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

Erman Ayday†, Jean Louis Raisaro†, Paul J. McLaren††, Jacques Fellay††, and Jean-Pierre Hubaux†

†School of Comp. and Comm. Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
††School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

***Abstract*—According to many scientists and clinicians, genomics is the "next big thing" in the field of medicine. On one hand, decreasing costs in genome sequencing has been paving the way to better preventive and personalized medicine. On the other hand, genomic data also raises serious privacy concerns, as it is the ultimate identifier of an individual and it contains privacy-sensitive data (e.g., disease predispositions, ancestry information). Thus, it is necessary to find ways of using genomic data without abusing the genomic privacy of individuals. To get a more comprehensive medical assessment, genomic information must be combined with other clinical and environmental data (such as demographic information, family history, disease history, laboratory test results, etc.) that are also privacy-sensitive (e.g., HIV status of an individual) and need to be treated as such. Focusing on disease risk tests, in this paper, we propose a privacy-preserving system for storing and processing genomic, clinical, and environmental data by using homomorphic encryption and privacy-preserving integer comparison. We implement the proposed system using real patient data and reliable disease risk factors. In particular, we use 23 genetic and 14 clinical and environmental risk factors to compute the risk of coronary artery disease in a privacy-preserving way. Finally, we show the practicality of the proposed system via a complexity evaluation.**

I. Introduction

The field of medicine is in the middle of a radical upheaval: As a result of the rapid evolution in genomic research and the dramatic decrease in the costs of sequencing, the paradigm of classic medicine has been shifting towards a more personalized approach. Genomic data provides opportunities for substantial improvements in diagnosis and preventive medicine. In particular, it has been shown that an individual's predisposition to a disease depends on genomic variations. Already now, some commercial companies (e.g., 23andMe [1] and Counsyl [2]) provide low-cost (genetic) disease risk tests to their customers for certain diseases. Even though the genome of an individual tells much about his disease risks, research also shows that the non-genomic attributes (described here as clinical and environmental data) of the individuals also contribute significantly to their disease risks. The clinical and environmental data of an individual can include his demographic information, his family history (e.g., diseases of his family members), the list of diseases that he carries, the results of his laboratory tests (e.g., cholesterol level), etc. Thus, such data should also be considered along with the individuals' genomic data when computing their risk for various diseases [3].

The use of individual genomic, clinical, and environmental data can be of interest for a large variety of healthcare stakeholders (here described as medical units), such as (i) a pharmacist checking if a given drug could be harmful (toxicity, interactions) for a patient, (ii) a pharmaceutical company categorizing people based on their risk for a particular disease in order to identify potential clinical trial participants, (iii) a regional health ministry determining the fraction of people at high risk for a particular disease in order to optimize a population-wide preventive medicine effort, (iv) an online direct-to-consumer service provider offering individual risk prediction for various diseases, considering genomic, clinical and environmental data, or (v) a physician, computing the risk of a patient for a particular disease for early diagnosis.

On one hand, for all the aforementioned examples (with the exception of the physician), in order to protect the privacy of his sensitive data, an individual might not want to directly provide his genomic data and clinical and environmental attributes to the medical unit. Furthermore, what is important for the medical unit (with the exception of the physician) is the end-result (i.e., risk of the patient for a disease or the compatibility of a person to a drug); not the individual attributes of the person that lead to the end-result. Even if the medical unit is embodied in a physician and the computed disease risk of a patient is low, the patient would not need to reveal any further (privacy-sensitive) information to the physician (if the disease risk is high, then the physician can extend the test and learn more about the cause of the high risk, with the consent of the patient). On the other hand, such (sensitive) data might play an important role in a disease risk test, hence the inaccuracy (or absence) of such data might cause incorrect (or misleading) results. Therefore, it is crucial to use the correct and complete data of the individuals for the accuracy of such disease risk tests, while still protecting their privacy.

The digitalization of health records has already become a fundamental modernization of the healthcare system to store the clinical and environmental data of the individuals. Furthermore, private storage techniques for such data have been intensively addressed and deployed (e.g., Indivo [4] or Microsoft Healthvault [5]). However, the same is not true for genomic data. Unfortunately, very little progress has been made for the protection of genomic information, and no

progress has been made for the privacy-preserving integration and processing of genomic, clinical, and environmental data.

Because of its extremely sensitive nature, genomic data has an unprecedented impact on privacy [6]. In particular, because the genome carries information about a person's genetic condition and predispositions to specific diseases, the leakage of such information could enable abuse and threats. For example, insurance companies might obtain the genomes of their clients, or employers might (indirectly) test their applicants; access to this information could lead to genetic discrimination or other abuses not yet fully understood. On the other hand, as we discussed before, genomic data includes invaluable medical information about individuals, hence it should be accessed and processed by authorized medical units for healthcare purposes. Thus, it is very important to protect individuals' privacy-sensitive genomic data, while enabling the access to the authorized parties.

In this work, we propose a system for protecting the privacy of individuals' sensitive genomic, clinical, and environmental information, while enabling medical units to process it in a privacy-preserving fashion in order to perform disease risk tests. We introduce a framework in which individuals' medical data (genomic, clinical, and environmental) is stored at a storage and processing unit (SPU) and a medical unit conducts the disease risk test on the encrypted medical data by using homomorphic encryption and privacy-preserving integer comparison. The proposed system preserves the privacy of the individuals' genomic, clinical, and environmental data from a curious party at the SPU and from a malicious party (e.g., a hacker) at the medical unit when computing the disease risk. We also implement the proposed system and show its practicality via a complexity evaluation.

The rest of the paper is organized as follows. In Section II, we summarize the existing work in genomic privacy and privacy of medical records. In Section III, we give a brief background on the tools we use in this paper. In Section IV, we introduce the system and threat models. In Section V, we describe the proposed solution in detail. In Section VI, we show the implementation of the proposed system and discuss its computational complexity. In Section VII, we conclude the paper.

II. Related Work

We first summarize the efforts for protecting genomic data and still enabling its functionality in some genetic tests. Private string searching (on the DNA sequence) by using a *finite state machine* is proposed by Troncoso-Pastoriza, Katzenbeisser, and Celik [7], and then re-visited by Blanton and Aliasgari [8]. To check the similarities of DNA sequences in a privacy-preserving way, Jha, Kruger, and Shmatikov propose using garbled circuits [9], while Bruekers *et al.* propose using homomorphic encryption [10]. Baldi *et al.* make use of private set intersection [11] for privacy-preserving similarity check on DNA sequences [12]. Furthermore, Eppstein and Goodrich propose a privacy-enhanced method for comparing two compressed DNA sequences [13] by using an invertible Bloom filter [14]. Different from the above string searching and comparison methods, Kantarcioglu *et al.* propose using homomorphic encryption to perform scientific investigations on integrated genomic data [15]. Canim, Kantarcioglu, and Malin propose securing the biomedical data by using cryptographic hardware [16]. Finally, in our preliminary work, we propose a privacy-preserving scheme for medical tests and personalized medicine methods that use patients' genomic data [17].[1]

There are also several efforts for protecting the privacy of clinical and environmental data. Many ad-hoc electronic health record (EHR) systems use cryptographic protocols to store medical information in a secure fashion and to define the access rights of the medical units. Both Narayan, Gagne, and Safavi-Naini [18] and Alshehri, Radziszowski, and Raj [19] propose encrypting EHRs based on healthcare providers' attributes or credentials. Benaloh *et al.* propose a system with patient controlled encryption that enables patients both to share partial access rights with others, and to perform searches over their records [20]. Few works also explore the possibility of directly processing the encrypted clinical and environmental data. For example, Barni *et al.* propose privacy-protecting protocols for the classification of medical data [21].

