PGuide: An Efficient and Privacy-Preserving Smartphone-Based Pre-Clinical Guidance Scheme

Guoming Wang, Rongxing Lu, and Cheng Huang
School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798
Email: wang0947@e.ntu.edu.sg; rxlu@ntu.edu.sg; huangcheng@ntu.edu.sg

***Abstract*—With the pervasiveness of smartphones, mobile e-Healthcare has attracted considerable attention in recent years. Disease risk prediction, as it can assist in predicting user's disease with big data analytics techniques, has become one of important topics in the field of e-Healthcare. However, if the privacy issue is not well addressed, disease risk predication cannot step into its flourish. Aiming at addressing this challenge, in this paper, we propose a new efficient and privacy-preserving pre-clinical guidance scheme, called PGuide, which offers self-diagnosis service to medical users in a privacy-preserving way. In specific, to motivate medical users to provide more detailed health profile for accurate disease risk prediction, we introduce a privacy-preserving comparison protocol PPCP in the PGuide scheme. As a result, with enough health profile information offered by the medical users, the accuracy of disease risk prediction can be improved. Detailed security analysis shows that our proposed PGuide scheme ensures the privacy-preservation for both medical users and service provider. In addition, the performance evaluation via extensive experiments also demonstrates that our proposed PPCP protocol is much efficient in terms of low computational cost and communication overhead.**

***Keywords:* e-Healthcare, pre-clinical guidance, privacy preserving comparison protocol**

I. Introduction

The most prominent problem in hospitals these days is its rareness of doctors relative to a large number of patients. Generally, a patient has to wait for more than 20 minutes in the USA [1] and more than one hour or much longer in China [2]. There are approximately 60 different types of doctors and specialists [3], so it seems that there is nothing more frustrating or upsetting to a patient for seeing a wrong doctor after waiting for a long time. However, most of the patients are not familiar with medical departments in hospitals and corresponding diseases. A large part of patients see doctors who are not specialized on their diseases, and this kind of mismatch makes the scarcity of the medical resource more serious.

As pervasiveness of the smartphones, many efforts have been devoted to convenient usages of the big data analytics results in combination with mobility. However, little progress has been made for commercial integration of processing clinical analysis and offering valuable functionality, because of its extremely sensitive medical information about individuals [4]. When privacy issues become a major concern of medical users while the sensitive medical data have to been exploited by the third-party companies other than the authorized parties like hospitals or government institutes, medical users are reluctant to share their data. Hence, we can imagine that, if the security and privacy challenges are not well addressed, medical users will not offer more detailed health information, which cannot ensure high accuracy of disease risk predication. In addition to the privacy requirements from medical users, information leakage is also a big concern for the service provider, as different disease risk predication models for different diseases are intellectual properties of service provider, which should be not allowed to be stolen and/or abused by others.

In order to address the above privacy challenges and improve the accuracy of disease risk prediction, we propose an efficient privacy-preserving pre-clinical guidance service scheme, called PGuide, for providing on-the-go medical guidance service, while preserving the health privacy for medical users and the intellectual properties of service provider. Specifically, the main contributions of this paper are threefold.

Firstly, we propose PGuide, a privacy-preserving pre-clinical guidance scheme. With PGuide, medical users can achieve privacy-preserving pre-clinical diagnosis by themselves based on their health profile [5]. In addition, it also enables the service provider to calculate the disease risk with disease prediction model in a privacy-preserving way.

Secondly, with detailed security analysis, we show that our proposed PGuide scheme can achieve the privacy-preservation for both the individual user and the service provider.

Finally, to validate the effectiveness of our proposed PGuide scheme, we also develop an Android user-end application on smartphone and a service application in Java for our experiments, and evaluation results via extensive experiments show that our proposed PGuide scheme is much efficient in terms of low computational cost and communication overhead.

