Privacy-preserving Polygenic Risk Scoring using Homomorphic Encryption

Shaedul Islam[1], Huseyin Demirci[2] and Gabriele Lenzini[2]
[1]University of Luxembourg, FSTC, Luxembourg,
[2]University of Luxembourg, SnT, Luxembourg huseyin.demirci@uni.lu

Abstract
The availability of direct-to-consumer genetic testing services and genome sequencing data bring novel opportunities for applications like genomic risk scoring where a polygenic disease risk score is calculated considering the statistical distribution of the disease associated SNPs. Nowadays, various websites are offering polygenic risk score estimations for various complex diseases. However, these services require the upload of the genomic data to their sites, which is a fairly sensitive personal data. Since, genome data uniquely identifies a person, anonymization is not sufficient alone and may become a threat in the long run. A potential solution is the use of cryptographic techniques along this goal. We propose to deploy homomorphic encryption, a technique which enables to do computation in encrypted data, for a web server providing polygenic risk score estimation. We implemented a proof-of-concept software to measure the performance of such a service with current technology. We also developed a GUI which facilitates the usage of homomorphic encryption for non-technical users. We conclude that recently developed homomorphic encryption libraries enable practical privacy-preserving genomic risk scoring services. Homomorphic encryption is becoming a strong alternative for practical secure privacy-preserving personalized medicine applications.

Keywords
Genomic privacy, Polygenic risk scoring, Homomorphic Encryption

1 INTRODUCTION

With the availability of direct-to-consumer genetic services, new types of web services are becoming available such as genotype imputation and polygenic disease risk prediction. A polygenic risk score (PRS) is an estimation of an individual's tendency to diseases which is calculated by considering statistical distribution of SNPs. Although, PRS cannot be used directly for diagnosis, it provides valuable information for risk stratification, prediction of the drug response or prognosis. To benefit from such services, the upload of the genomic data to the server side is required. On the other hand, genome data is strictly personal and sensitive. Uploading genomic data to a web server leads to privacy issues. There is a requirement for a new generation of genomic services which are privacy-preserving services and compatible with regulations like GDPR and CCPA.

Contribution.

In this work we demonstrate a proof-of concept implementation of a privacy- preserving web server which computes genomic disease risk scores from encrypted data. We accomplish this task by using new Homomorphic Encryption tools. In our use case scenario, the user sends her genomic variants encrypted to the insecure cloud server where all the operations are performed in the encrypted domain. The user gets back the results where she will decrypt the results on the fly. In this way, privacy preserving versions of the existing web applications become practical. This is of critical importance, considering thousands of genomes are now being sequenced each day. This opens the way to get genomic consultancy about complex diseases without revealing sensitive variants and can be considered as a step towards the goal private personalized medicine services.

The source code of the implementation of the case study is available in the GitHub repository:

https://github.com/Shaedul/GenomeAnalysis_PySEAL

2 RELATED WORK

It is well known that genomic data is sensitive: it contains personal and confidential information such as the ancestry of an individual and of his/her kin, and their susceptibility and predisposition to specific diseases such as Alzheimer's, schizophrenia, and cancer. Therefore, the leakage of genomic data leads to irreversible ethnic and social discrimination (e.g., see [12]). Genomic data should be stored, processed, and shared in a privacy preserving manner.

Several methods have been proposed to enable genomics privacy. The most common choice, anonymization, is however provably ineffective in this case [9]. Very little piece of information, like 100 independent SNPs are enough to identify a person uniquely [13] and other sources

The 18th Scandinavian Conference on Health informatics, Tromsø, Norway, August 22-24, 2022. Organized by UiT The Arctic University of Norway. Conference Proceedings published by Linköping University Electronic Press at https://doi.org/10.3384/ecp187.

of information such as social media can be used to link with personal attributes [9]. Therefore, the natural choice is to resort to cryptographic methods. A practice which is also compliant with legal requirements in directives such as the Europe's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), is to encrypt the databases. This step is not enough though to ensure protection from privacy violation: if, for the purpose of processing, encrypted data sets are shared but decrypted before use, the risks of privacy violation remain.

Alternative methods are using cryptographic protocols such as Differential Privacy, Secure Multiparty Computation (SCM) and Homomorphic Encryption (HE). Each method has its own pros and cons and has different use cases in genomics.

Secure multiparty computation allows to collaboratively evaluate a function without revealing information to the parties. Jagadeesh et al. proposed to use SMC to identify diagnosis for monogenic rare diseases while preserving privacy for the remaining variants [11]. Later Akgun et al. improved the performance of the protocol [2]. Cui et al. proposed to use secure function evaluation to implement a privacy-preserving Human Leukocyte Antigen (HLA) matching [6].

