Vol.2011-CSEC-54 No.25
Vol.2011-SPT-1 No.25
2011/7/13

プライバシーを保護した放射線疫学調査システム

佐藤智貴[†1]　菊池浩明[†1]　佐久間淳[†2]

放射線の健康への影響を明らかにすることが重要になってきている．しかし，既存の疫学調査では，地域がんセンターのがん罹患者リストと，放射線従事者中央登録センターの放射線事業従事者リストは各組織に管理されていて，プライバシー保護の壁に阻まれて照合ができないという問題がある．

厚生労働省の持つ死亡者リストと，放射線従事者者中央登録センターの放射線事業従事者リストは各組織に独立して管理されているが，プライバシーの関係で互いに照合することには問題がある．既存の疫学調査では，特殊な法律が必要であったり，従事者の同意が得られないという問題がある．

そこで，本研究では２つの独立した組織に分割された集合について，プライバシーを保護したまま２つの集合の交わりを計算するプロトコルを提案する．１つ目の提案方式は，放射線事業従事者リストと死亡者リストなどの，２つの集合の共通部分の要素数を求める．２つ目の提案方式は，２つの集合の共通部分の要素数を明らかにせずに仮説検定を行う．提案方式の性能は試験的に実装をして評価する．

Privacy-Preserving Protocol for Epidemiology in Effect of Radiation

Tomoki Sato,[†1] Hiroaki Kikuchi[†1] and Jun Sakuma[†2]

This paper studies privacy issues in epidemiologies that aims to clarify statistical significant of interested attribute that is distributed into two independent parties. The first proposed protocol computes the size of intersection of two subsets, e.g., set of workers in risk of radiation effect and set of dead people in a period of time, without revealing any of element of set. The second proposed protocol allows to perform a hypothesis test without revealing the size of intersections of two subsets. The performance of the proposed scheme are evaluated based on trial implementation.

1. はじめに

放射線事業従事者に対する低線量域での放射線の健康への影響を調査することが重要となってきている，(財) 放射線影響協会は 1990 年度から原子力発電施設等の放射線業務従事者を対象とした疫学的調査を実施しており，“原子力発電施設等放射線業務従事者等に係る疫学調査”[1] で調査報告を行っている．

この調査報告によると，低線量域での健康への影響は，中央登録センターの持つ放射線事業従事者リストと，厚生労働省の持つ国民の死因リストを照合することで求められる．しかし，各リストは各組織で個別に管理されており，プライバシーの問題で互いに照合することは難しい．

追跡調査される被験者の同意も必要であり，プライバシー保護を理由に長期の追跡を断ることも少なくない．現状の調査では，死亡したデータから比較調査をしているが，その前に様々な悪性新生物への発病や転移の状況も分からない．そこで本研究では，まずこの調査報告について検討を行い，疫学調査におけるプライバシーの課題を挙げて，要求条件を明らかにする．

この疫学調査の問題に対して，暗号プロトコルの適用を提案する．暗号技術により，被験者のプライバシーを守って，より頻度の高い，様々な要因との相互作用を考慮した，精度の高い疫学調査の実現を試みる．本研究では，過程に応じて 2 種類の問題設定を行い，それぞれに適したプロトコルを提案する．前者は，Agrawal らの提案した AES(Agrawal-Evfimievski-Srikant)03 プロトコル[4] を用いることで，データを秘匿したまま 2 つのリストを照合する．また，文献[1] で行われている，内部比較と外部比較の調査結果に基づいて，提案方式の実現可能性を評価する．この評価の為に，提案方式を Java を用いて試験実装した．その性能に基づいた実現可能性評価の結果について報告する．

†1 東海大学大学院工学研究科
Tokai University, Graduate School of Engineering
†2 筑波大学システム情報工学研究科
Guraduate Shool of Systems and Information Engineering, University of Tsukuba

Vol.2011-CSEC-54 No.25
Vol.2011-SPT-1 No.25
2011/7/13

2. 文献[1] に関する考察

2.1 概要 (文献[1] より引用)

2.1.1 調 査 目 的

(財) 放射線影響協会では，1990 年度から原子力発電施設等の放射線業務従事者等を対象とした疫学的調査を実施している．この放射線疫学調査では，未解明の点が多い低線量域放射線の健康影響について科学的知見を得ることを目的としている．

