Journal of Information Security and Applications 46 (2019) 271–280

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

An efficient and privacy-Preserving pre-clinical guide scheme for mobile eHealthcare

Guoming Wang[a,*], Rongxing Lu[b], Cheng Huang[c], Yong Liang Guan[a]

[a] The School of Electrical and Electronic Engineering, Nanyang Technological University, 639798 Singapore
[b] Faculty of Computer Science, University of New Brunswick, New Brunswick, Canada
[c] Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

ARTICLE INFO

Article history:
Available online 1 April 2019

Keywords:
Mobile e-Healthcare
Pre-clinical guidance
Privacy-preserving
Efficiency

ABSTRACT

With the increasing popularity of pervasive devices such as smartphones and Internet-of-Things devices, mobile e-Healthcare has become a research trend in recent years. Disease risk prediction using big data analytics techniques is one popular e-Healthcare research focus, and one associated research challenge is ensuring the privacy of user and patient data. In this paper, we propose a new efficient and privacy-preserving pre-clinical guidance scheme (hereafter referred to as PGuide) for mobile eHealthcare, designed to offer both self-diagnosis and hospital recommendation services to users in a privacy-preserving way. To provide users the capability to present a detailed health profile for accurate disease risk prediction, we introduce a privacy-preserving comparison protocol (PPCP) in PGuide, which will improve the accuracy of disease risk prediction. We also employ a single-attribute encryption technique to devise a privacy-preserving hospital recommendation service in PGuide, which can further guide users to choose a hospital appropriate for their visit after conducting a self-diagnosis. We then prove that PGuide can achieve the privacy-preservation requirements in both self-diagnosis and hospitals recommendation services. We also conduct a number of experiments, which demonstrate the efficiency of PGuide, in terms of computational cost and communication overhead.

1. Introduction

One of the most pressing issues facing hospital administration is the shortage of medical doctors, particularly with the increasing number of patients and in an ageing population. This has resulted in an increased waiting time for patients in many countries. For example, a patient has to wait on average of more than 20 minutes in U.S. [1], between 4 and 24 hours to be seen in a hospital in Canada [2], and significantly longer in China [3,4]. Patients are unlikely to be familiar with medical departments in hospitals and most are certainly not familiar with the symptoms associated with the different diseases. For example, there are approximately 60 different types of doctors and specialists [5] in a typical hospital, and it would be a frustrating and time-wasting exercise if a patient consults a doctor not trained or specialized in the particular disease. Some major hospitals have established manned inquiry counters to guide patients (e.g. which department or specialist the patient should go to), but this is not a viable solution due to a number of reasons. For example, a patient is unlikely or willing to share sensitive health-related issues, such as HIV and mental illness, publicly over the manned inquiry counters. Therefore, we need an effective solution to help ease the stretch on existing limited medical resources in hospitals.

Due to the increasing digitization of our society and popularity of pervasive devices (e.g. smartphones), where most of our data (including healthcare related data) are available electronically, using big data analytics to solve several healthcare related issues (e.g. disease risk prediction) has been the subject of recent research focus. However, little progress has been made in the commercial integration of processing clinical analytics while assuring the privacy of the sensitive medical information [6]. For example, how can we be assured that our sensitive medical information are not been made available and exploited by third-party companies, such as insurance companies and future employers? In addition to the privacy requirements of medical users, information leakage is a major concern for service providers as medical providers in countries such as U.S. are subject to exacting regulatory regime (e.g. HIPAA). Some recent works [7,8] discuss that the big data protection and privacy protection has become one of the hot issues in the researches about medical data analysis.

* Corresponding author.
E-mail addresses: wang0947@e.ntu.edu.sg (G. Wang), rlu1@unb.ca (R. Lu), cheng.huang@uwaterloo.ca (C. Huang), EYLGUAN@ntu.edu.sg (Y.L. Guan).

https://doi.org/10.1016/j.jisa.2019.03.009
2214-2126/

In order to address the above-mentioned privacy challenges, and improve the accuracy of disease risk prediction (and resulting in a shorter queue in hospitals), we propose an efficient privacy-preserving pre-clinical guidance service scheme (PGuide) designed to provide on-the-go medical guidance service while preserving user privacy. Different from our previous work [9], using the PGuide scheme proposed in this work, users can personally conduct privacy-preserving pre-clinical diagnosis based on their health profiles [10] and obtain recommendation from trusted sources (e.g. hospitals and medical service providers) based on the diagnosis. Specifically, the proposed PGuide ensures that an user is unable to learn the coefficients, the tercept and the threshold in the risk model for a disease, while pretects the user's health profile information being disclosed to the service provider. In addition, the information transmitted to the hospitals and other medical service providers to calculate the disease risk use a disease prediction model in a privacy-preserving way. Specifically, We then prove that our proposed PGuide scheme achieves the privacy-preservation for both the individual user and the medical service provider. To demonstrate the practicality of our scheme, we develop an Android app and a Java service application. Based on the findings from our evaluations, we show that our proposed PGuide scheme is efficient, in terms of computational cost and communication overhead.

The remainder of this paper is organized as follows. In Section 2, we introduce our system model, security model and design goal. In Section 3, we introduce the preliminaries required to understand our proposed scheme, prior to presenting our scheme in Section 4. The security analysis and performance evaluation are provided in Section 5 and Section 6, respectively. We also discuss the related work in Section 7. Section 8 concludes this paper.

2. Models and design goal

In this section, we formalize the system model, security model, and our design goal.

2.1. System model

In our system model, we focus on the disease risk calculation for users with the help of one or more medical service providers. In other words, our system model comprises four entities, namely: a group of m hospitals $H = \{H_1, H_2, \ldots, H_m\}$, a healthcare center as the trust authority (TA), a service provider (SP), and a number of users $\mathbb{U} = \{U_1, U_2, \ldots, U_N\}$. The system model is illustrated in Fig. 1, where N, m indicate the numbers of medical users and hospitals, respectively.

- Healthcare center (TA): Healthcare center is a trusted entity, mainly responsible for initializing the system key materials for hospitals and medical users.
- Hospitals $\mathbb{H} = \{H_1, H_2, \ldots, H_m\}$: Each hospital $H_j \in \mathbb{H}$ needs to have a one-off registration with the TA prior to processing user's disease queries, providing feedback to SP when it has the available resources (e.g. medical doctors) to treat the specific disease or not (i.e. "no"), etc.
- Service Provider (SP): SP is the core entity, and is responsible for information processing, building the disease risk prediction model, and providing disease risk calculation service to users based on the their health profiles. In addition, SP directs disease queries from users to the relevant hospitals, as well as recommending available hospitals to the users.
- Users $\mathbb{U} = \{U_1, U_2, \ldots, U_N\}$: Each user $U_i \in \mathbb{U}$ has installed our app on their smartphone(s). The app collects the user's health profile, sends them in a privacy-preserving way to the SP, and receives the recommendations from the SP.

2.2. Security model

In our security model, we consider both SP and $\mathbb{U} = \{U_1, U_2, \ldots, U_N\}$ are *honest-but-curious*. That is, SP will faithfully follow the disease risk diagnosis protocol, but also attempt to learn the users' sensitive health profile data. In addition, users $\mathbb{U} = \{U_1, U_2, \ldots, U_N\}$ are also *honest-but-curious*, i.e. each U_i will not report false data, but may attempt to learn SP's disease risk predication model, which is regarded as SP's intellectual property (IP). We also assume that due to vested interest (e.g. reputation risk and criminal sanctions), a SP is not in collusion with the hospitals. Therefore, the following security requirements should be satisfied in the pre-clinical system.

