社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

プライバシー保護確率検定システムの実装と評価

佐藤　智貴† 　菊池　浩明† 　佐久間　淳††

† 東海大学
†† 筑波大学，科学技術振興機構
E-mail: †kakiyan@cs.dm.u-tokai.ac.jp, ††kikn@tokai.ac.jp, †††jun@cs.tsukuba.ac.jp

あらまし　環境因子と疾病の因果関係を明らかにする疫学調査では，調査対象者の個人情報を扱うため，プライバシー保護に十分留意しなければならない．公開鍵暗号に基づく暗号技術である秘匿内積プロトコルと，秘匿関数計算プロトコルを用いることで，個人情報を秘匿したまま疫学調査を行うことができるが，秘匿関数計算プロトコルでは積や平方根などの計算は極めて遅い．そこで，本稿では秘匿関数計算プロトコルで評価可能な効率の良い確率検定プロトコルを提案し，試験実装に基づきその処理時間及び安全性について評価を行った．

キーワード　秘匿内積プロトコル，SFE，確率検定

Implementation and Evaluation of Privacy-Preserving Epidemic Analysis System

Tomoki SATO†, Hiroaki KIKUCHI†, and Jun SAKUMA††

† Tokai University
†† University of Tsukuba, JST
E-mail: †kakiyan@cs.dm.u-tokai.ac.jp, ††kikn@tokai.ac.jp, †††jun@cs.tsukuba.ac.jp

Abstract　This paper studies privacy issues about epidemiological study. Epidemiological study needs to preserve the privacy of subjects because of involved personal information. Privacy is preserved through the use of the secure scalar product protocol based on public key cryptsystem and the secure function evaluation. The secure function evaluation has limitation of the performance to evaluate a product and a squared root. Therefore, this paper proposes a new computationally-efficient scheme of privacy-preserving epidemiological analysis. The performance and security of the proposed scheme are evaluated based on trial implementation.

Key words　Secure Scalar Product, SFE, hypothesis test

1. はじめに

環境因子と疾病の因果関係を明らかにする疫学調査では，疾病の原因と考えられる因子と疾病の関係性を統計的に明らかにする [1] - [2]．例えば，喫煙者のがん罹患率と非喫煙者のがん罹患率を比較する事で，喫煙による健康への影響を明らかにすることができる．しかし，これらのデータは独立した組織によって管理されている事が多く，プライバシーと個人情報保護の観点から，照合が困難であった．そこで，本研究では統計的調査の手法に暗号プロトコルを適用し，プライバシーを保護したまま安全に確率検定を行うことを試みる．

秘匿内積プロトコル [3] と秘密関数計算プロトコル [5] を使用することで，理論的には安全に確率検定を行うことができる．しかし，現実的には，秘密関数計算プロトコルは処理が極めて遅く，積や平方根などの計算が困難であった．

そこで本稿では，秘密関数計算プロトコルでの処理効率を考慮して効率的に評価できる確率検定プロトコルを提案する．提案方式を Java を用いて実装し，その実現可能性と安全性を評価する．国立がん研究センターが行っている多目的コホート研究 [1] に使用されている約 14 万人のデータについて，実装したプログラムを適用した時の処理時間を見積もり，提案方式の実現可能性について検討する．

2. 要素技術

2.1 患者-対照調査

患者-対照調査とは，ある特定の疾病の患者群と非患者群について，ある要因に暴露していたか否かを調べ，因果関係を研究する調査である．表 1 の患者-対照調査のデータが与えられ

た時,

$$RR = \frac{a}{n_1} / \frac{c}{n_2} = \frac{a(c+d)}{(a+b)c} \approx \frac{ad}{bc} \quad (1)$$

を相対危険度 (relative risk) と呼ぶ．相対危険度とは，特定要因へ暴露した群が，暴露しなかった群に比べて，何倍の危険度を有するかを表す指標である [2]．相対危険度が大きいものほど因果関係が強い．この相対危険度 RR が 1 に等しいかどうかを検定する．推定された相対危険度の有意性は，統計量 $\chi = \frac{\sqrt{N-1}\{(ad-bc)\pm N/2\}}{\sqrt{n_1 n_2 m_1 m_2}}$ が $RR = 1$ の仮定の元，標準正規分布 $N(0,1)$ に従うか否かで検定することができる．本稿では，両側検定の有意水準 95%，すなわち χ が $Z(0.05/2) = 1.960$ を上回っているか否かで判定を行う．