As opposed to the aforementioned efforts, in this paper, we focus on the privacy-preserving storage and processing of genomic data, together with clinical and environmental attributes. More specifically, we show how specific disease risk tests can be done using genomic data, along with clinical and environmental data, while still preserving the privacy of the individuals.

III. Background

In this section, we briefly summarize the main concepts in genomics, statistics and cryptography that we use in this paper.

A. Genomic Background

The human genome is encoded in a double-stranded helical DNA molecule, as a sequence of nucleotides. Genome sequencing techniques record the nucleotides by using the letters A, T, G and C, and the whole human genome includes approximately 3 billion letters. Around 99.9% of the entire genome is identical between any two given individuals. The remaining part ($\sim 0.1\%$) is responsible for many of our inter-individual differences, for example, in physical appearance and in susceptibilities to diseases. Human genetic variation occurs on many levels from gross alterations in the karyotype to single nucleotide variants [22]. The latter are also called *single nucleotide polymorphisms* (SNPs) when they are found to be variable in at least 1% of the individuals in a population. For example, the two short sequences (i) *AAGTCG*, and (ii) *AATTCG* sampled from two different individual's genomes differ at the underlined SNP position.

[1]More information about our activities in this field can be found at: http://lca.epfl.ch/projects/genomic-privacy/.

In general, two different alleles (nucleotides found at a genomic position) are observed for each SNP (the alleles for the SNP in the aforementioned example are G and T, respectively). Furthermore, each individual carries two alleles at each SNP (one inherited from the mother and one from the father). If an individual receives the same allele from both parents, he is said to have a homozygous SNP. If, however, he inherits a different allele from each parent, he has a heterozygous SNP. So far, approximately 50 million SNPs have been identified in the human population [23].

Several studies have assessed both the evolutionary significance and medical applications of SNPs. In particular, Genome-Wide Association Studies (GWAS) have investigated the impact of SNPs on phenotypic traits, such as diseases, and have demonstrated associations between particular variants and disease risks. Each SNP has a different impact on the risk; some of them contribute to the development of the disease, whereas some are protective.

As we discussed before, two different alleles are observed for every SNP. In general, for a SNP that is associated with a disease, one of these alleles carries the risk for the corresponding disease and the other allele does not contribute. For example, assume that the SNP in the above example (with alleles G and T) is associated with a particular disease X. Also assume that out of these two alleles, G is the one carrying the risk for disease X. That is, the presence of G increases the risk for disease X. Then, the risk for disease X is the highest (due to the corresponding SNP) if an individual inherits G from both of his parents (i.e., if he has a homozygous SNP carrying two risk alleles). Whereas, the risk is weaker if he inherits one G and one T, and it is the lowest if he inherits T from both of his parents.[2] For simplicity, in this paper, we represent (i) an homozygous SNP carrying two non-contributing alleles as 0, (ii) an heterozygous SNP carrying one risk (or protective) allele and one non-contributing allele as 1, and (iii) an homozygous SNP carrying two risk (or protective) alleles as 2.[3] In short, each SNP can be in one of the states from $\{0, 1, 2\}$, and we let SNP_i^P represent the state (content) of SNP_i (SNP with ID i) for a patient P.

B. Computation of the Disease Risk

The strength of the association between each SNP and a disease is usually expressed by the *odds ratio* (OR), where the *odds* is the ratio of the probability of occurrence of the disease to that of its non-occurrence in a specific group of individuals. Thus, the OR is the ratio of *odds* in the group of individuals carrying a genetic variation (exposed) to that of those who do not carry it (unexposed). In other words, the OR illustrates by how much the risk of disease is multiplied in an individual carrying a genetic variation compared to another individual not carrying the same variation.

When multiple SNPs are associated with a disease, the overall genetic risk ($\mathbb{S}$) of an individual for the corresponding disease can be computed as a weighted average, based on the OR of each associated SNP by using a logistic regression model. This model is currently widely used among the geneticists and medical doctors for disease risk tests. In such a model, OR of a SNP_i (i.e., OR_i) is generally represented in terms of regression coefficient (β_i), where $OR_i = exp(\beta_i)$. Then, assuming $\Pr_g$ is the probability that an individual P will develop a disease X (only considering his genomic data), his overall genetic risk can be computed as below:[4]

$$\mathbb{S} = ln(\frac{\Pr_g}{1-\Pr_g}) = \alpha + \sum_{i \in \varphi_X} \beta_i \mathrm{p}_j^i(X), \qquad (1)$$

where $\mathrm{p}_j^i(X)$ is the contribution of the SNP_i to the genetic risk (for disease X) when $\mathrm{SNP}_i^P = j$ ($\mathrm{SNP}_i^P \in \{0, 1, 2\}$ as discussed in Section III-A), and α is the intercept of the model.

For clinical use, the genetic risk, computed in (1) should be categorized based on its risk group. For this purpose, generally, the distribution of the potential genetic scores (in a given population) is divided into smaller parts called *quantiles* (or risk groups) as in Fig. 1. In Fig. 1, there are 4 different risk groups, each with a different genetic regression coefficient. For example, if $\mathbb{S}$ is somewhere between b_1 and b_2, then we assign the genetic regression coefficient for the corresponding individual as β_2. For each individual, the genetic score is computed as in (1), and positioned into its risk group. We represent the genetic regression coefficient corresponding to the genetic risk $\mathbb{S}$ as β_g.

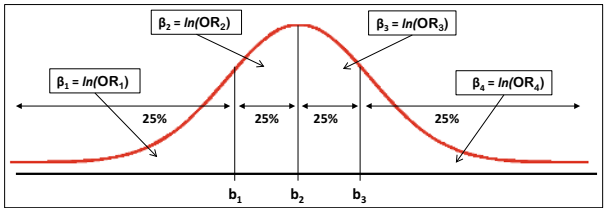

Fig. 1. Genetic score distribution partitioned in 4 genetic risk groups.

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

$$ln(\frac{\Pr}{1-\Pr}) = \beta_f = \beta_0 + \beta_g + \sum_{N_i \in \mathbb{N}} \bar{\beta}_i N_i, \qquad (2)$$

where β_0 is the new intercept, $\mathbb{N}$ is the set of clinical and environmental attributes associated with the disease, and $\bar{\beta}_i$ is the regression coefficient corresponding to the clinical or

[2]The same holds for protective SNPs.

[3]The number of alleles carrying the risk is usually called *genetic burden*.

