Federated Generalized Linear Mixed Model on Horizontally Partitioned Data

Wentao Li, Ph.D. Student

The School of Biomedical Informatics |The University of Texas Health Science Center at Houston

Please contact via email: Wentao.Li@uth.tmc.edu

ACKNOWLEDGEMENT:
Dr. Susan H. Fenton & Dr. Xiaoqian Jiang

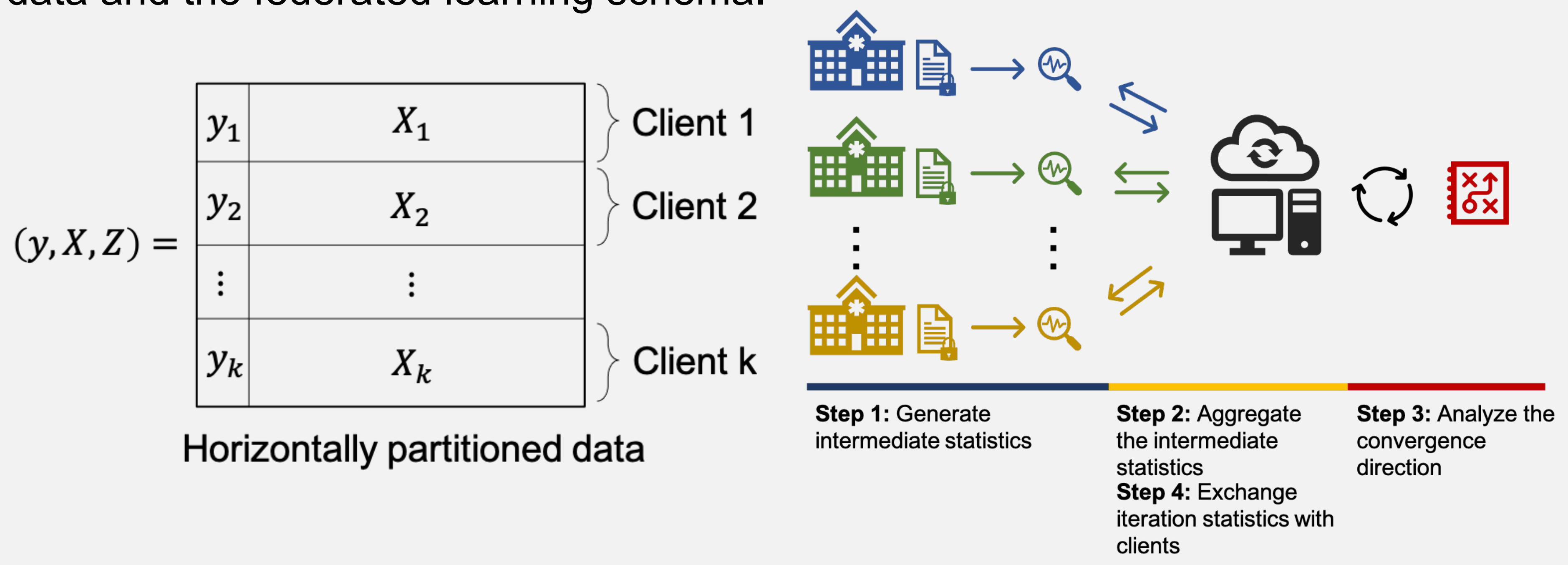
INTRODUCTION

Due to privacy protection regulations (i.e. HIPAA), data isolation and heterogeneity are prevalent phenomena in the medical field. But researchers and their models are eager for more data (Obermeyer & Emanuel, 2016). And this conflict may result in more privacy leaks and compromised analytic performance (Jin et al., 2019).

This research aims to develop a privacy-preserving machine learning technique that can bridge the gaps between isolated data holders and researchers. Instead of transmitting privacy data in traditional learning, such a technique communicates with the model intermediate data. Thus, analyses can conduct on many isolated data repositories without risking privacy and local information.