2.1.2 調 査 対 象

調査の対象者数は，1999 年 3 月 31 日までに原子力事業者等から (財) 放射線影響協会放射線従事者中央登録センターへ登録され，実際に放射線業務に従事した日本人の男女，合計約 27 万 7 千人である．生死の確認は，市区町村長から調査対象者の住民票の写し等の交付を受けて確認している．調査対象者のうち，2009 年 3 月 31 日まで，男女合計約 21 万 2 千人の生死を確認できており，残りの約 6 万 5 千人については住所情報を収集できなかった等の理由で生死を確認できていない．

2.1.3 外 部 比 較

「外部比較」では解析対象者の死亡率が，全日本人男性死亡率に比べて高いか否かを検討するため，標準死亡比（SMR = 観察死亡数 / 期待死亡数）を求めている．また，SMR が 1 に等しいかどうかについて両側検定を行い，p 値が 0.05 未満のときは有意であると判断している．表 1 は文献[1] から引用している．

表 1 表 3.3-1 死因別標準化死亡比 (SMR)
(前向き観察、最短潜伏期;0 年、年齢、暦年を調整)(文献[1] より引用)

死因	観察死亡数	期待死亡数	SMR	95%信頼区間	両側検定結果 p 値
全死因[⋆1]	14,224	14,086.9	1.01	(0.99 - 1.03)	0.250
食道	326	312.1	1.04	(0.93 - 1.16)	0.449
胃	1,002	989.4	1.01	(0.95 - 1.08)	0.700
肺	1,208	1,117.8	1.08	(1.02 - 1.14)	0.007

2.1.4 内 部 比 較

「内部比較」では解析対象者を年度別被ばく線量の累積値により 5 群に分類し，累積線量の増加に伴って死亡率が増加する傾向があるかについて片側検定を行い，p 値が 0.05 未満のときは有意であると判断している．表 2 は文献[1] から引用している．

表 2 表 3.4-1 死因別累積線量群別 O/E 比および傾向性の検定結果 (1)
(前向き観察、最短潜伏期; 白血病 2 年 その他の新生物 10 年、年齢、暦年、地域を調整)(文献[1] より引用)

	累積線量群 (mSv)					
	< 10	10−	20−	50−	100+	
死因	観察死亡数 期待死亡数 O/E 比 95%信頼区間	観察死亡数 期待死亡数 O/E 比 95%信頼区間	観察死亡数 期待死亡数 O/E 比 95%信頼区間	観察死亡数 期待死亡数 O/E 比 95%信頼区間	観察死亡数 期待死亡数 O/E 比 95%信頼区間	傾向性の 片側検定結果 p 値
全死因[⋆2]	10,315	1,408	1,434	639	428	
	10,515.5	1,287.5	1,343.6	652.7	424.8	
	0.98	1.09	1.07	0.98	1.01	0.136
	(0.96 - 1.00)	(1.04 - 1.15)	(1.01 - 1.12)	(0.90 - 1.06)	(0.91 - 1.11)	
全悪性新生物[⋆3]	3,822	494	526	245	124	
	3,902.6	475.0	488.9	225.3	119.1	
	0.98	1.04	1.08	1.09	1.04	0.032
	(0.95 - 1.01)	(0.95 - 1.14)	(0.99 - 1.17)	(0.96 - 1.23)	(0.87 - 1.24)	
食道	200	29	32	20	8	
	215.3	26.4	27.3	12.9	7.1	
	0.93	1.10	1.17	1.55	1.12	0.039
	(0.80 - 1.07)	(0.73 - 1.58)	(0.80 - 1.66)	(0.95 - 2.40)	(0.48 - 2.21)	
胃	669	85	85	41	18	
	674.4	81.3	83.8	38.2	20.3	
	0.99	1.05	1.01	1.07	0.89	0.532
	(0.92 - 1.07)	(0.84 - 1.29)	(0.81 - 1.25)	(0.77 - 1.45)	(0.53 - 1.40)	

