Privacy-Preserving Computation of Disease Risk by Using Genomic, Clinical, and Environmental Data

Reporter:Ximeng Liu
Supervisor: Rongxing Lu

School of EEE, NTU
http://www.ntu.edu.sg/home/rxlu/seminars.htm

May 1, 2014

1. Main References
2. Introduction
3. Background
4. The proposed scheme

Main References

1. Ayday E, Raisaro J L, McLaren P J, et al. Privacy-Preserving Computation of Disease Risk by Using Genomic, Clinical, and Environmental Data[J].
2. Damgård I, Geisler M, Krøigaard M. Efficient and secure comparison for on-line auctions[C]//Information Security and Privacy. Springer Berlin Heidelberg, 2007: 416-430.

Introduction

As a result of the rapid evolution in genomic research and the dramatic decrease in the costs of sequencing, the paradigm of classic medicine has been shifting towards a more personalized approach. The use of individual genomic, clinical, and environmental data can be of interest for a large variety of healthcare stakeholders (here described as medical units).

Introduction

Also, in order to protect the privacy of his sensitive data, an individual might not want to directly provide his genomic data and clinical and environmental attributes to the medical unit. Because of its extremely sensitive nature, genomic data has an unprecedented impact on privacy. In particular, because the genome carries information about a person's genetic condition and predispositions to specific diseases, the leakage of such information could enable abuse and threats.

Genomic Background

The human genome is encoded in a double-stranded helical DNA molecule, as a sequence of nucleotides. Genome sequencing techniques record the nucleotides by using the letters A, T, G and C, and the whole human genome includes approximately 3 billion letters.

Genomic Background

Around 99.9% of the entire genome is identical between any two given individuals. The remaining part (0.1%) is responsible for many of our interindividual differences. The latter are called *single nucleotide polymorphisms (SNPs)* when they are found to be variable in at least 1% of the individuals in a population.

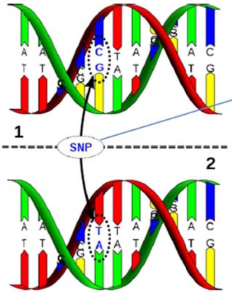
Genomic Background

Two different alleles are observed for every SNP. In general, for a SNP that is associated with a disease, one of these alleles carries the risk for the corresponding disease and the other allele does not contribute. For example, assume that the SNP in the above example (with alleles G and T) is associated with a particular disease X.

Genomic Background

In this paper, the author represent (i) an homozygous SNP carrying two noncontributing alleles as 0, (ii) an heterozygous SNP carrying one risk (or protective) allele and one non-contributing allele as 1, and (iii) an homozygous SNP carrying two risk (or protective) alleles as 2. In short, each SNP can be in one of the states from $\{0, 1, 2\}$, and we let SNP_i^P represent the state (content) of SNP_i (SNP with ID i) for a patient P.

Genomic Background



SNP associated with disease X

- Alleles: C and T
- Risk allele: C
- Genotypes: CC, TT, CT

$SNP_i^P = \{0,1,2\}$ based on the number of risk alleles it carries.

Computation of the Disease Risk

The strength of the association between each SNP and a disease is usually expressed by the odds ratio (OR), where the odds is the ratio of the probability of occurrence of the disease to that of its non-occurrence in a specific group of individuals. Thus, the OR is the ratio of odds in the group of individuals carrying a genetic variation (exposed) to that of those who do not carry it (unexposed).

Computation of the Disease Risk

When multiple SNPs are associated with a disease, the overall genetic risk ($\mathbb{S}$) of an individual for the corresponding disease can be computed as a weighted average, based on the OR of each associated SNP by using a **logistic regression model**.