FEDERATED LEARNING & HORIZONTALLY PARTITIONED

Federated learning can train a global model in multiple distributed local datasets by communicating local model intermediate data. This research will focus on the horizontally partitioned data scenario, which assumes the distributed data repositories are sharing the same observed/unobserved features and different samples. The following graph shows a demonstration of horizontally partitioned data and the federated learning schema.



$$(y, X, Z) = \begin{array}{|c|c|} \hline y_1 & X_1 \\ \hline y_2 & X_2 \\ \hline \vdots & \vdots \\ \hline y_k & X_k \\ \hline \end{array}$$

Client 1
Client 2
Client k

Horizontally partitioned data

Step 1: Generate intermediate statistics

Step 2: Aggregate the intermediate statistics

Step 4: Exchange iteration statistics with clients

Step 3: Analyze the convergence direction

GENERALIZED LINEAR MIXED MODEL

Isolated data repositories are considered to contain localized information in the data. For example, hospitals in different locations are serving their local patients with some commonly shared traits (i.e. incomes, education, race, etc.). Thus, conducting data analyses by ignoring local repositories' information is biased. And these effects will jeopardize the validity of data analyses research.

Generalized Linear Mixed Model (GLMM) is one solution to this problem. The method embedded mixed-effects to normal Generalized Linear Models, which will have biased local information into consideration. Each isolated data repository will "learn" unique local parameters while agreeing on other globally shared ones.

To achieve the goal, our research has utilized Laplace (LA) and Gauss-Hermite (GH) approximation on a non-tractable problem. Then, the model training process (updating parameters by calculating gradients and Hessian Matrices) can distribute to each local data repository. And no individual data are shared.

