Statistica Sinica Preprint No: SS-2024-0215

Title	Distributed Sequential Federated Estimation
Manuscript ID	SS-2024-0215
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0215
Complete List of Authors	Zhanfeng Wang, Xinyu Zhang and Yuan-chin Chang
Corresponding Authors	Yuan-chin Chang
E-mails	ycchang@sinica.edu.tw

Notice: Accepted author version.

Distributed sequential federated estimation

Zhanfeng Wang, Xinyu Zhang, Yuan-chin Ivan Chang

School of Management, University of Science and Technology of China, Hefei, China

Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan

Abstract: When analyzing data stored across multiple sites, concerns about data security and communication arise. Federated learning, which avoids centralizing data, offers a promising solution to address these concerns. However, integrating information from separate local sites in a statistically sound manner is crucial, as common averaging methods may lead to information loss due to data non-homogeneity and incomparable results among sites. By applying sequential methods in federated learning, integration can be facilitated and the analysis process can be accelerated, particularly within a distributed computing framework. We propose an efficient data-driven method that maintains the principles of classical sequential adaptive design. Numerical studies and an application to COVID-19 data from 32 hospitals in Mexico, using a regression model, illustrate the effectiveness of our approach.

Key words and phrases: Adaptive sampling; Data Communication; Random average; Sequential sampling.

*Yuan-chin Ivan Chang, ORCID ID: 0000-0002-4977-7721

1. Introduction

The centralization of data from multiple sites poses challenges in transport, communication, and security (Damiani et al., 2015; Huang et al., 2020). Federated learning enables decentralized model training but is often addressed from a technical perspective, overlooking key statistical challenges (Yan et al., 2013; Jordan et al., 2019; Li et al., 2018). A major issue is data heterogeneity, where site-specific variations make conventional aggregation methods like weighted averaging ineffective (McMahan et al., 2017). In addition, site-specific sample sizes are often ignored, affecting parameter estimation and prediction accuracy. To overcome these limitations, we propose a distributed sequential estimation framework that optimally determines sample sizes while ensuring statistical efficiency. By integrating sequential estimation into federated learning, our method enhances parameter accuracy and model performance across heterogeneous datasets.

Non-homogeneous data arise when the collected variables differ across sites beyond common variables of interest. This is common in large surveys, such as epidemiology and social sciences (Carlini et al., 2019), as seen in the COVID-19 data set used in this study. This variability creates uncertainty in variable selection and sample representativeness. Using COVID-19 data from 32 Mexican health sectors, we investigate whether diabetes or obesity increases the risk of infection, alongside other variables. To address the heterogeneity of the data that arises mainly from variations in site-specific characteristics, we propose a federated learning-based parameter estimation method that integrates distributed computing (Yu et al., 2022, 2025) and sequential estimation for improved accuracy.

To the authors' knowledge, existing literature lacks discussion on integrating results from multiple sites with random sample sizes, especially for prediction models. This study ensures precision and coverage probability while incorporating a prediction criterion to improve the accuracy of the model. Sequential analysis is applied locally to optimize data usage while preserving statistical properties. Unlike conventional weighted averaging, our approach determines sample sizes dynamically based on data quality and quantity, addressing key challenges in federated learning and handling non-homogeneous variables in sequential sampling. In addition, we employ statistical experimental design criteria to develop an adaptive sampling strategy for the proposed federated sequential learning method; For design-inspired subsampling methods, see Wang et al. (2018, 2019); Ai et al. (2021); He et al. (2024); Yao and Wang (2021); Yu et al. (2024).

The remainder of this paper is organized as follows: Section 2 introduces the distributed sequential federated estimation approach. Section 3 presents numerical results based on simulated data and COVID-19 data from 32 hospitals in Mexico. Finally, Conclusion section summarizes our key findings, with technical proofs and additional numerical results provided in Supplementary Materials.

2. Methodology

We illustrate the proposed method by applying a logistic regression model to COVID-19 data for classification. Consider M data sites, where the site j has n_j independent observations (y_{ji}, x_{ji}), with response y_j and covariate x_j. The data follow a generalized linear model (GLM, McCullagh and Nelder (1989)) with link function μ such that

$$E(y_j|x_j) = \mu(x_j^\top \beta_j), \tag{2.1}$$

$$\mathrm{Var}(y_j|x_j) = \nu(x_j^\top \beta_j) > 0, j = 1, \ldots, M, \tag{2.2}$$

where β_j is an unknown parameter vector. Let $x_j = (u^\top, v_j^\top)^\top$, and $\beta_j = (\theta^\top, \eta_j^\top)^\top$, where θ denotes the parameter of the common variable u at all sites, and v_j is a site-specific variable that may vary in length. Thus, for $j = 1, \ldots, M$,

$$E(y_j|x_j) = \mu(\theta^\top u + v_j^\top \eta_j) \tag{2.3}$$

$$\mathrm{Var}(y_j|x_j) = \nu(\theta^\top u + v_j^\top \eta_j) > 0. \tag{2.4}$$

For logistic regression, the mean and variance functions simplify to

$$\mu(\theta^\top u + v_j^\top \eta_j) = \frac{\exp\left(\theta^\top u + v_j^\top \eta_j\right)}{1 + \exp\left(\theta^\top u + v_j^\top \eta_j\right)}, \tag{2.5}$$

$$\nu(\theta^\top u + v_j^\top \eta_j) = \mu(\theta^\top u + v_j^\top \eta_j)(1 - \mu(\theta^\top u + v_j^\top \eta_j)). \tag{2.6}$$

This formulation enables robust estimation of the common parameter while accommodating data heterogeneity across distributed sites.

2.1 Federated sequential learning

Many classical "average-like" methods, such as voting schemes, weighted approaches, and robust statistical techniques, are widely used to integrate results from multiple sites when sample sizes are predetermined. However, in non-homogeneous data settings, fixed sample size strategies become impractical, leading to insufficient statistical information, especially when large variations exist due to site-specific data collection. Thus, conventional methods may not be suitable from a statistical perspective. Although sequential methods are commonly applied in scenarios like clinical trials where prefixed sample sizes are impractical, their sample efficiency and statistical robustness make them a strong alternative for integrating multi-site results. Instead of relying on predefined sample sizes, we prioritize statistical properties such as accuracy and coverage probability. In addition to these, we incorporate a prediction criterion in our sequential estimation procedure, tailored to the nature of the response variable.

For logistic regression models, we introduce the area under the receiver operating characteristic curve (AUC) as a classification performance metric in the sequential confidence set estimation. The stopping criterion is determined by the coverage probability, the precision of the confidence set, and the AUC, resulting in random stopping times and site-specific sample sizes. Naturally, variations in sample sizes increase as site heterogeneity increases. Following the notations above, we employ confidence set estimation to achieve a desired level of accuracy for the parameters $\boldsymbol{\theta}$, of interest in the context of generalized linear models. This study focuses on integrating results from M data sites to ensure final estimates with desirable properties, similar to conventional sequential procedures. We use a fixed-size confidence set estimation to illustrate this approach. By independently conducting M estimation procedures without a centralized data center, our method maintains key federated learning principles, such as preserving data privacy and reducing communication costs. We first describe the individual sequential procedure for data site j, followed by the integration of results across all M sites.

2.2 Sequential estimation with reserved parameter estimation precision and model prediction accuracy

Let $\mathcal{D}_j$ denote the data set of site j, and $C_{jk} = \{(y_{ji}, \boldsymbol{x}_{ji}), i = 1, ..., k\}$ be the subset of randomly recruited data of $\mathcal{D}_j$ up to the sampling stage k of the jth site, $j = 1, \cdots, M$. Then the maximum quasi-likelihood estimate (MQLE) for $\boldsymbol{\beta}_j$ at the kth stage (McCullagh and Nelder, 1989), say $\tilde{\boldsymbol{\beta}}_{jk} = (\tilde{\boldsymbol{\theta}}_{jk}^\top, \tilde{\boldsymbol{\eta}}_{jk}^\top)^\top$, is a solution to the estimation equation:

$$ln(\tilde{\boldsymbol{\beta}}_{jk}) \equiv \sum_{i=1}^{k} \dot{\mu}(\boldsymbol{x}_{ji}^\top \tilde{\boldsymbol{\beta}}_{jk}) w(\boldsymbol{x}_{ji}^\top \tilde{\boldsymbol{\beta}}_{jk})[y_{ji} - \mu(\boldsymbol{x}_{ji}^\top \tilde{\boldsymbol{\beta}}_{jk})]\boldsymbol{x}_{ji} = 0, \qquad (2.7)$$

where $\dot{\mu}(t) = d\mu(t)/dt$ is the first derivative of $\mu(t)$ and $w(t) = \nu^{-1}(t)$. Following the notations defined before, and let $\boldsymbol{L}_j$ be a $p_0 \times p_j$, $j = 1, \cdots, M$ diagonal matrix with diagonal elements $\text{diag}\{I_{j1}, \cdots, I_{jp_j}\}$, where $I_{j1} = \cdots = I_{jp_0} = 1$ and $I_{jk} = 0, k = p_0 + 1, \cdots, p_j$, and p_0 denotes the number of the common variables of interest among sites. Then $\tilde{\boldsymbol{\theta}}_{jk} = \boldsymbol{L}_j \tilde{\boldsymbol{\beta}}_{jk}$. Assume

(A1) $\sup_{i \le k} ||\boldsymbol{x}_{ji}||_2 < \infty$ for all j, and $E|\epsilon_{ji}|^\zeta < \infty$ with some $\zeta > 2$, where $\epsilon_{ji} = y_{ji} - \mu(\boldsymbol{x}_{ji}^\top \boldsymbol{\beta}_{j0})$ is the error term and $\boldsymbol{\beta}_{j0}$ is the true value of $\boldsymbol{\beta}_j$.