[4]In general, a logistic regression model is represented as $ln(\frac{\Pr}{1-\Pr}) = \alpha + \sum_i \beta_i X_i$. In our model, the explanatory variable X_i is p_j^i.

environmental attribute N_i. From (3), the probability (Pr) that the corresponding individual will develop disease X (considering all the genomic, clinical, and environmental data) can be computed as follows:

$$\Pr = \frac{e^{\beta_f}}{1 + e^{\beta_f}}. \tag{3}$$

C. Cryptographic Background

In this section, we briefly describe two cryptosystems along with their homomorphic properties: the modified Paillier cryptosystem (described in detail in [24] and [25]) and the DGK cryptosystem (described in detail in [26]).

1) Modified Paillier cryptosystem: The Paillier cryptosystem is a public key cryptosystem supporting some homomorphic operations. The public key is represented as $(n, g, h = g^x)$, where the strong secret key is the factorization of $n = zy$ (z, y are safe primes), the weak secret key is $x \in [1, n^2/2]$, and g of the order $(z-1)(y-1)/2$. By selecting a random $a \in \mathbb{Z}^*_{n^2}$, g can easily be computed as $g = -a^{2n}$.

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

  $$C_1 = g^r \mod n^2 \quad \text{and} \quad C_2 = h^r(1+mn) \mod n^2. \tag{4}$$

  For simplicity, in the rest of this paper, we represent the Paillier encryption of a message m as $[m]$.
- *Decryption*: The message m can be recovered from $[m]$ as follows:

  $$m = \Delta(C_2/C_1^x) \tag{5}$$

  where $\Delta(u) = \frac{(u-1) \mod n^2}{n}$, for all $u \in \{u < n^2 \mid u = 1 \mod n\}$.

- *Proxy re-encryption*: Assume we randomly split the secret key in two shares x_1 and x_2, such that $x = x_1 + x_2$. The modified Paillier cryptosystem enables an encrypted message (C_1, C_2) to be partially decrypted to a ciphertext pair $(\tilde{C}_1, \tilde{C}_2)$ using x_1 as below:

  $$\tilde{C}_1 = C_1 \quad \text{and} \quad \tilde{C}_2 = C_2/C_1^{x_1} \mod n^2. \tag{6}$$

  Then, $(\tilde{C}_1, \tilde{C}_2)$ can be decrypted using x_2 with the aforementioned decryption function to recover the original message.

2) DGK cryptosystem: The DGK cryptosystem is optimized for the secure comparison of integers. The key generation needs three parameters k, t and L where $k > t > L$. The parameter k represents the number of bits of the RSA modulus n, t is the size of two small primes v_p and v_q, and L is the message space size in bits. Assume that p and q are two distinct primes of equal bit length, such that $p-1$ is divisible by v_p and $q-1$ is divisible by v_q. Then, the public key is represented as (n, g, h, u), where u is a L-bit prime, $g \in \mathbb{Z}^*_n$ with order uv_pv_q, and h is an integer with order v_pv_q. Furthermore, the private key is represented as (p, q, v_p, v_q). For simplicity, in the rest of this paper we represent the DGK encryption of a message m as $\langle m \rangle$.

3) Homomorphic properties: Both modified Paillier and DGK cryptosystems support some computations in ciphertext domain. In particular, both cryptosystems have the following properties:

- The product of two ciphertexts is equal to the encryption of the sum of their corresponding plaintexts.
- A ciphertext raised to a constant number is equal to the encryption of the product of the corresponding plaintext and the constant.

These homomorphic operations are used in our proposed solution (in Section V) to compute the genetic risk and the overall disease risk in ciphertext domain.

IV. System and Threat Models

In this work, we propose a system for the privacy-preserving computation of disease risk by using both genomic data and clinical and environmental factors. In general, this type of a medical test involves a patient (P) and a medical unit (MU). As we discussed in Section I, the medical unit can be a pharmacist, a pharmaceutical company, a regional health ministry, an online direct-to-consumer service provider, or a physician (for early diagnosis).

We assume that the sequencing and the encryption of the genomic data of the patient are performed at a certified institution (CI), which is a trusted entity. We note that such a trusted entity is indispensable in such a system, as the sequencing has to be done at an institution to obtain the genetic variation profile of the patient. Furthermore, the clinical and environmental data of the patient is collected during his doctor visits (e.g., at the MU) or directly provided by the patient. As we discussed before, a patient might not be willing to reveal all his clinical and environmental data to an MU (e.g., his HIV status or family history). However, this privacy-sensitive data can play an important role in the accuracy of the computed disease risk. The proposed system allows the patient to choose what part of his clinical and environmental data to hide from an MU and it still involves such hidden data in the computation of the disease risk.

We assume that the storage and processing of genomic, clinical, and environmental data is done at a storage and processing unit (SPU) for efficiency and security. That is, instead of several MUs storing the same large amount of genomic data (tens to hundreds of gigabytes per patient), the genomic data of the patients is stored at a centralized SPU, and provided to the MUs upon request. Storing the genomic, clinical, and environmental data at the SPU also makes such data available to any MU at any given time (e.g., during

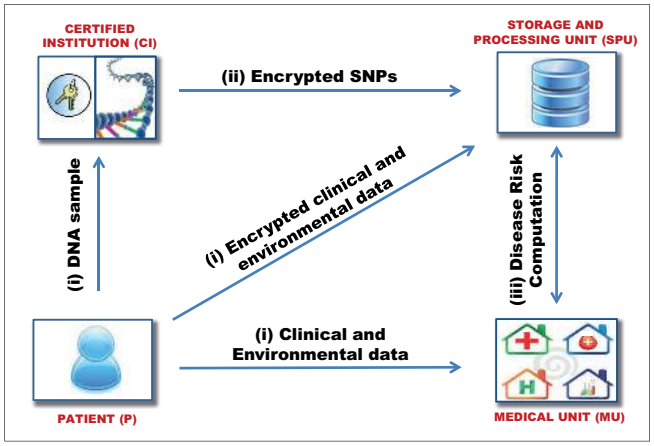

Fig. 2. Proposed system model for the privacy-preserving computation of the disease risk.

emergencies). Furthermore, as an MU can be embodied in several entities from a physician to an online service, it would be unrealistic to assume that all these different entities will pay high attention to the security of the data they store. It is easier therefore to provide the security of the genomic, clinical, and environmental data of the patients at the SPU.[5] We note that a private company (e.g., cloud storage service), the government, or a non-profit organization could play the role of the SPU.

The general architecture of the proposed system is illustrated in Fig. 2. In summary (it will be described in detail in Section V), the patient provides his sample for sequencing to the CI. Meanwhile, he also provides his clinical and environmental data to the SPU and the MU.[6] The CI is responsible for sequencing and encryption of the patient's genomic data. Then, the CI sends the encrypted genomic data to the SPU. Finally, the privacy-preserving computation of the disease risk takes place between the MU and the SPU.