- *Privacy Preservation.* Ensuring the privacy of user's sensitive health profile from SP is necessary, i.e. SP cannot learn user's medical history and other sensitive information (e.g. blood type). In addition, information about user's diagnosis should not be learned by SP, and we assume that SP does not collude with the hospitals. It is also necessary to protect SP's disease prediction model (i.e. IP).
- *Authentication.* Authenticating the hospitals' recommendations to the users is important. For example, if a non-registered hospital (i.e. a hospital that has not been vetted as having appropriate standard) is recommended to the user, this could result in fatalities and subsequent law suits. Therefore, in the hospi-

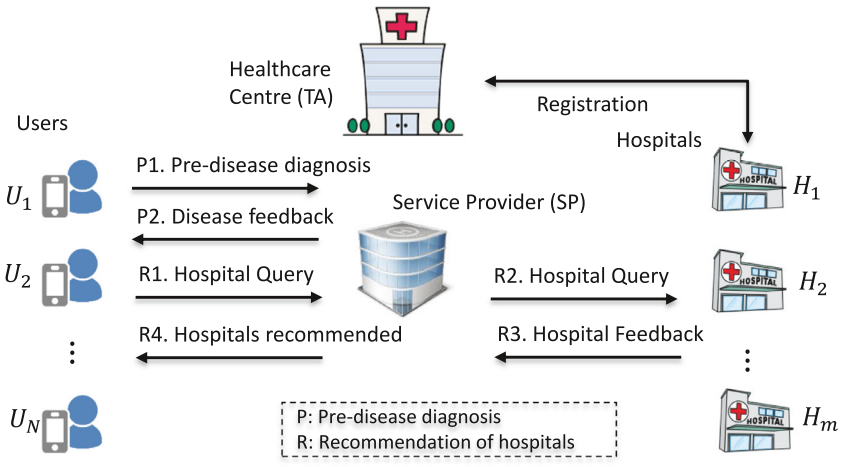


Fig. 1. System model under consideration.

tal recommendation system, only hospitals that have registered with the TA can receive the user request.

We acknowledge that there are several other security attacks, such as forgery attacks and background attacks, in a typical disease risk calculation system. Since our focus is on privacy-preserving disease risk prediction, attacks not targeting privacy issue are beyond the scope of this paper, and will be discussed in future work.

2.3. Design goal

Our design goal is to develop a privacy-preserving pre-clinical guidance scheme to provide disease risk prediction and hospital recommendation. A practical outcome of this scheme is to reduce the waiting time of a patient at a hospital. Using our scheme, a user seeking medical attention can have access to the disease risk predication services and hospital recommendation services from the medical service provider, without compromising on the privacy of user and SP. More specifically, the following two goals should be achieved.

- The proposed scheme should be effective, in terms of computation and communication. Despite the increase in computational resources available in smartphones, the storage and battery life are somewhat limited. Therefore, it is necessary to ensure that only lightweight computations are performed at the user-end and at the same time, do not overload the servers at SP or hospital.
- The security requirement should be guaranteed. If users are not assured that their health profiles are protected, then users will hesitate to (or not) use this service. Similarly, if the disease prediction model is not protected, then SP will not participate due to IP concerns. Therefore, the proposed scheme should also ensure that participating hospitals are authenticated.

3. Preliminaries

In this section, we revisit the disease risk model of Ayday, et al. [11], use the underlying disease risk threshold (S_{th}) and bilinear pairings [12] as the basis of our PGuide scheme.

3.1. Disease risk model

Many diagnosis prediction models combine patient characteristics and environmental data to predict the presence or absence of a certain diagnosis [13]. The association between each symptom and a disease is expressed by the *odds ratio* (OR), which is the ratio of odds in a group of individuals having the symptom to that of those who do not have. The OR OR_i of a disease Y_i for some symptom predictors $A_i = \{a_1, a_2, \ldots, a_m\}$, with each predictor value $a_j \in \{0, 1\}$ for $j = 1, 2, \ldots, m$, is generally represented in terms of regression coefficients $B_i = \{b_1, b_2, \ldots, b_m\}$ of the same length m. In this way, the predicted risk of the disease Y_i with regards to the symptom A_i can be calculated as:

$$P(Y_i = 1|A_i) = \frac{1}{1 + \exp(-(\gamma + \sum_{j=1}^{m} a_j \cdot b_j))}, \quad (1)$$

where γ is an estimated intercept in the model. This model has been widely used in the medicine and clinician fields for disease risk tests and predictions [13]. To simplify the risk score calculation, the overall disease risk score S corresponding to the risk $P = P(Y_i = 1|A_i) = \frac{1}{1+\exp(-(\gamma+S))}$ can be computed by

$$S = \ln \frac{P}{1-P} = \gamma + \sum_{j=1}^{m} a_j \cdot b_j \quad (2)$$

For a complete description of the logistic regression model, we refer the interested reader to [13].

3.2. Determination of disease risk threshold

In the above disease risk model, the regression coefficients $B_i = \{b_1, b_2, \cdots, b_m\}$ and the estimated intercept γ for predicting some disease Y_i can be derived from the logistic regression model with a large volume of real-world medical data. In order to determine whether a user $U_l \in \mathbb{U}$ with the symptom predictors $A_i = \{a_1, a_2, \cdots, a_m\}$ has the disease Y_i with a high probability, we can set a disease risk threshold S_{th}. If $\gamma + \sum_{j=1}^{m} a_j \cdot b_j \geq S_{\text{th}}$, then we can infer that user U_l has disease Y_i with a high probability. Otherwise, when $\gamma + \sum_{j=1}^{m} a_j \cdot b_j < S_{\text{th}}$, we infer that U_l has Y_i with a low probability. Because the disease risk model is an asset, the values $(B_i = \{b_1, b_2, \cdots, b_m\}, \gamma, S_{\text{th}})$ should be kept private (i.e. privacy-preserving requirement). In the next section, we will present our PGuide scheme, which utilizes the disease risk model in an efficient and privacy-preserving way to achieve pre-clinical guidance for medical user.

3.3. Bilinear pairings

Let $\mathbb{G}, \mathbb{G}_T$ be two multiplicative cyclic groups with the same prime order q. Suppose $\mathbb{G}$ and $\mathbb{G}_T$ are equipped with a pairing, i.e. a non-degenerated and efficiently computable bilinear map $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$, such that $e(g_1^a, g_2^b) = e(g_1, g_2)^{ab} \in \mathbb{G}_T$ for all $a, b \in \mathbb{Z}_q^*$, and any $g_1, g_2 \in \mathbb{G}$ in group $\mathbb{G}$, the Computational Diffie-Hellman (CDH) problem is hard. For the latter, given (g, g^a, g^b) for $g \in \mathbb{G}$ and unknown $a, b \in \mathbb{Z}_q^*$, it is intractable to compute g^{ab} in a polynomial time. However, the Decisional Diffie-Hellman (DDH) problem is easy. In other words, given (g, g^a, g^b, g^c) for $g \in \mathbb{G}$ and unknown $a, b, c \in \mathbb{Z}_q^*$, it is easy to determine whether $c = ab \mod q$ by checking $e(g^a, g^b) \stackrel{?}{=} e(g^c, g)$.

Definition 1. A bilinear parameter generator *gen* is a probabilistic algorithm that takes a security parameter k as input, and outputs a 5-tuple $(q, g, \mathbb{G}, \mathbb{G}_T, e)$, where q is a k-bit prime number, and $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$ is a non-degenerated and efficiently computable bilinear map.

4. Proposed PGuide scheme

In this section, we propose our PGuide scheme. The scheme consists of two main phases, namely: a system setting and privacy-preserving pre-clinical guidance, together with its correctness analysis. The pre-clinical guidance scheme has a privacy-preserving pre-disease diagnosis and a privacy-preserving hospital recommendation.

4.1. System setting

The TA located at the healthcare center will bootstrap the recommendation system. Specifically, given the security parameter k, TA generates the bilinear parameters $(q, g, \mathbb{G}, \mathbb{G}_T, e)$ by running $gen(k)$, and chooses a secure symmetric encryption algorithm $Enc()$ (i.e. AES) [14]. In addition, TA chooses two random numbers $(a, x) \in \mathbb{Z}_q^*$, as the master key, two random elements (h_1, h_2) in $\mathbb{G}$, and computes $b = H(a), A = g^a$, and $e(g, g)^x$. Finally, TA keeps the master key (a, b, x) secret, and publishes the system parameter $params = (q, g, \mathbb{G}, \mathbb{G}_{\mathbb{T}}, e, h_1, h_2, A, e(g, g)^x, Enc())$.

When each hospital H_j registers itself with the healthcare center, TA chooses two random numbers $(t_{j1}, t_{j2}) \in \mathbb{Z}_q^*$, and computes the access control key $ak_j = (g^{x+at_{j1}}, g^{t_{j1}}, g^{t_{j2}}, h_1^{t_{j1}} h_2^{t_{j2}})$ for the hospital H_j. The data acquisition process is crucial. For example, it was estimated that almost 80% of the time and effort is spent in cleaning and preparing real-world medical data before the data can be used by disease risk prediction model [6]. Therefore, in the system setting phase of PGuide, a trustworthy data analytic

Table 1
For each disease, we obtain the corresponding regression coefficients, the estimated intercept, the disease risk threshold, and the question set.