(A2) $\lim_{k\to\infty} \sum_{i=1}^k \boldsymbol{x}_{ji}\{\dot{\mu}(\boldsymbol{x}_{ji}^\top \boldsymbol{\beta}_{j0})^2 / \nu(\boldsymbol{x}_{ji}^\top \boldsymbol{\beta}_{j0})\}\boldsymbol{x}_{ji}^\top / k = \boldsymbol{\Sigma}_j$, where $\boldsymbol{\Sigma}_j$ is a positive definite matrix.

MQLE $\tilde{\boldsymbol{\beta}}_{jk}$ is shown to be a strong consistent estimate of $\boldsymbol{\beta}_j$(Chang, 1999), and $\sqrt{n}(\tilde{\boldsymbol{\theta}}_{jk} - \boldsymbol{\theta}_0) \longrightarrow N(0, \boldsymbol{L}_j \boldsymbol{\Sigma}_j^{-1} \boldsymbol{L}_j^\top)$ in distribution as $k \to \infty$, where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$. For classification purposes, we apply the proposed method to logistic regression models. For each j, let A_j be its corresponding AUC of the jth logistic model. Let $\hat{A}_j = \hat{A}_{jk}$ and $v_{Aj} = v_{Ajk}$ be strongly consistent estimates of A_j, and its variance, respectively. Let $\hat{y}_{ji}^k = \mu(\boldsymbol{x}_{ji}^\top \tilde{\boldsymbol{\beta}}_{jk})$ denote the fitted values of y_{ji} when using the data set C_{jk}. Denoted by $S_1 = S_{1jk} = \{\hat{y}_{ji}^k : y_{ji} = 1\}$ and $S_0 = S_{0jk} = \{\hat{y}_{ji}^k : y_{ji} = 0\}$. Let k_0 and k_1 be the sizes of S_0 and S_1, respectively. For a logistic regression model as in (2.5),

$$\hat{A}_j = \frac{1}{k_0 k_1} \sum_{v_1 \in S_1} \sum_{v_2 \in S_0} I(v_1 \geq v_2), \tag{2.8}$$

is an estimate of the AUC using the data set C_{jk}, where $I(\cdot)$ is an indicator function (see Zhou et al. (2009)). It follows that $(\hat{A}_j - A_j)/\sqrt{v_{Aj}}$ converges in distribution to $N(0, 1)$ as k tends to ∞.

2.2.1 Sequential procedure

Let C_{jk_0} be the initial data set of size $k_0 > 0$ for data site j, and let a be the square root of the $1 - \alpha$ quantile of a chi-square distribution with p_0 degrees of freedom, $\chi^2_{p_0}$. Let $\tilde{a}_j > 0$, for $j = 1, \cdots, M$, be a sequence of real numbers such that $\sum_{j=1}^M \tilde{a}_j^2 = a^2$. These $\tilde{a}_j$ values can be determined according to users and/or depending on other information for specific sites; for example, if jth site has a small sample size, then it can usually provide less information for our analysis purposes, and a small $\tilde{a}_j$ could be assigned to it. However, if there is no preference, then we can simply set $\tilde{a}_j^2 = a^2/M$. We show that different values of $\{\tilde{a}_j : j = 1, \cdots, M\}$ do not affect the statistical properties of the final parameter estimation.

Let $\mu_{jk} = \lambda_{max}[k\boldsymbol{L}_j\boldsymbol{\Sigma}_{jk}^{-1}\boldsymbol{L}_j^\top]$, where $\lambda_{max}(\boldsymbol{A})$ denotes the maximum eigenvalue of matrix $\boldsymbol{A}$, and $\boldsymbol{\Sigma}_{jk} = \sum_{i=1}^k \boldsymbol{x}_{ji}\{\dot{\mu}(\boldsymbol{x}_{ji}^\top\tilde{\boldsymbol{\beta}}_{jk})^2/\nu(\boldsymbol{x}_{ji}^\top\tilde{\boldsymbol{\beta}}_{jk})\}\boldsymbol{x}_{ji}^\top$. For $j = 1, \cdots, M$, define

$$\tilde{N}_j = N_{d_1,d_2} \equiv \inf\left\{k : k \geq k_0 \text{ and } \mu_{jk} \leq \frac{d_1^2 k}{\tilde{a}_j^2} \text{ and } v_{A_j} \leq \left(\frac{d_2}{a_p}\right)^2\right\}, \quad (2.9)$$

where v_{A_j} is a variance estimate of $\hat{A}_j$, d_1 and d_2 are two pre-chosen positive constants for the pre-specified estimation precision of $\boldsymbol{\theta}$ and AUC, respectively, a_p is the $1-\alpha/2$ quantile of the standard normal distribution, $N(0,1)$, and k_0 is a size of small initial data set for each sequential procedure. The condition $\mu_{jk} \leq (d_1^2 k)/\tilde{a}_j^2$ in (2.9) is to ensure the precision of parameter estimation, while $v_{Aj} \leq (d_2/a_p)^2$ is to preserve the classification prediction accuracy of the model. That is, $\tilde{N}_j$ denotes the stopping rule for site j, where the sampling procedure for site j stops when the stopping criterion of $\tilde{N}_j$ is satisfied. Thus, the sample size of site j depends on the estimates of the regression parameters and AUC of the jth logistic model via the included data, and therefore is random.

The initial sample size of k_0 may vary across sites, and its choice is a subject of ongoing debate. However, we also know that with smaller values of d_1 and d_2, we tend to have larger sample size in order to fulfill the inequalities in (2.9). And therefore stopping time $\tilde{N}_j$ tends to stop at a larger number, which also enables the jth sequential procedure to have a more precise estimation of $\tilde{\boldsymbol{\theta}}_{j\tilde{N}_j}$ and a better prediction accuracy with A_j. In general, we require the initial sample set that contains samples with both $y = 0$ and 1, which only concerns convergence of the numerical algorithm to calculate the estimate of the parameters in the logistic model. Because k_0 is usually small compared to the final samples used, the bias introduced by this initial set is not significant. Generally, the choices of d_1 and d_2 depend on demand of practical application, that is, how accurate estimation of the parameters is needed.

Suppose that we are at the $(k-1)$st stage, and have recruited $k-1$, $k > k_0$ samples. If the inequalities for $\tilde{N}_j$ are satisfied with data set C_{jk-1}, then we stop recruiting and save the current results. Otherwise, we select an additional sample from data site j, and update the estimates $\tilde{\boldsymbol{\beta}}_{jk} = (\tilde{\boldsymbol{\theta}}_{jk}^{\top}, \tilde{\boldsymbol{\eta}}_{jk}^{\top})^{\top}$, μ_{jk} and v_{Aj} using data in C_{jk}. And this recruiting procedure is repeated until the inequalities in $\tilde{N}_j$ are satisfied. Then following Chang (2011), we show that the parameter estimates for the generalized linear model have uniform continuity in probability (u.c.i.p.) property (Woodroofe, 1982). Moreover, the property u.c.i.p. implies that the estimates are asymptotically normally distributed as the sample size goes to infinity. Thus, for data site j, $\tilde{\boldsymbol{\theta}}_{j\tilde{N}_j}$ and $\hat{A}_j$ have the following asymptotic properties: as d_1 and $d_2 \to 0$, $\sqrt{\tilde{N}_j}(\tilde{\boldsymbol{\theta}}_{j\tilde{N}_j} - \boldsymbol{\theta}_0) \longrightarrow N(0, \boldsymbol{L}_j \boldsymbol{\Sigma}_j^{-1} \boldsymbol{L}_j^\top)$ in distribution , and $(\hat{A}_j - A_j)/\sqrt{v_{Aj}} \longrightarrow N(0,1)$ in distribution.