In this study, we consider the following two types of potential attackers: (i) an attacker at the MU, and (ii) a curious party at the SPU. The attacker can be represented by a careless or disgruntled employee at the MU or a hacker who breaks into the MU and aims to obtain private genomic, clinical, and environmental information about a patient (for which it is not authorized). We also assume that the SPU might be a curious entity (e.g., existence of a curious party or a disgruntled employee at the SPU), hence all genomic, clinical, and environmental data should be stored at the SPU in encrypted form (i.e., the SPU should not be able to access the contents of patients' data). Furthermore, patients' data is stored using pseudonyms (without revealing the real identities of the patients) at the SPU, hence SPU cannot associate a medical test to a patient. Other than being potentially curious, we assume that the SPU is an honest party. That is, it follows the protocol properly and does not change the integrity of the stored data. Finally, we assume that the MU and the SPU do not collude.

[5]For similar reasons, we prefer not to leave patients' genomic, clinical, and environmental data in their own hands (e.g., by storing it on their personal devices).

[6]As we discussed before, depending on the privacy-sensitivity of the clinical and environmental data, the patient can choose which clinical and environmental attributes to reveal to the MU, and which ones to encrypt and keep at the SPU.

V. Proposed Solution

A. Initialization

The cryptographic keys of each patient (for the modified Paillier cryptosystem) are generated and distributed to the patients during the initialization period. Furthermore, the patient's secret key x is randomly divided into x_1 and x_2 (such that $x = x_1 + x_2$ as discussed in Section III-C) and each share is distributed to the SPU and to the MU, respectively (i.e., x_1 is provided to the SPU and x_2 to the MU). Similarly, public and private keys of the SPU for the DGK cryptosystem are generated and the public key is shared with the MU. Finally, symmetric keys are established between the parties (to protect the communication between the parties from an eavesdropper). We note that the distribution, update and revocation of cryptographic keys are handled by a trusted entity.[7]

B. Sequencing and Clinical and Environmental Data Collection

The patient provides his sample for sequencing. The sample is sequenced by the CI with the consent of the patient. The sequence is further analyzed and the SNPs of the patient are extracted. We assume that the non-consented sequencing of the patients' genomes is forbidden by law. Furthermore, even if non-consented sequencing using collected samples of the patients were possible, this type of an attack would be both low-scale (in terms of the number of victims) and more importantly, very costly (due to the cost of sequencing machines) for the attacker.

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

In contrast with genomic data, some clinical and environmental data (e.g., cholesterol level) of the patient is subject to frequent changes over time.[8] From here on, we call such clinical and environmental data as the "variable data". Thus, upon collection, the variable data is accompanied with a date (e.g., date of the collection of the data). By doing so, the MU can decide whether any variable data should be updated by the patient before the computation of his risk for a disease.

[7]In this case, the trusted entity can be the certified institution (CI).

[8]Genomic data also rarely changes (e.g., via mutations). In such cases, the patient's genome might need to be sequenced again.

C. Encryption and Storage of Genomic, Clinical, and Environmental Data

Encryption of the contents of the SNPs are done at the CI by using the modified Paillier cryptosystem (Section III-C). We assume SNP_i represents the position (or ID) of a SNP (on the DNA sequence). We also assume SNP_i^P represents the content (or state) of SNP_i at patient P, where $\mathrm{SNP}_i^P \in \{0, 1, 2\}$ (as discussed in Section III-A). After the sequencing and the extraction of the SNPs of the patient, the CI encrypts the contents of all SNP positions of the patient (to obtain $[\mathrm{SNP}_i^P]$[9]) along with their squared values (to obtain $[(\mathrm{SNP}_i^P)^2]$). The squared values of the SNPs are required for the homomorphic operations (in Section V-D1). Eventually, the CI encrypts the contents of around 50 million SNP positions for the patient. Furthermore, the CI also individually encrypts each clinical and environmental attribute of the patient by using the modified Paillier cryptosystem (we will further discuss the contents of clinical and environmental attributes in Section V-D3).

After encryption, the CI sends the encrypted genomic, clinical, and environmental data (along with the pseudonym of the patient) to the SPU for storage. We note that only the contents of the SNPs are encrypted at the CI; not their positions (on the DNA). Thus, the SPU stores the positions (or the IDs) of the SNPs in plaintext, mainly to check the access rights of the MU for the requested SNPs. Similarly, the identifiers of the clinical and environmental attributes (e.g., "age" or "cholesterol level") and their dates (i.e., collection date of the variable data) are stored in plaintext at the SPU.

D. Privacy-Preserving Computation of Disease Risk

The computation of a patient's disease risk is done at the MU. In the remaining of this section, we will describe how the MU obtains the risk of patient P for a disease X. The MU requests the (encrypted) genomic, clinical, and environmental data of the patient that will be used for the computation of the disease risk from the SPU (request is done using the pseudonym of the patient). The SPU then verifies that the MU has the required access rights for the requested genomic, clinical, and environmental data for the corresponding computation and sends the requested (encrypted) data to the MU. Along with the encrypted genomic, clinical, and environmental data, the positions (or IDs) of the encrypted SNPs and the identifiers and the collection dates of the clinical and environmental data are also sent to the MU by the SPU. Looking at the collection date, MU can decide if the patient needs to update any of his clinical and environmental data before the computation of the disease risk.

The MU first computes the regression coefficient corresponding to the genetic risk of the patient for disease X (as discussed in Section III-B). Then, it combines this with the clinical and environmental factors and eventually obtains the overall risk of the patient to disease X. Next, we describe this process in detail.

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

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

where, χ, ψ, and μ are plaintext normalizing constants.

As we discussed in Section III-B, the MU needs to know on which genetic risk group (quantile) the above genetic risk is positioned in order to determine the regression coefficient (β_g) of the computed genetic risk (each risk group contributes the risk with a different regression coefficient). However, as the above computed genetic risk is encrypted, to find the regression coefficient corresponding to the computed genetic risk, we propose to use a privacy-preserving integer comparison algorithm [27] between the MU and the SPU.

2) Computing the genetic regression coefficient: The genetic risk distribution consists of ρ genetic risk groups (or quantiles), each with different regression coefficients (e.g., $\rho = 4$ in Fig. 1). We let b_i^l and b_i^u represent the lower and upper boundary of the ith risk group of the genetic risk scale, respectively. In short, MU compares $[\mathbb{S}]$ with the boundaries of the genetic risk scale in a privacy-preserving way, such that neither the MU nor the SPU learns the value of $\mathbb{S}$ or the result of any comparison. Assume that both $\mathbb{S}$ and b_i^j ($j \in \{l, u\}$) are L-bit numbers. We summarize the main steps of the privacy-preserving comparison algorithm below. The operations in these steps are also illustrated in Fig. 3.

Step 1 (@MU): The MU computes $[\mathrm{z}] = [2^L + \mathbb{S} - b_i^j]$. Let z_{L-1} represent the most significant bit of z. Then, (i) $z_{L-1} = 0$ if $\mathbb{S} < b_i^j$; and (ii) $z_{L-1} = 1$ if $\mathbb{S} \geq b_i^j$. Thus, the MU needs to compute $[z_{L-1}]$, where $[z_{L-1}] = [\mathrm{z} - (\mathrm{z} \mod 2^L)]$. However, the MU cannot compute $[\mathrm{z} \mod 2^L]$ using the homomorphic properties of the modified Paillier cryptosystem. Thus, the MU initiates a privacy-preserving comparison protocol with the SPU to compute $[\mathrm{z} \mod 2^L]$.

