Computers in Biology and Medicine 204 (2026) 111558

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Enhancing survival analysis through federated learning in non-IID and scarce data scenarios

Patricia A. Apellániz*, Juan Parras, Santiago Zazo

Information Processing and Telecommunications Center, ETS Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 28040, Madrid, Spain

HIGHLIGHTS

- FedSDS combines Synthetic Data Generation and Federated Learning for decentralized Survival Analysis while preserving privacy.
- Using SAVAE and VAE-BGM, FedSDS generates high-quality synthetic data to support cross-institution training without sharing raw data.
- FedSDS adds a biased aggregation strategy that aligns synthetic data with local distributions and outperforms FedAvg in non-IID settings.
- With a single communication round, FedSDS cuts communication overhead compared to traditional FL approaches like FedAvg.

ARTICLE INFO

Keywords:
Survival analysis
Federated learning
Synthetic data generation
Data scarcity
Data heterogeneity

ABSTRACT

Integrating Artificial Intelligence (AI) into Survival Analysis (SA) has advanced predictive modeling in healthcare, enabling precise and personalized predictions of time-to-event outcomes, such as patient survival. However, real-world SA datasets often suffer from data scarcity, heterogeneity, and privacy constraints, which limit the applicability of traditional and modern AI methods. To address these challenges, we propose the Federated Synthetic Data Sharing (FedSDS) framework, which integrates synthetic data generation with Federated Learning (FL). For SA, we leverage SAVAE, a state-of-the-art model for complex datasets. Using the Variational Autoencoder-Bayesian Gaussian Mixture model enhanced with artificial inductive bias, FedSDS generates high-quality synthetic data locally and shares them among nodes, enabling collaborative model training without direct data sharing. FedSDS introduces a *biased* aggregation strategy that aligns synthetic data with local distributions, outperforming traditional FL methods, such as Federated Average. Validated under independent and identically distributed (IID) and non-IID scenarios, FedSDS mitigates data imbalances and heterogeneity, showing significant performance improvements in scarce and heterogeneous data. The proposed framework offers a scalable and privacy-preserving solution for SA in decentralized environments. By enhancing model generalizability and robustness, FedSDS provides a promising path forward for collaborative analytics in healthcare, paving the way for improved patient outcomes and greater adoption of federated techniques in real-world applications.

1. Introduction

The integration of Artificial Intelligence (AI) into the medical domain has revolutionized healthcare, offering unprecedented opportunities for advancements in disease diagnosis, personalized treatment planning, and predictive analytics. AI-driven technologies have demonstrated their potential to significantly enhance clinical outcomes by enabling early detection of critical conditions through medical imaging, optimizing patient care pathways, and facilitating more accurate predictive models. Furthermore, these innovations promise to streamline healthcare processes, reducing costs while improving overall efficiency and accessibility.

Among the many applications of AI in healthcare, Survival Analysis (SA) is essential for understanding time-to-event outcomes such as patient survival, disease progression, or treatment efficacy. Historically, SA has relied on classical statistical methods. Non-parametric techniques, such as the Kaplan-Meier (KM) estimator [1], are widely used due to

* Corresponding author.
Email address: patricia.alonsod@upm.es (P.A. Apellániz).

https://doi.org/10.1016/j.compbiomed.2026.111558
Received 6 February 2025; Received in revised form 15 September 2025; Accepted 15 February 2026
Available online 19 February 2026

their simplicity in estimating survival probabilities without requiring distributional assumptions. Semi-parametric approaches, particularly the Cox proportional hazards (CoxPH) model [2], have traditionally dominated the field, offering flexibility by assuming proportional hazards while estimating hazard ratios. Over time, CoxPH has been extended to handle interactions, time-varying covariates, and stratified analyses, establishing its role as a gold standard for survival studies. However, these classical models are limited by their assumptions of linearity and proportionality, which restrict their applicability to high-dimensional, non-linear data structures.

Machine Learning (ML) has enabled more complex and flexible modeling in SA. Models such as Random Survival Forests [3] and gradient-boosted survival trees [4] have demonstrated superior predictive accuracy over traditional statistical models [5]. More recently, Deep Learning (DL) approaches have emerged, leveraging Neural Networks (NNs) to overcome the limitations of classical methods. Comprehensive reviews, such as [6], highlight the fast evolution of these methods. For example, DeepSurv [7] extends CoxPH by parameterizing the log-risk function with feed-forward NNs, maintaining the proportional hazards assumption. Cox-Time [8] introduces time-varying effects for greater flexibility, though at the expense of computational efficiency. Beyond Cox-based models, DL has introduced discrete-time methods like DeepHit [9] and Dynamic-DeepHit [10], which model survival probabilities directly without relying on proportional hazards assumptions. While achieving excellent performance, their reliance on discrete time introduces limitations in modeling continuous outcomes. Advanced parametric DL models, such as the Survival Analysis Variational Autoencoder [11] (SAVAE), overcome these constraints, offering robust and flexible solutions for modeling non-linear relations in survival outcomes. These advances are particularly valuable in precision medicine, where complex, high-dimensional data often yield critical insights into patient-specific outcomes.

Despite these advances, AI applications in SA remain significantly constrained by data scarcity and heterogeneity, particularly in real-world datasets. SA datasets often involve rare diseases, making them small, incomplete, and highly variable across institutions. High levels of censored data add further complexity, skewing distributions and reducing predictive accuracy. Geographic and demographic biases exacerbate data heterogeneity, creating significant challenges for building robust and generalizable models. Privacy regulations and institutional barriers further hinder data sharing, limiting opportunities for collaborative model training.

Synthetic Data Generation (SDG) has emerged as a promising solution to address data scarcity by augmenting small datasets with realistic, artificially generated data samples. Generative models based on Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) like CTGAN [12], TVAE [12], and VAE-BGM (Bayesian Gaussian Mixture model) [13] have demonstrated success in producing high-quality synthetic tabular data. These models can learn complex data distributions, enabling the generation of synthetic samples that closely resemble real-world datasets. However, these state-of-the-art SDG methods typically require large datasets with thousands of samples for effective training, which is often unattainable in SA scenarios. Consequently, their applicability in low-data scenarios remains limited.

To address this limitation, Apellániz et al. [14] proposed a novel methodology for SDG in low-data settings by introducing artificial inductive biases. Their approach uses the state-of-the-art VAE-BGM model and introduces artificial inductive biases through transfer learning and meta-learning techniques. This methodology incorporates pre-training, model averaging, and meta-learning approaches such as Model-Agnostic Meta-Learning [15] (MAML) and Domain Randomized Search [16] (DRS) to guide the generative model. These techniques improve the model's ability to generate realistic synthetic data, even from limited datasets, as demonstrated by enhanced divergence metrics such as Jensen-Shannon (D_{JS}) and Kullback-Leibler (D_{KL}) divergences. Yet, while the SDG methodology [14] offers a robust solution to data scarcity, it does not fully address the challenges posed by data heterogeneity across institutions. Geographic, demographic, and institutional differences often result in highly variable datasets, especially in rare diseases, where biases in covariates and event distributions are more pronounced [17–19].

Federated Learning (FL) presents a complementary approach to addressing the challenges of data scarcity and heterogeneity by fostering collaboration across institutions without requiring centralized data aggregation. FL ensures compliance with strict privacy regulations while enabling multi-institutional collaboration by allowing institutions to retain control over their local datasets. This approach is particularly well-suited for SA, where data heterogeneity is exacerbated by geographic, demographic, and institutional differences, especially in rare diseases. The decentralized nature of FL enables models to learn from diverse patient populations across institutions, mitigating biases introduced by localized datasets. By pooling insights from geographically and demographically diverse institutions, FL expands the scope of analysis, improving the robustness and generalizability of survival models. Existing FL approaches in SA predominantly focus on federating CoxPH models [20–22], which, as discussed, are constrained by their inherent limitations. Traditional FL techniques, such as Federated Averaging (FedAvg) [23], aggregate model parameters iteratively across institutions to produce a global model. However, FedAvg is designed for independent and identically distributed (IID) data and struggles in non-IID scenarios, which are prevalent in SA due to the inherent diversity in patient populations and healthcare settings. When data distributions vary significantly across institutions, FedAvg's reliance on parameter-sharing can lead to biased global models and suboptimal convergence. Several methods have been proposed to improve FL in non-IID settings, including FedProx [24], FedNova [25], Scaffold [26], and FedMA [27]. These techniques aim to stabilize updates, correct client drift, or normalize contributions to mitigate the effects of heterogeneity. However, they share a common objective: aligning clients toward a single global model. In clinical contexts, this alignment may not be ideal, as local data distributions often reflect meaningful and context-specific characteristics. Moreover, most of these methods still rely on multiple rounds of communication between clients and the central server, which can introduce additional latency and computational cost in practical deployments.

To address the dual challenges of data scarcity and heterogeneity, we propose integrating SDG into FL, resulting in the Federated Synthetic Data Sharing (FedSDS) framework. FedSDS replaces traditional parameter-sharing mechanisms by exchanging synthetic data generated locally at each institution. Each participating node employs the VAE-BGM model [13]—enhanced with artificial inductive bias techniques as described in [14]—to generate high-quality synthetic datasets that encapsulate the underlying characteristics of their local data. These synthetic datasets are then shared among institutions, enabling collaborative SA model training while preserving privacy. FedSDS offers several distinct advantages:

1. **Addressing Data Scarcity:** By generating synthetic data locally, FedSDS augments limited datasets, enabling effective model training even in resource-constrained environments.
2. **Handling Non-IID Data:** Sharing synthetic data instead of model parameters allows FedSDS to accommodate heterogeneous datasets, ensuring robust model performance across diverse institutions.
3. **Efficiency in Communication:** Unlike traditional FL techniques that require multiple rounds of communication, FedSDS operates in a single round of synthetic data exchange, significantly reducing communication overhead while maintaining performance.

4. **Privacy Preservation:** The exchange of synthetic data ensures that raw patient data remains private, fostering trust among participating institutions.

Thus, this paper introduces a novel framework for performing SA using real-world datasets characterized by scarcity, heterogeneity, and high levels of censoring. The framework addresses critical data availability and quality challenges by integrating SDG into FL while preserving privacy. The key contributions of this work are:

- **A Comprehensive Framework for SA using SAVAE:** We propose a novel framework for conducting SA on real-world datasets that are often small, incomplete, and heterogeneous. The framework employs SAVAE [11] as the core SA model, leveraging its flexibility and robustness in modeling non-linear relationships and time-to-event outcomes. This integration enables robust survival modeling in decentralized, privacy-sensitive environments, even with scarce and low-quality data.
- **Introducing FedSDS:** The FedSDS framework is proposed as a core component of this methodology. By incorporating SDG into FL, FedSDS enables collaborative training across institutions without requiring centralized data aggregation. This approach addresses data scarcity and heterogeneity, particularly in scenarios involving rare diseases and geographically biased datasets.
- **Leveraging VAE-BGM for SDG:** FedSDS employs the VAE-BGM model enhanced with artificial inductive bias techniques as described in [14]. This integration allows the generation of high-quality synthetic data in low-data scenarios, effectively augmenting datasets and enhancing the generalizability of SA models.
- **Addressing Non-IID Data in FL:** FedSDS introduces a novel approach to handling heterogeneous, non-IID datasets by replacing traditional parameter-sharing mechanisms with the exchange of synthetic data. This method ensures that model performance remains robust even when data distributions vary significantly across institutions. To achieve this, we take advantage of the encoder–decoder architecture of the SAVAE model, which enables latent space representation and facilitates the selection of relevant synthetic samples for biased aggregation.
- **Efficiency in Communication and Scalability:** Unlike traditional FL methods such as FedAvg or FedProx, which require multiple rounds of parameter sharing, FedSDS minimizes communication rounds by operating in a single round of synthetic data exchange. This design reduces communication overhead, making the framework scalable and suitable for resource-constrained institutions.
- **Evaluation on Realistic Scenarios:** We comprehensively evaluate FedSDS under both IID and non-IID scenarios using diverse SA datasets. This evaluation includes a direct comparison with the widely used FedAvg and FedProx techniques to assess the performance of our proposed method. The experiments demonstrate that FedSDS effectively overcomes data scarcity and heterogeneity while maintaining high model performance, highlighting its advantages over traditional FL approaches.

The remainder of this paper is organized as follows: Section 2 describes the methodology, including a detailed explanation of SA, the SAVAE model, and the FL techniques (FedAvg, FedProx, and FedSDS). Section 3 introduces the data, experimental settings, and evaluation scenarios, including both IID and non-IID configurations, and presents the results. Finally, Section 4 concludes the paper, summarizing the key findings, implications, and potential future research directions.