Disease	C oefficients	γ	S_{th}	Question set
Y_1	$\{b_{1,1}, b_{2,1}, \cdots, b_{m,1}\}$	γ_1	$S_{\text{th},1}$	$\{q_{1,1}, q_{2,1}, \cdots, q_{m,1}\}$
Y_2	$\{b_{1,2}, b_{2,2}, \cdots, b_{m,2}\}$	γ_2	$S_{\text{th},2}$	$\{q_{1,2}, q_{2,2}, \cdots, q_{m,2}\}$
Y_3	$\{b_{1,3}, b_{2,3}, \cdots, b_{m,3}\}$	γ_3	$S_{\text{th},3}$	$\{q_{1,3}, q_{2,3}, \cdots, q_{m,3}\}$
...				
Y_n	$\{b_{1,n}, b_{2,n}, \cdots, b_{m,n}\}$	γ_n	$S_{\text{th},n}$	$\{q_{1,n}, q_{2,n}, \cdots, q_{m,n}\}$

company first communicates with hospitals to construct the disease risk model, and determines the regression coefficients $B_i = \{b_1, b_2, \cdots, b_m\}$, the estimated intercept γ_i and the disease risk threshold $S_{\text{th},i}$ for each disease Y_i using some data mining methods [15,16]. For example, the disease predictors for Parkinson's Synucleinopathy-associated disease [10] include hyposmia, urinary dysfunction, specific sleep disturbances, depressive symptoms, and constipation.

According to each defined predictor $a_j \in A_i = \{a_1, a_2, \cdots, a_m\}$ of a disease Y_i, a corresponding question $q_j \in Q_i = \{q_1, q_2, \cdots, q_m\}$ is designed. For instance, as urinary dysfunction is an impact predictor of Parkinson's disease, we design the question "*Do you have increased urinary frequency and urgency?*", and the answer of each question q_i is in $a_j \in \{0, 1\}$. Specifically, the result of the setup is shown in Table 1. In addition to the above setting, a smartphone-based PGuide application is also developed for medical users to obtain medical self-diagnosis and hospital recommendation services.

4.2. Privacy-preserving pre-clinical guidance

Pre-Disease Diagnosis. In the system setting, for each disease Y_i, we design a question set ($q_j \in Q_i = \{q_1, q_2, \cdots, q_m\}$), and the answer for each question q_i can be mapped to a binary value $a_i \in \{0, 1\}$. The general procedure of pre-clinical disease diagnosis can be described as follows:

- With the smartphone-based PGuide application, a user $U_l \in \mathbb{U}$ chooses a disease Y_i that the user wishes to diagnose. A corresponding question set $Q_i = \{q_1, q_2, \cdots, q_m\}$ will be shown in the application. After the U_l has answered all these questions Q_i, the answers representing U_l's health profile will be mapped to a binary vector $A_i = \{a_1, a_2, \cdots, a_m\}$ and transmitted to SP.
- Upon receiving A_i, SP will query the database to obtain the weighted coefficients $B_i = \{b_1, b_2, \cdots, b_m\}$, the intercept γ_i and the threshold $S_{\text{th},i}$ for the chosen disease Y_i. Then, SP will compute $A_i \cdot B_i + \gamma_i = \sum_{j=1}^{m} a_j \cdot b_j + \gamma_i \geq S_{th,i}$ and determine whether U_l has the disease Y_i with a high possibility. Finally, the result will be returned to U_l.
- Upon receiving the result, U_l will decide whether a doctor or a specialist needs to be consulted.

Table 2
Parameters setting for PPCP.

PPCP Parameter	$\lvert p\rvert$	$\lvert\beta\rvert$	$\lvert\alpha\rvert$	$\lvert r_i\rvert$	$\lvert q\rvert$	$\lvert m\rvert$
Security Parameter	l_0	l_1	l_2	l_3	l_4	l_5

$|x|$ means the bit length of x.

However, the above general pre-clinical procedure does not ensure the user's privacy. Therefore, in the following, we introduce a privacy-preserving comparison protocol (PPCP) in PGuide to ensure that another user is unable to learn the coefficients $B_i = \{b_1, b_2, \cdots, b_m\}$, the intercept γ_i and the threshold $S_{\text{th},i}$ in the risk model for the disease Y_i, and will not result in the disclosure of the user's health profile information $A_i = \{a_1, a_2, \cdots, a_m\}$ to SP. The main steps of PPCP are summarized as follows, as shown in Fig. 2, where $\vec{a} = A_i = \{a_1, a_2, \cdots, a_m\} \in \mathrm{F}_2^m$ and $\vec{b} = B_i = \{b_1, b_2, \cdots, b_m\} \in \mathrm{F}_q^m$. We remark that in the pre-clinical model, all the original b_i ($b_i > 0$) are weighted coefficients, and each b_i is a small real number. For the efficient computation in PPCP, each b_i is expanded 10,000 times, such that all $\{b_1, b_2, \cdots, b_m\}$ are integer values lying in F_q^m with $q = 2^{16}$. To ensure correctness and security of the proposed PPCP protocol, the security parameters (l_0, l_1, l_2, l_3, l_4, l_5) are choosen first by the user U_l. For the reader's convenience, the relationship between the PPCP parameters to be used and these security parameters are summarized and shown in the Table 2.

The constraints of these security parameters are as follows: $l_1 < l_0$, $3l_2 + l_4 + l_5 < l_1$, $2l_2 + l_1 + l_3 + l_4 + l_5 < l_0 - 1$.

User: $(a1, a2, \cdots, a_m) \in \mathbf{F}_2^m$		SP: $(b1, b2, \cdots, b_m) \in \mathbf{F}_q^m, \gamma, S_{\text{th}}$
1. choose large primes α, β, p such that $\lvert\alpha\rvert = l_2$, $\lvert\beta\rvert = l_1$, $\lvert p\rvert = l_0$, and a large random $s \in \mathbf{Z}_p^*$; 2. **for** $i = 1, 2, \cdots, m$, choose random numbers x_i, y_i, r_i, such that $x_i + y_i = r_i \cdot \beta$; compute $c_i = \alpha \cdot a_i + x_i$, $c_i' = s \cdot y_i \bmod p$ **end for**		
choose a random y_0 such that $\lvert y_0\rvert < \lvert\alpha\rvert$, compute $c'_0 = s \cdot y_0 \bmod p$	$\xrightarrow{\alpha, p, \vec{c}, \vec{c}'}$	3.choose random numbers t_1, t_2, t_3 such that $\lvert t_2\rvert < \lvert\alpha\rvert, \lvert t_1 \cdot t_2\rvert > \lvert\alpha^2\rvert, \lvert t_3\rvert < \lvert\alpha\rvert$, and $\lvert t_1 + t_3 \cdot \alpha\rvert < \lvert\alpha^2\rvert$ **for** $i = 1, 2, \cdots, m$: $D_i = \alpha \cdot b_i \cdot c_i, , D_i' = \alpha \cdot b_i \cdot c_i' \bmod p$ **end for** $D_0' = t_3 \cdot c_0' \bmod p$, $D = t_2 \cdot \left(\sum_{i=1}^{m} D_i + \gamma \cdot \alpha^2 - S_{\text{th}} \cdot \alpha^2 + t_1\right)$
4. $E' = s^{-1} \cdot D' \bmod p$, $E = (D + E') \bmod \beta$ **if** $\lvert E\rvert = \lvert\beta\rvert$, then $\vec{a} \cdot \vec{b} + \gamma < S_{\text{th}}$; **else**: $\vec{a} \cdot \vec{b} + \gamma \geq S_{\text{th}}$	$\xleftarrow{D, D'}$	$D' = t_2 \cdot \left(\sum_{i=1}^{m} D_i' + D_0'\right) \bmod p$

Fig. 2. Description of PPCP protocol.

Step 1: The user U_l chooses three large primes, α, β, p such that $|\alpha| = l_2$, $|\beta| = l_1$, $|p| = l_0$, a random number $s \in \mathbf{Z}_p^*$, and computes $s^{-1} \bmod p$.

