3I2-OS-20-9

プライバシーを保護したピロリ菌疫学調査

Privacy-Preserving Epidemic Analysis on Risk of H. Pylori

菊池 浩明[*1] 佐久間 淳[*2] 三上 春夫[*3]
Hiroaki Kikuchi Jun Sakuma Haruo Mikami

[*1]東海大学 Tokai University
[*2]筑波大学, 科学技術振興機構 University of Tsukuba, JST
[*3]千葉県がんセンター Chiba Cancer Center

Helicobacter pylori is a microaerophllic bacterium, having 1 to 2% risk of aquiring stomach cancer. The paper proposes a privacy-preserving protocol for epidemic analysis of risk of H. pylori in terms of cancer. Given two datasets of patients of cancer and H. pylori, the proposed protocol determines the size of intersection of two sets without revealing any entries of datasets. With secure hash function, the proposed scheme identifies a patient from their personal attribute.

1. はじめに

ピロリ菌は胃がんの主要な原因の一つとして知られており，生活習慣や胃粘膜の抗体反応の強さなどがその発病の条件と考えられている．1994 年 WHO はピロリ菌が胃がんの原因であると発表し，その後の調査によりピロリ菌の感染者は胃がんになるリスクを 1,2%増加することが明らかになってきた [1]. ただし，ピロリ菌には単独でがんを作り出す力はなく，胃がん発症の原因となる胃炎，胃潰瘍を引き起こすための有害物質を発生する．それゆえ，ピロリ菌を除去することで胃がんのリスクを下げることが期待できる．しかし，ピロリ菌がどのように胃がんを引き起こすかの仕組みはまだよく解明されておらず，この除去をいつまでに行えば効果的なのか，疫学的な調査が続けられている．

こういった調査を可能として，がんの罹患の実態を把握してがん対策推進上の基礎資料とするために，多くの地方自治体ではがん患者およびがん死亡者の登録を実施している．千葉県でも，千葉県内の病院，診療所，がん検診実施機関の協力のもと，千葉がんセンターが 1975 年より医療機関から届出された悪性新生物登録票及び健康福祉センター（保健所）に提出された人口動態調査死亡票を収集している．ピロリ菌に関しても，厚生省の調査により，菌保有者の生活習慣や胃粘膜の条件などに関するデータベースが構築されている．そこで，これらの分散されたデータベースを照合すれば，ピロリ菌と胃がん発病の相関関係を明らかにし，感染のしくみや有効な対策に関する知見が抽出できることが期待できるのだが，個人情報保護を十分に考慮しなくてはならない．例えば，千葉県では「収集する情報はサーバー内で暗号化し保管するとともに，外部データの照合においても暗号化されたままの状態で照合を行う.」とその取扱いを定めている [2].

そこで，公開鍵暗号技術に基づく暗号プロトコルの一つである秘匿内積プロトコル (Secure Scalar Product)[3] を適用することで，二つのコホートを秘匿したままで複雑な条件について照合することを試みる．分散された二つのデータセットの照合を行うには，1. 準同型性を満たした公開鍵アルゴリズムによるセキュア内積プロトコル，2. 可換な一方向性関数によるプロトコル [4], 3. 秘匿多項式評価によるプロトコル [5] の 3 種類が代表的である．プロトコルの実行には，データベースのレコード毎の公開鍵暗号の計算が必要であり，その処理が実現可能なレベルにあるかどうかが大きな興味である．

セキュア内積プロトコルはこれらの中で最も高速であるが，二つの入力データは同一次元のベクトルで整合している条件がある．がん登録とピロリ菌患者のデータセットは，氏名，住所，生年月日などの個人識別の属性情報を有しているが，共通のユニーク ID は持たない．そこで，セキュアハッシュ関数により個人情報をハッシュ値に変換して ID とすることが考えられるが，ハッシュ関数の値域はセキュア内積プロトコルを適用するには大きすぎる．セキュア内積プロトコルはベクトルの次元に比例して計算コストがかかるので，例えば SHA256 の 256 ビットは 2^{256} 個の暗号文を生じさせて，現実的ではない為である．