2. Methodology

2.1. Survival analysis

SA, or time-to-event analysis, focuses on modeling the time until an event occurs, accounting for censored cases where the event has not occurred by the end of observation. SA datasets consist of triplets (x_i, t_i, d_i), where x_i is a covariate vector, t_i is the observed time, and $d_i \in \{0, 1\}$ indicates censoring. The survival function $S(t|x) = P(T > t|x)$ represents the central quantity of interest, often modeled alongside the hazard function $h(t|x)$ and the density function $p(t|x)$.

In this work, we adopt SAVAE [11], a generative model based on VAEs, which estimates the time-to-event distribution through latent variables. SAVAE flexibly handles censoring and covariate complexity without relying on assumptions about proportional hazards. It models both covariates and survival times through an Evidence Lower BOund (ELBO) objective, with a Weibull likelihood for time modeling. An overview of SAVAE's architecture is depicted in Fig. 1. Full mathematical details are provided in Appendix A.

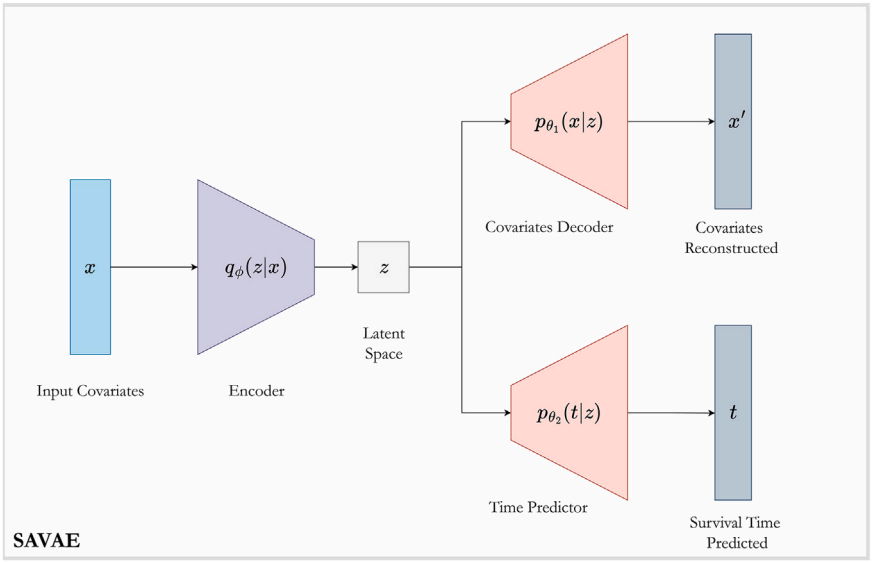


Fig. 1. Schematic representation of SAVAE. The encoder $q_\phi(z|x)$ maps the input covariates x into a latent Gaussian space z. This latent space is used by a decoder $p_{\theta_1}(x|z)$, which reconstructs the covariates x, and a predictor $p_{\theta_2}(x|z)$, which estimates the survival time t.

2.2. Federated learning

FL enables collaborative model training across decentralized institutions without sharing raw patient data, offering a compelling solution for privacy-preserving healthcare analytics. However, FL systems face significant challenges when data distributions are non-IID across clients, a common scenario in clinical settings due to demographic, procedural, and institutional differences. Most FL frameworks involve iterative local training followed by global aggregation via a central server. While effective under IID assumptions, this structure often suffers from degraded convergence and fairness in non-IID conditions.

2.2.1. State-of-the-art FL methods

FedAvg [23] remains the foundational FL algorithm. It performs multiple local update steps at each client and aggregates model parameters at a central server through weighted averaging. Despite its simplicity and scalability, FedAvg struggles when client data distributions are highly heterogeneous, leading to biased global models and slow convergence.

FedProx [24] extends FedAvg by adding a proximal term to the local loss function, discouraging local models from drifting too far from the global model. This stabilization technique improves convergence in non-IID environments but still relies on repeated communication rounds and enforces alignment to a global objective.

In this study, both FedAvg and FedProx are used as baseline methods for comparative evaluation against our proposed FedSDS framework. Full algorithmic descriptions of FedAvg and FedProx are provided in Appendix B.

2.2.2. FedSDS

FL has demonstrated its utility in enabling collaborative model training across decentralized institutions while preserving data privacy. However, traditional FL methods, such as FedAvg, encounter significant challenges in non-IID settings, where data distributions vary across nodes. These challenges often lead to imbalanced contributions from nodes, biased global models, and suboptimal convergence. Extensions like FedProx attempt to address these issues by constraining local model updates, improving stability under heterogeneity. Nevertheless, they still rely on repeated communication rounds and enforce alignment to a single global model, which may not be ideal in healthcare scenarios where preserving local data diversity is crucial. Recent works have explored synthetic data generation to alleviate non-IID challenges in FL [28]. These approaches typically generate auxiliary data to aid model generalization or simulate missing distributions across clients. However, they often lack fine-grained control over how synthetic samples are selected and integrated across heterogeneous nodes.

To address these issues, we propose FedSDS, an innovative FL strategy that leverages synthetic data generation to overcome the limitations of parameter-sharing methods in non-IID scenarios. FedSDS introduces a different paradigm to the state-of-the-art by leveraging synthetic data generation combined with latent-space-based filtering to selectively integrate synthetic samples that are most relevant to each local distribution. This targeted selection mechanism, to the best of our knowledge, has not been previously introduced in the context of FL. It allows FedSDS to respect and enhance the diversity of client distributions rather than forcing alignment toward a global average. Moreover, FedSDS shares conceptual similarities with one-shot FL [29], particularly in its goal to minimize communication overhead by exchanging external artifacts rather than iterative parameter updates. However, unlike one-shot FL, which aims to produce a global model from a single communication step, FedSDS focuses on iteratively improving local model performance through synthetic data augmentation without sacrificing the specificity of local datasets. Importantly, FedSDS also aligns with decentralized learning (DL) paradigms, where no central server is required and collaboration is achieved through peer-to-peer communication. Classical DL methods typically involve model sharing among nodes [30–32], whereas FedSDS departs from this approach by enabling data-centric collaboration. Nodes retain full autonomy over their models and exchange only synthetic data generated from local distributions, thereby preserving both model and data locality. This design provides an alternative approach to personalization and collaboration that eliminates the need for centralized coordination, potentially enhancing scalability and robustness in settings with stringent privacy or infrastructure constraints. Fig. 2 illustrates an overview of the system architecture.

At the core of FedSDS lies the VAE-BGM model (depicted in Fig. 3), a state-of-the-art approach for generating high-quality synthetic tabular data [13]. The VAE-BGM refines the latent space of a standard VAE by integrating a BGM, which models the latent representation as a mixture of multiple Gaussian components. Unlike traditional Gaussian priors, the BGM dynamically adjusts the number of components using a Dirichlet process, enabling the model to capture complex, multi-modal data distributions effectively. This flexibility allows the VAE-BGM to

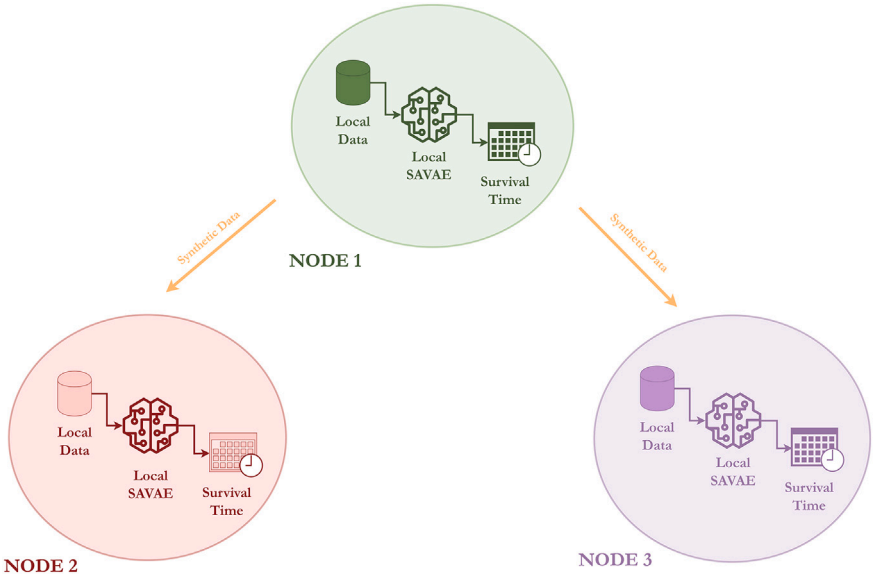


Fig. 2. Overview of the FedSDS framework. The best-performing node (Node 1) generates synthetic samples and shares them with the other nodes (Node 2 and Node 3). Each receiving node augments its local dataset with synthetic samples to improve survival model training.

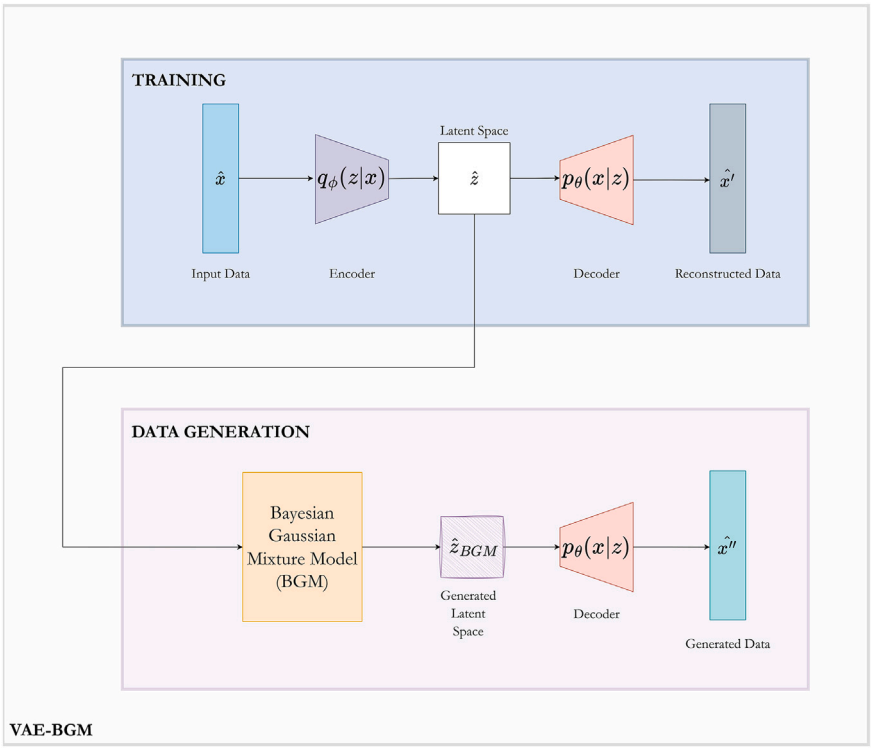


Fig. 3. Schematic representation of the two processes followed by the VAE-BGM. (1) VAE Training Process, where the input data $\hat{x}$ (covariates, time, and event) is encoded into a Gaussian latent space $\hat{z}$ and reconstructed to match the original data, optimizing the model's representation capabilities, and (2) Data Generation Process, where the latent space is refined using a BGM to generate a new latent representation $\hat{z}_{BGM}$, enabling the creation of high-quality synthetic data that reflects the original's data distributions.

handle mixed data types, including continuous, binary, and categorical variables, ensuring that synthetic datasets accurately reflect the underlying structure of real-world data. Additionally, in this framework, the latent space is crucial in enabling diversity and privacy preservation. Relying on a BGM introduces a more expressive distribution that captures multi-modal and heterogeneous latent structures, which is particularly important in medical datasets, where patient characteristics are rarely unimodal or homogeneous. The latent representations generated by the VAE-BGM reflect these complex structures. They are later used not only to reconstruct data but also to evaluate the similarity of synthetic samples to local data in the aggregation step. This approach brings two major benefits. First, the BGM enhances generative diversity, reducing the risk of generating samples that closely replicate real data. Second, it allows each node to filter incoming synthetic data based on their latent proximity to the node's real data, improving both personalization and robustness while adding a layer of implicit privacy protection. By leveraging this latent-space representation for filtering, FedSDS avoids direct reliance on raw features and instead compares samples in a compressed, abstracted space less susceptible to re-identification risks. In summary, this latent-space design not only enhances data utility but also provides an implicit safeguard against sample memorization, which is critical for privacy-sensitive applications like healthcare.