The remainder of this paper is organized as follows. In Section II, we introduce our system model, security model and design goal. In Section III, we introduce some preliminaries for our scheme. And in Section IV, we present our PGuide scheme, followed by its security analysis and performance evaluation in Section V and Section VI respectively. We also discuss the related works in Section VII. Finally, we draw our conclusions in Section VIII.

II. Models and Design Goal

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

A. System Model

In our system model, we focus on the disease risk calculation for medical users with the help of service provider. In

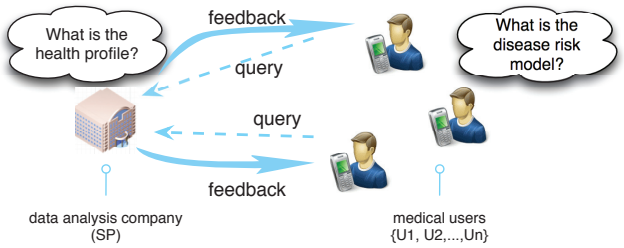


Fig. 1. System model under consideration

such a way, our system model mainly includes two entities: a service provider (SP) and a number of medical users $\mathbb{U} = \{U_1, U_2, \cdots, U_N\}$, as shown in Fig. 1, where N indicates the number of medical users.

- Service Provider (SP): SP is the core entity, which is responsible for information processing, building disease risk prediction model and providing disease risk calculation service to medical users with users' health profile.
- Medical Users $\mathbb{U} = \{U_1, U_2, \cdots, U_N\}$: Each user $U_i \in \mathbb{U}$ is equipped with a smartphone, on which our user-end application is installed. The smartphone collects the user's health profile, sends them to SP and receives the prediction result from SP.

B. Security Model

In our security model, we consider both the SP and the $\mathbb{U} = \{U_1, U_2, \cdots, U_N\}$ are *honest-but-curious*. That is, SP will faithfully follow the disease risk diagnosis protocol, but also attempt to know each individual user's sensitive health profile data once the condition is satisfied. In addition, medical users $\mathbb{U} = \{U_1, U_2, \cdots, U_N\}$ are also *honest-but-curious*, i.e., each U_i will not report false data, but also attempt to know the service provider's disease risk predication model, which is regarded as intellectual properties of SP.

Note that there are possible other attacks, i.e., forgery attack, in disease risk calculation system. Since our focus is on privacy-preserving disease risk prediction, those attacks are currently beyond the scope of this work, and will be discussed in future work.

C. Design Goal

Our design goal is to develop a secure and privacy-preserving pre-clinical guidance scheme to provide disease risk prediction. With our scheme, a medical user can obtain a disease risk predication services from service provider while without leaking out both the user and server provider's privacy.

III. Preliminaries

In this section, we recall the disease risk model [6] and identify the disease risk threshold (S_{th}) under the model, which will serve as the basis of our PGuide scheme.

A. Model of the Disease Risk

Many diagnostic prediction models, which combine patient characteristics and environmental data to predict the presence or absence of a certain diagnosis, have been well developed [7]. 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 $A_i = \{a_1, a_2, \cdots, a_m\}$, with each predictor value $a_j \in \{0, 1\}$ for $j = 1, 2, \cdots, m$, is generally represented in terms of regression coefficients $B_i = \{b_1, b_2, \cdots, b_m\}$ of the same length m. In this way, the predicted risk of the disease Y_i with regard to the symptom A_i can be calculated as:

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

where γ is an estimated intercept in the model. This model has been widely used by medicine and clinician field for disease risk test [7]. 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

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

For a more comprehensive description of logistic regression model, the reader can refer to [7].

B. 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 large volume of real medical data. In order to judge whether a medical 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 high probability, we can set a disease risk threshold S_{th}. If $\gamma + \sum_{j=1}^{m} a_j \cdot b_j \geq S_{\text{th}}$, we can estimate the medical user U_l has the disease Y_i with high probability. Otherwise, when $\gamma + \sum_{j=1}^{m} a_j \cdot b_j < S_{\text{th}}$, we suggest U_l has the disease Y_i with 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 privacy-preserving. In the next section, we will present our PGuide scheme, which can utilize the disease risk model in an efficient and privacy-preserving way to achieve pre-clinical guidance for medical user.