表 1　患者-対照調査における母集団のデータ

	死亡	生存	計
喫煙	a	b	n_1
非喫煙	c	d	n_2
計	m_1	m_2	N

2.2 秘匿内積プロトコル [3]

本稿では，秘匿内積プロトコルを使用してプライバシーを保護したまま患者-対照調査を行う．秘匿内積プロトコル [3] を Algorithm 1 に示す．2 つの組織がそれぞれ持つ 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．計算結果は，$s_A + s_B = |X_A \cap X_B|$ となるような 2 つの乱数 s_A と s_B に分散されるため，計算が終わっても秘匿されている．

Algorithm 1 秘匿内積プロトコル

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．ここで，暗号文の定義域を Z_n とする．

（1） Alice は準同型暗号の公開鍵対を作り，公開鍵を Bob に送る．

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

（3） Bob は s_B を Z_n からランダムに選び，

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

を計算し，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．

2.3 Fairplay [4]

提案方式では，秘匿内積プロトコルを使用して得られた 2 つに分散された値を互いに秘匿したまま有意水準を超えるか判定するために，Yao により提案された秘密関数計算 (Secure Function Evaluation(SFE) [5]) プロトコルを用いる．SFE は，AND や OR の論理ゲートレベルで，2 者間での分散評価を行うため，その回路サイズが小規模なものに制約されるが，任意の関数が秘密に評価できる．例えば，s_A と s_B を入力すると，$s_A + s_B > t$ を評価してその 1 bit の結果を出力する回路を構成することで，A と B のどちらもその和が分からないままで，有意水準を超えるか否かのみが分かる．

Fairplay は，Malkhi らによって開発された SFE の処理系である [4]．高級言語風のソースから，回路記述言語 SFDL を出力し，それに基づいて SFE を実行する．ただし，SFE の性質上，その機能には制約があり，例えば，乗算はプリミティブで用意されていないため，加算を組み合わせて書かなくてはならない．そのため，Fairplay は加算，減算，大小比較等の基本的な計算は高速にできるが，乗算，除算等の計算は極めて遅い．加算と乗算の処理時間の違いを図 1 に示す．加算は $t_A + t_B > \theta$ の比較，乗算は $t_A \cdot t_B > \theta$ を各々実行した時の処理時間である．それ故，Fairplay においては乗算の利用を極力避けなくてはならない．

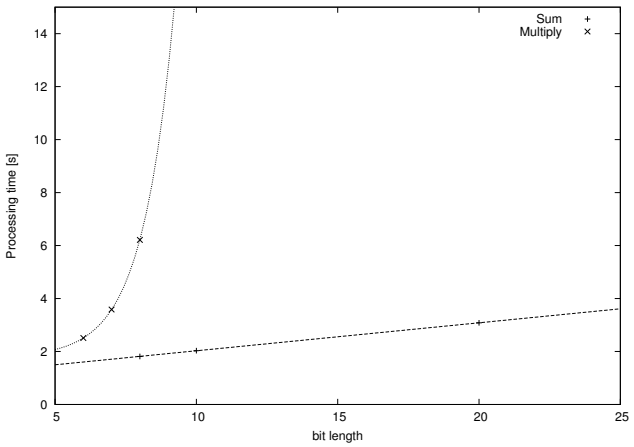

図 1　Fairplay における加算と乗算による分散比較の計算時間 [6]

3. 提案方式

3.1 問題定義

秘匿の必要がある集合 X_A を持つ組織 A と，X_B を持つ組織 B が協力して疫学調査を行う．例えば，組織 A は特定要因である喫煙者のデータを持っている組織，組織 B は死亡者のデータを管理する組織とする．すなわち，組織 A は表 1 の n_1，n_2 の情報を持ち，組織 B は m_1，m_2 の情報を持つ．それぞれの持っているデータの合計 N は公開するが，属性毎の合計値，n_1，n_2，m_1，m_2 は秘密とする．これらの条件の元で，効率的に疫学調査を行い，対象とする特定要因の相対危険度が有意かどうかを検定する．出力結果は，その検定結果のみとする．