FedSDS incorporates the model-averaging technique proposed in [14] to enhance the model's robustness. This method leverages multiple training instances of the VAE-BGM, initialized with different random seeds. Instead of discarding poorly performing seeds, the approach averages the parameters of all trained models to create a robust inductive bias:

$$\theta_0 = \frac{1}{S}\sum_{s=1}^{S}\theta_s, \tag{1}$$

where S is the number of seeds and θ_s represents the model's parameters trained with seed s. This computationally efficient strategy ensures that the final model captures diverse perspectives from all training instances, significantly improving the quality of synthetic data generated, even in low-data scenarios.

This study generates synthetic data exclusively at the 'best' node, defined as the client with the most complete and representative dataset. This node is selected based on data availability and quality, ensuring a sufficiently diverse local distribution for training the VAE-BGM model. Generating synthetic data from this node helps avoid relying on clients with few or highly biased samples, which can compromise the stability of the generative model and the quality of the synthesized data. This strategy mirrors principles from FL aggregation schemes, such as FedAvg, where nodes with more data exercise greater influence. 5,000 synthetic samples are shared with each node from the generated pool of synthetic samples. This ensures that all nodes, especially those with limited or skewed data, receive high-quality synthetic data to augment their local datasets. While synthetic generation is centralized at a single node for robustness in this work, future extensions could explore distributed synthetic generation across multiple clients, provided the generative models are adapted to operate reliably under data-scarce or heterogeneous conditions.

FedSDS replaces traditional parameter-sharing in FL by exchanging synthetic data generated locally at each node. This framework includes the following key components:

1. **Local Synthetic Data Generation:** The 'best' node independently trains a VAE-BGM model according to the model-averaging technique proposed in [14] and specified in Eq. (1). By leveraging multiple initializations and averaging their parameters, each node generates high-quality synthetic datasets that capture the complexity and nuances of their local data. This process ensures

that the generated synthetic data retains critical patterns and correlations while preserving privacy.

2. **Data Sharing and Aggregation:** Once synthetic data are generated, 5,000 samples are shared with each node in the network. This number was determined through empirical validation. We observed that model performance improved with more synthetic data up to approximately 5000 samples in preliminary experiments. Beyond this point, performance gains plateaued while computational and memory costs increased. Thus, 5000 samples offered a good trade-off between model enhancement and computational efficiency in the considered datasets. To address potential biases and ensure effective aggregation, FedSDS employs two distinct strategies:
   - **Random Aggregation (*naive* case):** Synthetic data from other nodes is randomly combined with the local dataset. This straightforward approach provides a baseline for comparison but may introduce biases if the synthetic data significantly differs from the local data.
   - **Similarity-Based Aggregation (*biased* case):** Synthetic samples are filtered based on their proximity to the local data in the latent space, ensuring that only the most relevant samples are integrated. This approach preserves the inherent characteristics of the local dataset while leveraging the diversity of the shared synthetic data. A more detailed explanation of this technique, including its implementation and underlying methodology, is provided in a subsequent section to highlight this key contribution.
3. **Training with Augmented Data:** Each node trains its local SA model, such as SAVAE [11], using the augmented dataset that combines real and synthetic data. This process improves the generalization ability of the local model by leveraging the additional diversity introduced through data sharing.

A key advantage of FedSDS is its robustness in non-IID scenarios. By sharing synthetic data instead of model parameters, FedSDS accommodates nodes with varying data distributions, including those with missing covariates or demographic biases. Synthetic data from one node can compensate for unavailable features in another, enabling collaborative training even in heterogeneous environments. Additionally, FedSDS significantly reduces communication overhead compared to traditional FL methods. Unlike FedAvg, which requires multiple rounds of parameter-sharing to achieve convergence, FedSDS operates in a single communication round of synthetic data exchange. This reduction in communication rounds enhances scalability and efficiency, making the framework particularly suitable for bandwidth-constrained and resource-limited nodes.

***Biased* Aggregation Strategy**

The *biased* aggregation strategy represents a key technical innovation in the FedSDS framework, addressing the challenge of ensuring that synthetic data shared between nodes aligns closely with local data distributions. This alignment is crucial in non-IID settings, where data heterogeneity can lead to suboptimal updates if irrelevant synthetic samples are integrated. To achieve this alignment, the following steps are implemented:

1. **Latent Representation Generation:** Each synthetic sample generated at the 'best' node is passed through the encoder of the local SAVAE model at the receiving node. This process maps the synthetic data into a latent space, resulting in a vector representation $z_{synthetic}$ for each sample. Similarly, the local dataset at each receiving node is encoded into its latent space representation z_{local}.
2. **Proximity Calculation:** To evaluate the relevance of each synthetic sample, the distance between $z_{synthetic}$ and z_{local} is computed. This metric quantifies the similarity between the synthetic sample and the local data. The distance is calculated using the Euclidean norm as defined below:

$$d(z^i_{synthetic}, z^j_{local}) = \|z^i_{synthetic} - z^j_{local}\|, \tag{2}$$

   where $z^i_{synthetic}$ and z^j_{local} represent the latent space vectors of the i-th synthetic sample and the j-th local sample, respectively. The norm $|\cdot|$ denotes the Euclidean norm. For each synthetic sample $z^i_{synthetic}$, the minimum distance to all local samples z^j_{local} is calculated.
3. **Relevance Filtering:** Based on the calculated distances, only the most relevant synthetic samples with the smallest distances to z_{local} are selected for integration into the local dataset. The synthetic samples are ranked based on their computed d_{min} distances. To filter the most relevant samples, the synthetic points are sorted in ascending order of distance, and the top M samples with the smallest distances are selected. This ensures that the aggregated synthetic data complements the local data characteristics rather than introducing noise or misaligned distributions.
4. **Dataset Augmentation:** The selected synthetic samples are then integrated into the local dataset, enhancing its size and diversity while maintaining consistency with its original distribution.

By aligning the shared synthetic data with local distributions, the *biased* aggregation strategy minimizes the risk of introducing distributional noise. This alignment accelerates model convergence and enhances predictive performance. Additionally, the use of the encoder–decoder architecture of the SAVAE model allows nodes to dynamically filter synthetic data based on latent space proximity, making this approach adaptable to varying levels of heterogeneity. Moreover, unlike the *naive* aggregation, which can be computationally wasteful, the *biased* approach optimizes the selection process, ensuring that only the most relevant samples are used. By implementing this strategy, FedSDS ensures that the synthetic data exchange preserves the unique characteristics of each node's data while benefiting from the diversity introduced by external samples.

Fig. 4 illustrates the process of *biased* aggregation in detail, showcasing the steps from latent space representation to the selection of the most relevant synthetic samples.

2.3. Survival analysis and federated learning integration

The integration of SA into FL provides a promising framework to address the data scarcity challenges that have long hindered the advancement of SA models. This integration offers a scalable, privacy-preserving solution for learning from scarce, incomplete, or heterogeneous datasets. To illustrate the workflow of our proposed method, Fig. 5 provides a detailed overview of the FedSDS system, highlighting the interaction between synthetic data generation, aggregation, and local model training.

A key advantage of FedSDS over FedAvg lies in its communication efficiency. Traditional FL methods like FedAvg require multiple iterative rounds of parameter updates and aggregations to achieve convergence, which can be resource-intensive in bandwidth-constrained environments. In contrast, FedSDS operates in a single communication round, wherein nodes can share synthetic data or their generative models (including decoders and BGM parameters) with others. Sharing the generative model allows nodes to generate as much synthetic data as needed locally, enabling greater flexibility in addressing data scarcity and enhancing local model training. This single-round approach drastically reduces communication overhead while preserving model performance, making it particularly advantageous in settings with limited communication resources. FedSDS accelerates model convergence and enhances scalability for practical FL deployments by eliminating the need for iterative exchanges. In addition, FedSDS is particularly effective in non-IID environments, where nodes may have different covariates or missing features. By sharing synthetic data tailored to each node's local

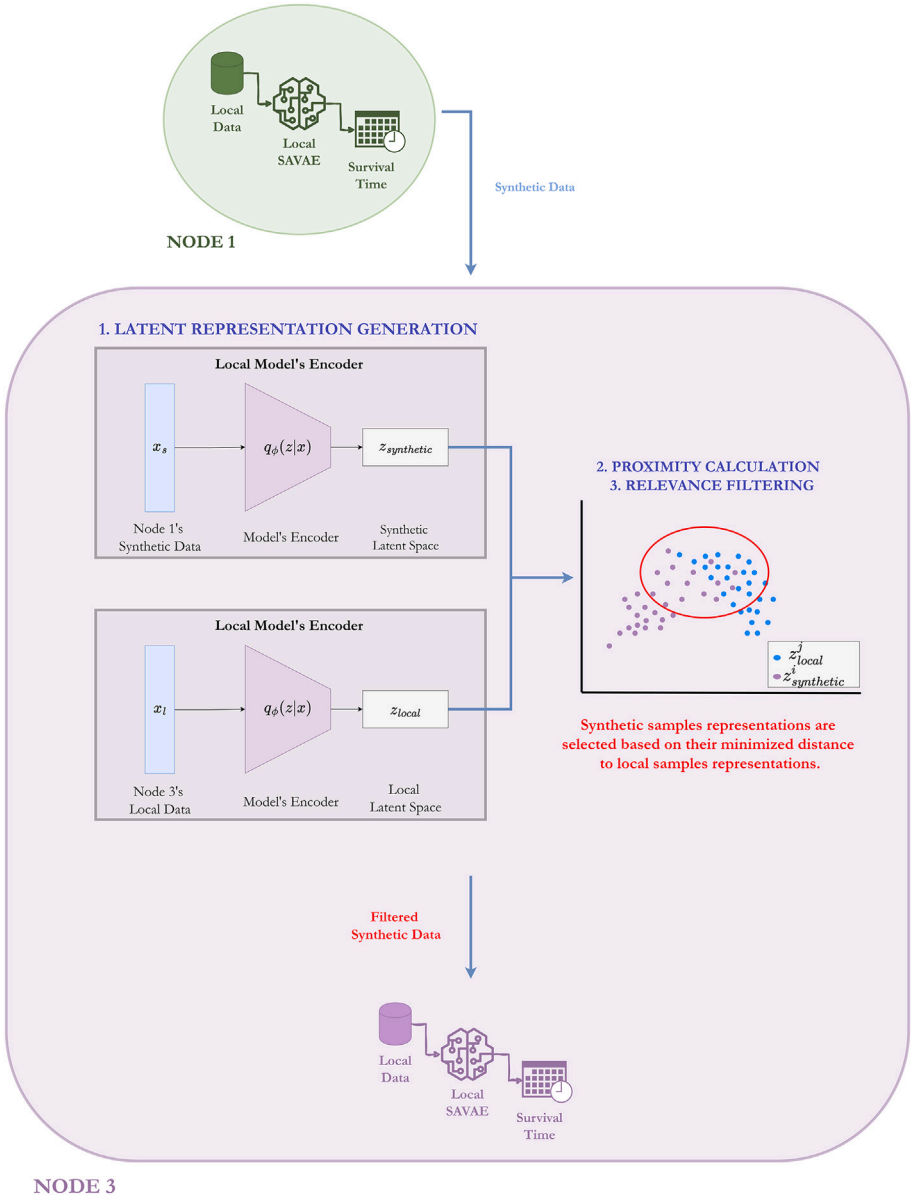


Fig. 4. Schematic representation of the *biased* aggregation process. Node 1 generates and shares synthetic data with Node 2, where latent representations $z_{synthetic}$ and z_{local} are generated using the SAVAE encoder. Relevant synthetic samples are selected based on minimal distance to z_{local} and integrated into the local dataset.

characteristics, FedSDS allows nodes to integrate data that complement their own, ensuring robust and unbiased updates to the global model. For example, synthetic data from one node can help another node predict unavailable covariates, enhancing its local model's performance. FedAvg, by contrast, cannot address these disparities, often resulting in suboptimal global models in heterogeneous settings.

3. Experiments

3.1. Survival data

This study used two well-established SA medical datasets to evaluate the proposed methodology: METABRIC and GBSG. These datasets encompass a wide range of disease domains, patient demographics, and event characteristics, providing a comprehensive framework to assess the performance and generalizability of the VAE-based models for SA and SDG, SAVAE, and VAE-BGM, respectively. Each dataset contributes unique features, including censored and uncensored observations, follow-up times, and clinical covariates, reflecting the diverse challenges commonly encountered in SA. The following section highlights the key attributes of each dataset:

- **METABRIC:** Derived from the Molecular Taxonomy of Breast Cancer International Consortium, this dataset includes genomic and clinical data from 1904 breast cancer patients [33]. Approximately 42.07% of the samples are censored, with an average event time of 125.03 (ranging from 0.0 to 355.20 days). The dataset comprises 9 covariates, excluding the survival-specific variables (event time and event indicator), making it highly suitable for modeling complex relationships between genetic mutations and survival outcomes.
- **GBSG:** Combining data from node-positive breast cancer patients and a chemotherapy trial, this dataset [34,35] includes 2232 samples and seven covariates. Approximately 43.23% of the samples are censored, with event times averaging 44.49 months(ranging from 0.26 to 87.36 months). This dataset is ideal for evaluating survival models in oncology settings.