2.2 既存の疫学調査における課題

既存の疫学調査には，プライバシーの関係で，

(1) 特殊な法律が必要
(2) 従事者の同意が必要
(3) 情報の粒度や鮮度が不十分

などの問題点がある．これらの問題を提案方式によって解決する．

⋆1 死因を同定できなかった 80 名を含む。
⋆2 死因を同定できなかった 80 名を含む。
⋆3 白血病を含め最短潜伏期 10 年とした。

Vol.2011-CSEC-54 No.25
Vol.2011-SPT-1 No.25
2011/7/13

3. 提 案 方 式

3.1 疫学調査の問題定義

秘匿の集合 $X_A \subset U$ を持つ組織 A，$X_B \subset U$ を持つ組織 B が協力して，疫学調査を行う．ここで，U は対象者の全体集合とする．例えば，放射線疫学調査の場合は，A は放射線従事者中央登録センターであり，B は (1) 人口動態調査死亡書「死亡テープ」を有する厚生労働省や，(2) 生存するがん患者のカルテを有する地域がんセンターが該当する．U は，調査全期間 (例えば 15 年間) の全人口から成る集合である．X_A には，年齢別に分割できる属性があり，例えば $X_A = X_{A,30} \cup X_{A,40} \cup \ldots \cup X_{A,80}$ と分割できるとする．この疫学調査の目的は，X_A における死亡率やがん罹患率が標準的な期待死亡率に対して，有意な差があるか判定することにある．

3.2 検 定 方 法

死亡や疾病などの様に，一定期間に独立に生じる事象の数は，ポアソン過程と見なせることがよく知られている．ある事象の発生数 X が期待値 λ へのポアソン分布に従う時，k 回生起する確率は

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{1}$$

で与えられる．

一定期間内に生じる事象の発生数，例えば死亡数の期待値 E は，

$$E = \sum_{j}^{m} q_j n_j \tag{2}$$

で与えられる．ここで，n_j は対象となる年令階級 j の人口であり，q_j は，j における死亡率を表す．n_j は組織 A が有しており，文献[1] の例では表 3 の様になる．一方，q_j の一般的な死亡率については，文献[2] における，「2-26 年齢別死亡数及び死亡率」のように，公開されているものも多いが，未知の病気の患者リストを有する病院の例の様に，組織 B のみが有することとする．[1] における A の年齢階級別期待死亡率を図 1 に示す．A の年齢分布は 30 から 60 歳に渡っており，45 歳台が最頻値だが，加齢に応じて死亡率が高いので，図 1 では年齢に応じて単調に増加した分布が示されている．

なお，疫学調査は 10 年以上に渡って何度も行われるが，この年齢の分布もそれに応じて右へシフトしていく．

表 3 A の年齢分布

年齢	人数	割合 (%)
30 - 34	29,264	14.4
35 - 39	42,791	21.0
40 - 44	37,039	18.2
45 - 49	43,907	21.5
50 - 54	33,804	16.8
55 - 59	17,099	8.4

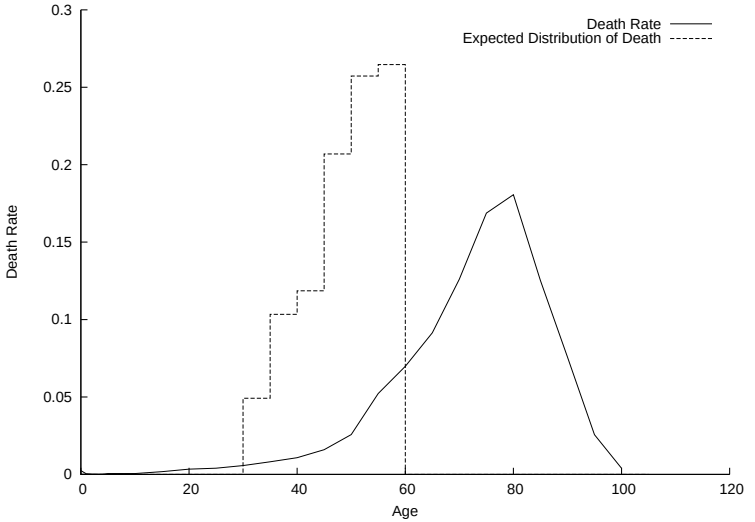

図 1 A における年齢階級別期待死亡率

発生 (死亡) 数 $X = O$ が観測された時，有意性を判断する評価尺度として，標準化死亡比：Standardized Mortality Ratio を

$$SMR = \frac{O}{E} \tag{3}$$

で定義する[⋆1]．

この SMR の期待値が 1 に等しいか否かで，有意性を判断する．すなわち，帰無仮説

⋆1 E：年齢以外には，性別，暦年，職業，地域など問題に応じて様々な層別が行われる．

Vol.2011-CSEC-54 No.25
Vol.2011-SPT-1 No.25
2011/7/13

$H_0 : \lambda = E$，対立仮説 $H_1 : \lambda \neq E$ を確率検定する．ポアソン分布は，E が 5 以上であれば，統計量

$$Z = \frac{O - E \pm 0.5}{\sqrt{E}} \tag{4}$$

により正規分布 $N(0,1)$ で近似できる．ここで，0.5 は連続修正項である．両側検定であれば，

$$Z = \frac{|O - E| - 0.5}{\sqrt{E}} > Z(\alpha/2) \tag{5}$$

の時に，有意水準 α で H_0 が棄却できる．

3.3 外部，内部比較

外部比較は，組織 A の死亡率が全日本人のそれと異なるかを比較し，内部比較では，A における集合を対象とする属性で分割し，それらの O/E 比の差を比較する．例えば，放射線従事者の累積放射線量 (10mSv 未満,10mSv 以上等) と死亡率の相関を検証する．