そこで本稿では，プロトコルの処理性能や抽出誤差の観点で，暗号プロトコルの適用可能性を明らかにして，現実的な条件に付いて考察する．

連絡先: 菊池浩明, 東海大学, 〒 108-8619 東京都港区高輪 2-3-23, TEL: 03-3441-1171 内線 1611, FAX: 03-3447-6005, Email: kikn@tokai.ac.jp

2. 要素技術

2.1 セキュア内積プロトコル

E, D を加法準同型性を満たす公開鍵の暗号化と復号の関数とする．加法準同型性暗号に関しては，拡張 El Gamal 暗号や Paillier 暗号が知られている．n 次元ベクトル $\boldsymbol{x}, \boldsymbol{y}$ の内積は，共に 1 のビットの総数を与え，アルゴリズム 1 で計算される．

2.2 相対危険度

相対危険度 (relative risk) は，特定要因に暴露した群における比率の，暴露しなかった群での比率に対する比で定義される．例えば，ピロリ菌保有者という要因ががんに罹患する件数が表 1 の分割表で与えられたとき，相対危険度 RR は

$$RR = \frac{a}{a+b} \cdot \frac{c}{c+d} \approx \frac{ad}{bc}$$

で与えられる．ここで，一般に罹患率は小さいので，$a+b=b$ とみなしている．

この相対危険度 $RR = 1$ かどうかを検定したい．その有意性の検定は，$RR = 1$ の仮説の元，統計量 χ が

$$\chi = \frac{\sqrt{N-1}\,((ad-bc) \pm N/2)}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

Algorithm 1 Secure Scalar Product

Input: Alice has n-dimentional vector $\boldsymbol{x} = (x_1, \ldots, x_n)$. Bob has n-dimentional $\boldsymbol{y} = (y_1, \ldots, y_n)$.
Output: Alice has s_A and Bob has s_B such that $s_A + s_B = \boldsymbol{x} \cdot \boldsymbol{y}$.

1. Alice generates a homomorphic public-key pair and sends public key to Bob.
2. Alice sends to Bob n ciphertexts $E(x_1), \ldots, E(x_n)$.
3. Bob chooses s_B at random, computes
$$c = E(x_1)^{y_1} \cdots E(x_n)^{y_n} / E(s_B)$$
and send c to Alice.
4. Alice decrypts c to get $s_A = D(c) = x_1 y_1 + \cdots + x_n y_n - s_B$.

が標準正規分布 $N(0,1)$ に従うことで判定を行う．

表 1: 患者-対照調査によるデータの分布

要因	がん罹患	対照 (無)	罹患率
ピロリ菌	a	b	$a/(a+b)$
未感染	c	d	$c/(c+d)$

3. 提案方式

3.1 名前 ID 変換

名前や住所などの属性情報の組みから成る集合 $A = \{a_1, \ldots, a_N\}$ がある．A の個数を N とする．A のデータには一意な ID がないので，ハッシュ関数 $h()$ を適用して ID に変換する．このハッシュ関数の値域の大きさを n と置くと，この名前・ID 変換は

$$h(A; n) = \{h(a_i) \bmod n \mid a_i \in A\}$$

と表される．ハッシュ関数のサイズ n によって衝突，すなわち，$h(a_i) = h(a_j)$ となる $a_i \neq a_j \in A$ が存在するので，変換された集合の大きさは必ずしも A の大きさにはならない．

3.2 最適なハッシュサイズの決め方

データセットの大きさ N に対して最適なビット長 n はいくらだろうか．n を挙げるほど，マッチングにおける精度は増加するが，比例して暗号処理速度と通信コストが増加する．明らかに，$n \geq N$ であるが，その大きさは自明ではない．

この問題は，$1/n$ の一様分布で生じる事象（ハッシュ値）を N 回繰り返し，全て異なっていることを意味しており，よく知られた誕生日パラドックス [8] と同値である．大きさ n のハッシュ関数で求めた N 個のハッシュ値が全て異なる確率は，

$$\begin{aligned} \prod_{j=1}^{N-1} \left(1 - \frac{j}{n}\right) &\approx \prod_{j=1}^{N-1} e^{-j/n} \\ &= e^{-N(N-1)/2n} \approx e^{-N^2/2n} \end{aligned}$$