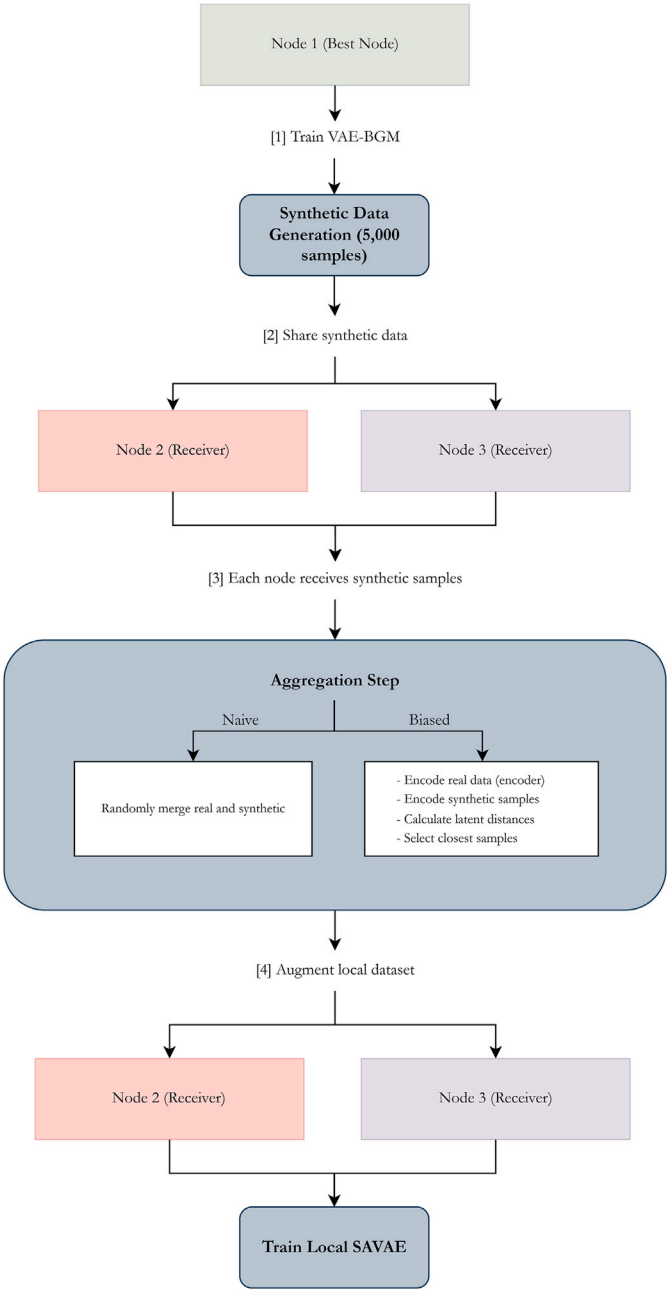


Fig. 5. Overview of the FedSDS workflow. Node 1 (best node) trains a VAE-BGM model and generates 5000 synthetic samples. These samples are shared with other nodes (Node 2 and Node 3). Upon reception, nodes perform either naive aggregation (random merging) or biased aggregation (latent-space filtering) to augment their local datasets before training their local SAVAE models.

These datasets illustrate the common challenges faced in real-world survival analysis, including data scarcity, heterogeneity, and high proportions of censored observations. Their diversity ensures that the proposed method's ability to handle such challenges is rigorously tested. The datasets' main characteristics are summarized in Table 1.

In addition to these two datasets, we have further incorporated an additional evaluation using an open real-world clinical dataset extracted from the publicly available TCGA Pan-Cancer resource [36], as described in Appendix D. This dataset includes information about the tissue source site, which allows us to partition the data across nodes, simulating a decentralized multi-institutional scenario. This additional experiment provides further evidence of the applicability of FedSDS under heterogeneous and realistic data distributions.

Table 1
Summary of the SA datasets used in this study. These datasets present diversity in sample size, number of features, proportion of censored observations, and data types. This variety ensures a robust assessment of the proposed methodology across different clinical and demographic settings, addressing key challenges such as data scarcity, heterogeneity, and censoring.

Dataset	Number of **samples**	Number of **features**	Number of **censored**	Data types
METABRIC	1904	9	801 (42.07%)	Binary and continuous
GBSG	2232	7	965 (43.23%)	Binary, continuous and discrete

3.2. Experimental design

This study evaluates the proposed methodology using distributed data across three nodes, simulating IID and non-IID scenarios (Table 2). The experimental design is structured to assess the performance of the proposed framework under diverse data distribution settings and varying degrees of heterogeneity. Given the small sample sizes of the original datasets, synthetic data was generated using the methodology proposed in [14] to ensure sufficient data to design the various nodes and scenarios. Below, we describe the distribution of data across nodes and the specific scenarios considered for each case.

Three nodes, each differing in terms of the quantity of data they possess, are used for the experiments. The study examines two primary cases: IID and non-IID data distributions. In each case, three distinct scenarios are defined to reflect varying data allocation strategies and the introduction of heterogeneity. Additionally, we include a special non-IID scenario designed to test the ability of the FedSDS framework to handle missing covariates.

3.2.1. IID data distribution

In the IID setting, the data across nodes follow similar distributions regarding covariates and event times. Three distinct scenarios are defined:

- **Scenario 1. Equal data distribution across nodes:** Each of the three nodes is allocated 3000 samples, with 2000 samples used for training and 1000 for validation. This scenario represents the most balanced and favorable case for FL, ensuring equal contributions from all nodes.
- **Scenario 2. Unequal data distribution across nodes:** The nodes have differing quantities of samples: Node 1 has 3000 samples, Node 2 has 1500 samples, and Node 3 has 1050 samples. Validation sets of 1000 samples are maintained for all nodes, leaving 2000, 500, and 50 samples, respectively, for training. This scenario introduces imbalances in data quantity between nodes, simulating real-world conditions.
- **Scenario 3. Unequal data distribution with missing data:** This scenario builds on Scenario 2's distribution of samples but introduces missing data in Node 3. Specifically, 50% of the data is missing. This scenario highlights the challenge of handling incomplete datasets in FL.

3.2.2. Non-IID data distribution

In the non-IID setting, the data across nodes are heterogeneous regarding covariate distributions. Three scenarios analogous to the IID case are defined but with added heterogeneity in the distribution of a key covariate, *regarding*, which is critical for SA.

- **Scenario 4. Equal data distribution with heterogeneous covariates:** Each node is allocated 2000 training and 1000 validation samples. However, heterogeneity is introduced in the *age* covariate:

Table 2
Overview of scenarios defined for the experimental design. Data quantity and quality across nodes under IID and non-IID conditions are detailed. Node 1, Node 2, and Node 3 refer to training data allocated across the nodes in each scenario.

Case	Scenario	Data Distribution	Node 1	Node 2	Node 3	Data Quality
IID	Scenario 1	Equal	2000	2000	2000	Homogeneous
IID	Scenario 2	Unequal	2000	500	50	Homogeneous
IID	Scenario 3	Unequal with missing data	2000	500	25 (50% missing)	Homogeneous
Non-IID	Scenario 4	Equal	2000	2000	2000	Node 2: 95% below median age, Node 3: 95% above median age
Non-IID	Scenario 5	Unequal	2000	500	50	Node 2: 95% below median age, Node 3: 95% above median age
Non-IID	Scenario 6	Unequal with missing data	2000	500	25 (50% missing)	Node 2: 95% below median age, Node 3: 95% above median age
Non-IID	Scenario 7	Two nodes, Node 2 lacks *age* covariate	3500	3500	N/A	Node 2: *age* covariate predicted from Node 1's synthetic data

  - Node 1: Uniform distribution of *age* samples.
  - Node 2: 95% of samples have ages below the median, with only 5% above the median.
  - Node 3: 95% of samples have ages above the median, with only 5% below the median.
- **Scenario 5. Unequal data distribution with heterogeneous covariates:** This scenario mirrors Scenario 2 regarding data quantity across nodes but incorporates the same heterogeneity in the age covariate as Scenario 4.
- **Scenario 6. Unequal data distribution with missing data and heterogeneous covariates:** This scenario extends Scenario 3 by introducing heterogeneity in the age covariate for Nodes 2 and 3, as defined in Scenario 4. Node 3 also retains 50% missing data, making this the most challenging non-IID scenario for both FL techniques.
- **Scenario 7. Addressing missing covariates:** This scenario cannot be evaluated using traditional parameter-sharing methods like FedAvg and FedProx, and is designed specifically to test the capabilities of FedSDS. Two nodes are used, each with 3500 training and 1000 validation samples. However, Node 2 lacks a crucial covariate—*age*—significantly impacting SA model performance. Using FedSDS, synthetic data generated by Node 1 is employed to train a predictor for the missing *age* column in Node 2. This predicted column is then integrated into Node 2's dataset, allowing the FedSDS framework to handle this non-IID scenario effectively and demonstrate its adaptability to situations with fully missing covariates.

3.3. Network architecture

3.3.1. Local SAVAE model at each node

The architecture of the SAVAE model implemented at each node follows the design outlined in [11]. The model comprises three different Deep NNs: one encoder and two decoders. The encoder maps the input data to a Gaussian latent space while the decoders reconstruct covariates and time parameters. The encoder features a simple architecture consisting of a single hidden layer with 50 neurons and a rectified linear unit activation function. The output layer applies a hyperbolic tangent activation function to generate the Gaussian latent space. The encoder processes covariate vectors from the training dataset, projecting them into a latent space of a fixed dimensionality ($d_z = 5$). The latent representation generated by the encoder serves as input to both decoders. Each decoder comprises two linear layers. The first layer employs 50 neurons, a ReLU activation function, and a dropout rate of 20% to mitigate overfitting. The second layer features activation functions tailored to the specific distributions of the covariates or time parameters, ensuring output suitability. Early stopping is applied during training to prevent overfitting, with a batch size of 250. Furthermore, due to the inherent variability in VAE initialization, the SAVAE model is trained using 10 different random seeds for robustness.

3.4. Synthetic data generation with VAE-BGM

We employ the VAE-BGM model for synthetic data generation as described in [13]. This model introduces a BGM in the VAE framework, enhancing the flexibility and expressiveness of the latent space. The VAE-BGM model architecture mirrors that of the SAVAE, featuring an encoder with a hidden ReLU layer of 50 neurons and a hyperbolic tangent output layer alongside a decoder that adapts activation functions to match the distributions of the covariates. The latent space dimensionality is set to $d = 5$, balancing feature representation and model complexity. Dropout with a 20% rate is incorporated to mitigate overfitting, and the model is trained for up to 10,000 epochs with early stopping, using a batch size of 250. A key enhancement of VAE-BGM is its treatment of the latent space as a mixture of Gaussian distributions. This is achieved using a Dirichlet process prior with a maximum of 20 components, allowing the model to adjust to the complexity of the underlying data dynamically. Each Gaussian component is parameterized with its covariance matrix, enabling the model to capture intricate dependencies and non-Gaussian structures in the data effectively. To further refine the VAE-BGM's performance in low-data scenarios, we integrate the model-averaging technique proposed in [14]. This approach leverages multiple training runs (10 seeds in our implementation) with different initializations to enhance robustness.

3.5. Federated learning framework

In the FL framework, the three techniques, FedAvg, FedProx, and FedSDS, are applied differently:

- **FedAvg and FedProx**: These techniques involve iterative training with five federated steps. During each step, nodes share their locally updated SAVAE model weights for aggregation, followed by a global update. This process ensures a progressively refined global model but incurs higher communication costs due to the multiple rounds of information exchange.
- **FedSDS**: In contrast, in FedSDS, each node trains the VAE-BGM locally to generate synthetic data. The generated synthetic datasets are shared only in the first round of FL, eliminating the need for iterative communication rounds. This approach significantly reduces communication overhead, making FedSDS a scalable and efficient solution for real-world FL applications.

