Research Proposal for Privacy-preserving Genome-Wide Association Studies (GWAS)

Wentao Li
BMI 7301 Grant Writing
The School of Biomedical Informatics | The University of Texas Health Science Center at Houston

Acknowledgment

This research is supported by Dr. Xiaoqian Jiang, Dr. Arif Harmanci and Dr. Han Chen at UTHealth.

Scan the QR code for the recorded presentation and Reference

Introduction

Unlike passwords or other forms of personal identification, an individual's DNA is something that cannot be changed. Any genetic information that is shared or leaked has the potential to remain public and accessible indefinitely. Therefore, Governments have recognized the importance of protecting genetic privacy and have implemented laws and regulations to govern the use and storage of genetic information (Table 1).

GWAS requires a large number of genetic data from different and diversified cohorts to draw significant conclusions (Hong, E. P., & Park, J. W., 2012). Hence, the dilemma of data utility and security needs to be addressed. One possible solution for balance is Federated Learning (Fig 1).

Aim 1: Literature reviews on privacy-preserving GWAS

A literature search was conducted in the PubMed database on March 09, 2023, at 11:50 CDT. MeSH terms include “Genome-Wide Association Study”, “Medical Informatics”, “public health informatics”, and “confidentiality”; and Text words include “federated learning” and “distributed learning”. Moreover, the query of the search is shown below.

- *((((Genome-Wide Association Study[MeSH Terms]) OR (Medical Informatics[MeSH Terms]) OR (public health informatics[MeSH Terms])) AND (privacy[MeSH Terms]) OR Confidentiality[MeSH Terms]) AND ((federated learning[Text Word])) OR (distributed learning[Text Word]))*

The query returned 215 papers and after a deliberate screening of the paper published within 5 years (2018-2023), I located 5 related works as the literature result (Table 2). However, these related researches did not produce statistically significant scores in terms of P-values to analyze the genetic data. Also, only a few of them can address the confounding effects caused by location bias or kinship relationship bias.

Aim 2: Address confounding effects with Federated Learning

There are two main types of confounding effects that can potentially draw down the performance of privacy-preserving GWAS: Location-wise bias and Family kinship relationship bias (Fig 2). The related methods do not consider these biases and consider all individuals are independent of each other.

Hence, we developed two federated learning methods, Federated Generalized Linear Mixed Models (FedGLMM) (Li, W, et al., 2022) and Federated Generalized Linear Mixed Model Association Tests (FedGMMAT). The results (P-values of SNPs) of the two federated methods showed little difference from a centralized method that “pooled” all data without the concerns of privacy.

FedGLMM was developed with two approximation methods: Laplace Approximation (LA) with less accuracy and higher speed, and Gauss-Hermite (GH) Approximation with higher accuracy but slower speed. The results are shown in Fig 4; FedGMMAT was developed based on the R package “GMMAT” (Chen, et al., 2016), and the performance results are shown in Fig 3.

Aim 3: Develop a Privacy-preserving Federated Platform

Develop a web-browser-based federated learning platform that contains common models in healthcare data analysis. The platform will be designed to facilitate the training of various models on decentralized data sources while ensuring the privacy and security of the data.

Tables

Table 1. **Enacted laws that protect genetic information in the US**

Laws and Regulations that protect individual Genetic data	Year enacted
Genetic Information Nondiscrimination Act (GINA)	2008
Health Insurance Portability and Accountability Act (HIPAA), Omnibus Rule amended HIPAA regulations on genetic information	2013

Table 2. **Related works in Privacy-preserving GWAS**

Methods	Year	Model protection	Statistical test(s)	Can address confounding effects
sPLINK	2020	Secret sharing	Chi-square, linear	No
Federated PCA	2021	Homomorphic encryption	No	No
SAFETY	2018	Homomorphic encryption	LD, HWE, **CATT**, **FET**	Yes (but not proved)
DyPS	2021	Trusted Execution Environment	Chi-square	No
cGLMM	2020	No	No statistical test	Yes

LD: Linkage Disequilibrium; **HWE**: Hardy-Weinberg Equilibrium; **CATT**: Cochran-Armitage Test for Trend; **FET**: Fisher’s Exact Test