Homomorphic encryption technique enables to perform computation inside encrypted data. In this way, computationally expensive operations can be privately outsourced to an insecure party like a public cloud. The existence of a fully homomorphic scheme is first proven by Gentry [8]. The first implementations of fully Homomorphic Encryption were far from being practical. It was considered as much slower than Secure Multiparty computation [11]. However, recent improvements in the implementation techniques and novel libraries enabled to perform practical applications.

Erman et al. used Paillier encryption system to implement a privacy-preserving cardiovascular disease risk analysis [3]. In [15] the authors developed a secure framework to conduct the rare variants analysis with a small sample size. Blatt et al. demonstrated that GWAS analysis of a real data set consisting of 25.000 individuals can be executed practically on encrypted data [4]. Recently, Harmanci et al. developed a secure imputation web server based on homomorphic encryption where untyped variant data is predicted from available genotype data with the help of a reference panel [10]. The applications of HE is becoming more available and practical in genomics area.

2.1. Polygenic risk scoring

Estimating the susceptibility to a disease is invaluable in medicine considering outcomes such as early diagnosis and prevention of common adult-onset conditions. Mendelian traits can point out significant outcomes such as the use of BRCA mutations, but they generally cover a small fraction of the population since they rely on rare variants. However, recent GWAS studies pointed out that, for most complex human diseases, joint consideration of common and low-frequency genetic variants that individually contribute small effects and provide much stronger predictions [14]. Generally, a polygenic risk score is calculated by computing the sum of risk alleles that are weighted by risk allele effect sizes. The effect sizes are estimated using GWAS studies. We refer the interested reader to [5] for a comprehensive tutorial of polygenic risk scoring.

In this paper we utilize the framework of Impute.me [7] for PRS scoring. For specific diseases such as Breast cancer, Type-1 Diabetes, Alzheimer disease and Celiac disease, we have considered the effect sizes provided in this work. For the top-SNP approach, where the authors consider the most significant SNPs, we benefit from the same list of SNPs. In this model the following scores were used:

$PopulationScore_{snp} = Frequency_{snp} \times \beta_{snp}$;
$ZerocenteredScore = \beta_{snp} \times Effect\text{-}allele\text{-}count_{snp} - PopulationScore_{snp}$;
$Z\text{-}score = ZerocenteredScore / \sigma_{population}$

where β is the reported effect size, $Frequency_{snp}$ is the allele frequency for the effect allele and *Effect-allele-countsnp* is the allele count of the genotype data (0,1 or 2) and $\sigma_{population}$ is the standard deviation of the population. This final Z-score is considered as the normalized metric for disease probability.

2.2. Homomorphic encryption

This cryptographic technique enables to perform computation on encrypted data with the help of a Homomorphic property, i.e. for every input plaintext pair x, y, we have:

$Enc_k(x + y) = Enc_k(x) \oplus Enc_k(y)$,

where Enc denote the encryption function, k the encryption key and $\oplus$ the homomorphic addition, respectively. The user encrypts the sensitive data by a public key and sends the encrypted data to an insecure party like a public cloud. Here, the insecure party performs operations in the encrypted data, without reaching the exact values of the data and sends the encrypted results to the user back. Here, the user decrypts the data with the secret key and obtains the desired result in his side. In this way, Homomorphic Encryption (HE) enables private outsourcing of computationally expensive operations to public clouds. This brings great flexibility from the security regulations (like GDPR) points of view since, even in the case of a data breach at the cloud side, no sensitive information is lost.

Although the idea of a HE system existed previously, the existence of a fully homomorphic system was only shown in 2009 by Gentry [8] using lattice algebra. There are different types of Homomorphic Encryption operations such as Partially, Somewhat and Fully Homomorphic systems. Partially Homomorphic systems preserve the operations for one single operation whereas Somewhat Homomorphic systems support homomorphism for two different types operations (i.e. addition and multiplication) for a bounded number of operations. Finally, Fully Homomorphic systems preserve the homomorphism for both types of operations unbounded number of times. We refer the reader to [1] for a comprehensive survey on the theory and implementations of Homomorphic Encryption schemes. The initial implementations of HE were not very practical. It was taking days to make simple calculations. It has been suggested that HE computation is about 5.000-10.000 times slower when compared to other privacy preserving techniques such as Secure Multiparty Computation [11]. Recently, new libraries and

implementations such as Microsoft's SEAL and IBM's Fully Homomorphic Encryption Toolkit For Linux have been made available for developing HE applications. These tools make HE applications more practical and available for a variety of cases including genomics analysis [15,10,4].