To evaluate the performance of the SA models in each node under the isolated case and the FL techniques, the final results are averaged over the three best runs with different random seeds. This approach accounts for the sensitivity of VAEs to initialization, ensuring a robust and comprehensive evaluation of the models.

3.6. Evaluation metrics

Each dataset is characterized by triplets $D = (x_i, t_i, d_i)_{i=1}^{N}$, where x_i represents the covariate vector, t_i denotes the time-to-event, and $d_i \in \{0, 1\}$ indicates the censoring status.

The Concordance Index (C-index) is a widely used metric in SA for evaluating a model's ability to rank predicted risks relative to observed times. It generalizes the ROC curve by measuring the rank correlation between predicted risk scores and observed event times. A model is considered effective if higher predicted risks correspond to shorter time-to-event. This analysis employs the time-dependent C-index [37], which extends the original formulation [38] by accounting for dynamic changes in risk over time. The time-dependent C-index is defined as:

$$C_{index} = P\Big(\hat{F}(t|x_i) > \hat{F}(t|x_j) | d_i = 1, t_i < t_j, t_i \leq t\Big), \tag{3}$$

where $\hat{F}(t|x_i)$ is the CDF estimated by the model at time t for covariates x_i. This probability is determined by comparing relative risks pairwise, emphasizing the model's ranking performance.

Finally, hypothesis testing was conducted to compare the mean C-index values of the FL cases against the isolated cases to ensure robust evaluation. The null hypothesis assumes that isolated training would yield worse C-index values (indicating worse performance of isolated training) than FL techniques. The validity of this hypothesis was assessed using p-values, with a significance threshold set at 0.05. A p-value below this threshold led to the rejection of the null hypothesis, indicating statistically significant superiority of the FL techniques. Conversely, a p-value exceeding 0.05 suggested no significant difference between the cases. This statistical approach ensured a comprehensive evaluation of model performance by accounting for variations in the metrics across different experiments.

Given the multiple hypothesis tests conducted, the risk of Type I errors (false positives) increases with the number of tests, as noted in [39–41]. To mitigate this, the Holm adjustment of the p-values [42] was applied, effectively controlling Family-Wise Error Rate (FWER) inflation and ensuring the reliability of the statistical conclusions.

3.7. Results

This section presents the C-index performance comparisons across various scenarios for IID and non-IID distributions on each dataset. Additional results using the Integrated Brier Score (IBS), which evaluates calibration and discrimination, are included in Appendix C. Furthermore, an empirical analysis of convergence behavior during training is provided in Appendix F, demonstrating the stability and robustness of the proposed framework across all experiments. The code for reproducing all experiments is available at https://github.com/Patricia-A-Apellaniz/fed_savae.

3.7.1. IID scenarios

The results below compare the performance of various approaches: isolated node training, FedAvg, FedProx, and the proposed FedSDS framework under both naive and biased synthetic data aggregation strategies. The results are shown for three distinct scenarios across three nodes, including centralized and isolated settings. Each cell reports the C-index values in the format (lower bound - mean - upper bound), reflecting the variability of results across multiple runs. Additionally, adjusted p-values are provided to evaluate the statistical significance of the differences between FL methods. Significant values (less than 0.05) are highlighted in bold. This analysis is replicated across all datasets used in the study to ensure a comprehensive evaluation.

The results for the METABRIC and GBSG datasets, presented in Tables 3 and 4, highlight distinct patterns across scenarios.

In METABRIC, Scenario 1, with equal data distribution, shows no significant improvement from FL methods. In contrast, in Scenario 2, disadvantaged nodes (Node 2 and Node 3) achieve substantial C-index gains. Both FedProx and FedSDS *biased* improve performance significantly. However, the gains are more pronounced with FedSDS *biased*. Scenario 3 highlights significant improvements for Node 2 with FedProx and FedSDS methods, particularly the *biased* strategy, and for Node 3 across all FL techniques, again favoring FedSDS *biased*.

In GBSG, Scenario 1 reveals significant gains in Node 3, attributed to using FedAvg and including synthetic data. Scenario 2 sees improvements for Node 2 with FedAvg, FedProx, and FedSDS *biased*, while Node 3 benefits from all FL techniques, with FedSDS *biased* yielding the strongest results. Scenario 3 follows a similar trend, with Node 2 improving significantly with FedProx and FedSDS methods, and Node 3 showing gains across all techniques, again led by FedSDS *biased*.

3.7.2. Non-IID scenarios

The following results are depicted as in the IID scenarios. C-index values are presented comparing isolated, FedAvg, FedProx, and FedSDS methods (*naive* and *biased*) for the different datasets under non-IID settings. In these scenarios, Nodes 2 and 3 exhibit biased distributions in the covariate *age* while maintaining the same number of samples as the corresponding IID scenarios. Specifically, the age distributions are deliberately skewed across nodes to introduce heterogeneity, creating a realistic challenge for FL techniques.

Tables 5 and 6 present the results for METABRIC and GBSG in non-IID scenarios, highlighting consistent patterns in the most disadvantaged nodes.

In METABRIC, Scenario 4, we observe a significant improvement at Node 3 with FedProx, reflecting its ability to handle mild heterogeneity. In Scenarios 5 and 6, Nodes 2 and 3 benefit significantly using FL techniques. Node 2 improves with FedAvg and FedSDS *biased* in Scenario 5, while Node 3 achieves gains across all FL techniques. A similar trend appears in Scenario 6, where Node 3 significantly improves

Table 3
C-index comparison of isolated, FedAvg, FedProx, and FedSDS (*naive* and *biased*) methods for the METABRIC dataset in IID scenarios. Average C-index results are shown with confidence intervals. Adjusted p-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Scenario	Nodes	Isolated	FedAvg	FedProx	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	Node 1	(0.619 - 0.647 - 0.673)	–	–	–	–	–
Scenario 1	Node 1	(0.606 - 0.637 - 0.666)	(0.611 - 0.638 - 0.666)	(0.615 - 0.643 - 0.670)	(0.613 - 0.641 - 0.668)	(0.609 - 0.640 - 0.671)	1.000/0.158/0.442/1.000
	Node 2	(0.621 - 0.648 - 0.676)	(0.625 - 0.652 - 0.681)	(0.624 - 0.653 - 0.682)	(0.617 - 0.646 - 0.676)	(0.623 - 0.651 - 0.679)	0.238/0.147/1.000/0.570
	Node 3	(0.621 - 0.649 - 0.677)	(0.620 - 0.648 - 0.677)	(0.624 - 0.651 - 0.679)	(0.619 - 0.647 - 0.677)	(0.620 - 0.652 - 0.681)	1.000/0.691/1.000/1.000
Scenario 2	Node 1	(0.613 - 0.641 - 0.669)	(0.608 - 0.635 - 0.662)	(0.615 - 0.642 - 0.670)	(0.611 - 0.638 - 0.666)	(0.612 - 0.640 - 0.669)	1.000/1.000/1.000/1.000
	Node 2	(0.576 - 0.608 - 0.644)	(0.589 - 0.621 - 0.658)	(0.597 - 0.626 - 0.656)	(0.583 - 0.628 - 0.673)	(0.611 - 0.640 - 0.673)	0.147/**0.023**/0.273/**0.001**
	Node 3	(0.587 - 0.621 - 0.651)	(0.598 - 0.629 - 0.660)	(0.613 - 0.641 - 0.670)	(0.595 - 0.634 - 0.675)	(0.614 - 0.645 - 0.678)	0.204 /**0.003**/0.421/**0.001**
Scenario 3	Node 1	(0.608 - 0.638 - 0.666)	(0.612 - 0.640 - 0.669)	(0.616 - 0.644 - 0.671)	(0.609 - 0.637 - 0.663)	(0.607 - 0.636 - 0.665)	0.771/0.052/1.000/1.000
	Node 2	(0.577 - 0.610 - 0.641)	(0.589 - 0.622 - 0.659)	(0.593 - 0.624 - 0.657)	(0.592 - 0.631 - 0.667)	(0.614 - 0.644 - 0.673)	0.273/**0.019**/**0.032**/**0.000**
	Node 3	(0.492 - 0.542 - 0.580)	(0.561 - 0.593 - 0.622)	(0.548 - 0.578 - 0.609)	(0.557 - 0.590 - 0.623)	(0.568 - 0.601 - 0.637)	**0.008**/**0.032**/**0.006**/**0.002**

Table 4
C-index comparison of isolated, FedAvg, FedProx, and FedSDS (*naive* and *biased*) methods for the GBSG dataset in IID scenarios. Average C-index results are shown with confidence intervals. Adjusted *p*-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Scenario	Nodes	Isolated	FedAvg	FedProx	FedSDS *naive*	FedSDS *biased*	Adjusted *p*-values
Centralized	Node 1	(0.660 - 0.688 - 0.714)	–	–	–	–	–
Scenario 1	Node 1	(0.643 - 0.673 - 0.702)	(0.648 - 0.674 - 0.699)	(0.648 - 0.675 - 0.701)	(0.646 - 0.673 - 0.698)	(0.646 - 0.673 - 0.700)	1.000/1.000/1.000/1.000
	Node 2	(0.648 - 0.677 - 0.706)	(0.654 - 0.681 - 0.707)	(0.652 - 0.680 - 0.706)	(0.649 - 0.681 - 0.711)	(0.650 - 0.684 - 0.715)	0.253/0.399/0.460/0.338
	Node 3	(0.658 - 0.686 - 0.714)	(0.666 - 0.693 - 0.721)	(0.662 - 0.692 - 0.723)	(0.661 - 0.691 - 0.720)	(0.664 - 0.693 - 0.719)	**0.021**/0.215/0.233/**0.019**
Scenario 2	Node 1	(0.646 - 0.672 - 0.697)	(0.647 - 0.674 - 0.701)	(0.646 - 0.673 - 0.700)	(0.646 - 0.673 - 0.699)	(0.645 - 0.673 - 0.701)	0.143/0.399/0.360/1.000
	Node 2	(0.620 - 0.650 - 0.680)	(0.628 - 0.658 - 0.686)	(0.632 - 0.661 - 0.690)	(0.624 - 0.656 - 0.686)	(0.634 - 0.664 - 0.695)	**0.039**/**0.009**/0.338/**0.010**
	Node 3	(0.563 - 0.601 - 0.649)	(0.627 - 0.666 - 0.710)	(0.633 - 0.667 - 0.703)	(0.633 - 0.668 - 0.696)	(0.650 - 0.683 - 0.714)	**0.000**/**0.001**/**0.001**/**0.000**
Scenario 3	Node 1	(0.646 - 0.673 - 0.698)	(0.647 - 0.674 - 0.700)	(0.647 - 0.675 - 0.702)	(0.643 - 0.672 - 0.699)	(0.644 - 0.671 - 0.697)	0.350/0.353/1.000/1.000
	Node 2	(0.617 - 0.649 - 0.679)	(0.625 - 0.656 - 0.685)	(0.635 - 0.662 - 0.690)	(0.625 - 0.662 - 0.692)	(0.633 - 0.666 - 0.701)	0.233/**0.010**/0.060 /**0.015**
	Node 3	(0.502 - 0.545 - 0.603)	(0.565 - 0.597 - 0.630)	(0.589 - 0.618 - 0.645)	(0.580 - 0.609 - 0.638)	(0.578 - 0.609 - 0.647)	**0.017**/**0.009**/**0.012**/**0.007**