Step 2: For each $a_i \in \vec{a}$, three random numbers (x_i, y_i, r_i) are chosen with the constraint $x_i + y_i = r_i \cdot \beta$, $\frac{r_i \cdot \beta}{2} < y_i < r_i \cdot \beta$ and $|r_i| = l_3 U_l$ computes the vectors $\vec{c} = \{c_1, c_2, \cdots, c_m\}, \vec{c}' = \{c_0', c_1', c_2', \cdots, c_m'\}$, where each (c_i, c_i') is

$$\begin{aligned} &c_i = \alpha \cdot a_i + x_i, \quad c_i' = s \cdot y_i \bmod p, \text{ for } i = 1, 2, \cdots, m \\ &c_0' = s \cdot y_0 \bmod p, \text{ where } y_0 < \alpha \text{ is a random number} \end{aligned} \tag{3}$$

and sends $(\alpha, p, \vec{c}_i, \vec{c}_i')$ to SP. Because of the large prime α, the random numbers x_i, y_i and mod p operation, SP is unable to determine whether $a_i \in \vec{a}$ is 1 or 0.

Step 3: After receiving $(\alpha, p, \vec{c}_i, \vec{c}_i')$, SP chooses three random numbers t_1, t_2, t_3 with the constraints $|t_2| < |\alpha|$, $|t_3| < |\alpha|$, $|t_1 \cdot t_2| > |\alpha^2|$, $|t_1 + t_3 \cdot \alpha| < |\alpha^2|$, and computes the vectors $\vec{D}, \vec{D}'$, where

$$\begin{aligned} &D_i = \alpha \cdot b_i \cdot c_i, D_i' = \alpha \cdot b_i \cdot c_i' \bmod p, \text{ for } i = 1, 2, \cdots, m \\ &D_0' = t_3 \cdot c_0' \bmod p \end{aligned} \tag{4}$$

Then, SP computes

$$\begin{aligned} D &= t_2 \cdot \left(\sum_{i=1}^{m} D_i + \gamma \cdot \alpha^2 - S_{\text{th}} \cdot \alpha^2 + t_1 \right) \\ D' &= t_2 \cdot \left(\sum_{i=1}^{m} D_i' + D_0' \right) \end{aligned} \tag{5}$$

SP will now return (D, D') to U_l, without revealing the values $b_i \in \vec{b}_i$ and the threshold of the disease risk S_{th} to U_l.

Step 4: Upon receiving the data (D, D'), U_l first computes

$$E' = s^{-1} \cdot D' \bmod p, \quad E = (D + E') \bmod \beta \tag{6}$$

Finally, U_l can determine the result from the bit length of E. If $|E| = |\beta|$, then U_l can determine $\vec{a} \cdot \vec{b} + \gamma < S_{\text{th}}$. Otherwise, $\vec{a} \cdot \vec{b} + \gamma \geq S_{\text{th}}$.

Correctness. The correctness of PPCP can be illustrated as follows: Given the constraints of the security parameters shown above, we obtain:

$$\begin{aligned} &2l_2 + l_1 + l_3 + l_4 + l_5 < l_0 - 1 \\ \Rightarrow\ &\alpha^2 \cdot m \cdot q \cdot r_i \cdot \beta < p/2 \end{aligned} \tag{7}$$

In addition, the relationship of α and p can be calculated as follows:

$$\begin{aligned} &3l_2 + l_4 + l_5 < l_1, \quad l_1 < l_0 \\ \Rightarrow\ &3l_2 + l_4 + l_5 < l_0 - 1 \Rightarrow \quad 3l_2 < l_0 - 1 \\ \Rightarrow\ &\alpha^3 < p/2 \end{aligned} \tag{8}$$

In step 3, SP receives the data and calculates:

$$\begin{aligned} &D_i = \alpha \cdot b_i \cdot c_i = \alpha \cdot b_i \cdot (\alpha \cdot a_i + x_i) = \alpha^2 \cdot a_i \cdot b_i + \alpha \cdot b_i \cdot x_i \\ &D_i' = \alpha \cdot b_i \cdot c_i' = \alpha \cdot b_i \cdot s \cdot y_i \bmod p, \text{ for } i = 1, 2, \cdots, m \\ &D_0' = t_3 \cdot c_0' = t_3 \cdot s \cdot y_0 \bmod p \end{aligned} \tag{9}$$

Then, SP can compute (D, D'), where

$$\begin{aligned} D &= t_2 \cdot \left(\alpha^2 \cdot \sum_{i=1}^{m} a_i \cdot b_i + \alpha \cdot \sum_{i=1}^{m} b_i \cdot x_i + \alpha^2 \cdot \gamma - \alpha^2 \cdot S_{\text{th}} + t_1 \right) \\ D' &= t_2 \cdot \left(\alpha \cdot \sum_{i=1}^{m} b_i \cdot s \cdot y_i + t_3 \cdot s \cdot y_0 \right) \bmod p \end{aligned} \tag{10}$$

In step 4, U_l removes the factor s from D' by multiplying $s^{-1} \bmod p$:

$$\begin{aligned} E' &= s^{-1} \cdot D' = s^{-1} t_2 \cdot \left(\alpha \cdot \sum_{i=1}^{m} b_i \cdot s \cdot y_i + t_3 \cdot s \cdot y_0 \right) \bmod p \\ &= t_2 \cdot \left(\alpha \cdot \sum_{i=1}^{m} b_i \cdot y_i + t_3 \cdot y_0 \right) \bmod p \\ &\xrightarrow{\because (t_2 \cdot \alpha \cdot \sum_{i=1}^{m} b_i \cdot y_i) < (\alpha^2 \cdot m \cdot q \cdot r_i \cdot \beta) < p/2 \quad Eq.(7)} \\ &\xrightarrow{\because and \quad (t_2 \cdot t_3 \cdot y_0) < \alpha^3 < p/2 \quad Eq.(8)} \\ &= t_2 \cdot \left(\alpha \cdot \sum_{i=1}^{m} b_i \cdot y_i + t_3 \cdot y_0 \right) \end{aligned} \tag{11}$$

In the last calculation, U_l obtains

$$\begin{aligned} E &= D + E' = t_2 \cdot \left(\alpha^2 \cdot \sum_{i=1}^{m} a_i \cdot b_i + \alpha \cdot \sum_{i=1}^{m} b_i \cdot x_i \right. \\ &\quad \left. + \alpha^2 \cdot \gamma - \alpha^2 \cdot S_{\text{th}} + t_1 + \alpha \cdot \sum_{i=1}^{m} b_i \cdot y_i + t_3 \cdot y_0 \right) \bmod \beta \\ &= t_2 \cdot \left[\alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) \right. \\ &\quad \left. + \alpha \cdot \sum_{i=1}^{m} b_i \cdot \beta + t_1 + t_3 \cdot y_0 \right] \bmod \beta \end{aligned} \tag{12}$$

Let $k_{gap} = |\beta| - |\alpha^2| - |q| - |t_2|$, and $k_{gap} > 200$. For example, $|\alpha| = 160$, $|\beta| = 700$, $|p| = 1024$, $|t_1| = 300$, $|t_2| = 100$, $|t_3| = 100$, $|q| = 16$. If $\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \geq 0$, Eq. (12) becomes

$$\begin{aligned} E &= t_2 \cdot \left[\alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) + t_1 + t_3 \cdot y_0 \right] \bmod \beta \\ &\xrightarrow{\because k_{gap} = |\beta| - |\alpha^2| - |q| - |t_2|, k_{gap} > 200,\ t_2 \cdot (t_1 + t_3 \cdot y_0) < \alpha^3 < \beta \text{ and } |\alpha^3| < |\beta|} \\ &= t_2 \cdot \left[\alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) + t_1 + t_3 \cdot y_0 \right] \end{aligned} \tag{13}$$

Because $|t_1 + t_3 \cdot \alpha| < |\alpha^2|$, the length of E is dominated by $t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right)$, that is, $|E| \approx |\alpha^2| + |q| + |t_2| \ll |\beta|$. It is easy to observe that the bit length of E is much less than that of β when $\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} >= 0$.