これを解くナイーブな方法は，X_A と X_B をベクトル表現し，2.2 節の秘匿内積プロトコルを適用することであるが，これには，SFE の実行に関する次の 2 つの問題点がある．

3.2 問題点

（1） 統計量を求めるための大きな計算量．秘匿内積プロトコルは 2 つに分散された値を出力する．これを秘匿したまま解くには Fairplay などのシステムが必要になるが，統計量 χ を式 (2.1) で求めるためには，SFE 上での平方根や積などの計算が必要になる．しかしながら，2.3 節で示した様に，Fairplay でこれらを計算するのは困難である．

（2） 分散値の定義域の大きさ．秘匿内積プロトコルでの STEP(3) 計算時に乱数 s_B を生成するが，この乱数は安全性の

為，準同型暗号の定義域 Z_n から選ぶ必要がある．しかし，公開鍵暗号の平文長の 2048bit の様な値は，Fairplay で計算するには大きすぎる．

3.3 アプローチ

問題点 (1) を解決するために，Fairplay での計算は加算や減算，比較のみに限定したい．そこで，次に示す様に，検定の条件を Fairplay で可能な等価な式に変形する．検定を行った際に有意になるか否かは，N，n_1，m_1 を固定した場合，a の大きさによって求めることができる．まず，式 (2.1) を a のみを用いるように変形すると，

$$\begin{aligned}\chi &= \frac{\sqrt{N-1}\{(ad-bc)\pm N/2\}}{\sqrt{n_1n_2m_1m_2}}\\ &= \frac{\sqrt{N-1}\{a(N-n_1-m_1+a)-(n_1-a)(m_1-a)-N/2\}}{\sqrt{n_1n_2m_1m_2}}\\ &= \frac{\sqrt{N-1}\{aN-n_1m_1-N/2\}}{\sqrt{n_1n_2m_1m_2}} \qquad (2)\end{aligned}$$

と，a の一次式になる．こうして，a を変化させた時の統計量 χ の変化を図 2 に示す．この値が有意水準 $Z(0.05/2)=1.960$ を

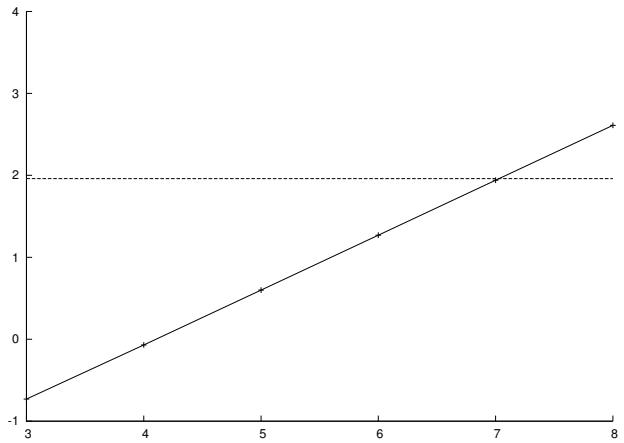

図 2　a を変化させた時の統計量 χ

超えたかを判断すれば良い．この境界の a を a^* とおくと，式 (2) より

$$\begin{aligned}a^* &= \left(\frac{\chi\cdot\sqrt{n_1n_2m_1m_2}}{\sqrt{N-1}}+n_1m_2+\frac{N}{2}\right)\cdot\frac{1}{N}\\ a^*N &= \frac{1.960\cdot\sqrt{n_1n_2m_1m_2}}{\sqrt{N-1}}+n_1m_2+\frac{N}{2} \qquad (3)\end{aligned}$$