Table 5
C-index comparison of isolated, FedAvg, FedProx, and FedSDS (*naive* and *biased*) methods for the METABRIC dataset in non-IID scenarios. Average C-index results are shown with confidence intervals. Adjusted *p*-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Scenario	Nodes	Isolated	FedAvg	FedProx	FedSDS *naive*	FedSDS *biased*	Adjusted *p*-values
Centralized	Node 1	(0.619 - 0.647 - 0.673)	–	–	–	–	–
Scenario 4	Node 1	(0.611 - 0.638 - 0.664)	(0.611 - 0.638 - 0.665)	(0.613 - 0.641 - 0.669)	(0.612 - 0.642 - 0.671)	(0.612 - 0.640 - 0.668)	1.000/0.173/0.334/0.623
	Node 2	(0.586 - 0.617 - 0.649)	(0.586 - 0.616 - 0.646)	(0.586 - 0.614 - 0.643)	(0.579 - 0.615 - 0.653)	(0.590 - 0.619 - 0.648)	1.000/1.000/1.000/1.000
	Node 3	(0.616 - 0.646 - 0.675)	(0.622 - 0.649 - 0.676)	(0.626 - 0.654 - 0.680)	(0.617 - 0.646 - 0.674)	(0.622 - 0.650 - 0.679)	0.564 /**0.029**/1.000/0.291
Scenario 5	Node 1	(0.612 - 0.640 - 0.670)	(0.610 - 0.638 - 0.665)	(0.614 - 0.641 - 0.669)	(0.610 - 0.639 - 0.669)	(0.613 - 0.641 - 0.672)	1.000/1.000/1.000/1.000
	Node 2	(0.582 - 0.616 - 0.646)	(0.600 - 0.631 - 0.661)	(0.595 - 0.622 - 0.649)	(0.582 - 0.625 - 0.673)	(0.602 - 0.635 - 0.667)	**0.005**/0.291/1.000/**0.005**
	Node 3	(0.502 - 0.545 - 0.589)	(0.566 - 0.599 - 0.634)	(0.566 - 0.600 - 0.632)	(0.559 - 0.590 - 0.622)	(0.563 - 0.594 - 0.632)	**0.003**/**0.003**/**0.007**/**0.003**
Scenario 6	Node 1	(0.607 - 0.638 - 0.667)	(0.611 - 0.639 - 0.667)	(0.615 - 0.643 - 0.673)	(0.611 - 0.640 - 0.669)	(0.609 - 0.636 - 0.665)	1.000/0.167/1.000/1.000
	Node 2	(0.589 - 0.620 - 0.653)	(0.593 - 0.625 - 0.655)	(0.600 - 0.630 - 0.666)	(0.576 - 0.620 - 0.660)	(0.612 - 0.643 - 0.671)	1.000/0.195/1.000/**0.002**
	Node 3	(0.496 - 0.533 - 0.567)	(0.523 - 0.557 - 0.592)	(0.520 - 0.553 - 0.586)	(0.525 - 0.563 - 0.603)	(0.544 - 0.573 - 0.603)	**0.003**/**0.007**/**0.007**/**0.001**

Table 6
C-index comparison of isolated, FedAvg, FedProx, and FedSDS (*naive* and *biased*) methods for the GBSG dataset in non-IID scenarios. Average C-index results are shown with confidence intervals. Adjusted *p*-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Scenario	Nodes	Isolated	FedAvg	FedProx	FedSDS *naive*	FedSDS *biased*	Adjusted *p*-values
Centralize	Node 1	(0.660 - 0.688 - 0.714)	–	–	–	–	–
Scenario 4	Node 1	(0.647 - 0.674 - 0.701)	(0.645 - 0.673 - 0.701)	(0.646 - 0.674 - 0.702)	(0.645 - 0.673 - 0.700)	(0.644 - 0.675 - 0.703)	1.000/1.000/1.000/1.000
	Node 2	(0.666 - 0.693 - 0.720)	(0.673 - 0.699 - 0.727)	(0.671 - 0.697 - 0.724)	(0.659 - 0.689 - 0.719)	(0.665 - 0.691 - 0.718)	**0.017**/0.116/1.000/1.000
	Node 3	(0.657 - 0.686 - 0.712)	(0.665 - 0.692 - 0.718)	(0.665 - 0.692 - 0.718)	(0.651 - 0.684 - 0.713)	(0.663 - 0.691 - 0.717)	**0.018**/**0.017**/1.000/**0.050**
Scenario 5	Node 1	(0.643 - 0.673 - 0.699)	(0.646 - 0.675 - 0.701)	(0.647 - 0.673 - 0.700)	(0.648 - 0.677 - 0.704)	(0.644 - 0.672 - 0.702)	0.989/1.000/0.149/1.000
	Node 2	(0.636 - 0.672 - 0.708)	(0.660 - 0.686 - 0.711)	(0.657 - 0.686 - 0.716)	(0.648 - 0.680 - 0.710)	(0.648 - 0.681 - 0.711)	0.222/0.166/0.823/0.662
	Node 3	(0.606 - 0.642 - 0.675)	(0.638 - 0.668 - 0.698)	(0.642 - 0.671 - 0.702)	(0.623 - 0.663 - 0.699)	(0.632 - 0.663 - 0.694)	**0.005**/**0.002**/0.050/**0.012**
Scenario 6	Node 1	(0.641 - 0.671 - 0.698)	(0.648 - 0.674 - 0.700)	(0.645 - 0.673 - 0.699)	(0.642 - 0.671 - 0.697)	(0.645 - 0.672 - 0.700)	0.330/1.000/1.000/1.000
	Node 2	(0.641 - 0.676 - 0.706)	(0.655 - 0.685 - 0.713)	(0.656 - 0.684 - 0.713)	(0.635 - 0.670 - 0.699)	(0.653 - 0.682 - 0.709)	0.254/0.293/1.000/0.517
	Node 3	(0.510 - 0.556 - 0.608)	(0.571 - 0.600 - 0.629)	(0.555 - 0.588 - 0.622)	(0.581 - 0.611 - 0.638)	(0.581 - 0.616 - 0.651)	**0.045**/0.106 /**0.021**/**0.008**

all methods, but FedSDS *biased* yields the most substantial gain, while Node 2 improves only with this approach.

For GBSG, Scenario 4 reveals improvements in Node 3 using FedAvg, FedProx, and FedSDS *biased*, but Node 2 sees improvement just using FedAvg. In Scenario 5, Node 3 presents significant gains across FedAvg, FedProx, and FedSDS *biased*. Finally, in Scenario 6, the most disadvantaged Node 3 benefits from all techniques except FedProx, achieving the largest performance improvement with FedSDS *biased*.

3.7.3. Special IID scenario

In Scenario 7, the evaluation focuses on a unique and challenging non-IID case where one of the two nodes lacks the *age* covariate. Due to this limitation, FedAvg and FedProx are not applicable, leaving FedSDS-based techniques as the only viable approach. Three distinct configurations are tested to assess the performance and adaptability of FedSDS, each leveraging synthetic data to compensate for the missing information and enhance the model's performance.

The first approach, labeled *Imputation*, involves generating synthetic data at the 'good' node (the one with complete information) and training a predictor to estimate the missing *age* column for the 'bad' node (the one lacking the *age* covariate). These imputed data are incorporated into the 'bad' node training process. The second configuration, *Imputation + Synthetic data naive*, extends the first by augmenting the 'bad' node's dataset with synthetic data generated by the 'good' node and aggregated using the *naive* technique. The third configuration, *Imputation + Synthetic data biased*, further refines the second by

Table 7
C-index comparison in Scenario 7 of the three different FedSDS settings. Average C-index results are shown with confidence intervals. Adjusted *p*-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Dataset	Isolated	Imputation	Imputation + Synthetic Data *naive*	Imputation + Synthetic Data *biased*	Adjusted *p*-values
METABRIC	(0.544 - 0.572 - 0.602)	(0.602 - 0.634 - 0.669)	(0.594 - 0.629 - 0.666)	(0.602 - 0.631 - 0.661)	**0.000/0.000/0.000**
GBSG	(0.598 - 0.625 - 0.652)	(0.629 - 0.663 - 0.697)	(0.629 - 0.660 - 0.692)	(0.639 - 0.670 - 0.697)	**0.000/0.000/0.000**

employing the *biased* aggregation technique to ensure that the synthetic data align closely with the local data distribution of the 'bad' node.

The results in Table 7 present the adjusted *p*-values for the C-index improvement in the 'bad' node across the three configurations. All datasets—METABRIC and GBSG—demonstrate significant improvements in the C-index for each configuration compared to the isolated case, with *p*-values consistently below 0.05. Notably, the datasets achieve robust improvements, with *p*-values of 0.000 across all configurations, showcasing the effectiveness of FedSDS even in the most challenging non-IID scenarios.

3.7.4. Discussion

The results of this study highlight the effectiveness of FL approaches, particularly the proposed FedSDS framework, in addressing challenges such as imbalanced data distributions, heterogeneity, and missing critical covariates. Across both IID and non-IID scenarios, FedSDS consistently outperforms traditional FL methods like FedAvg and also advanced non-IID methods such as FedProx, especially when using the *biased* aggregation strategy, which shows clear advantages over the *naive* approach. FedSDS *biased* consistently performs better than its *naive* counterpart due to its ability to align synthetic data more closely with the local distributions of each node. The *naive* strategy aggregates synthetic data randomly, which can introduce noise or misaligned distributions, particularly in nodes with highly skewed or heterogeneous data. This misalignment may lead to suboptimal model updates and hinder performance improvement. In contrast, the *biased* strategy filters synthetic samples based on their similarity to the node's local data in the latent space, ensuring that the aggregated synthetic data complements the node's unique characteristics. This targeted alignment enhances model convergence and improves the ability of FedSDS to handle nodes with extreme data imbalances or biases.

In IID scenarios, where data are assumed to be distributed identically across nodes, FedSDS shows clear advantages in cases of imbalance. While not too many significant improvements are observed in scenarios with equal sample distribution across nodes (Scenario 1), FedSDS, particularly the *biased* aggregation strategy, consistently improves performance in scenarios where nodes face data scarcity (Scenarios 2 and 3). In these cases, although FedProx also achieves some gains by stabilizing the training dynamics, FedSDS *biased* leads to more substantial improvements by augmenting scarce data without altering local data distributions. These improvements are most pronounced for nodes with fewer samples, reflecting the framework's ability to mitigate the impact of data imbalance by effectively leveraging synthetic data generated during training.

The performance gap becomes even more evident in non-IID scenarios, where nodes exhibit biased distributions in key covariates. FedSDS *biased* consistently provides the most significant improvements. Although FedProx improves performance by reducing client drift and enforcing proximity to a global model, it still aims for global consensus. In contrast, FedSDS is specifically designed to respect and enhance local data characteristics by selectively augmenting datasets with relevant synthetic samples. This distinction is crucial in healthcare and SA contexts, where preserving local data heterogeneity can be more important than achieving a perfectly aligned global model. In particular, FedSDS demonstrates notable superiority in the most challenging settings (Scenarios 5 and 6), where traditional FL techniques, such as FedAvg and FedProx, struggle to maintain performance.

Scenario 7, where one node lacks a critical covariate, is particularly challenging. In this setting, only FedSDS-based techniques are applicable. The results clearly show the effectiveness of leveraging synthetic data to compensate for missing information, with all configurations of FedSDS yielding significant improvements in model performance.

While FedSDS demonstrates consistent advantages, it is important to acknowledge that not all performance improvements observed across scenarios and clients are statistically significant. As shown in Tables 3–6, the most notable and statistically robust gains are observed in nodes with limited data availability, skewed distributions, or missing features. This behavior aligns with the core motivation of FedSDS: to enhance model performance precisely in challenging, heterogeneous federated environments where traditional parameter-averaging techniques struggle. Conversely, clients with ample and well-balanced data may exhibit more marginal improvements or none at all, a known challenge in FL under heterogeneity. We highlight this limitation as an opportunity for further refinement and client-specific adaptation of future FL strategies.

In addition to performance improvements, FedSDS achieves computational costs that are competitive with or lower than those of traditional FL methods. This is primarily due to its single-round communication protocol and lightweight generative models. A detailed comparison of training times across scenarios and methods is provided in Appendix G.

These findings underscore the importance of advanced FL techniques such as FedSDS for real-world applications, where data is often imbalanced, heterogeneous, or incomplete. By integrating synthetic data generation and aggregation tailored to local distributions, FedSDS provides a robust and scalable solution for FL. In particular, this approach is well-suited to privacy-sensitive domains such as healthcare, where sharing raw patient data is highly restricted, and addressing data heterogeneity is crucial for building effective predictive models.

Nonetheless, privacy concerns remain critical, even when synthetic data are used. It is well known that generative models such as VAEs can be prone to memorizing training data, which may pose risks of indirect information leakage. To mitigate this, FedSDS incorporates specific safeguards, including using the BGM and the latent-space filtering during the aggregation process, that promote the generation of diverse and non-identifiable samples. These mechanisms reduce the likelihood of synthetic samples replicating real individuals too closely. To empirically support this claim, Appendix E presents a nearest-neighbor distance analysis between real and synthetic samples, demonstrating the absence of duplication or near-duplication. While this work does not provide formal guarantees, such as Differential Privacy (DP), their integration is identified as a promising avenue for future research, particularly in highly regulated settings.