On the other hand, if $\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} < 0$, Eq. (12) becomes

$$\begin{aligned} E &= t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) + \beta \\ &\quad + \left(t_2 \cdot \alpha \cdot \sum_{i=1}^{m} b_i - 1 \right) \cdot \beta + t_2 \cdot (t_1 + t_3 \cdot y_0) \bmod \beta \\ &= t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) + \beta \\ &\quad + t_2 \cdot (t_1 + t_3 \cdot y_0) \bmod \beta \\ &\xrightarrow{\because t_2 \cdot \alpha^2 \cdot (\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}}) + \beta < \beta,\ t_2 \cdot (t_1 + t_3 \cdot y_0) < \alpha^3 < \beta \text{ and } |\alpha^3| < |\beta|} \\ &= t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) + \beta + t_2 \cdot (t_1 + t_3 \cdot y_0) \end{aligned} \tag{14}$$

Because $t_2 \cdot (t_1 + t_3 \cdot y_0) < \alpha^3 < \beta$, $t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) < 0$ and $|t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right)| \ll |\beta|$, we have the length of E dominated by β, that is, $|E| = |\beta|$.

From the above observations, U_l can distinguish the disease risk from the bit length of E. Thus, the correctness of the PPCP protocol is satisfied.

Recommendation of hospitals. After diagnosing several diseases $\mathcal{Y} = \{y_1, y_2, \ldots, y_w\}$ using the proposed PPCP protocol, the user U_l can choose a disease $y_i \in \mathcal{Y}$ that he/she wishes to seek treatment for. SP will query all hospitals within U_l's district or vicinity (e.g. whether these hospitals have medical doctors available to treat this disease) without SP knowing the specific disease. Then SP forwards the encrypted response from hospitals to U_l. Note that, the number of the hospitals within U_l's vicinity should not be large. It does not take much time for U_l to decrypt these responses and determine which hospital(s) had replied with a "yes". It is trivial to note that the privacy-preserving recommendation protocol helps to ensure that the privacy of the user's disease / treatment sought is preserved. In addition, only authenticated hospitals are recommended to the user.

Step 1: The app on U_l's smartphone first chooses a random number $s \in \mathbb{Z}_q^*$, computes $sk = e(g,g)^{xs}$ and $c = (c_1, c_2, c_3)$ as

$$c_1 = g^s, c_2 = A^s \cdot h_1^{-s}, c_3 = h_2^{-s} \tag{15}$$

Note that, in this scenario, the location privacy preservation for user is currently not considered, as it is not as critical as the user's sensitive health information. Thus, U_l sends $Enc_{sk}(y_i), c = (c_1, c_2, c_3)$ and the current location to SP. SP chooses hospitals within the user's district or vicinity and sends them the encrypted message $(Enc_{sk}(y_i), c_1, c_2, c_3)$.

Step 2: Upon receiving $(Enc_{sk}(y_i), c_1, c_2, c_3)$, each hospital H_j will perform the following steps:

- Uses the access control key $ak_j = (g^{x+at_{j1}}, g^{t_{j1}}, g^{t_{j2}}, h_1^{t_{j1}} \cdot h_2^{t_{j2}})$ to compute

$$\begin{aligned} sk &= \frac{e(c_1, g^{x+at_{j1}})}{e(g^{t_{t1}}, c_2) \cdot e(g^{t_{j2}}, c_3) \cdot e(h_1^{t_{j1}} h_2^{t_{j2}}, c_1)} \\ &= \frac{e(g^s, g^x g^{at_{j1}})}{e(g^{g_{j1}}, g^{as} \cdot h_1^{-s}) \cdot e(g^{t_{j2}}, h_2^{-s}) \cdot e(h_1^{t_{j1}} h_2^{t_{j2}}, g^s)} \\ &= \frac{e(g^s, g^x) e(g^s, g^{at_{j1}})}{e(g^{t_{j1}}, g^{as}) e(g^{t_{j1}}, h_1^{-s}) e(g^{t_{j2}}, h_2^{-s}) e(h_1^{t_{j1}} h_2^{t_{j2}}, g^s)} \\ &= \frac{e(g^s, g^x)}{e(g^s, h_1^{t_{j1}} h_2^{t_{j2}})^{-1} e(h_1^{t_{j1}} h_2^{t_{j2}}, g^s)} = e(g,g)^{xs} \end{aligned} \tag{16}$$

- Computes $y_i = Dec_{sk}(Enc_{sk}(y_i))$. The hospital information system (HIS) in each hospital H_j will return $Enc_{sk}(yes|timestamp)$ to SP if they have the medical doctor available to treat the patient y_i, otherwise they will return $Enc_{sk}(no|timestamp)$
- SP collects the encrypted responses from the participating hospitals and forwards them to U_l. As the number of the hospitals within U_l's vicinity is not large, it will not require U_l to spend much time in decrypting the responses and determining which hospitals had responded with a "yes". Therefore, the hospital recommendation system will reduce the waiting time due to a mismatch or going to a hospital that does not have an appropriate or available medical doctor to treat the patient.

5. Security analysis

In this section, we analyze the security of our proposed PGuide scheme, particularly focusing on how the proposed PPCP protocol can achieve the privacy-preservation of a user's health profile $(A = \vec{a} = \{a_1, a_2, \ldots, a_m\})$ and the SP's IP $(B = \vec{b} = \{b_1, b_2, \ldots, b_m\}, \gamma, S_{\text{th}})$ in the disease diagnosis phase. We will also examine the privacy-preservation of the disease information in the hospital recommendation phase.

• *Security of health profile in user query.* In a user query in PGuide, sensitive health profile information is encrypted by the PPCP scheme. A query consists of two vectors $\vec{c} = (c_1, \ldots, c_m)$, $\vec{c}' = (c_0', c_1', \ldots, c_m')$ and two large prime numbers α, p. As SP is honest-but-curious, it may attempt to recover $\vec{a} = (a_1, \ldots, a_m)$ using exhaustive attacks on $\vec{c}$ and $\vec{c}'$. However, the components $\vec{c}$ and $\vec{c}'$ can be viewed as an equation group of $2m$ equations with $4m+1$ unknowns s, (x_i, y_i, a_i, w_i), for $i = 1, 2, \ldots, m$, as below

$$\begin{cases} c_1 = \alpha \cdot a_1 + x_1 \\ \quad \vdots \\ c_m = \alpha \cdot a_m + x_m \\ c_1' = s \cdot y_1 \bmod p \\ \quad \vdots \\ c_m' = s \cdot y_m \bmod p \end{cases} \Rightarrow \begin{cases} c_1 = \alpha \cdot a_1 + x_1 \\ \quad \vdots \\ c_m = \alpha \cdot a_m + x_m \\ c_1' = s \cdot y_1 + w_1 \cdot p, w_1 \in \mathbf{Z}_{\geq 0} \\ \quad \vdots \\ c_m' = s \cdot y_m + w_m \cdot p, w_m \in \mathbf{Z}_{\geq 0} \end{cases}$$

Because the number of unknowns (i.e. $4m+1$) is more than those in the equations (i.e. $2m$), this equation group is not determined. That is, SP is unable to learn $\vec{a}$ by solving this equation group.

• *Security of SP's disease risk model.* The user is also honest-but-curious, and may seek to recover the coefficients of disease risk model and the threshold by generating and solving an over-determined polynomial equation group. If we do not include the random numbers t_1, t_2, t_3, the user may reveal the disease risk model $(\vec{b} = \{b_1, b_2, \ldots, b_m\}, \gamma, S_{\text{th}})$ as follows.

i) A user may attempt to reveal the disease risk model by attacking E. Without t_1, t_2, t_3, after issuing a query with $\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} > 0$, the user can obtain the encrypted value

$$\begin{aligned} E &= \alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right) + \alpha \cdot \sum_{i=1}^m b_i \cdot \beta \bmod \beta \\ &\xrightarrow{\because |\alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right)| < |\beta|} \\ &= \alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right) \end{aligned}$$

The unknowns in E are the coefficients $\{b_1, b_2, \cdots, b_m\}$, the intercept γ and the threshold S_{th} of the disease model. After issuing k queries with k different $\vec{a}$, a group of k equations and $m+2$ unknowns $\{b_1, b_2, \cdots, b_m\}$, γ, S_{th} can be generated.

$$\begin{cases} E_1 = \alpha^2 \cdot \left(\sum_{i=1}^m a_{i,1} \cdot b_i + \gamma - S_{\text{th}} \right) \\ \quad \vdots \\ E_k = \alpha^2 \cdot \left(\sum_{i=1}^m a_{i,k} \cdot b_i + \gamma - S_{\text{th}} \right) \end{cases}$$