IV. Proposed PGuide Scheme

In this section, we propose our PGuide scheme, which mainly consists of two phases: system setting and privacy-preserving pre-clinical guidance, together with its correctness analysis.

A. System Setting

The data understanding and data collection from the hospitals are very important steps for medical data mining in PGuide. It was estimated that almost 80% of the time and effort is spent in cleaning and preparing the real medical data for the disease risk prediction model [4]. Therefore, in the system setting phase of PGuide, a data analysis company, working as a service provider (SP), first communicates with hospitals to construct the disease risk model, determining 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 with some data mining methods [8], [9]. For example, the disease predictors for the Parkinson's Synucleinopathy-associated Disease [5] include: Hyposmia, Urinary dysfunction, Specific sleep disturbances, Depressive symptoms, and Constipation, etc.

According to each defined predictor $a_j \in A_i = \{a_1, a_2, \cdots, a_m\}$ of 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 the Parkinson's disease, we design the question "*Do you have increased urinary frequency and urgency ?*" accordingly, and the answer of each question q_i is supposed to be $a_j \in \{0, 1\}$. Specifically, the result of the setting up is shown in Table I. In addition to the above setting, a smartphone-based PGuide application is also developed for medical users to get medical self-diagnosis service.

TABLE I
FOR EACH DISEASE, WE GET THE CORRESPONDING REGRESSION COEFFICIENTS, THE ESTIMATED INTERCEPT, THE DISEASE RISK THRESHOLD AND THE QUESTION SET

Disease	coefficients	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	$S_{\text{th},n}$	$\{q_{1,n}, q_{2,n}, \cdots, q_{m,n}\}$

B. Privacy-preserving Pre-clinical Guidance

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

- With the smartphone-based PGuide application, a medical user $U_l \in \mathbb{U}$ chooses a disease Y_i that he or she wants to diagnose. A corresponding question set $Q_i = \{q_1, q_2, \cdots, q_m\}$ will be shown in the application. After the medical user U_l answers all these questions Q_i one by one, the answers representing the user's health profile will be mapped into a binary vector $A_i = \{a_1, a_2, \cdots, a_m\}$ and transmitted to the service provider (SP).
- After receiving A_i, SP first queries 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 judges $A_i \cdot B_i + \gamma_i = \sum_{j=1}^{m} a_j \cdot b_j + \gamma_i \geq S_{th,i}$ to decide whether U_l has the disease Y_i with high possibility. Finally, the result will be sent back to U_l.
- Upon receiving the result, the medical user U_l can decide whether he or she need to see this type of doctors or specialist.

However, the above general pre-clinical procedure does not address privacy issues. Therefore, in the following, we introduce a privacy-preserving comparison protocol (PPCP) in PGuide to ensure that a medical user cannot get to know 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 also not disclose his or her health profile information $A_i = \{a_1, a_2, \cdots, a_m\}$ to the service provider. The main steps of PPCP, as shown in Fig. 2, are summarized as follows, where $\vec{a} = A_i = \{a_1, a_2, \cdots, a_m\} \in \mathrm{F}_2^m$ and $\vec{b} = B_i = \{b_1, b_2, \cdots, b_m\} \in \mathrm{F}_q^m$. Noticing 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. Here for the efficient computation in PPCP, each b_i is expanded with $10,000$ times such that all $\{b_1, b_2, \cdots, b_m\}$ are integer values lying in F_q^m with $q = 2^{16}$.