The MU generates a random number r and computes $[\mathrm{d}] = [\mathrm{z} + \mathrm{r}]$. Then, the MU partially decrypts $[\mathrm{d}]$ using x_2 to obtain $[\tilde{\mathrm{d}}]$[10] and sends $[\tilde{\mathrm{d}}]$ to the SPU.

[9]We represent the encryption of a message m using modified Paillier cryptosystem as $[m]$.

[10]Partial decryption using a share of the patient's secret key is discussed in Section III-C.

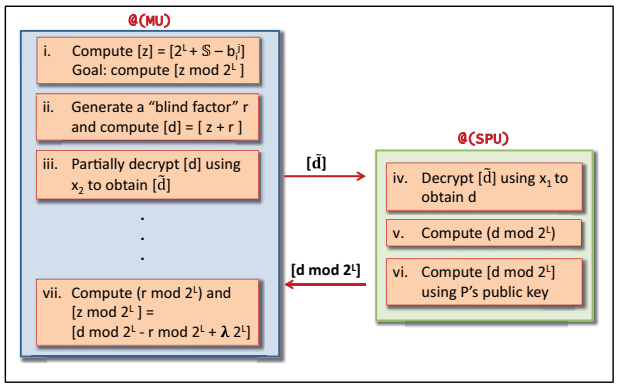

Fig. 3. Privacy-preserving comparison algorithm to determine the risk group of the genetic score $\mathbb{S}$.

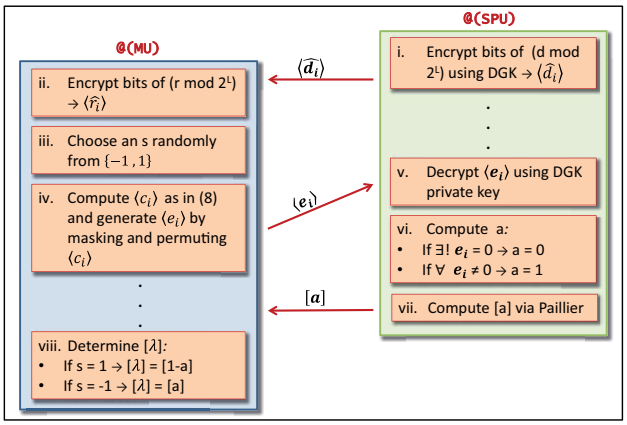

Fig. 4. Privacy-preserving comparison protocol to compute $[\lambda]$ (to determine whether there is an underflow in the computation of $[\mathrm{z} \mod 2^L]$ in Step 3).

Step 2 (@SPU): The SPU decrypts $[\tilde{\mathrm{d}}]$ using x_1 to obtain d. Then, the SPU computes $(\mathrm{d} \mod 2^L)$, encrypts it (via the patient's public key using the modified Paillier cryptosystem) to obtain $[\mathrm{d} \mod 2^L]$, and sends the encrypted value to the MU.

Step 3 (@MU): The MU computes $(\mathrm{r} \mod 2^L)$ and encrypts it (via the patient's public key, by using the modified Paillier cryptosystem) to obtain $[\mathrm{r} \mod 2^L]$. Then, it computes $\overline{[\mathrm{z} \mod 2^L]} = [\mathrm{d} \mod 2^L - \mathrm{r} \mod 2^L]$. We note that $\overline{[\mathrm{z} \mod 2^L]} = [\mathrm{z} \mod 2^L]$ if $(\mathrm{d} \mod 2^L) \geq (\mathrm{r} \mod 2^L)$. However, if $(\mathrm{r} \mod 2^L) > (\mathrm{d} \mod 2^L)$, an underflow occurs as the subtraction is done in modulo n (using the homomorphic properties of the modified Paillier cryptosystem). To avoid this underflow problem, the MU should compute $[\mathrm{z} \mod 2^L]$ as $[\mathrm{z} \mod 2^L] = [\overline{\mathrm{z} \mod 2^L} + \lambda 2^L]$, where $\lambda = 0$ if $(\mathrm{d} \mod 2^L) \geq (\mathrm{r} \mod 2^L)$, and $\lambda = 1$ otherwise. In the following we describe how the MU computes $[\lambda]$ with the help of the SPU.

Computing $[\lambda]$: For the efficiency of the protocol, the computation of $[\lambda]$ relies on the homomorphic encryption scheme proposed by Damgard *et al.* [26] (DGK cryptosystem, as described in Section III-C). Compared to the modified Paillier cryptosystem, the DGK cryptosystem allows for an efficient multiplicative masking, as it has a small plaintext space $\mathbb{Z}_u$ (where u is a prime number). Similar to modified Paillier, the DGK cryptosystem also supports the addition of two ciphertexts and the multiplication of a ciphertext with a plaintext constant. As discussed before, we represent the DGK encryption of a message m under the public key of the SPU as $\langle m \rangle$. In the following, we describe the computation of $[\lambda]$. We also summarize the main steps in the computation of $[\lambda]$ in Fig. 4.

Let $\hat{\mathrm{d}} = (\mathrm{d} \mod 2^L)$ and $\hat{d}_i$ represent the ith bit of $\hat{\mathrm{d}}$ (where $i \in \{0, 1, \ldots, L-1\}$). In Step 2 of the above protocol, the SPU also encrypts the bits of $\hat{\mathrm{d}}$ by its public key by using the DGK encryption to obtain $\langle \hat{d_0} \rangle, \ldots, \langle \hat{d_{L-1}} \rangle$ and sends these encrypted bits to the MU.

Similarly, let $\hat{\mathrm{r}} = (\mathrm{r} \mod 2^L)$ and $\hat{r_i}$ represent the ith bit of $\hat{\mathrm{r}}$ (where $r \in \{0, 1, \ldots, L-1\}$). In Step 3 of the above protocol, the MU encrypts the bits of $\hat{\mathrm{r}}$ under the public key of the SPU using the DGK encryption to obtain $\langle \hat{r_0} \rangle, \ldots, \langle \hat{r_{L-1}} \rangle$. Then, the MU randomly chooses an integer s from the set $\{1, -1\}$ and computes $\mathrm{C} = \{\langle c_0 \rangle, \ldots, \langle c_{L-1} \rangle\}$, where

$$\langle c_i \rangle = \Big\langle \hat{d}_i - \hat{r}_i + s + 3 \sum_{j=i+1}^{L-1} w_j \Big\rangle, \tag{8}$$

and $w_j = \hat{d}_j \oplus \hat{r}_j$. We note that $\langle \hat{d}_j \oplus \hat{r}_j \rangle = \langle \hat{d}_j + \hat{r}_j - (2\hat{r}_j)\hat{d}_j \rangle$, hence it can be easily computed at the MU (using the homomorphic properties of the DGK cryptosystem), as $\hat{r}_j$ values are available to the MU in plaintext. Next, for each $\langle c_i \rangle$, the MU selects a random number α_i (from $\mathbb{Z}_u$) and computes $\langle e_i \rangle = \langle c_i \alpha_i \rangle$ in order to multiplicatively mask the $\langle c_i \rangle$ values. The MU also permutes the ordering of $\langle e_i \rangle$ values and sends them to the SPU.

