c を求める.

$$c = |X \cap Y_1|$$

同様に，$p = 2, \ldots, q-1$ について求める.

$$a_p = |X \cap Y_p|$$

Step2. Alice は，c と Y_1 より d を求める.

$$d = |Y_1| - c$$

同様に，$p = 2, \ldots, q-1$ について求める.

$$b_p = |Y_p| - a_p$$

最後に，$p = q$ について求める.

$$a_q = |U| - (c + \textstyle\sum_{p=2}^{q-1} a_p)$$
$$b_q = |Y_q| - a_q$$

Step3. Alice は，表 5 に従い，RR と $\chi^2_{(p)}$ を求める.

結果の比較として，図 5 に異なる母数と導出方法で導出された身体活動量と結腸がんについての先行研究を示す [10]．いずれの場合も，身体活動量が最も低い集団に対して，他の集団の結腸がんのリスクは低下した．しかし，女性は男性に比較して強い相関がみられなかった．この理由として，本実験では，交絡因子に BMI 値のみを考慮したためと考えられる．実際のコホート研究において交絡因子は，飲酒，喫煙，年齢等を多面的に考慮されるため，BMI 値のみの考慮による相対危険度は，十分正確でない可能性がある．

表 6 身体活動量を曝露として結腸がんを疾病としたデータ分布（男性）

	X (178)	$\lvert Y_p - (X \cap Y_p) \rvert$ (41,108)	$\lvert Y_p \rvert$ (41,286)	RR	$\chi^2_{(p)}$
L(16,374)	79	13915	13994	1.00	Reference
S(9,594)	36	8229	8265	0.77	1.68
T(9,085)	25	7865	7890	0.56	6.54
H(11,184)	32	9830	9862	0.57	7.20

表 7 身体活動量を曝露として結腸がんを疾病としたデータ分布（女性）

	X (130)	$\lvert Y_p - (X \cap Y_p) \rvert$ (46,330)	$\lvert Y_p \rvert$ (46,460)	RR	$\chi^2_{(p)}$
L(17,404)	40	14347	14387	1.00	Reference
S(13,795)	32	11703	11735	0.98	0.01
T(11,865)	32	10283	10315	1.12	0.21
H(9,827)	19	8473	8492	0.80	0.61

4. 評価

4.1 AES03 のシステムパフォーマンス特性

14 万件の多目的コホートを用いて AES03 のシステムパフォーマンス特性を計測した．医療機関 Alice と Alice に協力する事業者 Bob を定義する．Alice は，件数の特定を避けるため常に全件を使用する．双方のシステムは，インターネット網上の秘匿通信によって接続される．表 8 の実験環境上で，図 A·1 のシーケンスに基づいて試験実装した