Computation of the Disease Risk

This model is currently widely used among the geneticists and medical doctors for disease risk tests. In such a model, OR of a SNP_i (i.e., OR_i) is generally represented in terms of regression coefficient (β_i), where $OR_i = exp(\beta_i)$. Then, assuming Pr_g is the probability that an individual P will develop a disease X (only considering his genomic data), his overall genetic risk can be computed as below:

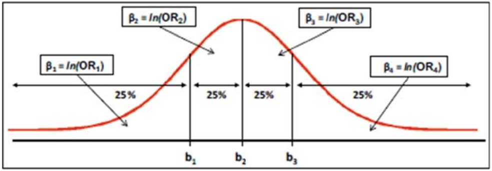
$$\mathbb{S} = ln(\frac{Pr_g}{1 - Pr_g}) = \alpha + \sum_{i \in \varphi_x} \beta_i p_j^i(X)(1)$$

where $p_j^i(X)$ is the contribution of the SNPi to the genetic risk (for disease X) when $SNP_i^P = j$ ($SNP_i^P \in 0, 1, 2$, and α is the intercept of the model.

Computation of the Disease Risk

For clinical use, the genetic risk, computed in (1) should be categorized based on its risk group. For this purpose, generally, the distribution of the potential genetic scores (in a given population) is divided into smaller parts called quantiles

Example


$\beta_2 = ln(OR_2)$
$\beta_3 = ln(OR_3)$
$\beta_1 = ln(OR_1)$
$\beta_4 = ln(OR_4)$
25%
25%
25%
25%
b_1
b_2
b_3

Example

There are 4 different risk groups, each with a different genetic regression coefficient. For example, if $\mathbb{S}$ is somewhere between b_1 and b_2, then we assign the genetic regression coefficient for the corresponding individual as β_2. For each individual, the genetic score is computed as in (1), and positioned into its risk group. We represent the genetic regression coefficient corresponding to the genetic risk $\mathbb{S}$ as β_g.

Example

The overall disease risk, the genetic information needs to be combined together with the clinical and environmental factors. For this purpose, assuming Pr is the probability of disease X (this time considering genetic, clinical, and environmental information), a second and final multi-variable logistic regression model is used to find the final (aggregate) regression coefficient β_f as below:

$$In(\frac{Pr}{1-Pr}) = \beta_f = \beta_0 + \beta_g + \sum_{N_i \in \mathbb{N}} \bar{\beta}_i N_i$$

Example

Where β_0 is the new intercept, $\mathbb{N}$ is the set of clinical and environmental attributes associated with the disease, and $\bar{\beta}_i$ is the regression coefficient corresponding to the clinical or environmental attribute N_i. The probability (Pr) that the corresponding individual will develop disease X as $\frac{e^{\beta_f}}{1+e^{\beta_f}}$.

Modified Paillier cryptosystem

The public key is represented as $(n, g, h = g^x)$, where the strong secret key is the factorization of $n = zy$ (z, y are safe primes), the weak secret key is $x \in [1, n^2/2]$, and g of the order $(z-1)(y-1)/2$.

Encryption: To encrypt a message $m \in Z_n$, we first select a random $r \in [1, n/4]$ and generate the ciphertext pair (C_1, C_2) as below:

$$C_1 = g^r \mod n^2$$

$$C_2 = h^r(1 + mn) \mod n^2$$

For simplicity, in the rest of this paper, we represent the Paillier encryption of a message m as [m].

Modified Paillier cryptosystem

Encryption: The message m can be recovered from [m] as follows:

$$m = \Delta(C_2/C_1^x)$$

where$\Delta(u) = \frac{(u-1) \mod n^2}{n}$
Proxy re-encryption: Assume we randomly split the secret key in two shares x_1 and x_2, such that $x = x_1 + x_2$. The modified Paillier cryptosystem enables an encrypted message (C_1, C_2) to be partially decrypted to a ciphertext pair (C_1', C_2') using x_1 as below:

$$C_1' = C_1$$

$$C_2' = C_2/C_1^{x_1} \mod n^2$$

Then, (C_1', C_2') can be decrypted using x_2 with the aforementioned decryption function to recover the original message.