となる．ここで，χ と N は与えられた公開情報であり，$\sqrt{n_1n_2m_1m_2}$ は n_1 と n_2 を持つ A と m_1，m_2 を持つ B が秘匿内積プロトコルを行えば分散した値が得られる．n_1m_1 も同様である．従って，秘匿内積プロトコルで X_A と X_B(喫煙者と死亡者) の積集合 a が，a^* を上回っていれば有意，下回っていれば有意ではないことを Fairplay で検査するには，

$$(s_A+s_B) > (t_A+t_B)+(u_A+u_B) \qquad (4)$$

を評価し，判定結果のみを出力すればよい．ここで，s_A, s_B, t_A, t_B, u_A, u_B は，

$$\begin{aligned}s_A+s_B &= aN = |X_A\cap X_B|N,\\ t_A+t_B &= \frac{\chi\cdot\sqrt{n_1n_2m_1m_2}}{\sqrt{N-1}} = \left(\frac{\chi\sqrt{n_1n_2}}{\sqrt{N-1}}\right)\cdot\sqrt{m_1m_2},\\ u_A+u_B &= n_1m_1+\frac{N}{2}\end{aligned}$$

で定義される値であり，Alg. 1 により効率的に求めることができる．

3.4 乱数 s_B の定義域

Alg. 1 の s_B を Z_n から選ぶと，s_A, s_B は $|Z_n|$ のサイズを持つ整数となり，SFE で求めるには大きすぎる．

そこで，Alg. 1 の STEP(3) を $E(s_B)$ で割る代わりに，小さな定義域から一様に選んだ s_B をかける，すなわち，

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

と変更し，

$$s_A - s_B = \boldsymbol{x}\cdot\boldsymbol{y}$$

となる s_A と s_B に分散する様にする．ただし，$s_A-s_B<0$ の時は，2 の補数表現で $|Z_n|$bit の整数が生じてしまうため，$s_A>s_B$ となる様に s_B を選ぶ．

$a=|X_A\cap X_B|\in[0,n]$ である時，(3) の乱数の定義域を μ 個の自然数，すなわち，$s_B\in[0,\mu-1]$ とする．μ は n に対して十分大きく，かつ，SFE で処理可能な大きさに定める必要がある．なぜならば，次に挙げる安全性の問題が生じるためである．

$s_A=s_B+a<\mu$ の時に，A が s_A を知ることにより真の a の大きさについて言えることは，

$$P(a=0|s_A)=\cdots=P(a=n|s_A)=\frac{1}{n+1}$$

であり，$[0,n]$ のどの値も一様に確からしい．しかし，$\mu<s_A$ の時，a の取り得る値は，

$$\begin{cases}P(a=0|s_A)=\cdots=P(a=s_A-\mu-1|s_A)=0\\ P(a=s_A-n|s_A)=\cdots=P(a=n+1|s_A)=\frac{1}{n+\mu+1-s_A}\end{cases}$$

となり，偏りが生じる．図 3 にこの s_A の危険な領域を示す．例

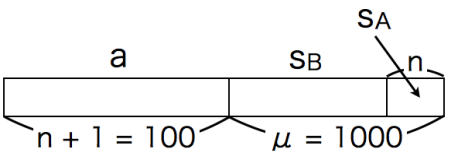

図 3　s_A の危険な領域

えば，$n+1=100$, $\mu=1000$ の時，$s_A=1005$ ならば，a は少なくとも 5 以上でなくてはならない．s_A がこの「危険な領域」に落ちる確率を次で与える．

定理 1 (乱数長からの安全性)　$a\in[0,n]$, $s_B\in[0,\mu-1]$ の一様分布から選んだ値とする．$s_A=s_B+a>\mu$ となる確率は，

$$P(\mu<s_A)=\frac{n-1}{2\mu}$$

である．

証明 1　$s_A = \alpha + \beta$ となる α, β を用いると，

$$\begin{aligned} P(\mu = s_A) &= P(a = \alpha) \cdot P(s_B = \beta) \\ &= \frac{1}{n+1} \cdot \frac{1}{\mu} \end{aligned}$$