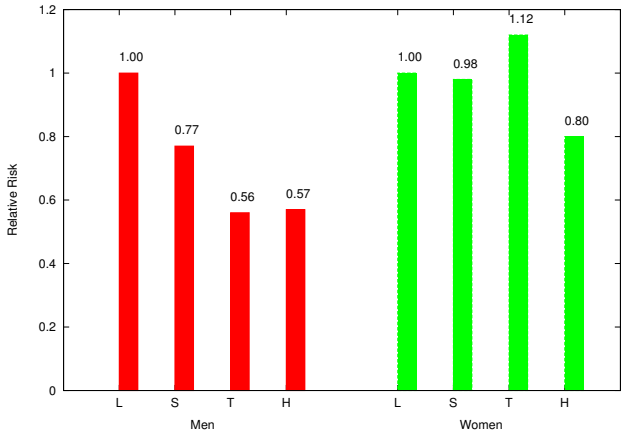

図 4 AES03 により導出した相対危険度による身体活動量と結腸がんの関係

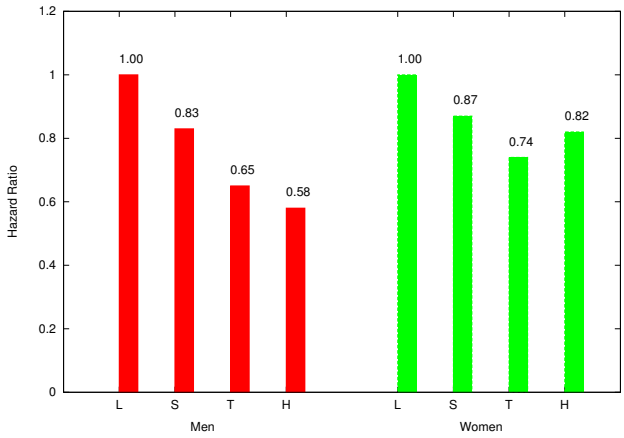

図 5 JPHC 研究事例のハザード比による身体活動量と結腸がんの関係 [8]

プログラムの処理時間を図 6, 7, に示す．Alice の 140,420 件固定に対して，Bob から 10,000 件，35,000 件，70,000 件，140,420 件の照合件数 $|P|$ をそれぞれ 10 回行い，処理時間の平均を取った．

Seq-a1 の値が，Bob 側のデータ量に関わらず一律に高い処理コストになっており，通信処理などの他要素は全体に占める割合は小さい．今回は測定が目的のため行っていないが，X が事前にランダム化されている前提で Seq-a2 を非同期に送信することで Seq-b2 開始までの時間を最小にすることで速度向上を行う余地がある．同様に Seq-b3 の通信についても部分的な非同期処理により対向側の待ち時間を圧縮できる余地がある．

4.2 本システムの想定するセキュリティプロトコル

本実験で用いた図 A·1 のシーケンスにおいて，通信上にデータが通過する Seq-a2，Seq-b3 に注目する．Seq-a2 のデータについては，SHA-1 で作成された Hash 値と乱数 u

表 8 実証実験環境

法 p の大きさ	2048bit
巡回群 G の位数 q	160bit
乱数 u，v	160bit
Application impl.	Scala
SHA-1	Java sphlib
modulo	Java Big integer
Data Store	csv text
Data Structure	Java HashSet Collection
OS	Ubuntu 12.10 amd64
CPU	Intel Celeron Processor G1610
Memory	4GB (DDR3 SDRM PC3-10600)
network speed	46Mbps (measured values average)