Once k is more than $m+2$, the number of unknowns is less than those in the equations. Consequently, this equation group is over-determined and the disease risk model can be revealed. In order to prevent such an attack, we configure two random numbers t_1 and t_2. For each user query, t_2 is a distinct random number, which increases the number of unknowns linearly with the number of queries. Thus, the equation group is not determined. However, if we only configure t_2 without t_1, a user may reveal $\alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right)$ based on two queries with the same vector $\vec{a}$, i.e., the common divisor of $E_j = t_{2,j} \cdot \alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right)$, for $j = 1, 2$. After obtaining $m+2$ components $\alpha^2 \cdot \left(\sum_{i=1}^m a_i \cdot b_i + \gamma - S_{\text{th}} \right)$, *for* $j = 1, 2, \ldots, m+2$, a user can also reveal the disease risk model as above. Therefore, to deal with this vulnerability, we add the random number t_1. Then, with two queries on the same $\vec{a}$, a user can get

$$E_j = t_{2,j} \cdot \alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) + t_{2,j} \cdot t_{1,j},\ j = 1, 2$$

If $|t_{1j} \cdot t_{2j}| < |\alpha^2|$, then $t_{2,j} \cdot \alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right)$ can be derived from $E_j - E_j \bmod \alpha^2$. Then again, the disease risk model can be revealed using the above attack. Therefore, we need the constraint $|t_1 \cdot t_2| > |\alpha^2|$ to prevent such an attack.

ii) A user may also attempt to reveal the disease risk model by attacking E'. If we configure t_1, t_2 without t_3, in step 4, the user can obtain the value

$$E' = t_2 \cdot \alpha \cdot \left(\sum_{i=1}^{m} b_i \cdot y_i \right) \bmod p$$

Then, the user can reveal $\alpha \cdot (\sum_{i=1}^{m} b_i \cdot y_i)$ from the two queries issued to the same vector $\vec{y} = (y_1, y_2, \ldots, y_m)$ (i.e. the common divisor of $E'_j = t_{2,j} \cdot \alpha \cdot (\sum_{i=1}^{m} b_i \cdot y_i),\ j = 1, 2$).

To prevent such an attack, we configure a random number t_3 on $c'_0 = s_0 \cdot y_0 \mod p$. Then, in step 4, the user obtains

$$E' = t_2 \cdot \alpha \cdot \left(\sum_{i=1}^{m} b_i \cdot y_i \right) + t_2 \cdot t_3 \cdot y_0 \bmod p$$

In addition, the constraint $|t_1 + t_3 \cdot \alpha| < |\alpha^2|$ is necessary, which prevents the random numbers from changing the result in E.

• *Security of disease information and authentication of hospitals in hospital recommendation protocol.* It is easy to see that $e(g, g)^{xs}$ can be recovered only by a registered hospital $H_j \in \mathbb{H}$ with its access key $ak_j = (g^{x+at_{j1}}, g^{t_{j1}}, g^{t_{j2}}, h_1^{t_{j1}} h_2^{t_{j2}})$ from $(c_1 = g^s, c_2 = A^s \cdot h_1^{-s}, c_3 = h_2^{-s})$, and the information about the disease y_i can only be obtained using the appropriate symmetric key $e(g, g)^{xs}$ and $Dec()$. There are many common diseases can be treated in different hospitals [17], and SP only receives encrypted feedback $Enc_{sk}(yes|timestamp)$ or $Enc_{sk}(no|timestamp)$ from the participating hospitals. Therefore, SP will not be able to learn the details of the disease. In addition, only the hospitals who have registered with TA will have the access control key $ak_j = (g^{x+at_{j1}}, g^{t_{j1}}, g^{t_{j2}}, h_1^{t_{j1}} \cdot h_2^{t_{j2}})$ to decrypt the encrypted user request. Such an authentication scheme prevents non-registered and non-vetted hospitals from participating.

Based on the above security analysis, we have shown that our proposed PGuide scheme provides privacy preservation for user health profile, disease and SP's disease risk prediction model, assuming that there is no collusion between hospitals and SP.

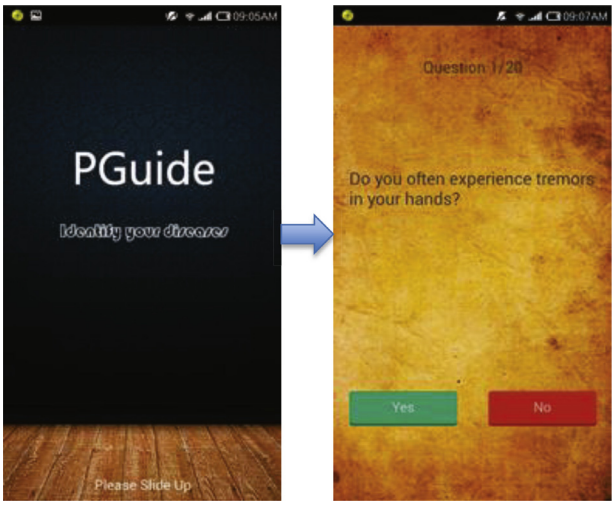


Fig. 3. Prototype: Android app for Pre-disease Diagnosis.

6. Performance evaluation

In this section, we evaluate our proposed PGuide scheme in two stages (i.e. Pre-disease Diagnosis; and Recommendation of Hospitals), in terms of computational cost and communication overhead.

6.1. Pre-disease diagnosis

Experimental setup. We design a PGuide Android app, as shown in Fig. 3, in addition to a server side application and a hospital side application on the Tomcat Apache server 8. To ensure repeatability, the detailed experimental settings are outlined in Table 3. In order to demonstrate the efficiency of PPCP in Pre-disease Diagnosis, we also develop a scheme with the same function of PPCP but built using a typical Paillier encryption with modulus $|n^2| = 2048$ [18] as a baseline comparison. We choose the length of vectors $\vec{a}, \vec{b}$ as $m = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ in the experiments, and run the experiments 100 times. The average results of these 100 experiments are reported below.

Computational cost for Pre-disease Diagnosis Fig. 4(a) and (b) plot the computational costs with varying vector lengths from 10 to 100. From Fig. 4(a), it is clear that our proposed PGuide scheme is significantly faster than the Pailliar-based scheme. As our proposed PGuide scheme does not employ computationally expensive operations, user's cost requirement is low, and significantly lower than the computational cost in the Paillier-based scheme. As shown in Fig. 4(b), the server's computational costs in the Paillier-based scheme are low with the vector length m because all the

Table 3
Experiment setup.

(a) Testbed setting

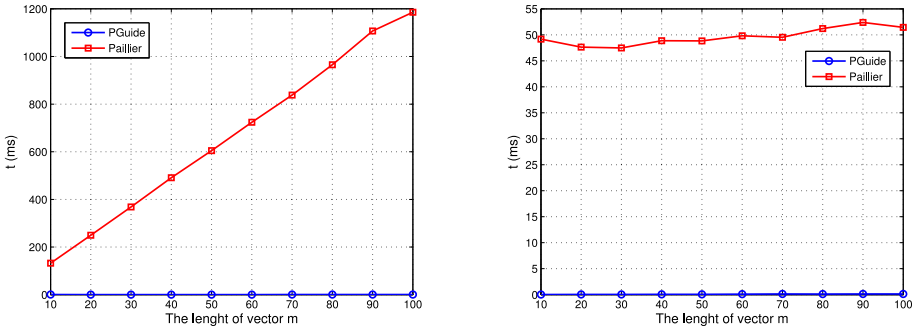
Role	Machine	Hardware & Software
SP	PC	3.1. GHz processor, 8GB RAM and Window 7 platform
Medical user	MI-ONE Plus	Android 4.1.2 system, dual core 1.5 GHz processor, and 1 GB RAM
Hospital	Mac Pro	2.9 GHz processor, 8GB RAM and OSX Yosemite platform

(b) Parameter setting for pre-disease diagnosis

Parameter	$\lvert\alpha\rvert$	$\lvert\beta\rvert$	$\lvert p\rvert$	$\lvert t_1\rvert$	$\lvert t_2\rvert$	$\lvert t_3\rvert$	$\lvert q\rvert$	$\lvert r_i\rvert$
Setting	160	700	1024	300	100	100	16	100

(c) Parameter setting for hospital recommendation

Parameter	Curve	q	$\lvert r\rvert$	$\lvert q\rvert$
Setting	$y^2 = x^3 + x$	$q = 3 \mod 4$	160	512



(a) Comparative Summary: Computational time at user-end with different vector lengths

(b) Comparative Summary: Computational time at server-end with different vector lengths

Fig. 4. Comparative Summary: Computational time, length of ciphertext with different vector lengths in the pre-disease diagnosis process.