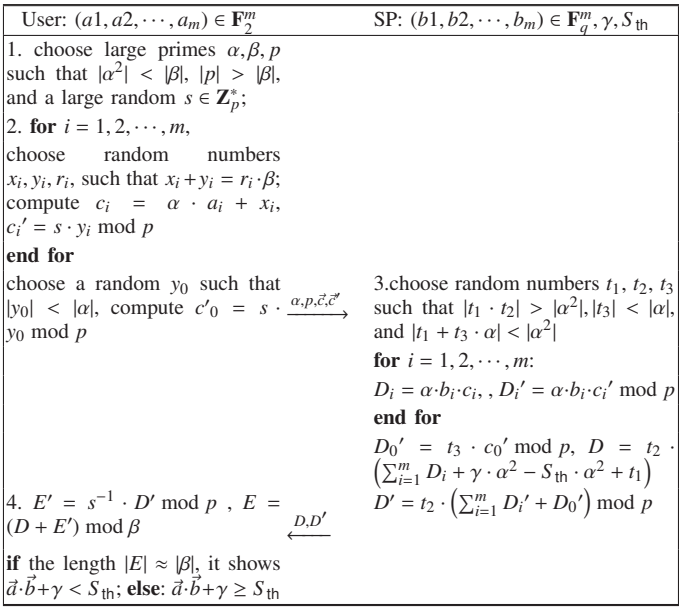


Fig. 2. The description of PPCP protocol

Step 1: The user U_l chooses three large primes, α, β, p such that $|\alpha^2| < |\beta|$, $|p| > |\beta|$, 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 $\alpha^2 \cdot q \cdot r_i \cdot \beta < p$. The user 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') are

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

and sends $(\alpha, p, \vec{c}_i, \vec{c}_i')$ to the service provider SP. Because of the large prime α, the random numbers x_i, y_i and mod$p$ operation, SP can not determine each $a_i \in \vec{a}$ 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

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

In the end, SP returns (D, D') back to the user U_l. In this way, the values $b_i \in \vec{b}_i$ and the threshold of the disease risk S_{th} are hidden to the user U_l obviously.

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 the length of E is close to $|\beta|$, i.e., $|E| \approx |\beta|$, U_l can judge $\vec{d} \cdot \vec{b} + \gamma < S_{\text{th}}$. Otherwise, $\vec{d} \cdot \vec{b} + \gamma \geq S_{\text{th}}$.

Correctness. The correctness of PPCP can be illustrated as follows: In the step 3, SP receives the data and calculates:

$$\begin{aligned} D_i &= \alpha \cdot b_i \cdot c_i = \alpha \cdot b_i \cdot (\alpha \cdot a_i + x_i) = \alpha^2 \cdot a_i \cdot b_i + \alpha \cdot b_i \cdot x_i \\ D_i' &= \alpha \cdot b_i \cdot c_i' = \alpha \cdot b_i \cdot s \cdot y_i \bmod p, \text{ for } i = 1, 2, \cdots, m \\ D_0' &= t_3 \cdot c_0' = t_3 \cdot s \cdot y_0 \bmod p \end{aligned} \tag{7}$$

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

$$\begin{aligned} D &= t_2 \cdot \left(\alpha^2 \cdot \sum_{i=1}^{m} a_i \cdot b_i + \alpha \cdot \sum_{i=1}^{m} b_i \cdot x_i + \alpha^2 \cdot \gamma - \alpha^2 \cdot S_{\text{th}} + t_1 \right) \\ D' &= t_2 \cdot \left(\alpha \cdot \sum_{i=1}^{m} b_i \cdot s \cdot y_i + t_3 \cdot s \cdot y_0 \right) \bmod p \end{aligned} \tag{8}$$

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

$$\begin{aligned} E' &= s^{-1} \cdot D' = s^{-1} t_2 \cdot \left(\alpha \cdot \sum_{i=1}^{m} b_i \cdot s \cdot y_i + t_3 \cdot s \cdot y_0 \right) \bmod p \\ &= t_2 \cdot \left(\alpha \cdot \sum_{i=1}^{m} b_i \cdot y_i + t_3 \cdot y_0 \right) \bmod p \\ &\xrightarrow{\because |t_2 \cdot \alpha \cdot \sum_{i=1}^{m} b_i \cdot y_i| < |\alpha^2 \cdot q \cdot r_i \cdot \beta| < |p|, \quad |t_2 \cdot t_3 \cdot y_0| < |\alpha^3| < |p|} \\ &= t_2 \cdot \left(\alpha \cdot \sum_{i=1}^{m} b_i \cdot y_i + t_3 \cdot y_0 \right) \end{aligned} \tag{9}$$