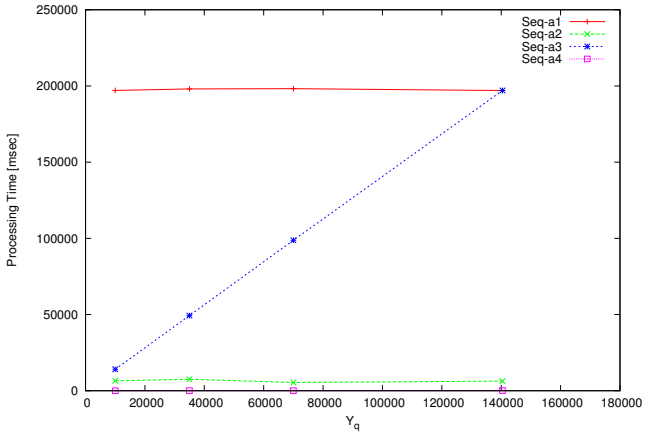

図 6 Alice のシステムパフォーマンス特性

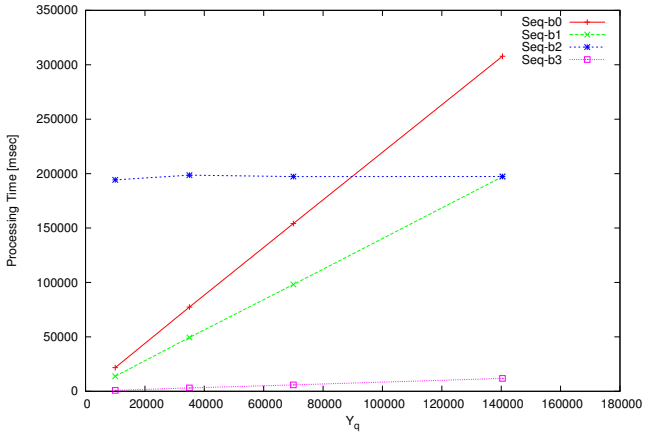

図 7 Bob のシステムパフォーマンス特性

で秘匿されているため総当たり攻撃の成功確率 S は以下の式で示すように十分に低い．

$$S = \frac{1}{|u|} = 2^{-160}$$

Seq-b3 は，同様の理由で安全である．さらに相手のデータを扱う A3, B5, B6 について考える．AES03 の安全性に

ついては，Agrawal らによって，AES03 を破り相手の持つ集合の要素を特定することのできる攻撃は，決定 DH 問題に帰着できることが証明されている [9]．よって，Seq-b2，Seq-a3，Seq-a4 のデータから秘匿された個人情報が漏れることはないと考える．

5. おわりに

氏名カナを含む個人情報を用いて個人を一意に特定するために，生年月日と住所が必要であることを PSO モデルの情報量と多目的コホートの実証を用いて示した．次に，AES03 を用いることで，互いに機微な情報を秘匿しながら相対危険度が導出できることを示した．今後の課題としては，住所の漢字コード利用回避のための，市区町村コード等の利用を検討，および並列処理などを用いた AES03 のパフォーマンス向上が挙げられる．

謝辞

本研究は，多くの関係者のご協力の下で推進させていただきました．特に，国立がん研究センターの津金昌一郎先生，株式会社サイバー・コミュニケーションズの小柳肇様，田口剛様，株式会社 ACCESS の加藤健二様，東海大学大学院の大久保成晃様には，物心両面で多大なご協力を賜りました．この場を借りて，御礼申し上げます．

参考文献

[1] 川村誠, 生路茂太, 小柳肇, 菊池浩明, “Hadoop を用いた大規模分散プライバシー保護システムと医療情報統合への応用”, 暗号と情報セキュリティシンポジウム（SCIS2013）, 1C1-5, pp. 1-6, 2013.
[2] 安井, 佐藤, 釘谷, 金井, 廣田, 谷本, “ブログにおける個人情報漏えいレベルの定量化”, IPSJ SIG Technical Report, Vol. 2009-EIP-43, pp. 9-16, 2009.
[3] 田畑 泉, “特定健診と特定保健指導の概要 －運動基準・運動指針（エクササイズガイド）との関連－”, 早稲田大学スポーツ科学学術院 スポーツ科学研究, Vol. 6, pp. 36-39, 2009.
[4] 田中康仁, “同姓同名の発生頻度”, 情報処理学会研究報告 自然言語処理 1977-NL-010, pp. 1-7, 1977.
[5] 千田, 間瀬, “日本人の名字の統計解析”, 日本統計学会誌, 第 35 巻 第 1 号, 2005.
[6] 榎並, “電子行政における外字問題の解決に向けて”, 富士通総研経済研究所 研究レポート, No. 400, 2013.
[7] 鈴木眞男, “2 × 2 分割表における chi-squre 検定と Yates の修正に関する最近の検討”, 愛知教育大学研究報告, 32(自然科学編), pp. 13-17, 1983.
[8] Inoue et al. Daily Total Physical Activity Level and Total Cancer Risk in Men and Women: Results from a Large-scale Population-based Cohort Study in Japan. Am J Epiderniol, 168, pp. 391-403, 2008.
[9] Rakesh Agrawal, Alexandre Evfimievski, and Ramakrishnan Srikant, “Information shareing across private databases”, in proc. of ACM SIGMOD International Conference on Management of Data, 2003.
[10] Report of a Joint WHO/FAO Expert Consultation, “Diet, nutrition and the prevention of chronic diseases”, WHO technical report series, 916, pp. 100, 2003.