ここで，$\mu = s_A$ の時，$s_A = \alpha + \beta$ となる (α, β) 組は，$(1, \mu-1), \cdots, (n+1, \mu-n+1)$ の $n-1$ 通り．一方，$n+\mu+1 = s_A$ となるのは，$(n+1, \mu)$ の 1 通り，よって，初項 $n-1$，公比 -1 の等比数列の和より，危険領域の条件を満たす組は，$(n-1)(n+1)/2$ 存在する．よって，

$$\begin{aligned} P(\mu < s_A) &= \sum_{\mu < s_A} P(s_A) \\ &= \sum_{\mu < s_A = \alpha+\beta} \frac{1}{n+1} \cdot \frac{1}{\mu} \\ &= \frac{(n+1)(n-1)/2}{(n+1)\mu} = \frac{n-1}{2\mu} \end{aligned}$$

で定理を得る．　$(Q.E.D.)$

例えば，$n+1 = 100$, $n = 10 \cdot n = 1000$ の乱数を選ぶと，A が a について一部の情報を得る確率は，$99/2 \cdot 1000 = 0.0495$ と十分に小さい．逆に，その確率を ϵ とすると，$\mu > \frac{n-1}{2\epsilon}$ を超える乱数を選べば良い．

最後に，a の情報が一部漏れた時の大きさを評価する．$\mu < s_A$ の時，$P(a|s_A) = \frac{1}{n+\mu+1-s_A}$ より，損なわれる条件付き確率の変化を図 4 に，そのエントロピーの減少を図 5 に示す．s_A が $\mu = 1000$ を超えてから，a について同定される程度を図示している．図 4 と図 5 では，どちらも s_A が 0 から 1000 までの間は，a が同定される確率は破線のように一様だが，s_A が 1000 を超えると，実線のように a が同定される確率が上がっていく．

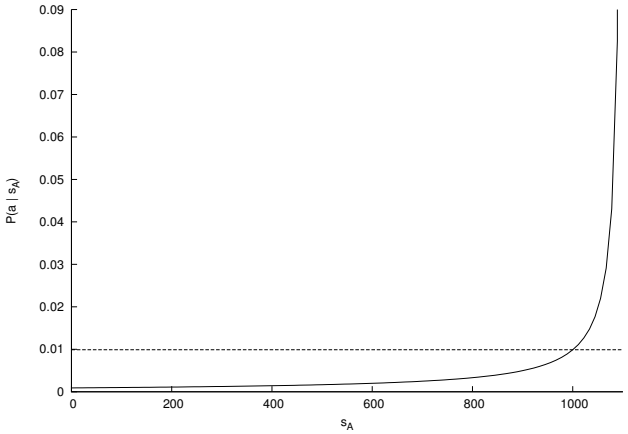

図 4　損なわれる条件付き確率 $P(a|s_A)$ の変化

3.5 提案方式

以上の提案方式を Algorithm 3 に示す．改良した秘匿内積プロトコルを Algorithm 2 に示す．出力される s_A と s_B が $[0, \mu-1]$ の値域に収まることに注意されたい．

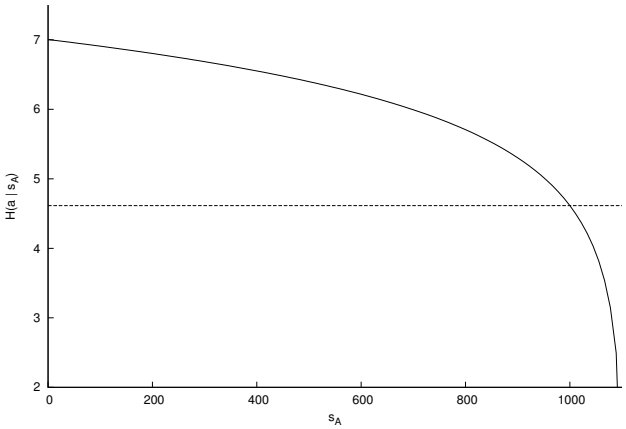

図 5　エントロピーの減少度合

Algorithm 2 改良した秘匿内積プロトコル

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対を作り，公開鍵を Bob に送る．

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

（3） Bob は s_B を $s_B \in [0, \mu-1]$ をランダムに選び，

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

を計算し，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．

Algorithm 3 提案方式

入力: $|X_A| = n_1$, $|X_B| = m_1$

出力: $|X_A \cap X_B| = a$ が 95%の水準で有意か

（1） Alg. 2 を用いて，$s_A - s_B = aN$ となる s_A を A が，s_B を B が得る．

（2） Alg. 2 を用いて，$t_A - t_B = \left(\frac{\chi\sqrt{n_1 n_2}}{\sqrt{N-1}}\right) \cdot \sqrt{m_1 m_2}$ となる t_A を A が，t_B を B が得る．