Remark 1. For a sequence of random variables, $\{z_m, m \geq 1\}$, if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $P\left\{\max_{0\leq k\leq m\delta} |z_{m+k} - z_m| \geq \varepsilon\right\} < \varepsilon$, for all $m \geq 1$, then the sequence $\{z_m, m \geq 1\}$ is uniform continuity in probability (u.c.i.p.). The u.c.i.p. (Woodroofe, 1982) is a sufficient condition such that the randomly stopped sequence has the same asymptotic distribution as the fixed sample size estimate.

By independently conducting M estimation procedures across M data sites, each sequentially recruits samples without replacement using local computing, constructing confidence sets of prefixed size for $\boldsymbol{\theta}$ in (2.5). This eliminates communication and security concerns from an IT perspective. Each estimation procedure ensures the pre-specified precision via sequential fixed-size estimation, enabling statistical integration into a final result with desired properties. Although naive averaging suffices for fixed sample sizes, integrating random sample sizes is not trivial, and classification performance must be preserved when combining results. We now describe the proposed federated learning procedure. In particular, $\tilde{a}_j^2$ depends only on p_0, the number of the common variables, and satisfies $\sum_{j=1}^{M} \tilde{a}_j^2 = a^2$. Leveraging this constraint, we control sample proportions across sites, ensuring proper allocation based on data quality and collection status.

2.2.2 Federated estimation

When all M sampling procedures stop, let $\hat{N}$ and $\hat{\boldsymbol{\theta}}$ denote the size of total samples and the estimate for the integrated procedure as follows:

$$\hat{N} = \sum_{j=1}^{M} \tilde{N}_j \text{ and } \hat{\boldsymbol{\theta}} = \sum_{j=1}^{M} \rho_j \tilde{\boldsymbol{\theta}}_{j\tilde{N}_j}, \tag{2.10}$$

where $\hat{N}$ is an integer-valued random variable, and $\hat{\boldsymbol{\theta}}$ is a weighted average estimate for $\boldsymbol{\theta}_0$ with “random weights” $\rho_j = \tilde{N}_j/\hat{N}$, $j = 1, \cdots, M$. Thus, the proposed “integrate procedure” focuses on variables of interest and allows non-homogeneity variables, while taking into account both the precision of the estimate and the precision of prediction of a model. (Note that in Chen et al. (2023), they only consider homogeneity data.)

Proposition 1. *Assume that* $\{(\boldsymbol{x}_{ji}, y_{ji}), i \geq 1\}$, *for each* $j = 1, \cdots, M$, *satisfies a GLM with mean and variance defined in* (2.1) *and* (2.2). *Suppose* *that Conditions* (A1) *and* (A2) *hold, and assume further that* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_M$*, then the estimate* $\hat{\boldsymbol{\theta}}$*, as defined in* (2.10)*, achieves the minimal covariance asymptotically in terms of the trace of the covariance matrix.*

Proposition 1 states that if all sites share the same variables and covariance matrix, the random weighted combination of estimates $\{\tilde{\boldsymbol{\theta}}_{j\tilde{N}_j}\}$ from M sites, $\hat{\boldsymbol{\theta}}$, is asymptotically efficient, effectively utilizing all available information. Unlike ensemble methods based on the "robust average" concept, which exclude estimates from certain sites, $\hat{\boldsymbol{\theta}}$ retains all site contributions, making it more "data-efficient" than a naive average.

Remark 2. *Suppose that* ρ_j *converges to* γ_j*, as* d_1 *tends to 0. Following the proof of Proposition 1, the optimal weights are*

$$w_j = \frac{\gamma_j tr(\boldsymbol{L}_j \boldsymbol{\Sigma}_j^{-1} \boldsymbol{L}_j^\top)^{-1}}{\sum_{k=1}^M \gamma_k tr(\boldsymbol{L}_k \boldsymbol{\Sigma}_k^{-1} \boldsymbol{L}_k^\top)^{-1}}, \quad j = 1, \cdots, M,$$

where $\mathrm{tr}(\cdot)$ *is the trace function. If the covariates from all sites are not homogeneous, then adopt the estimates of the optimal weights below:*

$$\hat{w}_j = \frac{\rho_j tr(\boldsymbol{L}_j \hat{\boldsymbol{\Sigma}}_j^{-1} \boldsymbol{L}_j^\top)^{-1}}{\sum_{k=1}^M \rho_k tr(\boldsymbol{L}_k \hat{\boldsymbol{\Sigma}}_k^{-1} \boldsymbol{L}_k^\top)^{-1}}, \quad j = 1, \cdots, M,$$

where $\hat{\boldsymbol{\Sigma}}_j = \sum_{i=1}^{\tilde{N}_j} \boldsymbol{x}_{ji} \{\dot{\mu}(\boldsymbol{x}_{ji}^\top \tilde{\boldsymbol{\beta}}_{j\tilde{N}_j})^2 / \nu(\boldsymbol{x}_{ji}^\top \tilde{\boldsymbol{\beta}}_{j\tilde{N}_j})\} \boldsymbol{x}_{ji}^\top / \tilde{N}_j$.

Remark 3. The proposed method can be directly extended to the case of partially overlapped variables for some of M sites, which is also interesting. Suppose we have a set of partially overlapped variables for data site $j = 1, \cdots, M_0$, besides the common variables of interest. Define $\boldsymbol{x}_j = (\boldsymbol{u}^\top, \boldsymbol{z}^\top, \boldsymbol{v}_j^\top)^\top$ for $j = 1, \cdots, M_0$, and $\boldsymbol{x}_j = (\boldsymbol{u}^\top, \boldsymbol{z}_j^\top, \boldsymbol{v}_j^\top)^\top$ for $j = M_0 + 1, \cdots, M$. Let length of $\boldsymbol{z}$ be p_1. The corresponding parameter $\boldsymbol{\beta}_j = (\boldsymbol{\theta}^\top, \boldsymbol{\zeta}^\top, \boldsymbol{\eta}_j^\top)^\top$ for $j = 1, \cdots, M_0$ and $\boldsymbol{\beta}_j = (\boldsymbol{\theta}^\top, \boldsymbol{\zeta}_j^\top, \boldsymbol{\eta}_j^\top)^\top$ for $j = M_0 + 1, \cdots, M$, where $\boldsymbol{\theta}$ is coefficient vector of the common $\boldsymbol{u}$, ζ is one of the partially overlapped variable $\boldsymbol{z}$ for all $j \in \{1, \cdots, M_0\}$ and $\boldsymbol{\zeta}_j$ is one of $\boldsymbol{z}_j$ for other j. For $\boldsymbol{\theta}$, the stopping times and combined estimate are defined in (2.9) and (2.10). For ζ, we only take datasets from $j = 1, \cdots, M_0$ into account. Similar to a and $\tilde{a}_j$, we denote b by the square root of the $1 - \alpha$ quantile of $\chi^2_{p_1}$, and $\sum_{j=1}^{M_0} \tilde{b}_j^2 = b^2$ with $\tilde{b}_j > 0$ for $j = 1, \cdots, M_0$. Replacing $\tilde{a}_j$ with $\tilde{b}_j$ in (2.9), and setting new stopping times $\tilde{N}_{zj}$ for data sites $j = 1, \cdots, M_0$, we obtain a set of estimates of $\boldsymbol{\zeta}$, $\tilde{\boldsymbol{\zeta}}_{j\tilde{N}_{zj}}$, for $j = 1, \cdots, M_0$. Then, a combined estimate of $\boldsymbol{\zeta}$, $\hat{\boldsymbol{\zeta}} = \sum_{j=1}^{M_0} \rho_{zj} \tilde{\boldsymbol{\zeta}}_{j\tilde{N}_{zj}}$, where $\hat{N}_z = \sum_{j=1}^{M_0} \tilde{N}_{zj}$ and $\rho_{zj} = \tilde{N}_{zj}/\hat{N}_z$. For data site $j \leq M_0$, we can simultaneously conduct these two sequential procedures, one for $\boldsymbol{\theta}$ and the other for ζ.