DGK cryptosystem

Random elements $g, h \in Z_n$ such that the multiplicative order of h is v modulo p and q, and g has order uv. The public key is now $pk = (n, g, h, u)$ and the secret key is $sk = (p, q, v)$. The ciphertext be $E_{pk}(m, r) = g^m h^r modn$. where m is the message. It is also possible to do a real decryption by noting that $E_{pk}(m, r)^v = (g^v)^m \mod n$ Clearly, g^v has order u, so there is a $1C1$ correspondence between values of m and values of $(g^v)^m$ mod n. Since u is very small, one can simply build a table containing values of $(g^v)^m \mod n$ and corresponding values of m.

DGK cryptosystem

random elements $g, h \in Z_n$ such that the multiplicative order of h is v modulo p and q, and g has order uv. The public key is now $pk = (n, g, h, u)$ and the secret key is $sk = (p, q, v)$. The ciphertext be $E_{pk}(m, r) = g^m h^r modn$. where m is the message. Decryption $E_{pk}(m, r)^v = (g^v)^m \mod n$ Clearly, g^v has order u, so there is a 1-1 correspondence between values of m and values of $(g^v)^m \mod n$. Since u is very small, one can simply build a table containing values of $(g^v)^m \mod n$ and corresponding values of m.

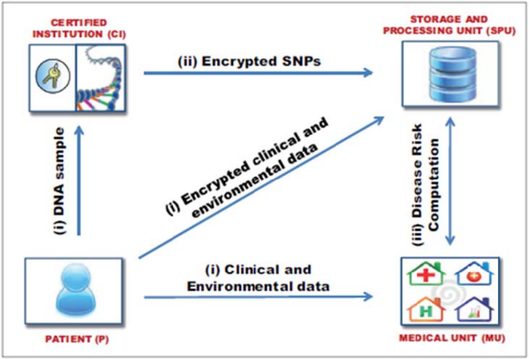
DGK cryptosystem

1)A and B compute, for $i = 1, \cdots, l$ sharings $[w_i]$ where

$$w_i = m_i + x_i - 2x_i m_i = m_i \oplus x_i$$

2) A and B now compute, for i $= i = 1, \cdots, l$ sharings $[c_i]$ where $c_i = x_i - m_i + 1 - \sum_{j=i+1}^{l} w_i$. Note that if $m > x$, then there is exactly one position i where $c_i = 0$.
3) A uses his secret key to decide, as described in the previous section, whether any of the received encryptions contain 0. If this is the case, he outputs $m > x$. Otherwise, $m \leq x$.

System Model


CERTIFIED INSTITUTION (CI)
(ii) Encrypted SNPs
STORAGE AND PROCESSING UNIT (SPU)
(i) DNA sample
(i) Encrypted clinical and environmental data
(iii) Disease Risk Computation
(i) Clinical and Environmental data
PATIENT (P)
MEDICAL UNIT (MU)

System Model

The encryption of the genomic data of the patient are performed at a certified institution (CI), which is a trusted entity. The medical unit can be a pharmacist, a pharmaceutical company, a regional health ministry, an online direct-to-consumer service provider, or a physician.

The storage and processing of genomic, clinical, and environmental data is done at a storage and processing unit (SPU) for efficiency and security. We note that a private company (e.g., cloud storage service), the government, or a non-profit organization could play the role of the SPU.

Initialization

The patient's secret key x is randomly divided into x_1 and x_2 (such that $x = x_1 + x_2$) and each share is distributed to the SPU and to the MU, respectively (i.e., x_1 is provided to the SPU and x_2 to the MU)

Sequencing and Clinical and Environmental Data Collection

Clinical and environmental data of the patient is collected during his doctor visits or directly provided by the patient. For example, data about his cholesterol level or his blood-sugar level is collected during his doctor visits. Whereas, data such as his age, weight, height, or family history is provided by the patient.

Encryption and Storage of Genomic, Clinical, and Environmental Data

After the sequencing and the extraction of the SNPs of the patient, the CI encrypts the contents of all SNP positions of the patient (to obtain $[SNP_i^P]$).