（3） Alg. 2 を用いて，$u_A - u_B = n_1 m_1 + \frac{N}{2}$ となる w_A を A が，w_B を B が得る．

（4） SFE を用いて，A は (s_A, t_A, u_A), B は (s_B, t_B, u_B) を入力し，(4) 式を判定する．

4. 評　　価

4.1 パフォーマンス

Java BigInteger クラスを用いて実装したプログラム (Alg. 3) の処理時間を図 6 に示す．実装したプログラムの処理時間は，要素数 N に対して線形に増加している．また，Fairplay で式 (4) の評価を行った時の処理時間を図 7 に示す．線形に増加しているが，1 の乗算にかかる計算時間と比較すると，十分早い．

入力する乱数 s_B の bit 長 $(=\mu)$ を変化させ，それぞれの bit で 5 回ずつ評価した時の処理時間の平均と分散である．

Alg. 3 の通信量を図 8 に示す．1 要素当たり 620byte の暗号文が生成されている．

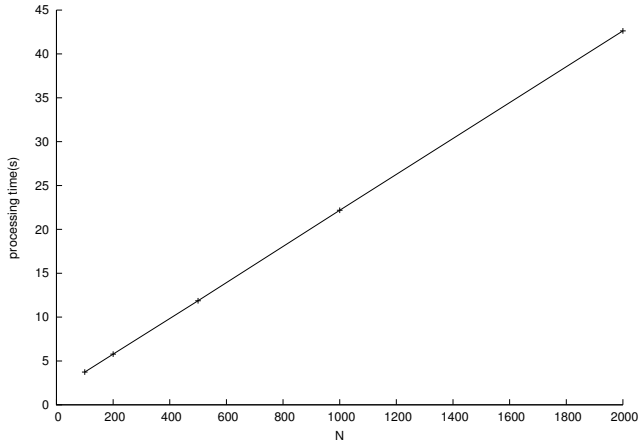

図 6　実装したプログラムの処理時間

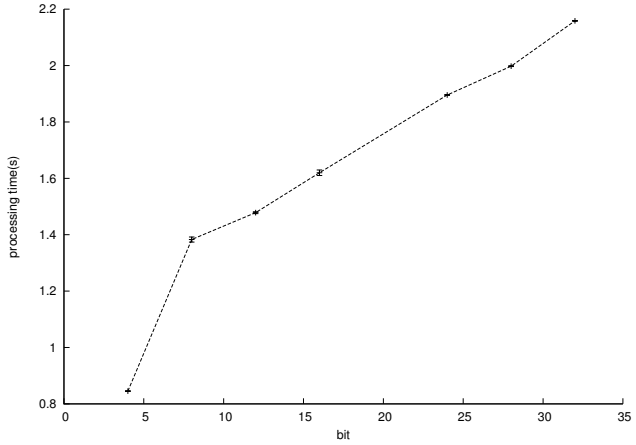

図 7　Fairplay を用いて式 (4) の評価を行った時の処理時間

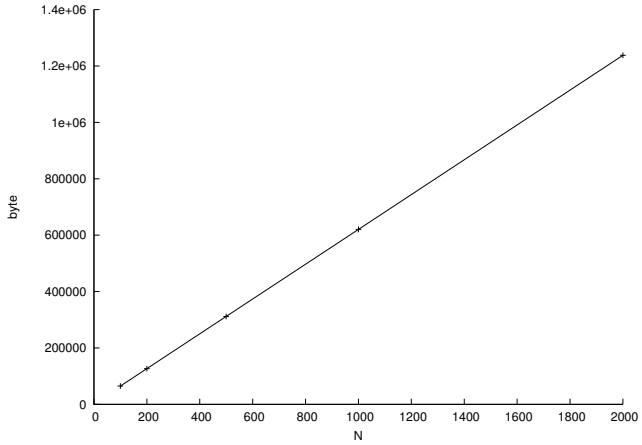

図 8　要素数 N に対する Alg. 3 の通信量

4.2　多目的コホート研究への適用

国立がん研究センターが行っている多目的コホート研究 [1] では，生活習慣病の科学的な予防法を明らかにすることを目的に，140,420 名のデータを元に，喫煙や飲酒のリスクなどを調査している．