内部比較では，死亡率は累積線量に依存して増加しているという仮説の下で，スコア検定統計量を用いて傾向性の片側検定 (Breslow-Day 検定) を行う．表 1 に外部比較，表 2 に内部比較の例を示す．

3.4 問 題 設 定

次の 2 つの問題を考える．

(問題 1) 公開死亡率における仮説検定

U の部分集合 X_A と X_B を有する A と B が互いの集合を秘匿したままで，

$$O = |X_A \cap X_B| \tag{6}$$

$$E = \sum^{m} d_i n_i \tag{7}$$

を求める問題．但し，X_A は，$X_A = X_{A1} \cup \ldots \cup X_{Am}$ の m 層に分割され，$n_i = |X_{Ai}|$ は A が知り，層 i における X_B の割合死亡率 d_i は公開されている．例えば，表 1 における食道がんの観察死亡数 326 は，全て公開して明らかになる．

(問題 2) 秘密死亡率における仮説検定

未知の疾病に対して，死亡率 (罹患率) を公開できない場合がある．そこで，この問題においては層 i における X_B の割合 d_i を B のみが持ち，A に秘匿したままで，両側検定の p 値

$$p = P[Z > \alpha/2] \tag{8}$$

のみを求める．よって，O と E は A と B 単体には秘密とする．

3.5 問題 1 への提案プロトコル

X_A と X_B を秘匿したまま，積集合の大きさ $|X_A \cap X_B|$ のみを求める暗号プロトコルには次の 3 つが知られている．

(1) AES03(可換一方向性関数)[4)]
X_A,X_B は集合である．AES のアルゴリズムを Algorithm1 に示す．

(2) VC02(セキュア内積プロトコル)[5)] $\boldsymbol{X}_A$ と $\boldsymbol{X}_B$ はベクトルであり，結果は $s_A + s_B = \boldsymbol{X}_A \boldsymbol{X}_B$ となり，二つの乱数 s_A と s_B に分散して得られる．VC02 のアルゴリズムを Algorithm2 に示す．

(3) FNP04(多項式評価)[6)]
X_A，X_B は集合である．加法準同型性に基づく秘匿多項式評価を応用し，X_A を根として持つ多項式 $f(x)$ を B が秘匿したままで $f(y)$ を計算する．

Algorithm 1 AES03[4)](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．

問題 1 は，各層 j について，これらのいずれかを適用して，Oj を求める．A は，公開 d_j を用いて，$E_j = \sum_j d_j n_j$ を公開する．4.3 節の方法で，外部比較，内部比較を実施し，p 値，信頼区間を求め，帰無仮説が有意に棄却できるか判断する．

3.5.1 問題 2 への提案プロトコル

A は，$\boldsymbol{X}_A$ と $n_1, \ldots, n_m$ を持ち，B は $\boldsymbol{X}_B$ と $q_1, \ldots, q_m$ を持つ．

(1) セキュア内積プロトコル VC02 を用いて

$$s_A + s_B = X_A X_B = O \tag{9}$$

となる s_A，s_B を得る．A は s_A^2 を，B は s_B^2 をそれぞれ計算する．

(2) A は加法準同型性を満たした公開鍵暗号を用いて，鍵対を作り，暗号文 $E(n_1), \ldots, E(n_m)$

Vol.2011-CSEC-54 No.25
Vol.2011-SPT-1 No.25
2011/7/13

Algorithm 2 VC02[5](セキュア内積プロトコル)

入力: Alice は n 次元ベクトル $\boldsymbol{x} = (x_1, \ldots, x_n)$ を持つ. Bob は n 次元の $\boldsymbol{y} = (y_1, \ldots, y_n)$ を持つ.

出力: Alice と Bob は $s_A + s_B = \boldsymbol{x} \cdot \boldsymbol{y}$ となるような，s_A，s_B を得る.