Privacy-Preserving Computation of Disease Risk

1) Computing the genetic risk: As before, let SNP_i represent the position (or ID) of a SNP, SNP_i^P represent the content of the corresponding SNP ($SNP_i^P \in 0, 1, 2$), and β_i represent the regression coefficient, thus the strength of the association between SNP_i and disease X. Also, let $p_j^i(X)$ be the contribution, depending on the content, of the SNP_i to the genetic risk (for disease X) when $SNP_i^P = j$. Then, the MU computes the (encrypted) genetic risk ($[\mathbb{S}]$) of patient P to disease X using the encrypted SNPs of the patient as below:

Privacy-Preserving Computation of Disease Risk

$$[\mathbb{S}] = \Bigg[\sum_{i \in \varphi_X} \beta_i \Bigg\{ \frac{\mathrm{p}_0^i(X)}{\chi} (\mathrm{SNP}_i^P - 1)(\mathrm{SNP}_i^P - 2) + \frac{\mathrm{p}_1^i(X)}{\psi} (\mathrm{SNP}_i^P)(\mathrm{SNP}_i^P - 2) + \frac{\mathrm{p}_2^i(X)}{\mu} (\mathrm{SNP}_i^P)(\mathrm{SNP}_i^P - 1) \Bigg\} \Bigg], \quad (7)$$

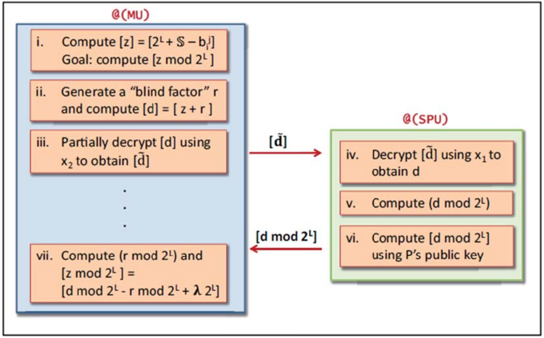
Privacy-Preserving Computation of Disease Risk

However, as the above computed genetic risk is encrypted, to find the regression coefficient corresponding to the computed genetic risk, we propose to use a privacy-preserving integer comparison algorithm between the MU and the SPU.

Privacy-Preserving Computation of Disease Risk

2)Computing the genetic regression coefficient: We let b_i^l and b_i^u represent the lower and upper boundary of the i-th risk group of the genetic risk scale, respectively. In short, MU compares $[\mathbb{S}]$ with the boundaries of the genetic risk scale in a privacy-preserving way, such that neither the MU nor the *SPU* learns the value of $\mathbb{S}$ or the result of any comparison.

Privacy-Preserving Computation of Disease Risk



i. Compute $[z] = [2^L + \mathbb{S} - b_i^j]$
Goal: compute $[z \bmod 2^L]$

ii. Generate a "blind factor" r and compute $[d] = [z + r]$

iii. Partially decrypt $[d]$ using x_2 to obtain $[\tilde{d}]$

iv. Decrypt $[\tilde{d}]$ using x_1 to obtain d

v. Compute ($d \bmod 2^L$)

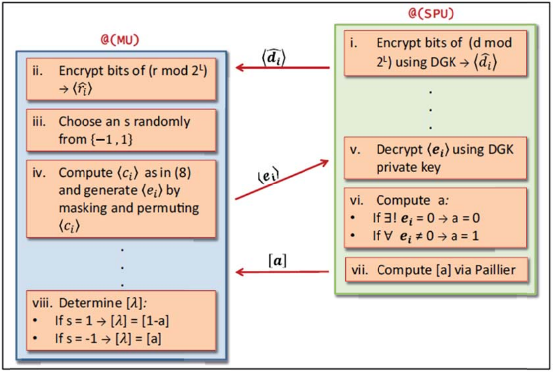
vi. Compute $[d \bmod 2^L]$ using P's public key

vii. Compute ($r \bmod 2^L$) and
$[z \bmod 2^L] =$
$[d \bmod 2^L - r \bmod 2^L + \lambda\, 2^L]$