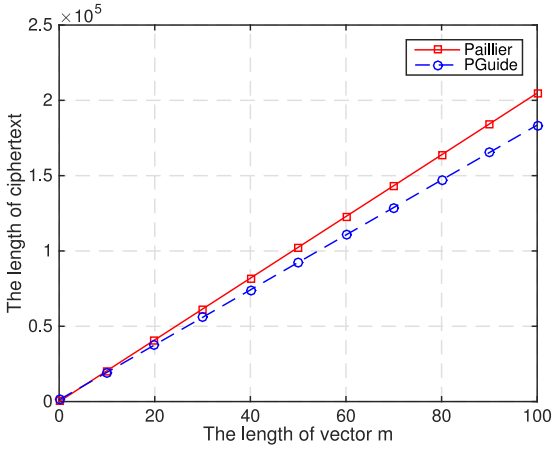


Fig. 5. Comparative Summary: Length of ciphertext with different vector lengths.

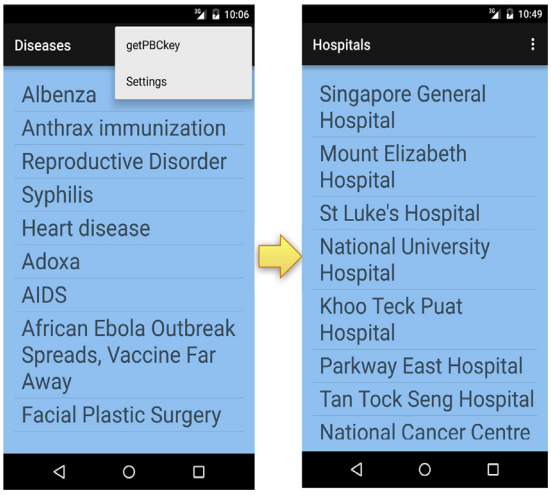


Fig. 6. Hospital recommendation Android app.

disease coefficients b_i are small integers and the exponent operation of ciphertext $Enc(a_i)^{b_i}$ does not require much time. Nonetheless, the computational cost of our proposed PGuide scheme at the server-end is still significantly less than that of the Paillier-based scheme at the server-end, as we do not require any expensive exponent operations.

Communication overhead for Pre-disease Diagnosis Fig. 5 plots the communication overhead with varying vector lengths (i.e. m from 10 to 100). Based on the above parameter settings, the length of $\vec{c}$ is the same as the length of $(\beta + x_i) \cdot m$, which is $(700 + 100)m$. The length of $\vec{c}'$ is at most the same as the length of $m \cdot p$, which is $1024m$. The lengths of α and p are $|\alpha| = 160$ and $|p| = 1024$. Therefore, the length of the ciphertext $(\vec{c}, \vec{c}', \alpha, p)$ in the proposed PGuide scheme is at most $1184 + 1824m$ bits. On the other hand, the length of m ciphertext of Paillier-based scheme is the same as the length of $n^2 \cdot m$ ($2048m$), which is larger than that of our proposed PGuide scheme.

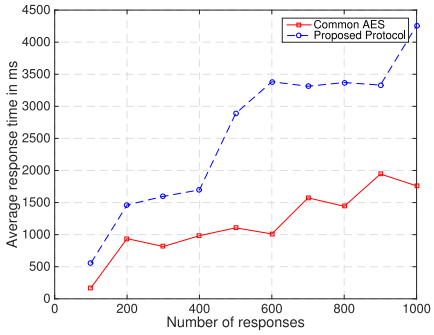
6.2. Recommendation of hospitals

Experimental setup. In our Hospital Recommendation Android app (see Fig. 6, the user can choose a disease in the disease list page as well as obtaining the corresponding recommendation. We adopt JPBC library [19] to implement the underlying pairing algorithm, where the detailed parameters are shown in Table 3(c). In addition, a 128-bit symmetric encryption algorithm is used to encrypt the user request. We used the SHA-256 hash function to hash the pairing element $e(g, g)^{xs}$ to a 256-bit length key, which is truncated to the 128-bit length symmetric encryption key.

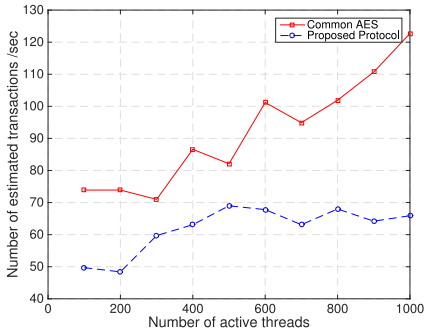
In our experiments, the pairing parameters were generated and stored in a file. Android users store this parameter file in the SD-card, and the hospital-side application stores the file on the local disk. For the user-end, it is not costly to compute the symmetric key sk and $c = (c_1, c_2, c_3)$ because users can conduct a one-off calculation of the keys offline. The time-consuming processes are the computations of the symmetric key sk with $c = (c_1, c_2, c_3)$ and the access control key ak for the hospital-side because of the bilinear pairing calculations. In addition, we developed a common AES encryption as a reference for comparison. For a fair comparison of computation complexity (i.e. response time and throughput), we used Apache JMeter to simultaneously send n user requests. $n = \{100, 200, \ldots, 1000\}$. The results are reported below.

Computational cost for Recommendation Fig. 7(a) and (b) show the average response time and throughput with a varying number of user requests. The average response time of our proposed single-attribute encryption protocol is nearly twice of that of common symmetric encryption-based protocol. Our recommendation protocol is based on AES and Bilinear Pairing encryption. This is more time-consuming than common symmetric encryption based protocol with a given key. The gaps illustrated in Fig. 7(a) and (b) are acceptable for practical applications, while our protocol offers convenience in terms of key management.

Fig. 8(a) shows that the message encryption time for the user-side grows linearly over the number of hospitals for common symmetric encryption based protocol while that of our proposed single-attribute encryption protocol remains constant. This is because the common symmetric encryption based protocol generates

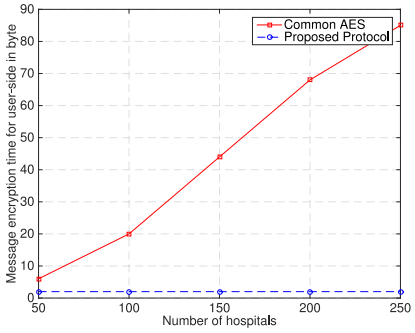


(a) Comparative Summary: Average response time with different number of responses

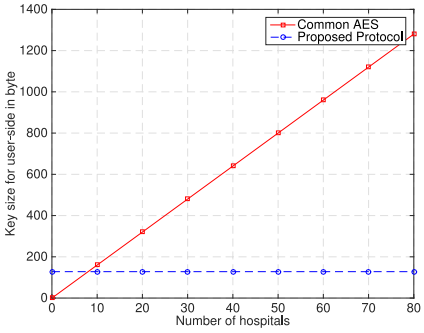


(b) Comparative Summary: Throughput with different number of active threads

Fig. 7. Recommendation of hospitals Android app.



(a) Comparative Summary: Message encryption time for user-side



(b) Comparative Summary: Key sizes for user-side

Fig. 8. Comparative Summary: Key generation time, key sizes for user-side and communication cost with varying number of hospitals.