本プログラムをこの 140,420 人のデータに対して使うと，3,191 秒，約 1 時間で計算を行うことができ，十分実用的な時間と考えられる．

また，多目的コホート研究のような，ある属性に対して複数の要素の相対危険度を評価する必要がある場合，Alice は一回だけ全ての要素を暗号化するだけで済み，それに対して Bob が複数回計算を行うことで実現できる．計算にかかる時間は，ほぼ STEP(1) の暗号化処理なので，多属性について評価する場合は，より本プロトコルが有効になると考えられる．要素数に対する処理時間を表 2 に示す．

表 2　要素数に対する処理時間

	100	200	500	1000	2000
(1)Alice	67%	79%	90%	94%	97%
(2)Bob	17%	11%	5%	3%	2%
(3)Alice	16%	10%	5%	3%	1%

4.3　安　全　性

[3] では Z_n から乱数を選ぶことが提案されているが，Fairplay で計算するため，乱数の大きさを抑える必要がある．そのため, [3] の秘匿内積プロトコルよりも安全性は落ちる．しかし，3.4 で示したように，Fairplay で計算可能な範囲で適切な乱数を選ぶことで，リスクを抑えることができる．

提案方式の安全性は，要素技術である秘匿内積プロトコル，SFE の安全性に依存する．セミオネストモデルの仮定の下，Alg. 3 は検定結果以外の情報を漏らさない．

5.　おわりに

秘匿内積プロトコルと Fairplay を使用し，2 つのデータセットを秘匿したまま，確率検定を行うシステムを実装した．実装する際にネックとなる Fairplay の制約に対して，従来の式を変形し加算や減算，比較のみで計算を行い，乱数の大きさを抑えることで現実的な時間で処理を行えるようになった．

本稿の主要な結論は次の通りである．

（1） SFE で評価可能な，効率の良い確率検定プロトコルを提案した．その処理時間は，$N = 1000$, $\mu = 2^{31}$ の時，約 26 秒である．

（2） 秘匿内積プロトコルの結果を SFE で比較する際の乱数長の問題を指摘し，適切な乱数長 μ と，その安全性やエントロピーの損失を明らかにした．

今後の課題としては，実際の疫学調査で必要とされる多値や連続値等も扱えるシステムの拡張である．

文　　献

[1] 独立行政法人 国立がん研究センター, ”多目的コホート研究の成果”, pp. 1-18, 2011.
[2] 古川俊之，丹後俊郎, ” 新版 医学への統計学”, 朝倉書店, 1993.
[3] Bart Goethals, Sven Laur, Helger Lipmaa and Taneli Mielikainen, ”On Private Scalar Product Computation for Privacy-Preserving Data Mining”, The 7th Annual International Conference in Information Security and Cryptology (ICISC 2004), Vol. 3506 of LNCS, pp. 104-120, 2004.
[4] Dahlia Malkhi, Noam Nisan, Benny Pinkas, and Yaron Sella, ”Fairplay - A Secure Two-Party Computation System”, Usenix Security Symposium, pp. 1-17, 2004.
[5] A. C. Yao. ”How to generate and exchange secrets”. In Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, pages 162-167, 1986.

[6] Hiroaki Kikuchi, Daisuke Kagawa, Anriban Basu, Kazuhiko Ishii, Masayuki Terada, Sadayuki Hongo,“ Scalable Privacy-Preserving Naive Bayes Classification over Asynchronously Partitioned Datasets”, Proceedings of the 26th International Information Security Conference (IFIPSEC), pp. 1-16, 2011.