4. Conclusion

This study demonstrates the effectiveness of the FedSDS framework in addressing the challenges of data scarcity, heterogeneity, and missing critical covariates in SA. By integrating SDG with FL, FedSDS offers a robust, scalable, and privacy-preserving approach for collaborative model training in decentralized settings. Across both IID and non-IID scenarios, FedSDS, particularly the *biased* aggregation strategy, consistently outperforms traditional methods such as FedAvg and more advanced techniques designed for non-IID settings like FedProx. While FedProx

successfully stabilizes training by preventing local models from drifting far from the global model, FedSDS *biased* goes beyond by directly benefiting each local node, augmenting its data without distorting its intrinsic distribution. This local personalization is crucial in SA and healthcare contexts, where client-specific heterogeneity carries meaningful clinical implications. The experimental results consistently demonstrate that FedSDS achieves significant improvements, especially in nodes with limited data availability, biased distributions, or missing features. Notably, FedSDS outperforms FedProx even in the most challenging non-IID scenarios, highlighting the benefits of aligning synthetic data with local distributions instead of enforcing rigid global model consensus. Furthermore, FedSDS achieves competitive or even lower computational costs compared to traditional FL methods, thanks to its single-round communication protocol and lightweight generative models.

The potential of FedSDS to transform SA methodologies is evident, but it also paves the way for further exploration.

A key direction involves enhancing the fidelity of synthetic data, with future efforts potentially leveraging advanced generative models to better capture complex distributions in low-data settings. Moreover, the framework's flexible architecture allows for experimentation with alternative SA models. These models must incorporate the latent space representation characteristic of encoder–decoder architectures, which is essential for the *biased* aggregation strategy. By incorporating different methodologies beyond SAVAE, the generalizability and applicability of FedSDS could be further validated.

While our study focuses on unimodal tabular data and a single-task setting (survival estimation), many healthcare applications are inherently more complex. Clinical scenarios often involve multi-task learning (e.g., jointly predicting survival and disease subtype) and multi-modal inputs (e.g., imaging, genomics, clinical notes). Extending FedSDS to handle such data jointly would unlock richer representations and improve its applicability in diverse real-world settings. However, we note that multi-modal and multi-task SA remains an emerging field, with most current approaches designed for centralized environments [6]. Initial efforts such as SAMVAE [43] illustrate the potential of combining survival objectives with modality fusion, but these solutions have yet to be extended to federated or privacy-constrained scenarios. We therefore view the extension of FedSDS to these settings as a promising research avenue, particularly given the variability of available modalities across decentralized healthcare institutions.

Another important avenue is adapting FedSDS to dynamic SA environments, where real-world data evolves due to changes in demographics, treatment protocols, or institutional practices. Ensuring adaptability over time would enable the framework to maintain its relevance in an ever-changing healthcare landscape.

From a systems perspective, while FedSDS currently operates under a single-round communication scheme to ensure computational efficiency, extending it to a multi-round setup represents an interesting future direction. In a multi-shot FedSDS variant, synthetic data generation and survival model training could co-evolve across multiple communication rounds, allowing models to refine synthetic datasets based on local model feedback dynamically. Such an extension could enhance the personalization and effectiveness of collaborative learning, particularly in highly heterogeneous environments.

In terms of privacy, although synthetic data sharing inherently improves protection by decoupling model training from raw patient data, we acknowledge that this does not guarantee formal protection against potential information leakage. Integrating formal privacy guarantees, such as DP, is a critical next step in highly regulated domains like healthcare. Although this work focuses on architectural safeguards (e.g., latent-space filtering and BGM-based diversification), future research should explore how these mechanisms can be combined with DP to provide provable privacy bounds. For instance, adding noise in the latent space during sample generation or applying DP mechanisms at the filtering or aggregation stages could strengthen the framework's robustness. Such extensions would not only ensure regulatory compliance but also enhance the trustworthiness and adoption of FedSDS in practical deployments.

Methodologically, alternative proximity calculation techniques in the *biased* aggregation strategy could also be explored, such as using Kullback-Leibler or Jensen-Shannon divergences. Since these metrics compare probability distributions, they could be well-suited to the latent space representations used in FedSDS, potentially improving the alignment of synthetic data with local distributions.

Finally, while this study employed controlled simulations using publicly available datasets to ensure reproducibility and systematic evaluation, we acknowledge that they may not fully reflect the complexity of real-world federated healthcare environments. In particular, scenarios involving a larger number of clients or higher-dimensional data distributions, such as those found in intensive care units or multi-department hospital systems, remain unexplored. Datasets such as MIMIC [44], where care units could naturally partition data to simulate real-world institutional decentralization, offer a promising benchmark for validating FedSDS in institutional-level federated scenarios. Additionally, future studies should explore deployments involving a larger number of nodes to better reflect realistic federated networks. This would allow a more thorough assessment of the scalability and robustness of FedSDS under complex, real-world conditions. In parallel, integrating FedSDS within broader DL architectures, potentially leveraging peer-to-peer topologies or gossip-based protocols to propagate synthetic data across clients. Such adaptations could enhance resilience, reduce single points of failure, and enable FedSDS to scale in more dynamic and distributed environments.

In conclusion, FedSDS represents a significant advance in FL for SA. It addresses the critical limitations of traditional methods while enabling effective collaboration in privacy-sensitive settings. By continuing to refine and expand its capabilities, FedSDS has the potential to become a cornerstone methodology for decentralized healthcare analytics, ultimately improving patient outcomes and advancing precision medicine.

CRediT authorship contribution statement

Patricia A. Apellániz: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Juan Parras:** Writing – review & editing, Supervision, Software, Conceptualization. **Santiago Zazo:** Writing – review & editing, Supervision, Project administration.

Funding

This work was funded by the GenoMed4All and SYNTHEMA projects from European Union's Horizon 2020 Research and Innovation Program under Grant 101017549 and Grant 101095530. However, views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Patricia A. Apellaniz reports that financial support was provided by Horizon Europe. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Mathematical foundations of SA and SAVAE

An SA dataset, D, consists of N observations represented by triplets $D = (x_i, t_i, d_i)_{i=1}^{N}$, where:

- $x_i \in$ is a covariate vector encompassing all the relevant features for the i-th individual;
- t_i denotes the observed time-to-event;

- $d_i \in \{0, 1\}$ is an indicator variable, where $d_i = 1$ signifies the event occurrence (uncensored data) and $d_i = 0$ indicates censoring. A distinct feature of SA models is their ability to handle censored data cases where the event of interest is not observed within the study's time frame.

The primary objective of SA is to model the probability of event occurrence over time. A survival model typically predicts the survival function, which represents the probability that the event occurs after a given time t, conditional on the covariates x and is defined as

$$S(t|x) = P(T > t|x) = 1 - F(t|x), \tag{4}$$

where $F(t|x)$ is the CDF of the time. Additionally, SA involves related functions such as:

- The hazard rate function, $h(t|x)$, quantifying the instantaneous event occurrence rate at time t.
- The time probability density function, $p(t|x)$, providing the probability density of event times.

These functions are interconnected as follows:

$$p(t|x) = h(t|x)S(t|x), \tag{5}$$

where $h(t|x)$ and $S(t|x)$ jointly describe the event occurrence dynamics.

SAVAE [11] offers a generative approach to SA by leveraging VAEs to model time-to-event data. Unlike traditional SA methods, SAVAE does not assume proportional hazards, enabling a more flexible representation of survival outcomes, particularly in the presence of censoring and intricate covariate correlations.

SAVAE aims to estimate the predictive distribution $p\left(t^*|x^*, \{x_i, t_i\}_{i=1}^N\right)$, where t^* is the time-to-event for a given covariate vector x^*. This distribution is expressed as:

$$p\left(t^*|x^*, \{x_i, t_i\}_{i=1}^N\right) = \int p\left(t^*|z, \{x_i, t_i\}_{i=1}^N\right) p\left(z|x^*, \{x_i, t_i\}_{i=1}^N\right) dz, \tag{6}$$

where z represents the latent variable that encodes dependencies within the data. Since direct computation of the true posterior $p(z|x)$ is intractable, SAVAE employs a variational approximation $q_\phi(z|x)$ to estimate it.

The model is trained by maximizing the ELBO, defined as:

$$\mathscr{L}(x, \theta, \phi) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x, t|z)\right] - D_{KL}(q_\phi(z|x)||p(z)), \tag{7}$$

where D_{KL} denotes the Kullback-Leibler divergence. The joint likelihood $p_\theta(x, t|z)$ is factorized as:

$$p_\theta(x, t|z) = p_{\theta_1}(x|z)p_{\theta_2}(t|z), \tag{8}$$

where θ_1 and θ_2 parameterize distributions for covariates x and survival times t, respectively. Then, the ELBO can be expanded as:

$$\mathscr{L}(x, \theta_1, \theta_2, \phi) = -D_{KL}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_{\theta_1}(x|z) + \log p_{\theta_2}(t|z)\right]. \tag{9}$$

For censored data, the survival likelihood adapts to incorporate the hazard $h(t)$ and the survival $S(t)$ functions, as defined in [45]:

$$p_{\theta_2}(t|z) = \begin{cases} h(t|z)S(t|z) & \text{if uncensored,} \\ S(t|z) & \text{if censored.} \end{cases}, \tag{10}$$

SAVAE models survival times using the Weibull distribution $p(t; \alpha, \lambda)$, widely validated in [46]. This distribution allows explicit parameterization of the survival, hazard, and probability density functions:

$$\begin{cases} p(t; \alpha, \lambda) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right) \\ S(t; \alpha, \lambda) = \exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right) \\ h(t; \alpha, \lambda) = \frac{p(t;\alpha,\lambda)}{S(t;\alpha,\lambda)} = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \end{cases}. \tag{11}$$

This parameterization enables SAVAE to effectively model both censored and uncensored data, ensuring an accurate representation of time-to-event distributions.

Additionally, SAVAE assigns distributions to covariates in the reconstruction term of the ELBO, providing flexibility to capture their underlying structure. Common distributions include Gaussian for continuous variables, Bernoulli for binary variables, and Categorical for discrete variables. This design enhances SAVAE's adaptability to complex datasets while supporting other differentiable distributions for broader applicability. By integrating these elements, SAVAE ensures robust performance in modeling survival outcomes under diverse and challenging conditions.

Appendix B. Details of FL baselines

FL is a decentralized ML paradigm enabling collaborative model training across multiple devices or institutions (nodes) without requiring direct real data sharing. Unlike traditional centralized methods that require aggregating data into a single repository, FL allows participants to retain their data locally, addressing privacy concerns and ensuring compliance with strict real data-sharing regulations. This architecture is particularly beneficial in domains like healthcare, where data sensitivity, privacy requirements, and regulatory constraints significantly hinder centralized data collection.

The FL framework involves training local models on individual datasets and aggregating their learned parameters or updates to produce a global model. This approach is well-suited to handling challenges such as data heterogeneity (non-IID distributions), limited communication bandwidth, and scalability in multi-institutional environments. FL enhances the resulting models' security and robustness by leveraging local computation and employing privacy-preserving protocols.

FL has demonstrated considerable promise in healthcare applications such as disease diagnosis, risk prediction, and SA. It facilitates collaborative learning among nodes, enabling them to build robust predictive models while preserving the confidentiality of sensitive patient data. However, FL still faces several challenges, including communication inefficiency, difficulties in achieving model convergence, and reduced performance in environments characterized by non-IID data distributions.

To evaluate the effectiveness of our proposed technique (FedSDS), we compare it against two widely used FL baselines: FedAvg [23] and FedProx [24].

B.1. FedAvg

FedAvg, first introduced by McMahan et al. [23], is a foundational and widely adopted method in FL. FedAvg enables decentralized collaborative model training across multiple nodes, each possessing a portion of the overall data, without requiring direct real data sharing. The method operates iteratively, with the following key steps:

1. **Initialization:** A unified model architecture, such as the SAVAE model, is designed and shared among all the nodes. While the initial model weights can differ across nodes, the architecture (e.g., number of layers) must remain consistent to ensure compatibility during the aggregation step. Each node initializes its model and begins training on its local dataset.
2. **Local Training:** Each node, denoted as $i = 1, 2, \ldots, L$, performs model training using its local dataset D_i. The training process involves optimizing the model weights W_i locally, typically over

multiple epochs, using standard optimization techniques such as stochastic gradient descent. The resulting weights W_i capture the knowledge learned from each node's unique data distribution.

3. **Model Aggregation:** All nodes transmit their update model parameters W_i to a central server after local training. The server aggregates these parameters through a weighted averaging process, where the contribution of each node is proportional to the number of samples M_i in its dataset. The aggregated global model W is computed as:

$$W = \sum_{i=1}^{L} \frac{M_i}{N} W_i, \tag{12}$$

where $N = \sum_{i=1}^{L} M_i$ represents the total number of samples across all nodes.

4. **Iterative Training Process:** The aggregated global model W is distributed back to the nodes, further refining their local datasets. This iterative process of local training and global aggregation continues for a predefined number of iterations or until the model achieves the desired level of accuracy. Each iteration improves the global model by incorporating knowledge from the distributed data.
5. **Stopping Criteria and Convergence:** The FedAvg training process finishes after a fixed number of iterations or when the global model converges, indicated by diminishing improvements in performance metrics.