In the last calculation, U_l obtains

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

Let $k_{gap} = |\beta| - |\alpha^2| - |q| - |t_2|$, and $k_{gap} > 200$. For example, $|\alpha| = 160$, $|\beta| = 700$, $|p| = 1024$, $|t_1| = 300$, $|t_2| = 100$, $|t_3| = 100$, $|q| = 16$. Then, if $\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \geq 0$, Eq. (10) becomes

$$\begin{aligned} E =& t_2 \cdot \left[\alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) + t_1 + t_3 \cdot y_0 \right] \bmod \beta \\ & \xrightarrow{\because k_{gap} = |\beta| - |\alpha^2| - |q| - |t_2|, k_{gap} > 200,\ t_2 \cdot (t_1 + t_3 \cdot y_0) < \alpha^3 < \beta \text{ and } |\alpha^3| < |\beta|} \\ =& t_2 \cdot \left[\alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) + t_1 + t_3 \cdot y_0 \right] \end{aligned} \tag{11}$$

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. (10) becomes

$$\begin{aligned} E =& t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) + \beta \\ & + \left(t_2 \cdot \alpha \cdot \sum_{i=1}^{m} b_i - 1 \right) \cdot \beta + t_2 \cdot (t_1 + t_3 \cdot y_0) \bmod \beta \\ =& t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) + \beta \\ & + t_2 \cdot (t_1 + t_3 \cdot y_0) \bmod \beta \\ & \xrightarrow{\because t_2 \cdot \alpha^2 \cdot (\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}}) + \beta < \beta,\ t_2 \cdot (t_1 + t_3 \cdot y_0) < \alpha^3 < \beta \text{ and } |\alpha^3| < |\beta|} \\ =& t_2 \cdot \alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right) + \beta + t_2 \cdot (t_1 + t_3 \cdot y_0) \end{aligned} \tag{12}$$

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 is dominated by β, that is, $|E| \approx |\beta|$.

From the above observations, user U_l can distinguish the disease risk from the bit length of E. As a result, the correctness of the PPCP protocol is satisifed.

V. Security Analysis

In this section, we analyze the security of our proposed PGuide scheme. Especially, we focus on how the proposed PPCP protocol can achieve the privacy-preservation of user's health profile ($A = \vec{a} = \{a_1, a_2, \cdots, a_m\}$) and service provider's intellectual property ($B = \vec{b} = \{b_1, b_2, \cdots, b_m\}, \gamma, S_{\text{th}}$).

• *Security of health profile in user query.* In PGuide, when a user queries, the sensitive health profile information is encrypted in PPCP scheme concretely. A query consists of two vectors $\vec{c} = (c_1, ..., c_m)$, $\vec{c}' = (c'_0, c'_1, ..., c'_m)$ and two large prime numbers α, p. As service provider is curious, he may attempt to recover $\vec{a} = (a_1, ..., a_m)$ through 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 $i = 1, 2, \cdots, m$, 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 that of equations, i.e., $2m$, this equation group is not determined. That is, the service provider can not reveal $\vec{a}$ through solving this equation group.

• *Security of the disease risk model of service provider.* Alternatively, user is also curious, he or she may attempt 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 random numbers t_1, t_2, t_3, user may reveal the disease risk model ($\vec{b} = \{b_1, b_2, \cdots, b_m\}, \gamma, S_{\text{th}}$) as follows.

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

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

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 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.