The SPU decrypts the $\langle e_i \rangle$ values using its private key and checks for the following two cases: (i) If all e_i values are non-zero, the SPU sets $a = 1$, and (ii) if exactly one e_i value is zero, the SPU sets $a = 0$. Then, the SPU encrypts a using the modified Paillier cryptosystem under the patient's public key to obtain $[a]$, and sends this encrypted value to the MU.

We note that if $a = 1$ and $s = 1$ (the number randomly selected by the MU), then $\hat{\mathrm{d}} \geq \hat{\mathrm{r}}$ (i.e., $\lambda = 0$). Similarly, if $a = 0$ and $s = 1$, then $\hat{\mathrm{r}} > \hat{\mathrm{d}}$ (i.e., $\lambda = 1$).[11] Thus, if $s = 1$, the MU sets $[\lambda] = [1 - a]$ and if $s = -1$, it sets $[\lambda] = [a]$. Using $[\lambda]$, the MU can compute $[\mathrm{z} \mod 2^L]$, and hence $[z_{L-1}]$ as we discussed before.

Let $[G(\mathbb{S}, b_i^u)] = [z_{L-1}]$ represent the (encrypted) result of the comparison between $\mathbb{S}$ and b_i^u. Then, (i) $G(\mathbb{S}, b_i^u) = 0$ if $\mathbb{S} < b_i^l$; and (ii) $G(\mathbb{S}, b_i^u) = 1$ if $\mathbb{S} \geq b_i^l$. Given there are ρ risk groups in the genomic risk scale, using the above privacy-preserving comparison algorithm, the MU can determine the genetic regression coefficient (β_g) corresponding to $\mathbb{S}$ as

[11] The opposite holds when $s = -1$.

SNP	Chr	Allele	Risk allele	OR
rs3798220	6	T>C	C	1.51
rs4977574	9	A>G	G	1.29
rs9982601	21	C>T	T	1.18
rs17114036	1	A>G	A	1.17
rs17465637	1	C>A	C	1.14
rs6725887	2	T>C	C	1.14
rs1122608	19	G>T	G	1.14
rs964184	11	C>G	G	1.13
rs12413409	10	G>A	G	1.12
rs2306374	3	T>C	C	1.12
rs599839	1	A>G	A	1.11
rs579459	9	T>C	C	1.10
rs12526453	6	C>G	C	1.10
rs11556924	7	C>T	C	1.09
rs1746048	10	C>T	C	1.09
rs12190287	6	C>G	C	1.08
rs3825807	15	A>G	A	1.08
rs216172	17	C>G	G	1.07
rs12936587	17	A>G	G	1.07
rs4773144	13	A>G	G	1.07
rs17609940	6	G>C	G	1.07
rs2895811	14	T>C	C	1.07
rs46522	17	T>C	T	1.06

TABLE I

GENOMIC VARIANTS (SNPS) USED FOR THE COMPUTATION OF THE GENETIC RISK FOR THE CORONARY ARTERY DISEASE (CAD) [3]. "CHR" IS THE CHROMOSOME ON WHICH THE CORRESPONDING SNP IS LOCATED. "ALLELE" IS THE SET OF NUCLEOTIDES THAT ARE OBSERVED FOR THE CORRESPONDING SNP (THE FREQUENTLY OBSERVED ALLELE IS THE ONE ON THE LEFT). "RISK ALLELE" IS THE ALLELE WHICH CARRIES THE RISK FOR CAD. "OR" IS THE ODDS RATIO OF THE CORRESPONDING SNP PER RISK ALLELE (E.G., IF THE PATIENT CARRIES TWO RISK ALLELES OF A PARTICULAR SNP, THE OR BECOMES TWO TIMES THE VALUE INDICATED IN THE TABLE). AS WE DISCUSSED IN SECTION III-B, THE RELATION BETWEEN THE OR AND THE REGRESSION COEFFICIENT (β_i) OF A SNP i CAN BE REPRESENTED AS $OR_i = exp(\beta_i)$. THE ASSOCIATIONS OF THESE SNPS TO CAD WERE IDENTIFIED THROUGH A META-ANALYSIS OF MULTIPLE GENOME-WIDE ASSOCIATION STUDIES (GWAS).

below:

$$[\beta_g] = \Big[\beta_1(1 - G(\mathbb{S}, b_1^u)) + \sum_{i=2}^{(\rho-1)} \beta_i(G(\mathbb{S}, b_{i-1}^u) - G(\mathbb{S}, b_i^u)) + \beta_\rho G(\mathbb{S}, b_{\rho-1}^u)\Big], \quad (9)$$

where β_i is the genetic regression coefficient of the ith quartile (risk group). We note that the above computation can be easily conducted using the homomorphic properties of the modified Paillier cryptosystem.

3) Computing the final disease risk: To compute the final disease risk of the patient, the MU combines $[\beta_g]$ with the patient's clinical and environmental regression coefficients to obtain the aggregate regression coefficient β_f. Even though some clinical and environmental data entries, such as smoking behavior, can be represented binary (e.g., 1 is the patient smokes and 0 if he does not), some entries are the results of medical tests (e.g., cholesterol level) or demographic data (e.g., age of the patient). To compute the aggregate regression coefficient of the patient for a particular disease, these entries should be categorized following the requirements of the medical test. For example, when computing the cardiovascular disease risk [3] of a patient, the regression coefficient of the patient's age takes 2 different values: (i) it gets a higher value if the patient is over 45; and (ii) it gets a lower value if the patient is 45 or younger. In this particular example, the encrypted age of the patient can be compared with 45 to represent this attribute as a binary value (i.e., 1 if the patient is over 45, and 0 otherwise). Therefore, similar to before, the MU use a privacy-preserving integer-comparison algorithm [27] to make such comparisons for the encrypted clinical and environmental data, before computing the aggregate regression coefficient of the patient.[12]

Let $\mathbb{N} = \{[N_1], [N_2], \ldots, [N_m]\}$ be the set of encrypted clinical and environmental attributes of the patient (that are required for the computation of the risk for disease X), where $N_i \in \{0, 1\}$ for the simplicity of the presentation. That is, $N_i = 1$ if the patient has the corresponding clinical or environmental attribute, and $N_i = 0$ otherwise. As we discussed before, even if N_i is non-binary, it can be transformed to a binary number using the privacy-preserving comparison algorithm (in Section V-D2). Then, the (final) aggregate regression coefficient of the patient (for disease X) is computed as below:

$$[\beta_f] = \Big[\beta_0 + \beta_g + \sum_{i=1}^{m} \bar{\beta}_i N_i\Big], \quad (10)$$

where $\bar{\beta}_i$ is the regression coefficient of the ith clinical or environmental attribute, and β_0 is the intercept (as discussed in Section III-B). This encrypted aggregate regression coefficient is then sent back to the SPU, where is it partially decrypted by using x_1 to obtain $[\tilde{\beta_f}]$. Then, the SPU sends $[\tilde{\beta_f}]$ back to the MU, where is it decrypted using x_2 to obtain β_f. Finally, the MU computes the final disease risk of the patient for disease X as $\frac{e^{\beta_f}}{1+e^{\beta_f}}$.