(1) Alice は準同型暗号の公開鍵を作り，公開鍵を Bob に送る

(2) Alice は Bob に暗号化した $E(x_1), \ldots, E(x_n)$ を送る.

(3) Bob は s_B をランダムに選び,

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

を計算し，c を Alice に送る.

(4) Alice は c を復号し，$s_A = D(c) = x_1 y_1 + \cdots + x_n y_n - s_B$ を得る.

を B へ送り，B は乱数 t_B を作って,

$$y = \prod_i^m E(n_i)^{q_i} / E(t_B) \tag{10}$$

を計算し，A は $t_A = D(y) = t_B + \sum n_i q_i$ を作る．$t_A + t_B = E$ である.

(3) 再び，セキュア内積プロトコル VC02 を用いて,

$$w_A + w_B = s_A s_B \tag{11}$$

となる w_A と w_B を求め，A と B で分散管理する.

(4) A は，s_A,${s_A}^2$,t_A,w_A を，B は s_B,${s_B}^2$,t_B,w_B を万能秘密関数計算プロトコル SFE(Secure Function Evaluation)[7] へかけて，次の p 値を求める.

$$z^2 = \frac{{s_A}^2 + 2(w_A + w_B) + {s_B}^2}{t_A + t_B} - 2(s_A + s_B) + (t_A + t_B) = \frac{O^2}{E} - 2O + E \tag{12}$$

(5) z^2 が自由度 1，有意水準 α の χ^2 値未満ならば，帰無仮説 H_0 を棄却する．(外部比較)

(6) (内部比較) 累積線量によって k 個に分割された $X_{A1}, \ldots, X_{Ak}$ の各々について，1 から 5 を実行し，求めた k 個の統計量の和が

$${Z_1}^2 + \cdots + {Z_k}^2 < \chi^2_\alpha(k-1) \tag{13}$$

ならば，H_0:死亡率は累積線量に依らず一定である，とする帰無仮説を棄却する.

3.6 評　　価

3.6.1 安　全　性

方式 1 は用いる要素技術 AES03，VC02，FNP04 の安全性に基づいて，各集合 X_A と X_B を秘匿する．AES03 プロトコルは，入力 x と $H(x)^{uv}$ が一対一対応する．このため，入力が特定の分布に従う場合は，$H(x)^{uv}$ の分布から x を統計的に推定する攻撃方法が存在する．しかし本稿の事例のように，入力集合の要素が取り得る値が，ユニークな ID であったり，常に一つ以下しか存在しないことが保証されている属性値であるような場合には，このような統計的攻撃は成立せず，問題にならない.

方式 2 は，A と B が正直に振舞うセミオネストモデルの下で，正しく検定量 Z^2 を計算し，ベクトル $\boldsymbol{X}_A$,$\boldsymbol{X}_B$ に加えて，観察数 $O = X_A, X_B$ と期待値 $E = \sum g_i n_i$ を秘匿する．安全性は加法準同型性を満たす暗号の識別不能性に帰着する.

3.6.2 提案方式の比較

要素技術と提案方式との関係を表 4 に整理する.

要素の同定とは，積集合の数だけではなく，積集合そのものを同定するプロトコルが可能なのは，AES03 と FNP04 のみである．処理性能は，リストの大きさ n と定義域の大きさ N に依存する．AES03 のパフォーマンスは，4 章の実装に基づいた値である．これらの要素技術は互換ではなく，本論文の提案方式 2 はセキュア内積プロトコルしか適用できないことに注意が必要である.

表 4 暗号プロトコルの比較

	AES03[4]	VC02[5]	FNP04[6]
要素の同定可能	yes	no	yes
入力	集合	ベクトル	集合
処理性能	$O(n)$	$O(N)$	$O(n^2)$
パフォーマンス	360 件/s	10 件/s	-
提案方式 2 への利用	no	yes	no

4. 実 装 評 価

4.1 試験実装したプログラム

問題 1 に対する提案プロトコルについて，その実現可能性を検討するために，Agrawal らの提案した AES03 プロトコル[4] を Java により実装し，その性能を評価した．BigInteger クラスとリストの照合に Collection フレームワークの Map クラスを用いた.

Vol.2011-CSEC-54 No.25
Vol.2011-SPT-1 No.25
2011/7/13

使用するデータの例を表 5 に示す[⋆1].