で与えられる．従って，N 個のハッシュ値がユニークになる確率を ϵ とすると，

$$\frac{N^2}{2n} = \ln \epsilon^{-1}$$

表 2: 千葉がんセンター登録データセット CAN

登録年	男性	女性	総数
2003	2,330	1,134	3,464
2004	2,610	1,242	3,852
2005	2,559	1,205	3,764
計	7,500	3,581	11,081

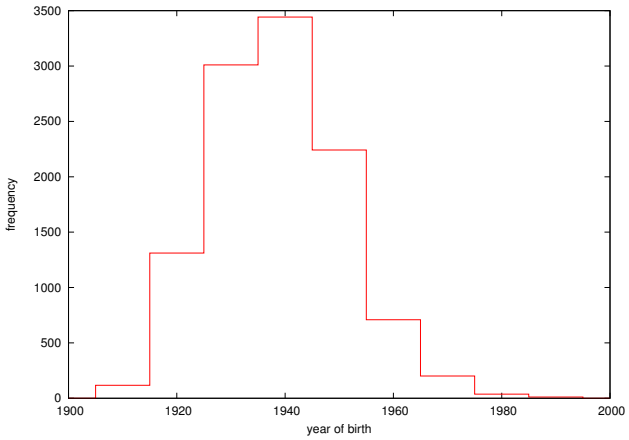

図 1: データセット CAN の人口分布

の関係があり，N と精度 ϵ を与えた時のハッシュ関数のサイズ N は $n = N^2/2 \ln \epsilon^{-1}$ で与えられる．例えば，がん患者サンプル数 $N = 7,000$ の時，ハッシュ値 $n = 4.7 \times 10^8$ は 95%の確率で一意になる．

3.3 プロトコルの高速化

内積プロトコルにおいて，暗号文の種類は 2 種類 $E(1)$ と $E(0)$ しかないことに着目すれば，ベクトルの全要素について暗号化をする必要はなく，入力ベクトルの要素におうじて，$E(1)$ か $E(0)$ を選んで出力することで高速化を計る．ただし，第三者にどちらの暗号文か識別不能にするためには，準同型性を応用して再暗号化の処理を行う必要がある．

4. 実験

4.1 実験データ

本実験では，表 2 のがん登録データセット CAN と表 4 のピロリ菌保有者のデータセット PYL の 2 種類の異なる実験データを用いる．CAN は，千葉がんセンター研究所予防疫学研究部によって収集されている，1975 年より千葉県内のがん登録から成るデータである．本実験では，表 2 の 2003 年から 3 年間の登録データを用いた．この登録の年齢分布を図 1 に示す．登録されている属性情報の一覧を表 3 に示す．

一方，データセット PYL は，厚生省によって調査された 2001 年から 2002 年に，千葉県の一部の地域在住者の検診データから，ピロリ菌の保有者を抽出したデータである．

表 4: 厚生省ピロリ菌データセット PYL

登録年	男性	女性	総数
2001–2002	2,671	5,206	7,877

表 3: データセット CAN の属性

属性	値
基本情報	氏名，性別，生年月日，住所，読み，住所コードなど
管理情報	受付番号，診断日，死亡日
診断部位	胃 (境界，小弯，大湾，底部)，幽門，幽門前庭など
組織	悪性リンパ腫，血管，腫瘍など
分化度	0 から 9

表 5: データセット CAN の重複数

重複種別	件数	割合 [%]
一意	7,413	98.9
2 名重複（同性同名）	86	11.5
3 名重複	3	0.04
4 名重複	3	0.04
男性全体	7,499	100

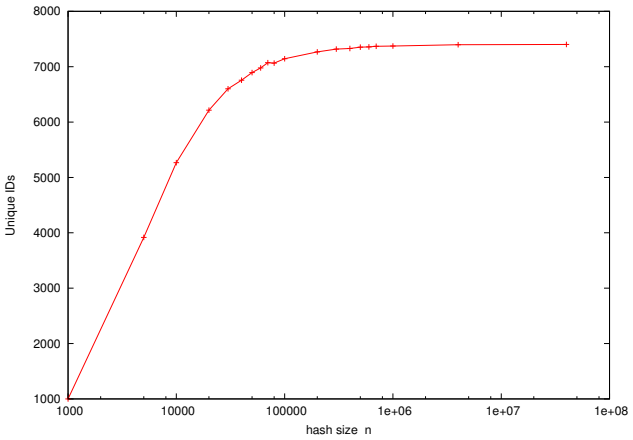

図 2: ハッシュ長 n に対するユニーク ID 数

4.2 名前の一意性

データセット CAN の男性を対象として氏名だけで一意に識別可能かどうか検証した．1 名異なる部位で 2 回登録されており，それ以外に表 5 に示される同姓同名の登録データがあった．名前だけでは識別は困難であり，生年月日や住所などの付帯属性と合わせて用いる必要がある．