3 PRIVACY PRESERVING GENOMIC RISK SCORING

We briefly explain our privacy preserving scheme. When a user, Alice, who owns her genomic variants (as a .vcf file) decides to benefit from the PPPRS (Privacy- preserving Polygenic Risk Score) web server, she carries out the following tasks, respectively:

1. Alice generates a pair of Public and Private keys, denoted by PK and SK, respectively. She sends a copy of the Public Key to the cloud server and keeps the secret key to herself.

2. She requests a PRS for a specific disease from the web server.

3. The web server sends the list of related SNPs which had previously been determined with GWAS studies.

4. Alice encrypts the list of related variants and sends the encrypted list to the cloud.

5. The cloud server performs PRS calculations using encrypted values according to Algorithm and sends the encrypted PRS result to the user.

6. Alice uses her secret key to decrypt her PRS score.

The general outline of this application is depicted in Figure 1.

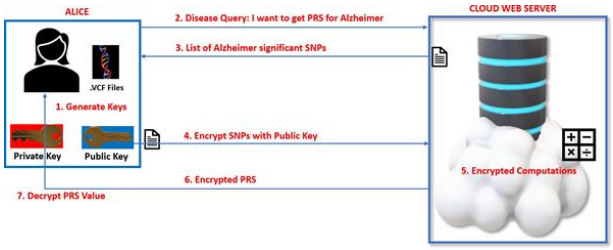

Figure 1. Privacy Preserving PRS.

4 IMPLEMENTATION

4.1. User side

To facilitate the use of the system, we have developed a Graphical User Interface (GUI). The structure of the GUI can be observed in Figure 2. The main aim of the GUI is to facilitate the use privacy preserving operations for the disease risk scoring. The operation of the client site is completed in a few steps. First, the user must register and log in with an anonymous user id and email address. Then, she generates the Public-Secret key pair, where the Public key will be shared with the cloud server to enable to encrypt the data and send to cloud. The user needs to store the secret key privately. Later, the user browses the genome variants data in VCF file format and selects the trait to interpret the disease risk score. After this step, the genomic variants data are encrypted, and the status field shows the progress of the encryption process. Finally, the encrypted data is sent to the cloud for further analysis. After the analysis is performed inside the cloud, the user receives the encrypted results and decrypts in this GUI platform using the Secret key. A the end of the process the user could observe both the calculated normalized Z-score as well as the graphical interpretation of this Z-score.

4.2. Cloud side

We have implemented the homomorphic operations in the cloud site by using Amazon Web Service (AWS), EC2 environment. The cloud server has GWAS reference data in unencrypted form to support the PRS calculation. The reference data obtains the related Chromosome Number, SNPs ID, Effect size, Minor/Allele Frequency (MAF). When the cloud receives any encrypted genome variants data with the request of PRS for a specific disease, then reference data related to the disease will also be encrypted using the user public key. Afterwards, the cloud performs the computation for the risk score and sends the result to the client in encrypted form. We setup the Microsoft PySEAL (Python version of Simple Encrypted Arithmetic Library) library to implement the homomorphic encryption environment inside the cloud.

Computational Performances.

The time performance of the HE experiments is provided in Table 1. These experiments are carried out with a laptop with i-7 1.8- 2.3 GHz. CPU and 16 GB RAM with 64-bit Windows Enterprise operating system. We used Python version 3.7.4 on the Spyder Environment for the implementation.

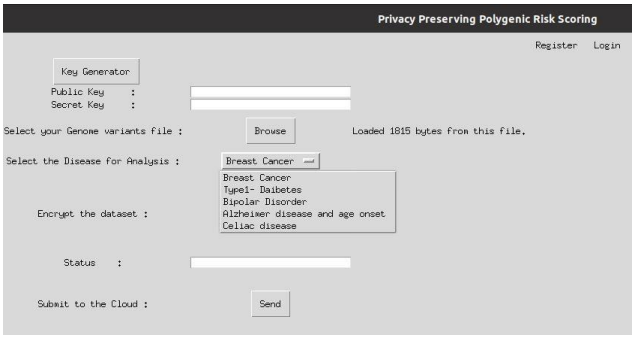

Figure 2. GUI of the Privacy-Preserving PRS Software.

Disease	# SNPs	Enc.	Dec.	Cloud
Celiac	19	1.557	0.007	1.88
Alzheimer	33	2.803	0.012	3.27
T I Diabetes	45	3.564	0.016	4.47
Breast Cancer	635	255.422	0.23	262.83
Whole Genome	3800000	1528	1.406	376.10

Table 1. Time performance of HE operations (seconds).