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

Once k is more than $m + 2$, the number of unknowns is less than that of equations. As a result, this equation group is over-determined and the disease risk model can be revealed. In order to prevent this attack, we configure two random numbers t_1, t_2. For each user query, t_2 is a distinct random number, which keeps the number of unknowns increases linearly with the number of queries. Thus, the equation group is not determined. However, if we only configure t_2 without t_1, user may reveal $\alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right)$ from 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 $\alpha^2 \cdot \left(\sum_{i=1}^{m} a_i \cdot b_i + \gamma - S_{\text{th}} \right)$, $for\ j = 1, 2, \cdots, m+2$, 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 $\vec{a}$, 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 $|t_{1,j} \cdot t_{2,j}| < |\alpha^2|$, 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, disease risk model can be revealed by the above attack. Therefore, we need the constraint $|t_1 \cdot t_2| > |\alpha^2|$ to prevent this attack.

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

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

Then, user may reveal $\alpha \cdot (\sum_{i=1}^{m} b_i \cdot y_i)$ from two queries with the same vector $\vec{y} = (y_1, y_2, \cdots, y_m)$, 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 this attack, we configure a random number t_3 on $c_0' = s_0 \cdot y_0 \bmod p$. Then, in step 4, user obtains

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

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

Based on the above security analysis, we ensure that the privacy preservation can be achieved in our proposed PGuide scheme.

VI. Performance Evaluation

In this section, we evaluate our proposed PGuide scheme in terms of computational cost and communication overhead.

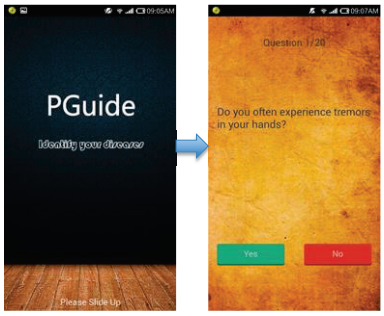
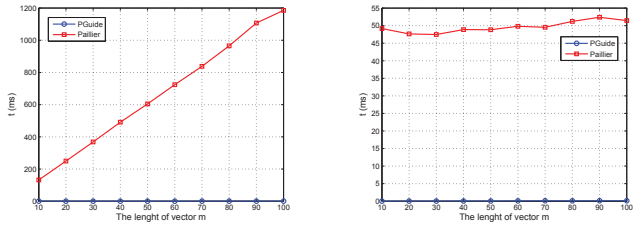


Fig. 3. Android application for PGuide Scheme

In specific, in our experiments, we develop our PGuide Android application, as shown in Fig. 3, and run it on a smartphone with android 4.1.2 system, dual core 1.5 GHz processor, and 1 GB RAM, together with a server-side application run on a Laptop with 3.1. GHz processor, 8GB RAM, and Window 7 platform. The detailed parameter settings are as follows: $|\alpha| = 160$, $|\beta| = 700$, $|p| = 1024$, $|t_1| = 300$, $|t_2| = 100$, $|t_3| = 100$, $|q| = 16$, and $|r_i| = 100$. At the same time, we also develop a scheme with the same function but built upon

typical Paillier encryption with modulus $|n^2| = 2048$ [10] as a reference for comparison. For the clear comparison of computation complexity, we choose the length of vectors $\vec{a}, \vec{b}$ as $m = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ in experiments. We run our experiments 100 times, and the average results are reported below.

Computation cost. Fig. 4 plots the computational costs varies with the length of vector m from 10 to 100, with increasement of 10 on both user-end and server-side. From the figure, it is easy to see our proposed PGuide scheme is much faster, around 200 times, than the Pailliar-based scheme. As our proposed PGuide scheme does not employ time-consuming operations, the computational cost grows slowly, while the computational cost of Pailliar-based scheme, due to the involvement of time-consuming exponential operations, climbs significantly with the vector length m, where m is the length of vector.



(a) User side (b) Server side

Fig. 4. Computational costs of user-end and server-side varying with m

Communication overhead. Based on the above parameter setting, the length of the ciphertext $(\vec{c}, \vec{c}', \alpha, p)$ in the proposed PGuide scheme is at most $1184 + 1824m$ bits, which is less than $2048m$ bits in Pailliar-based scheme.