TABLES & CHARTS

Figure 1: Precision

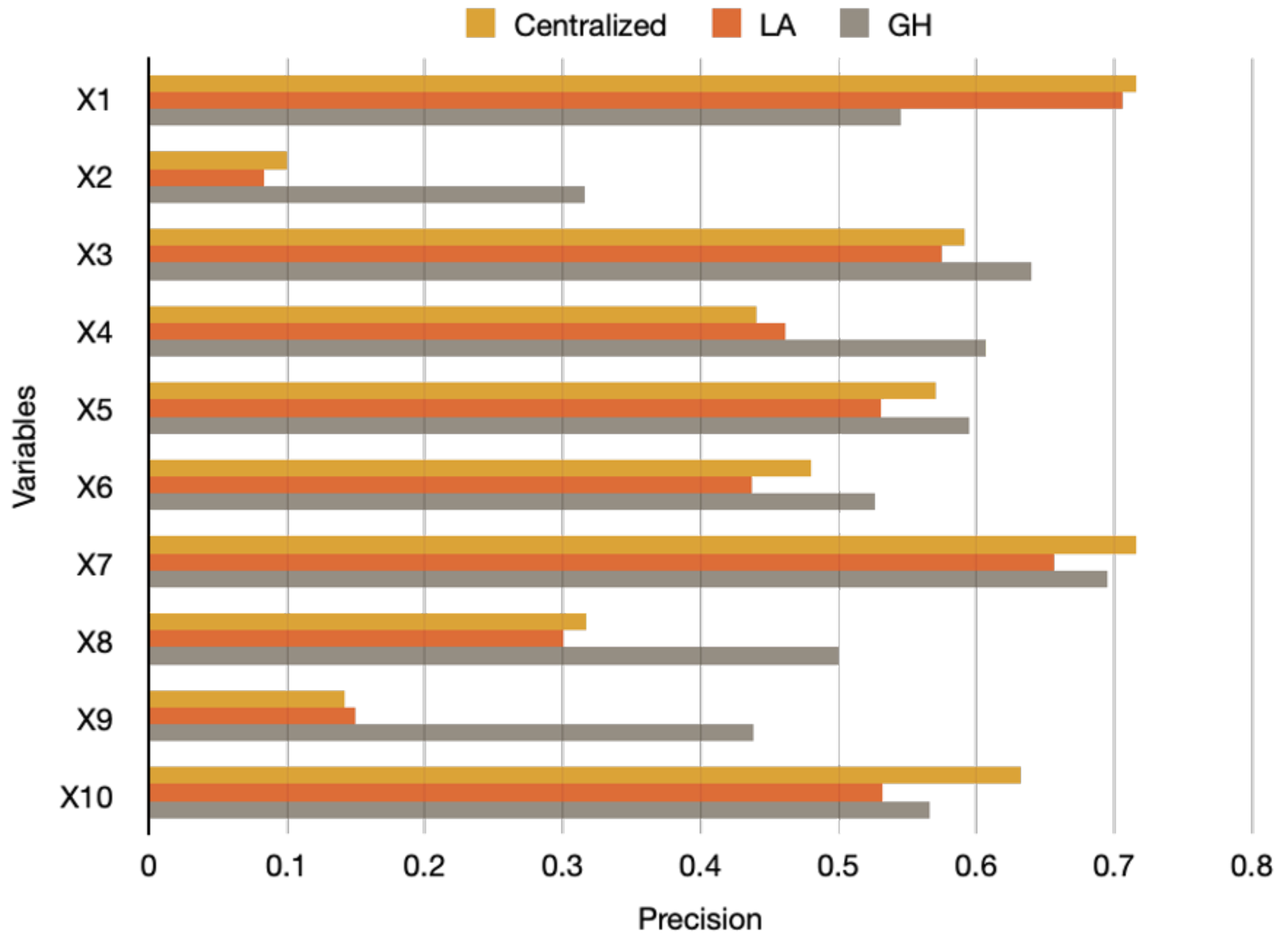


Figure 2: Recall

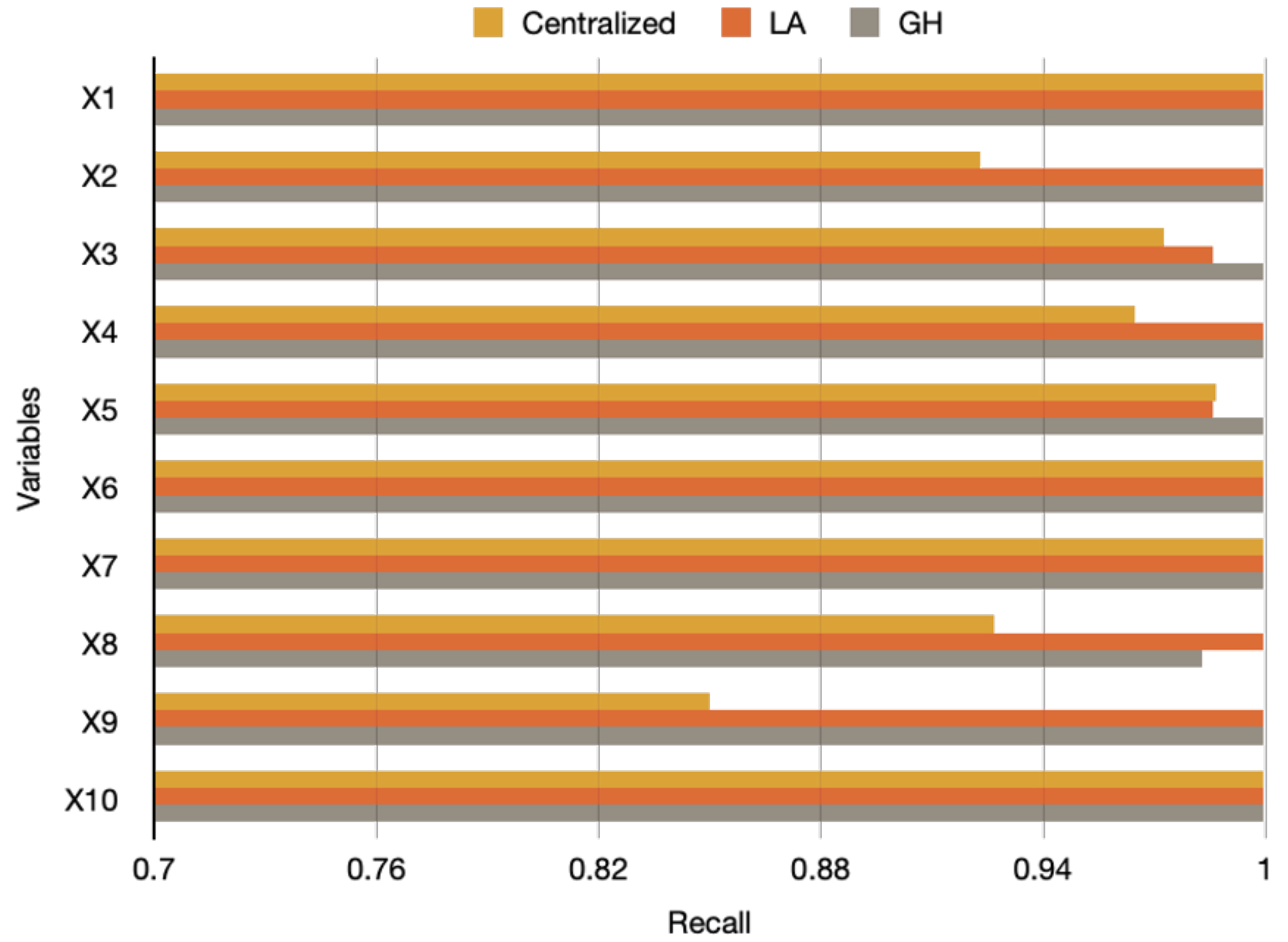


Table 1: The summary of synthetic data

Setting	Number of local sites	Sample size in each site	Variance
1	2	500	small
2	2	500	large
3	10	500	small
4	10	500	large
5	2	30	small
6	2	30	large
7	10	30	small
8	10	30	large

Table 2: The convergence rate

	LA		GH	
Setting	Steps	Runtime (s)	Steps	Runtime (s)
1	22.875 (21.623)	47.953 (20.513)	34.850 (9.213)	104.460 (10.614)
2	21.500 (21.977)	40.947 (36.466)	35.000 (8.711)	100.940 (19.940)
3	29.867 (31.719)	108.931 (65.486)	34.900 (6.138)	1259.285 (231.956)
4	27.846 (24.034)	84.343 (76.502)	36.650 (6.310)	1342.695 (250.603)
5	59.722 (42.057)	10.631 (3.945)	33.750 (10.146)	12.568 (2.116)
6	67.188 (48.994)	10.499 (4.054)	31.400 (11.081)	11.430 (3.064)
7	96.286 (53.635)	96.501 (38.632)	37.450 (3.818)	369.165 (41.998)
8	116.083 (46.479)	91.304 (62.410)	37.150 (4.295)	309.693 (36.621)

METHODS & EXPERIMENTS

To measure the performance of the proposed methods, we are interested in the significance of features in GLMM. So we generated 8 different settings of synthetic data with details in Tab1. The number of local sites will imitate the isolated local data repositories. The sample size is all the same in each setting across data repositories. And the variance is also introduced to test the robustness of the results.

In each data setting, two federated methods (LA approximation and GH approximation) will compare with a non-federated method (denoted as Centralized) on the significance of 10 features in the dataset. The goal of the experiment is to see which federated method will have better performance and to have the non-federated model as a benchmark. Fig 1 shows the precision of each model, and Fig 2 shows the recall.

CONCLUSION & DISCUSSION

Both the federated methods achieve the baseline set by the non-federated setting (denoted in Centralized).

Under federated learning settings, performance results showed Gauss-Hermite approximation overperformed the Laplace approximation both in precision and recall. However, the Gauss-Hermite method took much longer times to converge compared with the Laplace method (Tab 2). That is because the Gauss-Hermite approximation requires more computations.

REFERENCES

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine, 375(13), 1216.

Jin, H., Luo, Y., Li, P., & Mathew, J. (2019). A review of secure and privacy-preserving medical data sharing. IEEE Access, 7, 61656-61669.