2.3 Adaptive sampling strategy

When site-specific data exhibit non-homogeneity, estimating regression parameters becomes uneven, making random sampling inefficient. Instead, adaptive sampling, which selects data based on its contribution, offers a more effective approach, particularly in sequential analysis. Leveraging statistical experimental design criteria, such as D-optimality (Deng et al., 2009; Smucker et al., 2018; Chen et al., 2020; van Sluijs et al., 2022) and A-optimality (Woods et al., 2006; Montgomery, 2009; Limmun et al., 2018; Hassanein and Seyam, 2019; López-Fidalgo et al., 2007), enhances data selection before analysis. Applications include D-optimality for adaptive variable selection in GEE methods (Chen et al., 2020), A-optimal vs. D-optimal screening design comparison (Jones et al., 2021), and a weighted A-optimality criterion for robust mixture designs (Limmun et al., 2018). Thus, federated learning with adaptive sampling emerges as an effective strategy for optimizing data analysis in non-homogeneous settings.

For each j, let $\{\boldsymbol{x}_{ji} : i = 1, ..., k\}$ be the set of selected samples up to the kth stage, called an active set as in Settles (2010) (see also Chen et al., 2020; Li et al., 2020), and let $\boldsymbol{U}_{jk}$ be its inactive counter part, a set of data that are not yet included in the analysis until stage k. If we adopt an A-optima criterion for sample selection, then we select a new sample $\boldsymbol{x}_j^*$ from $\boldsymbol{U}_{jk}$, such that $\boldsymbol{x}_j^* = \mathrm{argmin}_{\boldsymbol{x}\in\boldsymbol{U}_j}\mathrm{tr}\{(\boldsymbol{O}_j + \boldsymbol{x}\{\dot{\mu}(\boldsymbol{x}^\top\boldsymbol{\beta}_{j0})^2/\nu(\boldsymbol{x}^\top\boldsymbol{\beta}_{j0})\}\boldsymbol{x}^\top)^{-1}\}$, where $\boldsymbol{O}_j = \sum_{i=1}^k \boldsymbol{x}_{ji}\{\dot{\mu}(\boldsymbol{x}_{ji}^\top\boldsymbol{\beta}_{j0})^2/\nu(\boldsymbol{x}_{ji}^\top\boldsymbol{\beta}_{j0})\}\boldsymbol{x}_{ji}^\top$. We then repeat this selection scheme until the stopping criterion (2.9) is satisfied. Figure 1 in Supplementary Materials A2 illustrates the computation procedure for distributed sequential federated estimation. The A-optimal criteria are used for recruiting samples, while stopping rules govern procedures for each site.

We perform M estimation procedures separately, using data from their corresponding sites. Moreover, it is known that as $d_1 \to 0$,

$$\sqrt{\hat{N}}\left(\sum_{j=1}^{M} \rho_j \boldsymbol{L}_j \boldsymbol{\Sigma}_j^{-1} \boldsymbol{L}_j^{\top}\right)^{-1/2} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \to N(0, \boldsymbol{I}_{p_0}) \text{ in distribution}, \quad (2.11)$$

$\boldsymbol{I}_{p_0}$ is an identity matrix with rank p_0. By (2.11), we have

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\top} \tilde{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \longrightarrow \chi^2_{p_0}, \text{ as } d_1 \to 0, \quad (2.12)$$

where $\tilde{\boldsymbol{\Sigma}} = \sum_{j=1}^{M} \rho_j^2 \boldsymbol{L}_j \boldsymbol{\Sigma}_{j\tilde{N}_j}^{-1} \boldsymbol{L}_j^{\top}$. Let $\boldsymbol{Z} = (z_1, \cdots, z_{p_0})^{\top}$, then we have

$$R_{\hat{N}} = \left\{ \boldsymbol{Z} \in R^{p_0} \colon \frac{S_{\hat{N}}}{\hat{N}} \leq \frac{d_1^2}{\mu_{\hat{N}}} \right\} \quad (2.13)$$

is a confidence set for $\boldsymbol{\theta}_0$, where $S_{\hat{N}} = (\boldsymbol{Z} - \hat{\boldsymbol{\theta}})^{\top} \tilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{Z} - \hat{\boldsymbol{\theta}})$ and $\mu_{\hat{N}} = \sum_{j=1}^{M} \tilde{a}_j^2 \mu_{j\tilde{N}_j} / a^2$. When all M sequential procedures stop recruiting new samples, we then integrate the results. Then we have Theorem 1 below, and its proof is given in Supplementary Materials A1.

Theorem 1. *Suppose that the* $\{(\boldsymbol{x}_{ji}, y_{ji}), i \geq 1\}$, *for site* $j = 1, \cdots, M$, *satisfy a GLM with mean and variance defined in* (2.1) *and* (2.2), *and Conditions* (A1) *and* (A2) *hold. Then* $(i) \lim_{d_1 \to 0} \frac{d_1^2 \hat{N}}{a^2 \mu} = 1$, almost surely, $(ii) \lim_{d_1 \to 0} P(\boldsymbol{\theta}_0 \in R_{\hat{N}}) = 1 - \alpha$, $(iii) \lim_{d_1 \to 0} \frac{d_1^2 E(\hat{N})}{a^2 \mu} = 1$, *where* $\mu =$

$\sum_{j=1}^{M} \tilde{a}_j^2 \mu_j / a^2$, μ_j is the maximum eigenvalue of matrix $\boldsymbol{L}_j \boldsymbol{\Sigma}_j^{-1} \boldsymbol{L}_j^{\top}$, and a^2 is the $1-\alpha$ quantile of $\chi^2_{p_0}$.

Note that Theorem 1 holds even if the variables $p_j - p_0$, for all j, are not the same and does not assume that there is the same covariance matrix for all sites as in Proposition 1. It shows that the proposed method has the properties that ratio of the (random) total sample size to the (known) optimal one is equal to 1, and the coverage probability of $1-\alpha$, asymptotically, which are named as "asymptotic consistency" and "asymptotic efficiency" in Chow and Robbins (1965). Note that using a simple random sampling method at each stage to select a new observation can be viewed as a special case, where Theorem 1 still holds. From (2.9) and (2.10), we know that the maximum axis of the confidence set $R_{\hat{N}}$ is not greater than $2d_1\{\lambda_{max}(\hat{N}\tilde{\boldsymbol{\Sigma}})/\mu_{\hat{N}}\}^{1/2}$, which converges to $\{\lambda_{max}(\sum_{j=1}^{M} \rho_{j0} \boldsymbol{L}_j \boldsymbol{\Sigma}_j^{-1} \boldsymbol{L}_j^{\top})/\mu\}^{1/2}$, with $\rho_{j0} = \tilde{a}_j^2 \mu_j / \sum_{j=1}^{M} \tilde{a}_j^2 \mu_j$. Hence, $\{\lambda_{max}(\sum_{j=1}^{M} \rho_{j0} \boldsymbol{L}_j \boldsymbol{\Sigma}_j^{-1} \boldsymbol{L}_j^{\top})/\mu\}^{1/2} \leq 1$, if $\mu_1 = \cdots = \mu_M$. This implies that the length of the maximum axis of $R_{\hat{N}}$ is less than $2d_1$.

Remark 4. *The proposed method is highly adaptable to various computing frameworks, allowing its implementation in distributed computing and efficient analysis of large-scale datasets. Even in a centralized data pool, partitioning into M sub-datasets enables independent model fitting. This* *flexibility extends beyond traditional federated learning, making it applicable in diverse computing environments.*

Remark 5. *Sites with smaller sample sizes may not meet the stopping criterion, especially for small d_1. Adjusting d_1 and d_2 or modifying $\{\tilde{a}_j, j = 1, \ldots, M\}$ based on sample sizes is an efficient solution without affecting the estimation of the final parameters. Ignoring very small sites, as their contribution is minimal, is also viable. Beyond the fully sequential method, a multistage sequential approach can improve the analysis of sites with limited prior information (see Park and Chang (2016) and the references therein). This allows for incremental data collection to enhance estimation.*

3. Numerical studies

In this section, we present the numerical results of the proposed method based on the synthesized data, and the COVID-19 data set from Mexican health authorities.