VII. Related Works

In this section, we briefly discuss some related works on disease risk prediction and privacy-preserving secure comparison algorithms. As early diagnosis of disease can minimize the side-effects, safety risks and the financial costs, *disease risk prediction* has attracted considerable attention. For example, in 2012, Anooj et al. [11] develop a fuzzy rule-based decision support system for prediction of heart disease. In 2013, Bouwmeester et al. [7] use the multivariate logistic regression technique for developing the 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. Actually, to flourish the medical industry, more and more data analysis companies are encouraged to mine new knowledge for efficient medical service.

Based on the Paillier encryption, Ayday et al. [6] introduce a privacy-preserving disease prediction scheme. However, due to its time-consuming exponential operations, it is not quite efficient in calculating the privacy-preserving comparison result. Different from the above works, our proposed PPCP protocol does not require time-consuming operations, and as shown in the above performance evaluation part, it is much efficient in terms of computational cost and communication overhead.

VIII. Conclusions

In this paper, we have proposed an efficient and privacy-preserving pre-clinical guidance scheme, called PGuide. The proposed PGuide is characterized by employing an efficient privacy-preserving comparison protocol (PPCP), which enables a medical user to obtain a disease risk predication services from service provider without leaking both the user and server provider's privacy out. Detailed security analysis shows the PPCP really achieves the privacy requirements in PGuide. In addition, extensive experiments are also conducted, and the results demonstrate the efficiency of the PPCP protocol, which subsequently shows the practicality of our proposed PGuide scheme. In future work, we aim at dealing with other practical security and privacy issues in disease risk predication model.

Availability

The user-end android application and server-side java jar package for the proposed PGuide scheme can be downloaded at `http://www.ntu.edu.sg/home/rxlu/project/`.

References

[1] R. Mattio. Shortest average wait time for doctors in major cities increased one minute year over year. [Online]. Available: http://www.reuters.com/article/2014/03/26/ny-vitals-idUSnBw265955a+100+BSW20140326

[2] jessie. Waiting all night outside a hospital hoping to see a doctor. [Online]. Available: http://www.chinasmack.com/2009/pictures/chinese-waiting-hospital-crowds.html

[3] ——. List of different types of doctors and what they do. [Online]. Available: http://mynamein.wordpress.com/2011/03/28/list-of-different-types-of-doctors-and-what-they-do/

[4] K. J. Cios and G. William Moore, "Uniqueness of medical data mining," *Artificial intelligence in medicine*, vol. 26, no. 1, pp. 1–24, 2002.

[5] J. Winkler, R. Ehret, T. Büttner, U. Dillmann, W. Fogel, M. Sabolek, J. Winkelmann, and J. Kassubek, "Parkinsons disease risk score: moving to a premotor diagnosis," *Journal of neurology*, vol. 258, no. 2, pp. 311–315, 2011.

[6] E. Ayday, J. L. Raisaro, P. J. McLaren, J. Fellay, and J. Hubaux, "Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data," in *2013 USENIX Workshop on Health Information Technologies, HealthTech '13, Washington, D.C., August 12, 2013*, 2013.

[7] W. Bouwmeester, J. W. Twisk, T. H. Kappen, W. A. Klei, K. G. Moons, and Y. Vergouwe, "Prediction models for clustered data: comparison of a random intercept and standard regression model," *BMC medical research methodology*, vol. 13, no. 1, p. 19, 2013.

[8] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial intelligence in medicine*, vol. 34, no. 2, pp. 113–127, 2005.

[9] M. Brameier and W. Banzhaf, "A comparison of linear genetic programming and neural networks in medical data mining," *Evolutionary Computation, IEEE Transactions on*, vol. 5, no. 1, pp. 17–26, 2001.

[10] R. Lu, X. Lin, and X. Shen, "Spoc: A secure and privacy-preserving opportunistic computing framework for mobile-healthcare emergency," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 3, pp. 614–624, 2013.

[11] P. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," *Journal of King Saud University-Computer and Information Sciences*, vol. 24, no. 1, pp. 27–40, 2012.