4.3 ユニークハッシュ

図 2 は，データセット CAN の男性を対象として，ハッシュ長 n を変化させた時に生成されるユニークな ID の総数の変化を表している．ハッシュ関数のサイズ n に対して生成されるユニークな ID の個数は増加していき，$n = 4 \times 10^6$ の時にオリジナルのデータ数 7,500 に達する．

4.4 マッチング数

ハッシュされた ID について，2 つのデータセット CAN と PYL の照合を行い，ハッシュサイズ n に対するマッチングの数の変化を図 3 に示す．n の増加に対して偽りのマッチング数は減っていき，$n^* = 5 \times 10^7$ の時に真のマッチング数に達する．一方，CAN と PYL のユニークな ID の数は減っていき，それぞれ，n^* の前後で真のデータセット数に収束する．

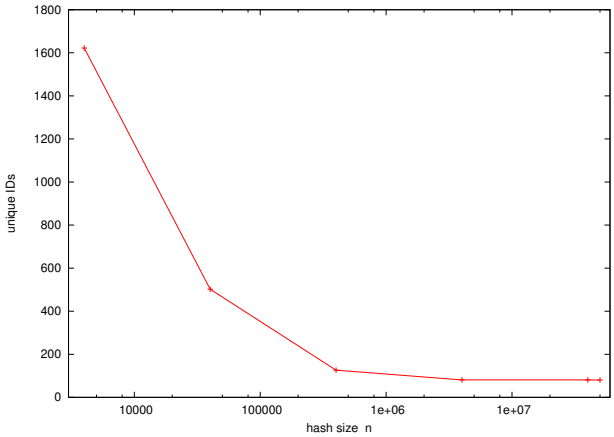

図 3: ハッシュ長 n に対するデータセット CAN, PYL のマッチング数

4.5 パフォーマンス

図 5 にハッシュ長 n を変化させた時のセキュア内積プロトコルの実行処理時間を示す．n に対して線形に増加している．図 6 は，Paillier 暗号のビット長 (剰余 n の大きさ) に対する計算時間の変化を示す．ビット長の 3 乗のオーダーで処理時間が増加している．

4.6 仮説検定

2 つのデータセットの照合結果から，ピロリ菌のがん罹患率に対する相対危険度を求める．実験結果より，2 つのデータセットの照合数とそれぞれの母集団の大きさは表 6 のとおりである．ただし，ピロリ菌の母集団の 23 万人は調査対象とした 5 つの市町村の登録当時の人口分布から算出し，千葉県がん登録数の母集団は千葉県の 2003 年の男性人口から与えている．