3.1 Simulation studies

Let $\boldsymbol{\beta}_j = (\beta_{j0}, \boldsymbol{\theta}^\top, \boldsymbol{\eta}_j^\top)^\top \in R^{p_j}$ be the parameter vector, as the notations used in Section 2, where $\boldsymbol{\theta}$ is the coefficient vector of the common variables of interest, and $(\beta_{j0}, \boldsymbol{\eta}_j^\top)$, with length $p_j + 1$, are the remain-ing variables of the site j, for $j = 1, \cdots, M$. Let $M = 5$ be the number of sites, and $\boldsymbol{\theta} = (2.0, 1.0)^\top$ be the fixed value of the common parameters. The other regression parameters for the following two scenarios are (1) **B1**: $\beta_{j0} = -2.0$ and $\boldsymbol{\eta}_j = (1.0, 0.5)^\top$, for $j = 1, \cdots, 5$; and (2) **B2**: $\beta_{10} = -2.0$, $\boldsymbol{\eta}_1 = (1.0,\ 0)^\top$, $\beta_{20} = -2.0$, $\boldsymbol{\eta}_2 = (1.0,\ 0.5)^\top$, $\beta_{30} = -2.0$, $\boldsymbol{\eta}_3 = (1.0,\ 0.5,\ 0)^\top$, $\beta_{40} = -1.5$, $\boldsymbol{\eta}_4 = (1.0,\ 0)^\top$, $\beta_{50} = -2.5$, $\boldsymbol{\eta}_5 = (1.0,\ 1.0)^\top$. In scenario **B1**, the parameter vectors at all five data sites (i.e., for all j) are identical. In contrast, scenario **B2** introduces heterogeneity by allowing some parameters to differ across the five sites, as described previously. Notably, the parameter vector $\boldsymbol{\eta}_3$ in scenario B2 has a different dimension compared to the others. Specifically, for $p_j = 5$ with $j \neq 3$, the covariate vector $\boldsymbol{x}$ follows a multivariate normal distribution:

$$\boldsymbol{x} \sim N\left(0, \text{diag}(\phi_{ji}),\ i = 1, \ldots, p_j - 1\right).$$

Two setups for ϕ_{ji} are considered: (1) **H1**: $\phi_{ji} = 1$ for all $i = 1, \ldots, p_j - 1$ and $j = 1, \ldots, 5$; here, $p_3 = 5$; (2) **H2**: $\phi_{23} = \phi_{24} = 4$, $\phi_{43} = \phi_{44} = 2$, and $\phi_{53} = \phi_{54} = 4$, while all other ϕ_{ji} values are set to 1; in this case, $p_3 = 6$. We set the significance level at $\alpha = 0.05$ for all studies. The simulation study varies two key parameters: (i) $d_1 \in 0.2, 0.3$, which controls the size of the confidence set for $\boldsymbol{\theta}$, and (ii) $d_2 \in 0.04, 0.05$, related to the AUC estimation criterion. To investigate the effect of site heterogeneity, we define $\gamma_j = \tilde{a}_j^2/a^2$, where $\tilde{a}_j$ represents the local scale parameter at site j, and a is a global reference. We consider two configurations: (**G1**): $\gamma_j = 1/5$ for all $j = 1, \ldots, 5$, representing uniform site contributions; and (**G2**): $\gamma_1, \ldots, \gamma_4 = 1/10$, $\gamma_5 = 6/10$, simulating a setting where site 5 dominates. For each parameter combination, 200 replications are performed to ensure stable estimates. Simulation data $(y_j, \boldsymbol{x}_j)$ are generated from logistic regression models:

$$P(y_j = 1 \mid \boldsymbol{x}_j) = \mu(\boldsymbol{x}_j^\top \boldsymbol{\beta}_j) = \frac{\exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}_j)}{1 + \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}_j)}, \quad j = 1, \ldots, M, \tag{3.14}$$

where $y_j \in \{0, 1\}$ is a binary response variable, and $\boldsymbol{x}_j$ is a covariate vector. Each y_j is drawn from a Bernoulli distribution with success probability $p = P(y_j = 1 \mid \boldsymbol{x}_j)$, conditional on a given $\boldsymbol{x}_j$. That is, each $\boldsymbol{x}_j$ produces one corresponding y_j.

Table 1 presents the stopping times, coverage frequency (CF), and average AUC for adaptive (**A**) sample selection under covariate setup **H1** and parameter configuration **B1**. The corresponding results for random selection (**R**) are provided in Table T1 of Supplementary Materials A2. As expected, the stopping time N increases and the coverage frequency (CF) converges to 0.95 as d_1 approaches 0. In the equal γ case (**G1**), stopping times are similar across all sites. In contrast, under **G2**, sites 1–4 exhibit significantly smaller sample sizes than site 5, illustrating that appropriate selection of γ values can effectively control the distribution of sample sizes across sites.

Compared to selection **R** (random sampling), selection **A** (adaptive sampling based on the A-optimal design) results in smaller stopping times, indicating that adaptive sampling prioritizes efficiency and reduces sample-related costs. Although AUC values under random sampling are slightly higher, and CFs are marginally closer to 0.95, these outcomes are largely attributable to the larger sample sizes obtained through random selection.

We also evaluate the performance of the estimates of the parameter vector $\boldsymbol{\theta} = (\beta_1, \beta_2)^\top$. Under the setup of selection **A**, covariate configuration **H1**, and parameter setting **B1**, Table 2 reports the absolute bias, $|\hat{\beta}_i - \beta_i|$ for $i = 1, 2$, of the estimates obtained using the proposed method (RW). For comparison, the table also includes estimates obtained by combining data from the five sites with equal weights (EW). The corresponding results for selection **R** are provided in Table T2 of Supplementary Materials A2.

These findings indicate that estimates obtained from individual sites exhibit significantly larger biases and standard deviations compared to both ensemble estimators: RW (the proposed method) and EW (the equal-weighted method). Under scenario **G1**, even for small values of d_1, the RW method performs comparably to the EW method in terms of both bias and

Table 1: Stopping times, AUC and coverage frequency (CF) of the adaptive selection case with covariate set **H1** and parameter set **B1**.

d_2	d_1			N	N_1	N_2	N_3	N_4	N_5	AUC	CF
0.05	0.3	**G1**	Est.	1203.88	238.29	237.21	244.67	240.44	243.28	0.893	0.925
			Sd	83.54	39.10	34.79	39.00	41.67	38.20	0.005	-
		G2	Est.	1328.32	166.10	168.32	170.98	168.41	654.50	0.898	0.930
			Sd	89.67	24.50	24.84	28.02	24.98	72.76	0.005	-
	0.2	**G1**	Est.	2447.69	489.66	482.05	491.93	492.50	491.56	0.885	0.970
			Sd	140.46	63.91	64.85	65.95	61.80	64.90	0.004	-
		G2	Est.	2551.05	262.87	269.95	272.19	271.17	1474.87	0.890	0.945
			Sd	149.14	44.50	44.02	47.97	40.91	115.20	0.005	-
0.04	0.3	**G1**	Est.	1420.68	284.45	284.25	282.26	284.25	285.46	0.896	0.955
			Sd	84.56	36.41	39.53	35.24	35.64	39.36	0.005	-
		G2	Est.	1667.64	253.01	259.55	257.76	254.88	642.43	0.896	0.975
			Sd	132.20	53.75	54.18	49.87	56.94	78.94	0.006	-
	0.2	**G1**	Est.	2463.70	495.77	495.56	488.95	492.86	490.57	0.886	0.930
			Sd	150.92	68.15	66.41	64.06	65.26	65.25	0.004	-
		G2	Est.	2657.90	295.45	300.86	299.95	301.33	1460.31	0.891	0.950
			Sd	139.94	32.09	38.71	32.01	32.32	115.93	0.004	-

G1 and **G2** denote two different sets of γ_j's, $j = 1, \cdots, 5$. d_1 and d_2 are the sizes of confidence set and prefixed parameters for AUC, respectively.

Table 2: Absolute bias of estimate of $\boldsymbol{\theta} = (\beta_1, \beta_2)$ with the adaptive selection strategy, covariate setup **H1** and parameter set **B1**.

d_2	d_1			RW	EW	Site 1	Site 2	Site 3	Site 4	Site 5
0.05	0.3	**G1**	β_1	0.10(0.07)	0.09(0.06)	0.19(0.14)	0.18(0.12)	0.20(0.14)	0.20(0.15)	0.18(0.15)
			β_2	0.06(0.04)	0.06(0.04)	0.13(0.09)	0.12(0.09)	0.13(0.10)	0.13(0.09)	0.12(0.10)
		G2	β_1	0.10(0.08)	0.13(0.09)	0.24(0.17)	0.26(0.19)	0.28(0.22)	0.25(0.18)	0.11(0.09)
			β_2	0.06(0.05)	0.07(0.06)	0.15(0.12)	0.16(0.12)	0.16(0.14)	0.16(0.12)	0.07(0.05)
	0.2	**G1**	β_1	0.06(0.04)	0.06(0.04)	0.13(0.10)	0.14(0.10)	0.14(0.10)	0.12(0.09)	0.13(0.10)
			β_2	0.04(0.03)	0.04(0.03)	0.09(0.07)	0.09(0.07)	0.09(0.07)	0.09(0.06)	0.09(0.07)
		G2	β_1	0.07(0.05)	0.08(0.06)	0.20(0.14)	0.19(0.14)	0.20(0.14)	0.17(0.13)	0.08(0.06)
			β_2	0.04(0.03)	0.05(0.04)	0.13(0.09)	0.11(0.08)	0.13(0.09)	0.11(0.09)	0.05(0.04)
0.04	0.3	**G1**	β_1	0.10(0.07)	0.09(0.07)	0.19(0.14)	0.19(0.14)	0.17(0.13)	0.19(0.13)	0.19(0.14)
			β_2	0.06(0.04)	0.05(0.04)	0.11(0.09)	0.11(0.09)	0.11(0.09)	0.12(0.09)	0.12(0.09)
		G2	β_1	0.09(0.07)	0.12(0.08)	0.20(0.16)	0.22(0.20)	0.22(0.17)	0.23(0.18)	0.12(0.10)
			β_2	0.06(0.04)	0.07(0.05)	0.14(0.11)	0.13(0.11)	0.13(0.12)	0.15(0.12)	0.08(0.06)
	0.2	**G1**	β_1	0.07(0.05)	0.06(0.05)	0.14(0.11)	0.14(0.11)	0.13(0.10)	0.15(0.10)	0.13(0.10)
			β_2	0.04(0.03)	0.04(0.03)	0.09(0.06)	0.09(0.07)	0.08(0.06)	0.08(0.06)	0.09(0.07)
		G2	β_1	0.06(0.04)	0.07(0.06)	0.16(0.12)	0.18(0.14)	0.18(0.14)	0.17(0.12)	0.07(0.06)
			β_2	0.04(0.03)	0.05(0.04)	0.11(0.08)	0.11(0.08)	0.12(0.09)	0.11(0.09)	0.06(0.04)