Privacy-Preserving Computation of Disease Risk

The MU computes $[z] = [2^L + \mathbb{S} - b_i^j]$. Let z_{L-1} represent the most significant bit of z. Then, (i) $z_{L-1} = 0$ if $S < b_i^j$; and (ii) $z_{L-1} = 1$ if $S \geq b_i^j$. Thus, the MU needs to compute $[z_{L-1}]$, where $[z_{L-1}] = [z - (zmod2^L)]$.

Computing $[\lambda]$



@(SPU)

i. Encrypt bits of (d mod 2^L) using DGK → $\langle \widehat{d_i} \rangle$

$\langle \widehat{d_i} \rangle$

@(MU)

ii. Encrypt bits of (r mod 2^L) → $\langle \widehat{r_i} \rangle$

iii. Choose an s randomly from $\{-1, 1\}$

iv. Compute $\langle c_i \rangle$ as in (8) and generate $\langle e_i \rangle$ by masking and permuting $\langle c_i \rangle$

$\langle e_i \rangle$

v. Decrypt $\langle e_i \rangle$ using DGK private key

vi. Compute a:

- If $\exists !\ e_i = 0 \rightarrow a = 0$
- If $\forall\ e_i \neq 0 \rightarrow a = 1$

vii. Compute [a] via Paillier

$[a]$

viii. Determine $[\lambda]$:

- If s = 1 → $[\lambda]$ = [1-a]
- If s = -1 → $[\lambda]$ = [a]

Computing $[\lambda]$

Where $< c_i > = < \hat{d}_i - \hat{r}_i + s + 3\sum_{j=i+1}^{L-1} w_j >$.
If $a = 1$ and $s = 1$ (the number randomly selected by the MU), then $\hat{d} \geq \hat{r}$ (i.e., $\lambda = 0$). Similarly, if $a = 0$ and $s = 1$, then $\hat{r} > \hat{d}$ (i.e.,$\lambda = 1$). Thus, if $s = 1$, the MU sets $[\lambda] = [1 - a]$ and if $s = -1$, it sets $[\lambda] = [a]$. Using $[\lambda]$, the MU can compute $[zmod2^L]$.

Computing $[\lambda]$

Let $[G(\mathbb{S}, b_i^u)] = [z_{L-1}]$ represent the (encrypted) result of the comparison between $\mathbb{S}$ and b_i^u. Then, (i) $G(\mathbb{S}, b_i^u) = 0$ if $S < b_i^l$; and (ii) $G(\mathbb{S}, b_i^u) = 1$ if $\mathbb{S} \geq b_i^l$.

Computing $[\lambda]$

$$[\beta_g] = \Big[\beta_1(1 - G(\mathbb{S}, b_1^u)) + \sum_{i=2}^{(\rho-1)} \beta_i(G(\mathbb{S}, b_{i-1}^u) - G(\mathbb{S}, b_i^u)) + \beta_\rho G(\mathbb{S}, b_{\rho-1}^u)\Big],$$

Computing the final disease risk

$$[\beta_f] = [\beta_0 + \beta_g + \sum_{i=1}^{m} \bar{\beta}_i N_i]$$

$N_i = 1$ if the patient has the corresponding clinical or environmental attribute, and $N_i = 0$ otherwise. As we discussed before, even if N_i is non-binary, it can be transformed to a binary number using the privacy-preserving comparison algorithm. Finally, the MU computes the final disease risk of the patient for disease X as $\frac{e^{\beta_f}}{1+e^{\beta_f}}$

Thank you

Rongxing's Homepage:

http://www.ntu.edu.sg/home/rxlu/index.htm

PPT available @: http://www.ntu.edu.sg/home/rxlu/seminars.htm

Ximeng's Homepage:

http://www.liuximeng.cn/