付　　録

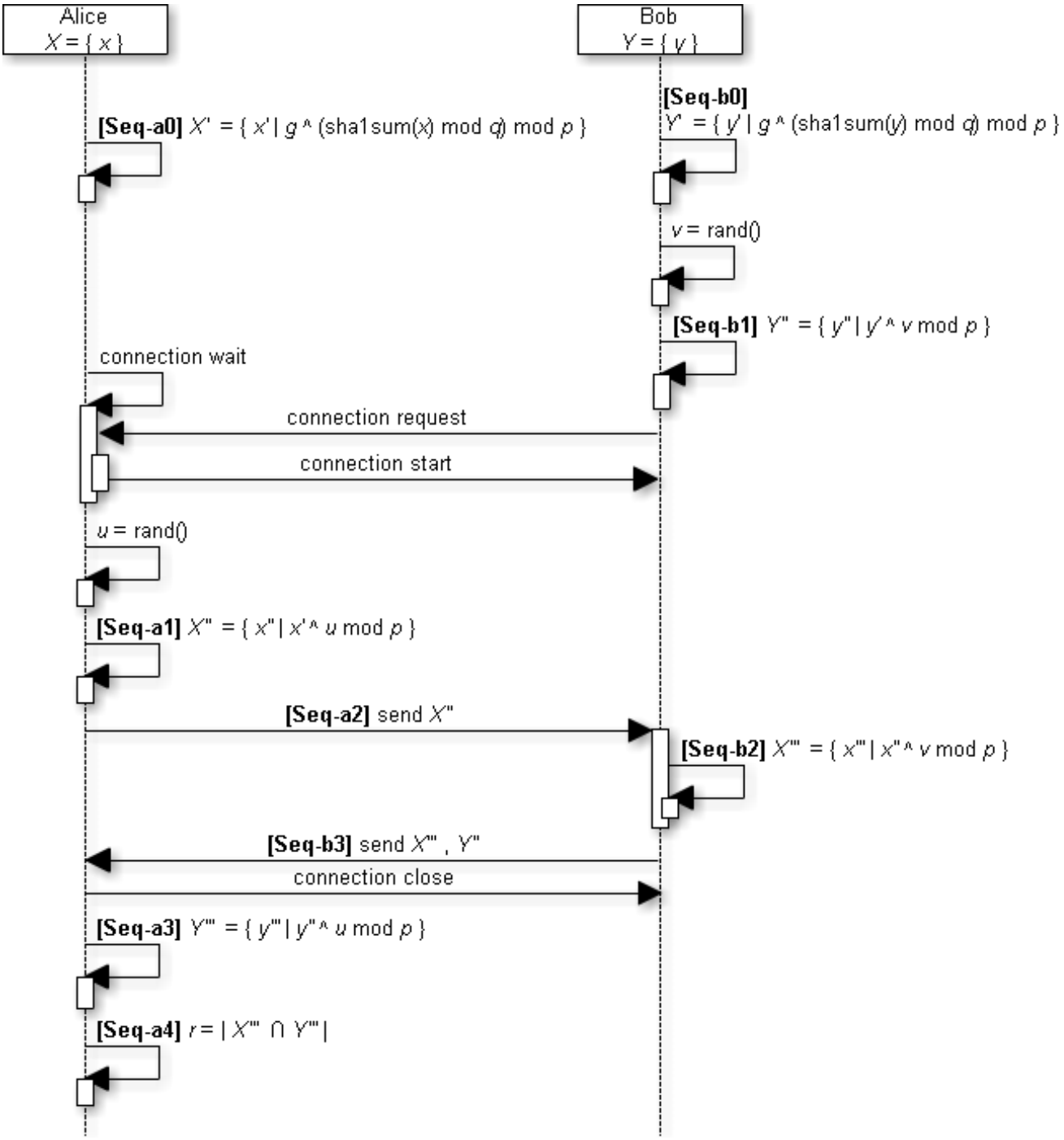

図 A·1　AES03 を用いた実装シーケンス