Standard deviations are in parentheses.

standard deviation. However, under scenario **G2**, the RW estimator yields smaller biases and standard deviations, with the advantage becoming more pronounced when selection **A** (adaptive sampling) is employed.

To evaluate the proposed method under non-homogeneity and varying regression dimensions (**B2**), we generate data under covariate settings **H1** and **H2** across four scenarios: **S1** (**G1**, **H1**), **S2** (**G2**, **H1**), **S3** (**G1**, **H2**), and **S4** (**G2**, **H2**). Table 3 reports the stopping times, coverage frequency (CF), and average AUC across five sites for $d_1 = 0.2$ and $d_2 = 0.05$, while Table 4 presents the absolute biases of $\boldsymbol{\theta} = (\beta_1, \beta_2)^\top$. The proposed method (RW) consistently achieves lower biases and standard deviations than single-site estimators, particularly in **S2** and **S4**, and performs comparably or better than the equal-weighted (EW) method. These results confirm the effectiveness and robustness of the sequential federated approach in accurately estimating $\boldsymbol{\theta}$ across all settings.

3.2 Case Study: COVID-19 Data from Mexico

We apply the proposed method to the publicly available COVID-19 dataset released by the Mexican Ministry of Health. Although this dataset can be centrally pooled, we use it to emulate a realistic federated learning environment, where data are distributed across multiple sites and cannot be

Table 3: Simulation results about stopping times, AUC and coverage frequency with non-homogeneous covariate setup **B2**, with $d_1 = 0.2$ and $d_2 = 0.05$.

		S1	**S2**	**S3**	**S4**
R	N	3973.98(276.33)	4218.45(282.24)	4215.86(261.60)	4579.10(284.17)
	N_1	792.76(118.45)	420.24(83.69)	796.76(110.02)	405.35(86.78)
	N_2	798.20(117.62)	413.23(82.10)	867.04(141.63)	453.95(87.49)
	N_3	781.18(113.77)	409.25(82.80)	799.68(120.93)	409.57(83.00)
	N_4	717.27(112.10)	370.62(69.01)	771.29(103.10)	392.57(74.39)
	N_5	884.57(141.99)	2605.11(250.40)	981.08(139.14)	2917.65(231.65)
	AUC	0.902(0.006)	0.903(0.007)	0.916(0.005)	0.916(0.006)
	CF	0.955	0.955	0.945	0.970
A	N	2479.91(132.46)	2574.33(160.42)	2446.55(149.30)	2513.76(149.53)
	N_1	503.40(63.22)	267.14(46.49)	494.80(64.30)	268.94(48.17)
	N_2	488.56(67.06)	263.54(44.75)	476.62(63.51)	260.39(44.64)
	N_3	509.74(59.49)	274.19(45.69)	500.51(64.79)	267.37(42.22)
	N_4	483.30(57.85)	262.87(41.63)	482.09(58.73)	251.97(45.75)
	N_5	494.92(64.28)	1506.59(121.94)	492.52(61.96)	1465.10(113.49)
	AUC	0.885(0.004)	0.888(0.005)	0.889(0.004)	0.893(0.005)
	CF	0.965	0.905	0.935	0.965

Standard deviations are in parentheses. **R** and **A** stand for Random and Adaptive samplings, respectively. **S1** to **S4** denote 4 different combination of simulation parameter setups.

Table 4: Absolute bias of estimate of $\boldsymbol{\theta} = (\beta_1, \beta_2)$ with with non-homogeneous covariate setup **B2**, $d_1 = 0.2$ and $d_2 = 0.05$.

			RW	EW	Site 1	Site 2	Site 3	Site 4	Site 5
R	**S1**	β_1	0.07(0.05)	0.06(0.05)	0.14(0.10)	0.14(0.10)	0.13(0.10)	0.14(0.10)	0.15(0.11)
		β_2	0.05(0.04)	0.05(0.04)	0.10(0.07)	0.11(0.08)	0.10(0.08)	0.10(0.08)	0.11(0.08)
	S2	β_1	0.06(0.05)	0.07(0.05)	0.20(0.14)	0.19(0.14)	0.20(0.14)	0.17(0.13)	0.08(0.06)
		β_2	0.05(0.03)	0.06(0.04)	0.15(0.11)	0.14(0.10)	0.13(0.10)	0.15(0.12)	0.07(0.05)
	S3	β_1	0.06(0.05)	0.06(0.04)	0.13(0.10)	0.15(0.11)	0.14(0.09)	0.14(0.10)	0.13(0.10)
		β_2	0.05(0.03)	0.04(0.03)	0.10(0.07)	0.10(0.08)	0.10(0.08)	0.09(0.07)	0.11(0.08)
	S4	β_1	0.06(0.05)	0.07(0.05)	0.20(0.14)	0.18(0.13)	0.18(0.15)	0.18(0.13)	0.07(0.05)
		β_2	0.04(0.03)	0.06(0.04)	0.16(0.13)	0.15(0.10)	0.15(0.11)	0.15(0.11)	0.05(0.04)
A	**S1**	β_1	0.06(0.04)	0.05(0.04)	0.13(0.11)	0.13(0.11)	0.13(0.09)	0.13(0.09)	0.13(0.11)
		β_2	0.04(0.03)	0.04(0.03)	0.09(0.06)	0.08(0.07)	0.08(0.06)	0.08(0.06)	0.08(0.06)
	S2	β_1	0.07(0.05)	0.08(0.07)	0.20(0.13)	0.19(0.14)	0.18(0.15)	0.18(0.13)	0.08(0.06)
		β_2	0.04(0.03)	0.05(0.04)	0.12(0.09)	0.12(0.09)	0.12(0.09)	0.12(0.09)	0.06(0.04)
	S3	β_1	0.06(0.05)	0.06(0.05)	0.13(0.10)	0.14(0.10)	0.13(0.11)	0.12(0.09)	0.13(0.10)
		β_2	0.04(0.03)	0.04(0.03)	0.08(0.07)	0.07(0.06)	0.08(0.06)	0.08(0.06)	0.08(0.06)
	S4	β_1	0.06(0.05)	0.08(0.06)	0.22(0.14)	0.19(0.13)	0.17(0.13)	0.19(0.14)	0.07(0.06)
		β_2	0.04(0.03)	0.05(0.03)	0.12(0.09)	0.12(0.09)	0.12(0.09)	0.12(0.09)	0.05(0.04)

Standard deviations are in parentheses. **R** and **A** stand for Random and Adaptive samplings, respectively. RW = random weight via the proposed method. EW = equal weight. **S1** to **S4** denote 4 different combination of simulation parameter setups.

shared directly due to privacy, legal, or institutional constraints. This data were collected from 32 health sectors, which includes 6,659,184 records of suspected cases, distinguishing outpatients and inpatients according to clinical diagnoses. The dataset, subject to updates, was downloaded in April 2021. The dataset includes personal and health information such as gender, age, and medical history (e.g., pneumonia, diabetes, COPD, asthma, immunosuppression, hypertension, cardiovascular disease, and chronic renal failure). Additional factors include obesity, smoking, exposure to SARS-CoV-2 cases (EOC), and COVID-19 status (positive/negative). We use the COVID-19 status as the response variable and the others as covariates. Except for age, all variables are binary: "1" for "Y" and "0" for "N"; for gender, "1" represents females and "0" males. Our analysis explores whether diabetes or obesity increases the likelihood of COVID-19 infection using logistic regression. After excluding missing values, the dataset includes 5,816,861 subjects across 32 sites, with sample sizes ranging from 21,746 to 2,396,133. Sites 4, 6, 7, and 18 each have fewer than 30,000 subjects. Due to varying sample sizes, we use two sampling strategies: (**C1**) equal proportional allocation ($\gamma_j = 1/32$ per site) and (**C2**) allocating $1/100$ of samples to sites 4, 6, 7, and 18, and $6/175$ to others, ensuring unequal γ's. We fit a logistic regression model to the full dataset to serve as a centralized baseline, and compare its parameter estimates and AUC with those obtained from our proposed distributed method. This case study highlights how the proposed method can provide accurate inference and robust prediction while respecting data locality and site heterogeneity—key challenges in federated healthcare analytics.