この結果より，相対危険度の推定値 (RR: Relative Risk) は，

$$RR = \frac{80 \cdot 106,988}{2,549 \cdot 346} = 9.70$$

であり，ピロリ菌の保菌者は，癌になるリスクが通常の 10 倍近く上がることを示している．この時の，優位性検定は

$$\chi \quad = \quad 17.81$$

と与えられる．母集団はピロリ菌の検診対象の市町村に限定されており，十分な有意水準に達していることを表している．

5. おわりに

プライバシーを保護して疫学調査を実施する方法について検討し，セキュア内積プロトコルを適用することで，二つのデータベースの照合を実現する方法を提案した．

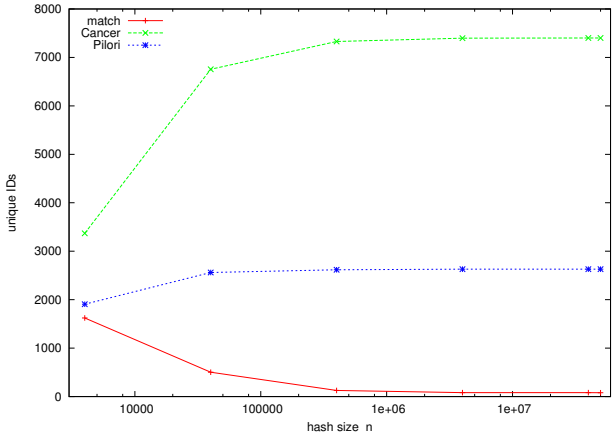

図 4: ハッシュ長 n に対するデータセット CAN, PYL の変化とマッチング数

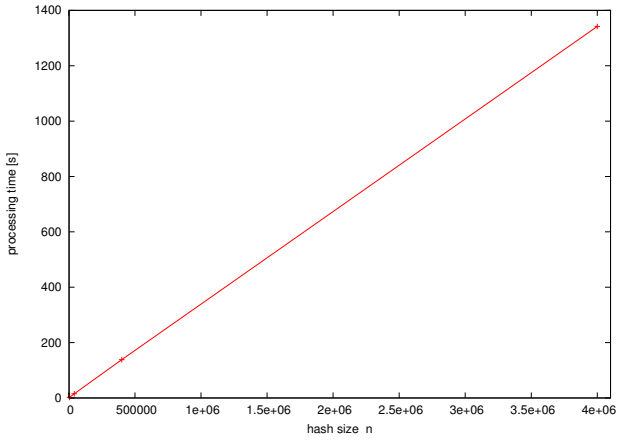

図 5: ハッシュ長 n に対する処理時間

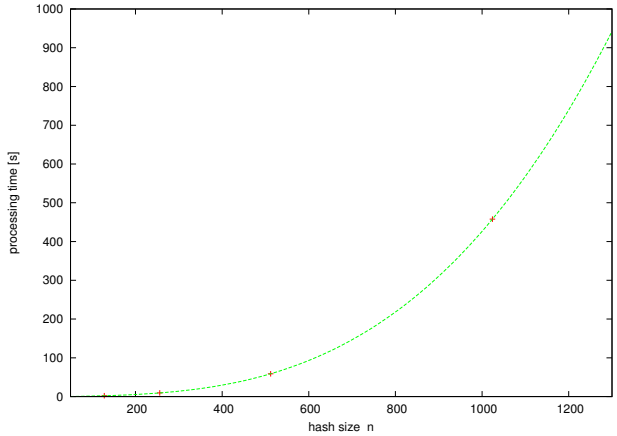

図 6: 暗号鍵長 (bit) に対する処理時間

謝辞

本研究の遂行にあたって，千葉がんセンターの高山喜美子氏にがん登録に関する有益な助言を頂いた．感謝致します．

参考文献

[1] 矢沢サイエンスオフィス, 久保田 哲朗, 米村 豊, 吉田 和彦, “胃ガンのすべてがわかる本”, 学研，2005.

[2] 千葉県健康福祉部健康づくり支援課，“がん登録の手引き”, pp. 26-27, 2011.

[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] Rakesh Agrawal, Alexandre Evfimievski, and Ramakrishnan Srikant, “Information shareing across private databases”, in proc. of ACM SIGMOD Intl. Conf. on Management of Data, 2003.

[5] M. J. Freedman, K. Nissim, and B. Pinkas, “Efficient private matching and set intersection”, EUROCRYPT 2004, LNCS 3027, pp. 1?19, Springer-Verlag, 2004.

[6] Dahlia Malkhi, Noam Nisan, Benny Pinkas, and Yaron Sella, “Fairplay - A Secure Two-Party Computation System”, Usenix Security Symposium, 2004.

[7] 古川俊之，丹後俊郎，“新版 医学への統計学”, 朝倉書店, pp. 133-147, 1993.

[8] Michael Mitzenmacher and Eli Upfal, “Probability and Comuting – Randomized Algorithms and Probabilistic Analysis”, Cambridge University Press, 2005.

表 6: ピロリ菌とがん罹患率の間の関係

男性	登録がん患者数	非登録者数	計
ピロリ菌保有者	80	2,549	2,629
非保有者	346	106,988	107,334
計	426	2,999,574	3,000,000