We note that this proposed scheme preserves the privacy of patients' genomic data relying on the security strength of modified Paillier cryptosystem and the DGK cryptosystem. The extensive security evaluation of the modified Paillier and DGK cryptosystems can be found in [24] and [26], respectively.

VI. IMPLEMENTATION AND COMPLEXITY EVALUATION

To evaluate the practicality of the proposed privacy-preserving algorithm, we implemented it, and assessed its storage requirement and computational complexity. We evaluated the proposed system using real genomic data. In particular, we encrypted a real individual's SNP profile from [28], and we computed the coronary artery disease (CAD) risk by using real data (i.e., OR values and genetic risk distribution) from [3]. In summary, CAD risk computation includes (i) 23 SNPs associated with cardiovascular risk (in Table I) for the

[12]There might be more than 2 categories for some clinical or environmental attributes. In this case, the MU needs to determine under which category the patient's (encrypted) attribute falls using a similar technique discussed in Section V-D2.

Variable	Odds Ratio
Age (>45 years)	3.21
Current smoking	2.25
Family history of CAD	2.00
Lopinavir (> 1 year)	1.74
Diabetes	1.81
Current abacavir exposure	1.62
Past smoking	1.51
Indinavir (> 1 year)	1.28
High cholesterol	1.61
On ART	1.51
Hypertension	1.44
Low HDL cholesterol	1.11
CD4	0.99
HIV RNA	1.00
Genetic score quantile 2 vs. quantile 1	1.12
Genetic score quantile 3 vs. quantile 1	1.33
Genetic score quantile 4 vs. quantile 1	1.62

TABLE II
CLINICAL, ENVIRONMENTAL, AND GENETIC RISK FACTORS USED FOR COMPUTATION OF THE CORONARY ARTERY DISEASE (CAD) RISK [3].

computation of the genetic risk ($\mathbb{S}$), and (ii) 14 clinical and environmental factors (in Table II) to compute the overall CAD risk. Furthermore, the genetic score distribution is partitioned into 4 genetic risk groups or quantiles (i.e., $\rho = 4$ in Section V). We note that in [3], the contributions of genetic, clinical, and environmental factors to the disease risk are computed via a logistic regression model on a population of 2078 individuals.

We implemented the proposed system on an Intel Core i7-2620M CPU with 2.70 GHz processor under Windows 7 Operating System. We set the size of the security parameter (n in Paillier cryptosystem in Section III-C) to 4096 bits. The security parameters of the DGK cryptosystem are set to the following values: $L = 16$, $t = 160$, $k = 1024$. Our implementation relies on a MySQL 5.5 database managed by the open source tool MySQL Workbench. To provide a platform-independent implementation, we used the Java programming language along with the open-source Integrated Development Environment, NetBeans IDE 7.1.1., for the implementation of the Java code.[13] In Table III, we summarize the computational and storage complexities of the proposed solution.

We emphasize that the encryption of the variants at the CI (using the modified Paillier cryptosystem) is a one-time operation and is significantly faster than sequencing and analysis of the sequence. Further, this encryption can be conducted much more efficiently by pre-computing some parameters, such as (g^r, h^r) pairs, for various r values, for each patient. Indeed, by pre-computing (g^r, h^r) pairs, we observed that the encryption takes only 0.168 ms. per attribute at the CI. All these numbers show that our privacy-preserving algorithm is very realistic and could be implemented with current computing technology.

Finally, in Fig. 5, we illustrate two screen shots from our implementation in which we illustrate the operations conducted at the patient (P) and medical unit (MU), respectively.[14]

[13]We note that our code for the implementation is not optimized, and better results can be expected with an optimized implementation.

VII. CONCLUSION

In this paper, we have proposed a framework in which patients' genomic, clinical, and environmental data is securely stored at a storage and processing unit, and in which a medical unit conducts disease risk tests on this encrypted data by using homomorphic encryption and privacy-preserving integer comparison. We have shown that the proposed system preserves the privacy of the patients against a curious party at the storage and processing unit and a malicious party at the medical unit. We have also implemented the proposed solution and shown its practicality. We believe that the proposed privacy-preserving disease risk test would encourage the use of genomic, clinical, and environmental data in medical tests by ensuring the patients that the privacy of their sensitive data will be preserved.

VIII. ACKNOWLEDGEMENTS

We would like to thank Dr. Amalio Telenti, Dr. Philip E. Tarr, and Dr. Jacques Rougemont for their useful comments and suggestions. We also would like to thank Dr. Vincent Mooser, Dr. Didier Trono and Dr. Martin Vetterli for their encouragements in this research endeavor.

REFERENCES

[1] https://www.23andme.com/welcome/, Visited on 22/Apr/2013.
[2] https://www.counsyl.com, Visited on 22/Apr/2013.
[3] M. Rotger and *et al.*, "Contribution of genetic background, traditional risk factors and HIV-related factors to coronary artery disease events in HIV-positive persons," *Clinical Infectious Diseases*, Mar. 2013.
[4] K. D. Mandl, W. W. Simons, W. C. Crawford, and J. M. Abbett, "Indivo: A personally controlled health record for health information exchange and communication," *BMC Medical Informatics and Decision Making*, p. 25, Sep. 2007.
[5] https://www.healthvault.com, Visited on 22/Apr/2013.
[6] E. Ayday, E. D. Cristofaro, G. Tsudik, and J. P. Hubaux, "The chills and thrills of whole genome sequencing," *arXiv:1306.1264*, 2013. [Online]. Available: http://arxiv.org/abs/1306.1264
[7] J. R. Troncoso-Pastoriza, S. Katzenbeisser, and M. Celik, "Privacy preserving error resilient DNA searching through oblivious automata," *CCS '07: Proceedings of the 14th ACM Conference on Computer and Communications Security*, pp. 519–528, 2007.
[8] M. Blanton and M. Aliasgari, "Secure outsourcing of DNA searching via finite automata," *DBSec'10: Proceedings of the 24th Annual IFIP WG 11.3 Working Conference on Data and Applications Security and Privacy*, pp. 49–64, 2010.
[9] S. Jha, L. Kruger, and V. Shmatikov, "Towards practical privacy for genomic computation," *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pp. 216–230, 2008.
[10] F. Bruekers, S. Katzenbeisser, K. Kursawe, and P. Tuyls, "Privacy-preserving matching of DNA profiles," Tech. Rep., 2008.
[11] R. Agrawal, A. Evfimievski, and R. Srikant, "Information sharing across private databases," *Proceedings of SIGMOD Conference*, 2003.
[12] P. Baldi, R. Baronio, E. De Cristofaro, P. Gasti, and G. Tsudik, "Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes," *CCS '11: Proceedings of the 18th ACM Conference on Computer and Communications Security*, pp. 691–702, 2011.
[13] D. Eppstein, M. T. Goodrich, and P. Baldi, "Privacy-enhanced methods for comparing compressed DNA sequences," *CoRR*, vol. abs/1107.3593, 2011. [Online]. Available: http://arxiv.org/abs/1107.3593

[14]We skip the illustration of the intermediate steps, and only present the key steps of the algorithm.