A logistic regression model fitted on the full dataset serves as the baseline, with its parameter estimates and AUC used as references. To illustrate the proposed method, three variable sets are considered: (1) **All** — all available variables; (2) **P1** — five key variables (pneumonia, COPD, asthma, CRF, EOC); and (3) **P2** — ten key variables (gender, age, diabetes, asthma, hypertension, other diagnoses, cardiovascular disease, obesity, CRF, smoking). Under configuration **C1** (equal site proportions), Table 5 presents parameter estimates from the adaptive sampling, while Table T3 in the Supplementary Materials A2 shows those from the random sampling. Both approaches yield results closely aligned with the baseline. For configuration **C2** (unequal site proportions), similar conclusions hold based on Tables T4 and T5 in Supplementary Materials A2.

Table 6 presents the stopping times and AUC values for the three variable sets: **All**, **P1**, and **P2**. Table 7 reports stopping times for sites 4, 6, 7, and 18, each with fewer than 3000 samples. As shown in Table 6, adaptive

Table 5: Parameter estimate for COVID-19 data with $d_2 = 0.05$, adaptive selection and equal proportion **C1**.

	d_1		GE	PN	AG	DI	CO	AS	IM	HY	OT	CA	OB	CR	SM	EO
A **All**	0.3	Est.	-0.19	1.03	0.01	0.10	-0.27	-0.10	-0.35	0.04	-0.26	-0.36	0.26	-0.23	-0.19	0.45
		Sd	0.03	0.04	0.00	0.04	0.05	0.05	0.05	0.04	0.05	0.05	0.04	0.05	0.05	0.03
	0.2	Est.	-0.17	0.91	0.01	0.14	-0.21	-0.01	-0.26	0.05	-0.17	-0.30	0.27	-0.20	-0.20	0.49
		Sd	0.03	0.03	0.00	0.03	0.04	0.04	0.04	0.03	0.04	0.04	0.03	0.04	0.03	0.03
P1	0.3	Est.	-	1.10	-	-	-0.12	-0.07	-	-	-	-	-	-0.19	-	0.45
		Sd	-	0.05	-	-	0.07	0.07	-	-	-	-	-	0.07	-	0.04
	0.2	Est.	-	1.06	-	-	-0.26	-0.08	-	-	-	-	-	-0.25	-	0.44
		Sd	-	0.05	-	-	0.06	0.05	-	-	-	-	-	0.06	-	0.03
P2	0.3	Est.	-0.17	-	0.01	0.10	-	-	-	0.03	-0.25	-0.37	0.23	-0.29	-0.17	-
		Sd	0.03	-	0.00	0.05	-	-	-	0.04	0.06	0.06	0.04	0.06	0.05	-
	0.2	Est.	-0.18	-	0.01	0.09	-	-	-	0.04	-0.24	-0.34	0.28	-0.21	-0.22	-
		Sd	0.03	-	0.00	0.04	-	-	-	0.04	0.05	0.05	0.04	0.05	0.04	-
B		Est.	-0.11	1.32	0.01	0.18	-0.17	-0.08	-0.19	0.09	0.06	-0.20	0.34	-0.26	-0.24	0.06
		Sd	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.00

A and **B** stand for the adaptive sampling and baseline model, respectively. GE: gender; PN: Pneumonia; AG: age; DI: Diabetes; CO: Chronic obstructive pulmonary; AS: asthma; IM: immunosuppression; HY: Hypertension; OT: Other diseases; CA: cardiovascular; OB: obesity; CR: Chronic renal failure; SM: smoke; EO: Exposed to other cases diagnosed as SARS CoV-2.

Table 6: Stopping times and AUC for COVID-19 data with $d_2 = 0.05$.

			Stopping time			AUC			
		d_1	**All**	**P1**	**P2**	**All**	**P1**	**P2**	Baseline
R	**C1**	0.3	199380	84280	91580	0.625	0.629	0.627	0.598
		0.2	432780	170680	183380	0.622	0.625	0.625	0.598
	C2	0.3	203280	84580	92680	0.626	0.632	0.629	0.598
		0.2	434780	172880	185380	0.622	0.626	0.627	0.598
A	**C1**	0.3	18610	16480	16990	0.668	0.672	0.672	0.598
		0.2	27270	18020	20480	0.662	0.670	0.666	0.598
	C2	0.3	18750	16550	17100	0.670	0.672	0.672	0.598
		0.2	26920	18230	20720	0.663	0.670	0.668	0.598

R and **A** stand for Random and Adaptive samplings, respectively. Baseline denotes the model built with all data. **All**, **P1** and **P2** stand for all variables, five key variables (PN, CO, AS, CR, EO), and ten key variables (GE, AG, DI, AS, HY, OT, CA, OB, CR, SM), respectively.

sampling requires substantially fewer samples, while both sampling strategies yield comparable parameter estimates (Table 5 and Table T3 in the Supplementary Materials A2).

Random sampling requires more than 100,000 samples to meet the threshold $d_2 = 0.05$, which is infeasible for small sites using only local data. In contrast, Table 7 shows that the proposed distributed sequential feder-

Table 7: Stopping times of sector 4, 6, 7, and 18 with data size less than 30000 for COVID-19 data.

			All				**P1**				**P2**			
		d_1	Site 4	Site 6	Site 7	Site 18	Site 4	Site 6	Site 7	Site 18	Site 4	Site 6	Site 7	Site 18
R	**C1**	0.3	5815	4615	6715	3315	3715	1615	3815	1015	3815	1815	2615	1215
		0.2	14715	12215	12215	7615	5215	3215	6815	2215	8415	3915	5715	2715
	C2	0.3	2515	1815	2615	1015	915	615	1115	515	1915	815	915	515
		0.2	4215	3615	5115	2015	2515	1215	2615	715	2915	1415	1915	1015
A	**C1**	0.3	605	455	555	465	515	445	385	465	545	445	385	465
		0.2	1025	805	1025	705	585	445	455	465	735	505	595	495
	C2	0.3	515	445	385	465	515	445	385	465	515	445	385	465
		0.2	545	445	425	465	515	445	385	465	515	445	385	465

R and **A** stand for Random and Adaptive samplings, respectively. **All**, **P1** and **P2** stand for all variables, five key variables (PN, CO, AS, CR, EO), and ten key variables (GE, AG, DI, AS, HY, OT, CA, OB, CR, SM), respectively.

ated learning satisfies the criteria under both sampling methods. Moreover, our method consistently achieves higher AUCs than the baseline (Table 6), confirming its effectiveness in producing accurate estimates, maintaining high classification performance, and preserving data at small sites.

As shown in Table 6, including more variables increases stopping times:

models with **P1** require the fewest samples, while the full model requires the most. Smaller d_1 values also lead to longer stopping times. Although the total sample sizes under **C1** and **C2** are similar, Table 6 shows that under **C2**, smaller sites contribute fewer samples, illustrating that unequal allocation reduces their sampling burden. This early stopping effect, especially when combined with adaptive sampling, improves overall efficiency.

Despite design constraints, the COVID-19 results confirm known risk factors: both **P2** and the baseline model indicate a higher risk of infection for elderly individuals with diabetes or obesity, while females with cardiovascular disease or a smoking history have a lower risk. Pneumonia and EOC significantly increase infection risk, whereas asthma and CRF are associated with lower risk, consistent with previous studies Hernández-Garduño (2020); Rashedi et al. (2020); Louis et al. (2020); Liu et al. (2020); Memon and Biswas (2022).