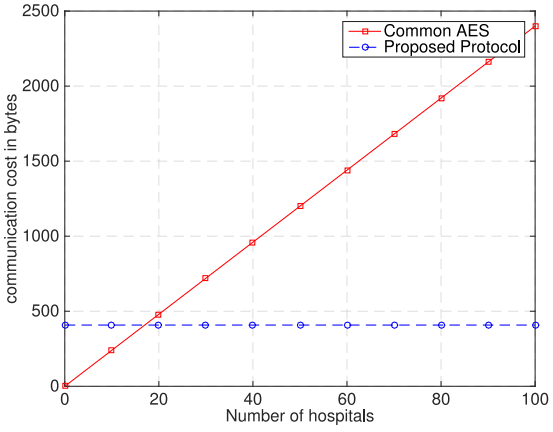


Fig. 9. Comparative Summary: Communication costs.

as many messages as the number of hospitals, while our proposed protocol performs a one-off offline calculation of $c = (c_1, c_2, c_3)$, and encrypts the message online only once in every request. In addition, Fig. 8(b) illustrates that the key size of user-side grows linearly with the number of hospitals for the common symmetric encryption based protocol. The large number of keys could be leaked if users are required to maintain the keys of all hospitals. Our proposed protocol solves this key management issue.

Communication overhead for Recommendation Fig. 9 illustrates the computational overhead comparative summary between our proposed protocol and the common symmetric encryption based protocol. In our protocol, the length of $\mathbb{G}$ is 128 bytes, and we assume that the length of the disease name Y_i is 24 bytes, then the length of the payload for medical user request c_1, c_2, c_3, $Enc_{sk}(Y_i)$ is $128 * 3 + 24$ bytes. On the other hand, for the common symmetric encryption based protocol, the user has to encrypt as many pieces of disease information as the number of hospitals just for one recommendation. Therefore, our proposed protocol is clearly more practical when compared with the common symmetric encryption based protocol, which has the linearly incremental message size.

7. Related work

In this section, we briefly discuss existing literature on disease risk prediction [20–23] and privacy-preserving secure comparison algorithms. As early diagnosis of disease can minimize the side-effects, safety risks, financial costs, etc, *disease risk prediction* has attracted the attention of medical and bioinformatics researchers. For example, in 2012, Anooj et al. [20] develop a fuzzy rule-based decision support system for the prediction of heart disease. In 2013, Bouwmeester et al. [13] use the multivariate logistic regression technique to develop a risk prediction model, in which a linear combination of predictors associated with multiple symptoms and environmental data are used to fit a logarithmic transformation of the probability of the tested disease. This is no doubt a topical research area, particularly in the use of big data analytics and ensuring that the privacy of user and healthcare-related data is preserved.

Based on the Paillier encryption, Ayday et al. [11] introduce a privacy-preserving disease prediction scheme. However, due to its time-consuming exponential operations, the proposed scheme is not efficient in calculating the privacy-preserving comparison results. In our proposed PPCP protocol, however, the protocol does not require any time-consuming operations, and as evident in the

above performance evaluations, the protocol is significantly more efficient, in terms of computational cost and communication overhead.

Katzenbeisser et al. [24] introduce privacy-preserving recommendation systems, which use homomorphic public-key encryption schemes such as Paillier cryptosystm. We have illustrated that the public-key encryption scheme is very time-consuming. Common symmetric encryption based systems (e.g. [25]) are very efficient, but key management becomes a challenging issue. In this work, we address the key management issue as each user only needs to store a key.

In comparison to the above privacy-preserving disease risk prediction models which use time-consuming homomorphic encryption system, our disease risk prediction model is very efficient due to our PPCP protocol. In addition, we modify the original symmetric encryption system in our hospital recommendation protocol to increase the efficiency of key generation and storage, which results in the proposed pre-clinical guide scheme being practical for real-world deployment.

8. Conclusions

In this paper, we have proposed an efficient and privacy-preserving pre-clinical guidance scheme (PGuide). The scheme has two key phases. Firstly, it employs an efficient privacy-preserving comparison protocol (PPCP), which enables a user to obtain disease risk predication services from a service provider without compromising the privacy of the user and the server provider. Secondly, it employs a single-attribute encryption technique to conduct an efficient privacy-preserving hospital recommendation service. Our security analysis demonstrated that the PPCP and hospital recommendation service achieve the privacy-preserving requirements. Evaluations using our Android app prototype demonstrated the efficiency and practicality of real-world deployment of our scheme. Future work will include extending the scheme to cover a wider range of attacks, as well as collaborating with a hospital to roll out the scheme.

References

[1] Mattio R. Shortest average wait time for doctors in major cities increased one minute year over year. 2014. http://www.reuters.com/article/2014/03/26/ny-vitals-idUSnBw265955a+100+BSW20140326.
[2] Linton M., Agency Q. Doctor's office waiting times increasingly frustrating. 2011. http://www.torontosun.com/2011/08/17/doctors-office-waiting-times-increasingly-frustrating.
[3] Huang E. It isn't getting any easier to get a doctor's appointment in china. 2013. http://www.theatlantic.com/china/archive/2013/05/it-isnt-getting-any-easier-to-get-a-doctors-appointment-in-china/276400/.
[4] Jessie. Waiting all night outside a hospital hoping to see a doctor. 2009. http://www.chinasmack.com/2009/pictures/chinese-waiting-hospital-crowds.html.
[5] Pardhu S. List of different types of doctors and what they do. 2009. http://mynamein.wordpress.com/2011/03/28/list-of-different-types-of-doctors-and-what-they-do/.
[6] Cios KJ, William Moore G. Uniqueness of medical data mining. Artif Intell Med 2002;26(1):1–24.
[7] Zhang D. Big data security and privacy protection. In: 8th International Conference on Management and Computer Science (ICMCS 2018). Atlantis Press; 2018.
[8] Dimitrov DV. Medical internet of things and big data in healthcare. Healthc Inform Res 2016;22(3):156–63.
[9] Wang G, Lu R, Huang C. Pguide: An efficient and privacy-preserving smartphone-based pre-clinical guidance scheme. IEEE Globecom'15, San Diego, CA, USA, December 6, - 10; 2015.
[10] Winkler J, Ehret R, Büttner T, Dillmann U, Fogel W, Sabolek M, et al. Parkinson disease risk score: moving to a premotor diagnosis. J Neurol 2011;258(2):311–15.
[11] Ayday E, Raisaro JL, McLaren PJ, Fellay J, Hubaux J. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. 2013 USENIX Workshop on Health Information Technologies, HealthTech '13, Washington, D.C, August 12, 2013; 2013.
[12] Zhang F, Safavi-Naini R, Susilo W. An efficient signature scheme from bilinear pairings and its applications. In: Public Key Cryptography–PKC 2004. Springer Berlin Heidelberg; 2004. p. 277–90.
[13] Bouwmeester W, Twisk JW, Kappen TH, Klei WA, Moons KG, Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC Med Res Methodol 2013;13(1):19.
[14] Daemen J, Rijmen V. The design of Rijndael: AES-the advanced encryption standard. Springer Science & Business Media; 2013.
[15] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34(2):113–27.
[16] Brameier M, Banzhaf W. A comparison of linear genetic programming and neural networks in medical data mining. IEEE Trans Evol Comput 2001;5(1):17–26.
[17] Nate. Top 10 most common diseases found in hospitals. 2013. http://www.nursingschoolhub.com/top-10-most-common-diseases-found-in-hospitals/.
[18] Lu R, Lin X, Shen XS. SPOC: A secure and privacy-preserving opportunistic computing framework for mobile-healthcare emergency. IEEE Trans Parallel Distrib Syst 2013;24(3):614–24.
[19] De Caro A, Iovino V. jpbc: Java pairing based cryptography. In: Proceedings of the 16th IEEE Symposium on Computers and Communications, ISCC 2011. Kerkyra, Corfu, Greece, June 28, - July 1: IEEE; 2011. p. 850–5.
[20] Anooj P. Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. J King Saud Univ-ComputInf Sci 2012;24(1):27–40.
[21] Liu X, Lu R, Ma J, Chen L, Qin B. Privacy-preserving patient-centric clinical decision support system on naive bayesian classification. J Biomed Health Inform to appear.
[22] Yigzaw KY, Bellika JG. A communicable disease prediction benchmarking platform. In: IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2014. IEEE; 2014. p. 564–8.
[23] Sheng J, Li F, Wong S. Optimal drug prediction from personal genomics profiles. J Biomed Health Inform 2015;19(4):1264–70.
[24] Katzenbeisser S, Petković M. Privacy-preserving recommendation systems for consumer healthcare services. In: Third International Conference on Availability, Reliability and Security, 2008. ARES 08. IEEE; 2008. p. 889–895.
[25] Elminaam DSA, Abdual-Kader HM, Hadhoud MM. Evaluating the performance of symmetric encryption algorithms. Int J Netw Secur 2010;10(3):216–222.