表 5 垂直分割データセットの例

A			*B*	
名前	年齢	累積線量 [mSv]	名前	死因
佐藤	20	12	田中	肺がん
菊池	30	51	鈴木	前立腺がん
佐久間	30	33	佐藤	外因死
鈴木	70	46	後藤	肺がん

プログラムの入力は放射線事業従事者の ID が入ったリスト X_a とがん罹患者の ID が入ったリスト X_b で，出力は標準的な日本人全体のがん罹患率と比べて，放射線事業従事者のがん罹患率が高いか低いかの判定結果とを出力する．

4.2 試験実装 (提案方式 1) の処理性能

法のサイズ 1024 ビットの時の，試験実装したプログラムの処理時間を図 2 に示す．

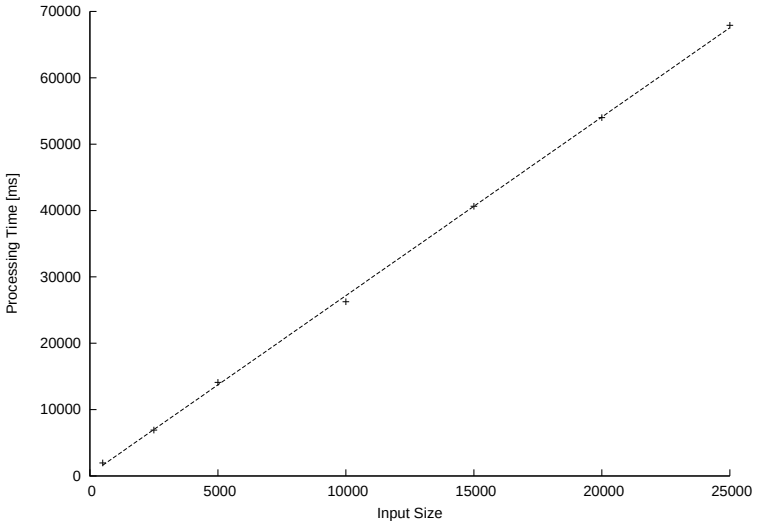

図 2 試験実装 (提案方式 1) の処理時間

⋆1 この例では，2つのデータセットが同期されていないため，VC02 を使うことはできない．

1 秒当たり約 360 個のデータについて照合を行っている計算になり，もう一方のリストも 20 万人と仮定した場合，文献[1] で行っている約 20 万人のデータだと 550 秒，約 9 分かかる．

4.3 結　　論

疫学におけるプライバシー問題について，既存の文献を基に検討し，課題と要求条件を定義した，被験者のプライバシーを考慮することと，より粒度の高い詳細な調査が必要であることが矛盾する要求である．

この問題に対して，暗号プロトコルの適用を提案し，試験実装に基づいて十分に実現可能であることを示した．

今後の課題には，試験実装した提案方式 1 のプログラムの高速化や，提案方式 2 の試験実装が挙げられる．

謝　　辞

本研究は，平成２２年度科学研究費補助金，基盤研究 (B)「組織間でのプライバシー保護疫学調査技術の研究」及び，最先端研究開発支援プログラム (FIRST) の支援を受けている．

参 考 文 献

1) 放射線影響協会, 原子力発電施設等放射線業務従事者等に係る疫学的調査, 2010.
2) 統計局政策統括官・(統計基準担当) 統計研修所, 厚生労働省大臣官房統計情報部人口動態・保健統計課「人口動態統計」
3) 菊池浩明, 香川大介, 石井一彦, 寺田雅之, 本郷節之, ”組織間プライバシー保護データマイニングの考察”, SCIS2010.
4) Rakesh Agrawal, Alexandre Evfimievski, and Ramakrishnan Srikant, “Information shareing across private databases”, in proc. of ACM SIGMOD International Conference on Management of Data, 2003.
5) Vaidya, J. C. Clifton, Privacy preserving association rule mining in vertically partitioned data, in‘The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining’, SIGKDD, ACM Press, Edmonton, Canada, pp. 639—644, 2002.
6) M. J. Freedman, K. Nissim, and B. Pinkas, ”Efficient private matching and set intersection”, EUROCRYPT 2004, LNCS 3027, pp. 1—19, Springer-Verlag, 2004.
7) Dahlia Malkhi, Noam Nisan, Benny Pinkas, and Yaron Sella,“Fairplay — A Secure Two-Party Com- putation System”, Usenix Security Symposium, 2004.

頁	該当箇所	誤	正
6	謝辞の最後に 1 行追加	～の支援を受けている.	～の支援を受けている. また, 疫学調査についての情報をご教示頂いた放射線影響協会, 工藤伸一氏に感謝致します.