Remark 6. Supplementary Materials A2 also examine the performance of the proposed method under partially overlapping parameters (Tables T6–T8) and model misspecification (Table T9). The results show that the proposed method (RW) achieves smaller or comparable biases in parameter estimates compared to the equal-weight method (EW). Under mild misspecification and $d_1 = 0.2$, the biases remain close to zero. In summary, both the numerical studies and the COVID-19 analysis demonstrate that the adaptive approach yields more accurate parameter estimation and prediction than naive averaging (EW), and greater efficiency than conventional subsampling (selection $\mathbf{R}$).

4. Conclusion

We propose a novel approach that integrates distributed sequential estimation into the federated learning framework, while preserving its original computational structure. This enables independent, site-level sequential inference, reducing communication costs and enhancing both robustness and efficiency (Lindell, 2005; Feigenbaum et al., 2001; Carlini et al., 2019). Via sequential analysis, the proposed method provides precise parameter estimates at data-driven stopping times, offering improved stability compared to conventional aggregation techniques. The adaptive sampling strategy, inspired by principles of experimental design and information theory, efficiently selects informative observations—particularly beneficial for large-scale datasets such as those arising in pandemic surveillance. While this work focuses on parameter estimation, the proposed framework offers a foundation for broader inferential tasks in federated settings, with potential applications in privacy-preserving analytics and real-time, data-driven decision-making.

Supplementary Materials

Supplementary material contains a detailed proof of the main results and additional numerical results.

Acknowledgements

This research is supported in part by research grants from National Natural Science Foundation of China (No. 12371277, 12231017), and National Science and Technology Council of Taiwan (111-2118-M-001-003-MY2). Xinyu Zhang and Yuan-chin Ivan Chang are co-corresponding authors.

References

Ai, M., J. Yu, H. Zhang, and H. Wang (2021). Optimal subsampling algorithms for big data regressions. *Statistica Sinica 31*(2), 749 – 772.

Carlini, N., C. Liu, U. Erlingsson, J. Kos, and D. Song (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Neural Information Processing Systems (NeurIPS)*, pp. 267–284.

Chang, Y.-c. I. (2011). Sequential estimation in generalized linear models when covariates are subject to errors. *Metrika 73*, 93–120.

Chang, Y. I. (1999). Strong consistency of maximum quassi-likelihood eistimate in generalized linear models via a last time. *Statistics & Probability Letters 45*, 237–246.

Chen, Z., Z. Wang, and Y. Chang (2020). Sequential adaptive variables and subject selection for gee methods. *Biometrics 76*(2), 496–507.

Chen, Z., Z. Wang, and Y. Chang (2023, March). Distributed sequential estimation procedures. *Canadian Journal of Statistics 52*(1), 271 – 290.

Chow, Y. and H. Robbins (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Ann. Math. Statist. 36*, 457–462.

Damiani, A., M. Vallati, R. Gatta, N. Dinapoli, A. Jochems, T. Deist, J. v. Soest, A. Dekker, and V. Valentini (2015). Distributed learning to protect privacy in multicentric clinical studies. In *Conference on artificial intelligence in medicine in europe*, pp. 65–75. Springer.

Deng, X., V. Joseph, A. Sudjianto, and C. F. Wu (2009). Active learning through sequential design, with applications to detection of money laundering. *Journal of the American Statistical Association 104*, 969–981.

Feigenbaum, J., Y. Ishai, T. Malkin, K. Nissim, M. J. Strauss, and R. N. Wright (2001). Secure multiparty computation of approximations. In *International Colloquium on Automata, Languages, and Programming*, pp. 927–938. Springer.

Hassanein, W. A. and M. M. Seyam (2019). Construction of some compound criteria via a-optimality. *Communications in Statistics-Theory and Methods 48*(22), 5559–5570.

He, L., W. Li, D. Song, and M. Yang (2024). A systematic view of information-based optimal subdata selection: algorithm development, performance evaluation, and application in financial data. *Statistica Sinica 34*, 611 – 636.

Hernández-Garduño, E. (2020). Obesity is the comorbidity more strongly associated for covid-19 in mexico. a case-control study. *Obes Res Clin Pract. 14*(4), 375–379.

Huang, L., Y. Yin, Z. Fu, S. Zhang, H. Deng, and D. Liu (2020). Loadaboost: Loss-based adaboost federated machine learning with reduced computational complexity on iid and non-iid intensive care data. *Plos one 15*(4), e0230706.

Jones, B., K. Allen-Moyer, and P. Goos (2021). A-optimal versus d-optimal design of screening experiments. *Journal of Quality Technology 53*(4), 369–382.

Jordan, M., J. Lee, and Y. Yang (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681.

Li, C., P. Zhou, L. Xiong, Q. Wang, and T. Wang (2018). Differentially private distributed online learning. *IEEE Transactions on Knowledge and Data Engineering 30*(8), 1440–1453.

Li, J., Z. Chen, Z. Wang, and Y. Chang (2020). Active learning in multiple-class classification problems via individualized binary models. *Computational Statistics & Data Analysis 145*, 106–119.

Limmun, W., J. J. Borkowski, and B. Chomtee (2018). Weighted a-optimality criterion for generating robust mixture designs. *Computers & Industrial Engineering 125*, 348–356.

Lindell, Y. (2005). Secure multiparty computation for privacy preserving data mining. In *Encyclopedia of Data Warehousing and Mining*, pp. 1005–1009. IGI global.

Liu, S., Y. Zhi, and S. Ying (2020). Covid-19 and asthma: Reflection during the pandemic. *Clin Rev Allergy Immunol. 59*(1), 78–88.

López-Fidalgo, J., M. J. Rivas-López, and B. Fernández-Garzón (2007). A-optimality standardized through the coefficient of variation. *Communications in Statistics—Theory and Methods 36*(4), 781–792.

Louis, R., D. Calmes, A. Frix, and F. Schleich (2020). Covid-19 and asthma. *Rev Med Liege. 75*(S1), 130–132.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models, 2nd Edition.* Chapman & Hall, New York.

McMahan, B., E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas (2017). Communication-efficient learning of deep networks from decentralized data. *Artificial intelligence and statistics. PMLR 54*, 1273–1282.

Memon, S. and D. Biswas (2022). Covid-19 and diabetes mellitus: from pathophysiology to clinical management. *Cureus 14*(11), e31895.

Montgomery, D. C. (2009). *Design and Analysis of experiments* (7th ed.). Hoboken, NJ, USA: JohnWiley&Sons.

Park, E. and Y. Chang (2016). Multiple-stage sampling procedure for covariate-adjusted response-adaptive designs. *Statistical Methods in Medical Research 25*(4), 1490–1511.

Rashedi, J., P. B. Mahdavi, V. Asgharzadeh, M. Pourostadi, K. H. Samadi, A. Vegari, H. Tayebi-Khosroshahi, and M. Asgharzadeh (2020). Risk factors for covid-19. *Infez Med. 28*(4), 469–474.

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison 52*(55-66), 11.

Smucker, B., M. Krzywinski, and N. Altman (2018). Optimal experimental design. *Nature Methods 15*(8), 559–560.

van Sluijs, B., R. J. M. Maas, A. J. van der Linden, T. F. A. de Greef, and W. T. S. Huck (2022). A microfluidic optimal experimental design platform for forward design of cell-free genetic networks. *Nature Communications 13*(1), 3626.

Wang, H., M. Yang, and J. Stufken (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association 114*(525), 393 – 405.

Wang, H., R. Zhu, and P. Ma (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association 113*(522), 829 – 844.

Woodroofe, M. (1982). *Nonlinear renewal theory in sequential analysis.* CBMS-NSF regional conference series in applied mathematics.

Woods, D., S. Lewis, J. Eccleston, and K. Russell (2006). Designs for generalized linear models with several variables and model uncertainty. *Technometrics 48*(2), 284–292.

Yan, F., S. Sundaram, S. Vishwanathan, and Y. Qi (2013). Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering 25*(11), 2483–2493.

Yao, Y. and H. Wang (2021). A review on optimal subsampling methods for massive datasets. *Journal of Data Science 19*(1), 151 – 172.

Yu, J., M. Ai, and Z. Ye (2024). A review on design inspired subsampling for big data. *Statistical Papers 65*, 467 – 510.

Yu, J., H. Wang, and M. Ai (2025). A subsampling strategy for aic-based model averaging with generalized linear models. *Technometrics 67*(1), 122 – 132.

Yu, J., H. Wang, M. Ai, and H. Zhang (2022). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association 117*(537), 265 – 276.

Zhou, X., D. McClish, and N. Obuchowski (2009). *Statistical methods in diagnostic medicine*. John Wiley & Sons.

School of Management, University of Science and Technology of China, Hefei, China

E-mail: zfw@ustc.edu.cn

Academy of Mathematics and Systems Science, Chinese Academy of Sciences

E-mail: xinyu@amss.ac.cn

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan

E-mail: ycchang@as.edu.tw