5 DISCUSSIONS

What we have exposed so far shows that is feasible to conduct secure genomic polygenic risk scoring using homomorphic encryption. If the number of SNPs associated with the disease is around 100, then whole operation can be executed in seconds. We note that another possibility is the encryption of the whole variant set (around 3.5-5 million variants) in the .vcf file once and then carry

out the required operations. In this case, the dominant part of the computation is the encryption part.

6 CONCLUSION

In this work we have presented a proof-of-concept secure web server for the estimation of polygenic risk scores by implementing an algorithm using Homomorphic Encryption (HE). With the recently deployed secure imputation this study fills the gap for privacy preserving genomic analysis. Our study demonstrates that recent HE libraries enable to do risk score estimations without revealing the sensitive genome information. We have conducted experiments for various diseases and number of SNPs.

7 REFERENCES

[1] Aksu, H., Uluagac, A.S., Conti, M.: "A survey on homomorphic encryption schemes: Theory and implementation" in *ACM Computing Surveys* (CSUR) 51(4), 1–35, 2018.

[2] Akgün, M., Ünal, A.B., Ergüner, B., Pfeifer, N., Kohlbacher, O.: "Identifying disease-causing mutations with privacy protection" in Bioinformatics, 2020.

[3] Ayday, E., Raisaro, J.L., McLaren, P.J., Fellay, J., Hubaux, J.P.: "Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data", in: *2013 USENIX Workshop on Health Information Technologies* (HealthTech 13), 2013.

[4] Blatt, M., Gusev, A., Polyakov, Y., Goldwasser, S.: "Secure large-scale genome-wide association studies using homomorphic encryption", in "*Proceedings of the National Academy of Sciences*, 117(21), 11608–11613, 2020.

[5] Choi, S.W., Mak, T.S.H., O'Reilly, P.F.: "Tutorial: a guide to performing polygenic risk score analyses", in *Nature Protocols*, 15(9), 2759–2772, 2020.

[6] Cui, J., Li, H., Yang, M.: "Privacy-preserving computation over genetic data: Hla matching and so on", IACR Cryptol. ePrint Arch. 2019, 1305, 2019.

[7] Folkersen, L., Pain, O., Ingason, A., Werge, T., Lewis, C.M., Austin, J.: "Impute. me: an open-source, non-profit tool for using data from direct-to-consumer genetic testing to calculate and interpret polygenic risk scores" in *Frontiers in genetics* 11, 578 2020.

[8] Gentry, C.: "Fully homomorphic encryption using ideal lattices" in: *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 169–178, 2009.

[9] Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: "Identifying personal genomes by surname inference" in *Science* 339(6117), 321–324, 2013.

[10] Harmanci, A.O., Kim, M.,Wang, S., Li, W., Song, Y.,Lauter, K., Jiang, X.: "Open imputation server provides secure imputation services with provable genomic privacy" *bioRxiv*, 2021.

[11] Jagadeesh, K.A., Wu, D.J., Birgmeier, J.A., Boneh, D., Bejerano, G.: "Deriving genomic diagnoses without revealing patient genomes" in *Science* 357(6352), 692–695, 2017.

[12] Naveed, M., Ayday, E., Clayton, E.W., Fellay, J., Gunter, C.A., Hubaux, J.P., Malin, B.A., Wang, X.: "Privacy in the genomic era" in *ACM Comput. Surv.* 48(1) , Aug 2015.

[13] Pakstis, A.J., Speed, W.C., Fang, R., Hyland, F.C., Furtado, M.R., Kidd, J.R., Kidd, K.K.: "Snps for a universal individual identification panel" in *Human genetics* 127(3), 315–324, 2010.

[14] Torkamani, A., Wineinger, N.E., Topol, E.J.: "The personal and clinical utility of polygenic risk scores" in *Nature Reviews Genetics* 19(9), 581–590 2018.

[15] Wang, S., Zhang, Y., Dai, W., Lauter, K., Kim, M., Tang, Y., Xiong, H., Jiang, X.: "Healer: homomorphic computation of exact logistic regression for secure rare disease variants analysis in GWAS" in *Bioinformatics* 32(2), 211–218, 2016.

8 ACKNOWLEDGEMENT

This work has been supported by the EU 956562, MSCA-ITN-2020 - Innovative Training Networks, "Legality Attentive Data Scientists" (LeADS) project. The authors wish to thank Dr. Patrick May for valuable discussions on the topic of polygenic risk calculation.