Complexity of the Proposed System				
Encryption	**Storage**	**Computation of disease risk**		
380 ms./attribute (with pre-computed values: 0.168 ms./attribute)	51.2 GB per patient	*Computation of the genetic risk*	*Privacy-preserving integer comparison*	*Computation of the final risk*
		230 sec (23 SNPs)	3.390 sec (3 comparisons)	140 sec (14 environmental factors)
		Total: 373.432 sec		

TABLE III

COMPUTATIONAL AND STORAGE COMPLEXITIES OF THE PROPOSED SYSTEM. THE KEY SIZE FOR THE MODIFIED PAILLIER CRYPTOSYSTEM IS SET TO 4096 BITS. IN THE "ENCRYPTION" PHASE, BOTH GENETIC ATTRIBUTES (I.E., SNPS) AND CLINICAL AND ENVIRONMENTAL ATTRIBUTES ARE ENCRYPTED. THE DATA IS STORED AT THE SPU. THE TOTAL TIME FOR THE COMPUTATION OF THE DISEASE RISK ALSO INCLUDES MINOR OPERATIONS SUCH AS PROXY RE-ENCRYPTION AND PAILLIER DECRYPTION.

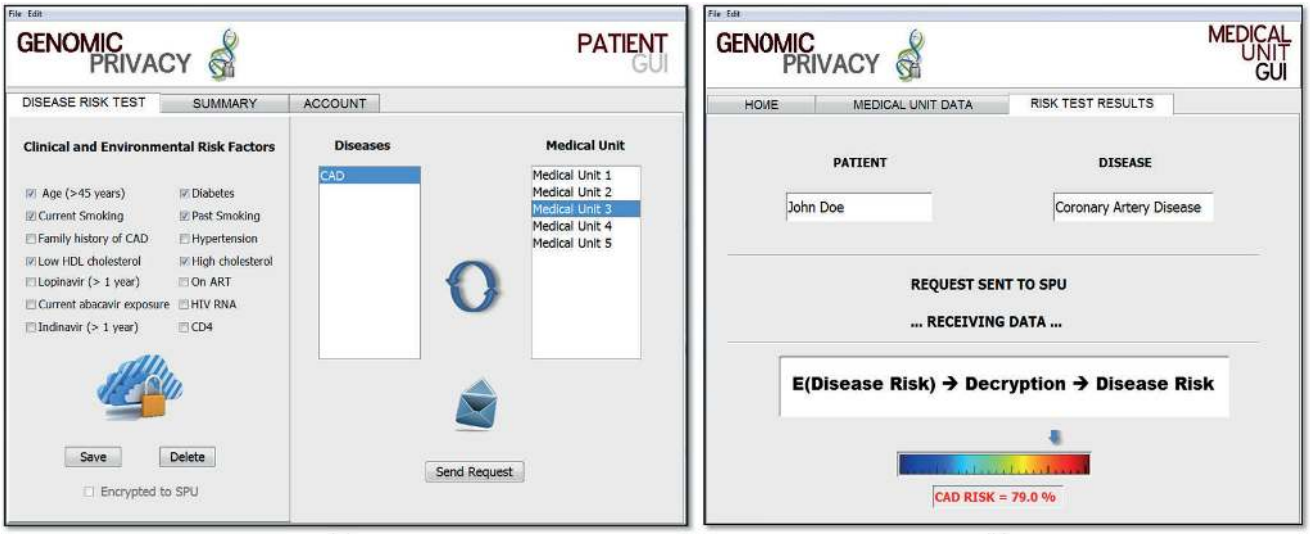

Fig. 5. Implementation of the proposed system at patient P and the medical unit (MU) for the computation of coronary artery disease (CAD) risk. In Fig. 5(a), P specifies his clinical and environmental risk factors either directly to the MU, or these attributes are encrypted and stored at the SPU. In Fig. 5(b), the MU receives the encrypted SNPs and clinical and environmental factors of P (that are associated with CAD) from the SPU, and computes CAD risk. That is, the MU recovers the probability that P will develop "CAD" in the future based on his genetic variations and clinical and environmental risk factors.

[14] D. Eppstein and M. T. Goodrich, "Straggler identification in round-trip data streams via Newton's identities and invertible Bloom filters," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 2, pp. 297–306, 2011.

[15] M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin, "A cryptographic approach to securely share and query genomic sequences," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 5, 2008.

[16] M. Canim, M. Kantarcioglu, and B. Malin, "Secure management of biomedical data with cryptographic hardware," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 1, 2012.

[17] E. Ayday, J. L. Raisaro, and J. P. Hubaux, "Privacy-enhancing technologies for medical tests using genomic data," *(short paper) in 20th Annual Network and Distributed System Security Symposium (NDSS)*, 2013.

[18] S. Narayan, M. Gagne, and R. Safavi-Naini, "Privacy preserving EHR system using attribute-based infrastructure," *Proceedings of the 2010 ACM Workshop on Cloud Computing Security*, p. 4752, 2010.

[19] S. Alshehri, S. Radziszowski, and R. Raj, "Secure access for healthcare data in the cloud using ciphertext-policy attribute-based encryption," *Proceedings of the 28th International Conference on Data Engineering Workshops (ICDEW)*, pp. 143–146, 2012.

[20] J. Benaloh, M. Chase, E. Horvitz, and K. Lauter, "Patient controlled encryption: Ensuring privacy of electronic medical records," *Proceedings of the 2009 ACM Workshop on Cloud Computing Security*, p. 103114, 2009.

[21] M. Barni, P. Failla, V. Kolesnikov, R. Lazzeretti, A.-R. Sadeghi, and T. Schneider, "Secure evaluation of private linear branching programs with medical applications," *Proceedings of the 14th European Conference on Research in Computer Security ESORICS 2009*, pp. 424–439, 2009.

[22] J. M. Kidd and *et al.*, "Mapping and sequencing of structural variation from eight human genomes," *Nature*, vol. 453, no. 7191, pp. 56–64, May 2008.

[23] http://www.ncbi.nlm.nih.gov/projects/SNP/, Visited on 22/Apr/2013.

[24] E. Bresson, D. Catalano, and D. Pointcheval, "A simple public-key cryptosystem with a double trapdoor decryption mechanism and its applications," *Proceedings of Asiacrypt 03*, pp. 37–54, 2003.

[25] M. Pirretti, P. Traynor, P. McDaniel, and B. Waters, "Secure attribute-based systems," *Proceedings of the 13th ACM Conference on Computer and Communications Security*, pp. 99–112, 2006.

[26] I. Damgård, M. Geisler, and M. Krøigaard, "Efficient and secure comparison for on-line auctions," *Information Security and Privacy*, vol. 4586, pp. 416–430, 2007.

[27] Z. Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft, "Privacy-preserving face recognition," *Proceedings of Privacy Enhancing Technologies*, pp. 235–253, 2009.

[28] The 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 2010.