Figures

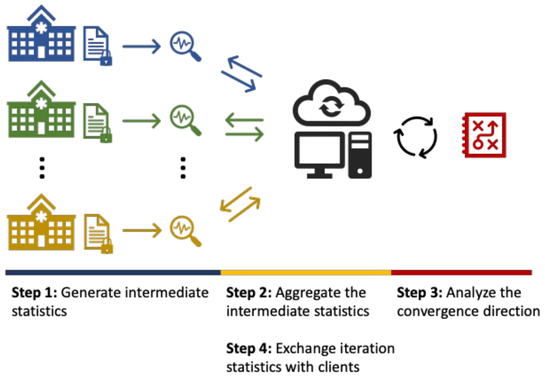

Figure 1. **Federated Learning.** The Federated Learning technique can analyze data from different repositories without sharing the original data. Only model information is being communicated during the training process.

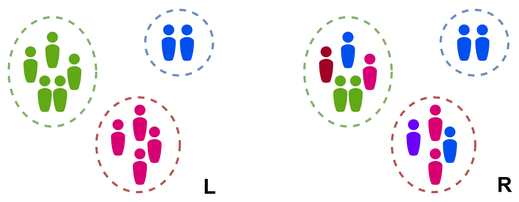

Figure 2. **Two types of confounding effects.** Left (L): Location-wise bias, individuals within the same location are likely to share common traits; Right (R): Family kinship relationships bias, family members are scattered.

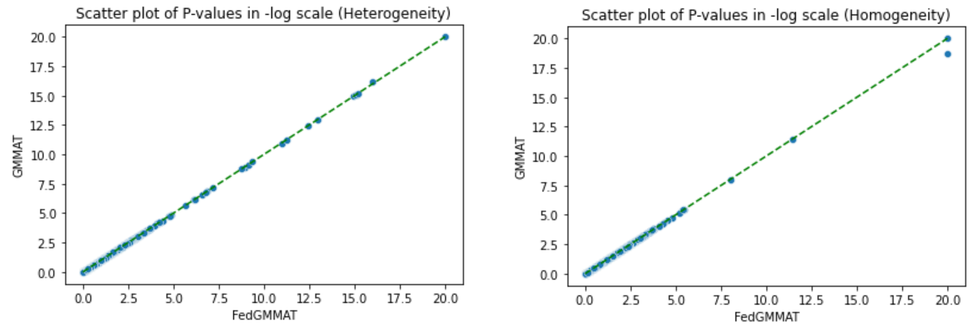

Figure 3. **Performance of FedGMMAT.** Scatter plots of P-values between privacy-preserving method FedGMMAT and centralized method GMMAT. Tested on homogeneous populations and heterogeneous population datasets from dbGAP.

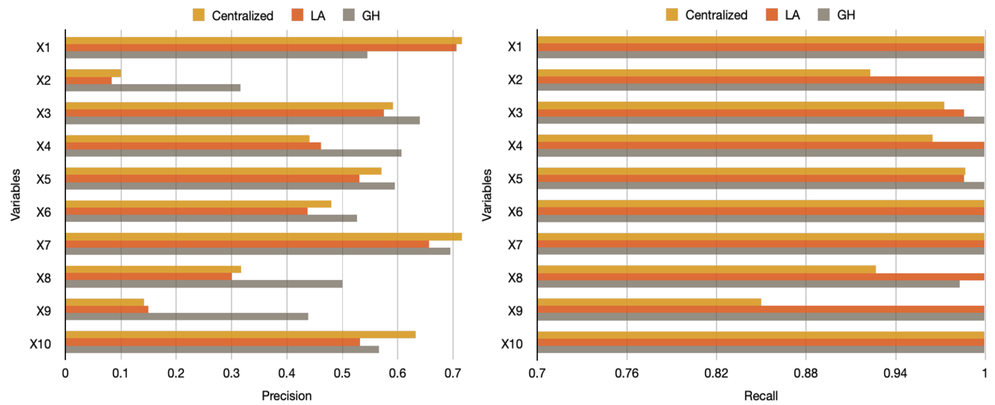

Figure 4. **Performance of FedGLMM.** Left (L): Precision of SNPs’ P-values among centralized methods and LA/GH-based FedGLMM methods; Right (R): Recall of SNPs’ P-values among centralized methods and LA/GH-based FedGLMM methods.

Please contact via email: wentao.li@uth.tmc.edu