FedAvg's key innovation lies in its aggregation mechanism, which ensures that the global model encapsulates distributed knowledge without compromising the privacy of local datasets. The weighted averaging step prioritizes contributions from nodes with larger datasets (i.e., higher M_i), balancing the influence of nodes with varying data quantities. This approach mitigates the risk of overfitting to specific nodes while promoting robust generalization across the collective data.

B.2. FedProx

FedProx, proposed by Li et al. [24], extends FedAvg by addressing its limitations under non-IID data. It introduces a proximal term in the local objective function to limit the divergence between local models and the current global model.

Specifically, each client solves the following modified local optimization problem:

$$\min_{W_i} \; f_i(W_i) + \frac{\mu}{2} \|W_i - W\|^2, \tag{13}$$

where $f_i(W_i)$ is the local empirical loss at node i, W is the global model, and $\mu > 0$ is the proximal coefficient controlling how strongly local models are regularized toward the global model.

The training process mirrors FedAvg, with the following differences:

- Local objectives explicitly penalize deviation from the global model.
- This stabilization mitigates client drift and improves convergence in heterogeneous settings.

While FedProx improves robustness under non-IID distributions, it still relies on iterative communication rounds and enforces alignment to a central model, which can be limiting in highly diverse clinical datasets.

Appendix C. Additional validation metric

We have incorporated an additional validation metric for the proposed experiments and the C-index. Below, we present result tables for the Integrated Brier Score (IBS), structured in the same format as the original manuscript, to ensure consistency and facilitate comparison. This supplementary metric provides further insights into the models' performance.

The Brier Score (BS), grounded in predictive accuracy metrics [47,48], serves as the second evaluation measure in this study. BS calculates the squared prediction error, adjusted for censored data using Inverse Probability of Censoring Weighting (IPCW) [49]. IPCW assigns higher weights to uncensored samples, compensating for the presence of censored data. The BS at a specific time t is defined as:

$$BS(t) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(S(t|x_i))^2}{G(t_i)} \cdot \mathbb{I}(t_i < t, d_i = 1) + \frac{(1 - S(t|x_i))^2}{G(t)} \cdot \mathbb{I}(t_i \geq t) \right], \tag{14}$$

where $G(t)$ is the survival function for censoring. The BS assesses both discrimination (the ability to rank risks accurately) and calibration (the alignment of predicted risks with observed event probabilities). To provide a time-agnostic evaluation, the Integrated Brier Score (IBS) is utilized, calculated as follows:

$$IBS(t_{max}) = \frac{1}{t_{max}} \int_0^{t_{max}} BS(t) dt. \tag{15}$$

Adjusted p-values for the statistical significance of the differences between FL methods were calculated based on Holm's method [42]. This ensures robust comparisons and highlights significant differences between the evaluated methods and scenarios. Including IBS alongside C-index enhances the comprehensiveness of the evaluation, providing a more holistic view of model performance.

C.1. IID scenarios

The following tables present the IBS results under the IID setting, where the number of samples and the percentage of missing values vary between nodes across different scenarios. The IBS provides a complementary perspective to the C-index by evaluating discrimination and calibration. It makes it more stringent to improve upon due to its sensitivity to model miscalibrations and errors in ranking survival probabilities.

In the METABRIC dataset (Table C.1), we observe the consistent pattern of significant improvement in IBS for Node 3 in Scenarios 2 and 3, corresponding to the nodes with the least data and a high percentage of missing information. These improvements highlight the ability of federated approaches to mitigate the impact of data scarcity. Furthermore, the mean IBS values for Node 3 in Scenario 2 show a greater relative decrease when using FedSDS methods, particularly the *naive* approach, compared to the isolated case.

For the GBSG dataset in Table C.2, the IBS improvements are more limited under the IID scenarios. In Scenario 1, significant reductions are observed for Nodes 2 and 3 using the FedAvg technique, similar to the C-index case. Significant reductions are observed only for Node 3 in Scenario 2, and for Node 2 in Scenario 3, both achieved using FedAvg. These results suggest that while federated approaches can enhance IBS performance, the nature of the dataset, such as its structure and survival distribution, influences the degree of improvement.

C.2. Non-IID scenarios

In the non-IID case, where heterogeneity has been introduced in the distribution of the *age* covariate across Nodes 2 and 3, the IBS results highlight distinct improvement patterns. These are notably different from those observed for the C-index, underscoring the complementary nature of these metrics.

For the METABRIC dataset (Table C.3), significant improvements for IBS are observed in Scenarios 5 and 6, particularly for Node 2. This contrasts with the C-index results, where improvements are mainly concentrated in Node 3. In Scenario 5, FedAvg and FedSDS *biased* techniques demonstrate improvements, with the latter showing the most pronounced enhancements. In Scenario 6, significant improvements are achieved using FedProx and FedSDS *biased*, the latter demonstrating its robustness in handling data heterogeneity, achieving better performance.

Table C.1
IBS comparison of isolated, FedAvg, FedProx, and FedSDS (*naive* and *biased*) methods for the METABRIC dataset in IID scenarios. Average IBS results are shown with confidence intervals. Adjusted p-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Scenario	Nodes	Isolated	FedAvg	FedProx	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	Node 1	(0.127 - 0.146 - 0.166)	–	–	–	–	–
Scenario 1	Node 1	(0.134 - 0.153 - 0.175)	(0.132 - 0.152 - 0.173)	(0.134 - 0.153 - 0.174)	(0.133 - 0.153 - 0.174)	(0.133 - 0.153 - 0.173)	0.656/1.000/1.000/1.000
	Node 2	(0.144 - 0.164 - 0.185)	(0.144 - 0.163 - 0.184)	(0.143 - 0.164 - 0.186)	(0.141 - 0.165 - 0.190)	(0.143 - 0.164 - 0.188)	1.000/1.000/1.000/1.000
	Node 3	(0.135 - 0.154 - 0.175)	(0.133 - 0.153 - 0.173)	(0.135 - 0.154 - 0.175)	(0.133 - 0.157 - 0.182)	(0.134 - 0.154 - 0.177)	0.103/1.000/1.000/1.000
Scenario 2	Node 1	(0.133 - 0.153 - 0.173)	(0.134 - 0.153 - 0.174)	(0.134 - 0.153 - 0.173)	(0.135 - 0.154 - 0.176)	(0.134 - 0.153 - 0.175)	1.000/1.000/1.000/1.000
	Node 2	(0.151 - 0.174 - 0.199)	(0.150 - 0.173 - 0.196)	(0.151 - 0.171 - 0.193)	(0.144 - 0.170 - 0.200)	(0.152 - 0.172 - 0.195)	1.000/0.692/1.000/1.000
	Node 3	(0.161 - 0.190 - 0.223)	(0.143 - 0.167 - 0.191)	(0.146 - 0.168 - 0.192)	(0.131 - 0.154 - 0.180)	(0.139 - 0.160 - 0.184)	**0.017**/**0.022**/**0.001**/**0.006**
Scenario 3	Node 1	(0.134 - 0.153 - 0.174)	(0.132 - 0.152 - 0.173)	(0.133 - 0.152 - 0.173)	(0.134 - 0.153 - 0.174)	(0.133 - 0.153 - 0.175)	0.692/1.000/1.000/1.000
	Node 2	(0.151 - 0.173 - 0.195)	(0.151 - 0.173 - 0.198)	(0.150 - 0.171 - 0.193)	(0.144 - 0.168 - 0.193)	(0.151 - 0.172 - 0.195)	1.000/0.654/0.692/1.000
	Node 3	(0.180 - 0.212 - 0.250)	(0.147 - 0.172 - 0.196)	(0.147 - 0.172 - 0.196)	(0.135 - 0.161 - 0.188)	(0.149 - 0.173 - 0.198)	**0.008**/**0.009**/**0.002**/**0.008**

Table C.2
IBS comparison of isolated, FedAvg, FedProx, and FedSDS (*naive* and *biased*) methods for the GBSG dataset in IID scenarios. Average IBS results are shown with confidence intervals. Adjusted p-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Scenario	Nodes	Isolated	FedAvg	FedProx	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	Node 1	(0.174 - 0.196 - 0.219)	–	–	–	–	–
Scenario 1	Node 1	(0.163 - 0.186 - 0.211)	(0.163 - 0.185 - 0.207)	(0.165 - 0.186 - 0.209)	(0.166 - 0.187 - 0.210)	(0.165 - 0.188 - 0.212)	1.000/1.000/1.000/1.000
	Node 2	(0.159 - 0.181 - 0.204)	(0.151 - 0.173 - 0.196)	(0.155 - 0.177 - 0.201)	(0.160 - 0.183 - 0.211)	(0.156 - 0.181 - 0.212)	**0.002**/0.079/1.000/1.000
	Node 3	(0.150 - 0.173 - 0.198)	(0.145 - 0.166 - 0.188)	(0.148 - 0.169 - 0.194)	(0.150 - 0.172 - 0.199)	(0.152 - 0.172 - 0.194)	**0.014**/0.413/1.000/1.000
Scenario 2	Node 1	(0.167 - 0.188 - 0.211)	(0.165 - 0.187 - 0.210)	(0.166 - 0.187 - 0.209)	(0.165 - 0.186 - 0.209)	(0.166 - 0.188 - 0.211)	0.536/0.083/0.073/1.000
	Node 2	(0.165 - 0.190 - 0.217)	(0.162 - 0.185 - 0.213)	(0.165 - 0.186 - 0.210)	(0.171 - 0.197 - 0.229)	(0.164 - 0.191 - 0.220)	1.000/1.000/1.000/1.000
	Node 3	(0.172 - 0.197 - 0.230)	(0.158 - 0.180 - 0.204)	(0.156 - 0.177 - 0.200)	(0.161 - 0.189 - 0.217)	(0.163 - 0.188 - 0.216)	**0.027**/**0.021**/0.642/0.445
Scenario 3	Node 1	(0.166 - 0.187 - 0.210)	(0.163 - 0.186 - 0.209)	(0.165 - 0.186 - 0.209)	(0.164 - 0.186 - 0.211)	(0.163 - 0.186 - 0.210)	1.000/1.000/1.000/1.000
	Node 2	(0.160 - 0.186 - 0.215)	(0.162 - 0.187 - 0.210)	(0.154 - 0.183 - 0.209)	(0.165 - 0.191 - 0.220)	(0.164 - 0.192 - 0.223)	1.000/1.000/1.000/1.000
	Node 3	(0.184 - 0.208 - 0.236)	(0.178 - 0.203 - 0.236)	(0.183 - 0.206 - 0.229)	(0.187 - 0.217 - 0.246)	(0.190 - 0.215 - 0.242)	1.000/1.000/1.000/1.000

Table C.3
IBS comparison of isolated, FedAvg, FedProx, and FedSDS (*naive* and *biased*) methods for the METABRIC dataset in non-IID scenarios. Average IBS results are shown with confidence intervals. Adjusted p-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Scenario	Nodes	Isolated	FedAvg	FedProx	FedSDS *naive*	FedSDS *biased*	Adjusted p-values
Centralized	Node 1	(0.127 - 0.146 - 0.166)	–	–	–	–	–
Scenario 4	Node 1	(0.133 - 0.152 - 0.173)	(0.133 - 0.152 - 0.174)	(0.134 - 0.153 - 0.174)	(0.134 - 0.153 - 0.174)	(0.132 - 0.153 - 0.173)	1.000/1.000/1.000/1.000
	Node 2	(0.154 - 0.175 - 0.197)	(0.153 - 0.174 - 0.196)	(0.154 - 0.175 - 0.197)	(0.154 - 0.176 - 0.203)	(0.154 - 0.175 - 0.197)	0.666/1.000/1.000/1.000
	Node 3	(0.135 - 0.156 - 0.180)	(0.134 - 0.153 - 0.174)	(0.136 - 0.155 - 0.176)	(0.132 - 0.155 - 0.179)	(0.134 - 0.153 - 0.174)	0.603/1.000/1.000/0.784
Scenario 5	Node 1	(0.134 - 0.154 - 0.178)	(0.133 - 0.153 - 0.175)	(0.134 - 0.153 - 0.174)	(0.134 - 0.153 - 0.174)	(0.134 - 0.154 - 0.178)	1.000/1.000/1.000/1.000
	Node 2	(0.153 - 0.174 - 0.198)	(0.148 - 0.168 - 0.190)	(0.148 - 0.170 - 0.193)	(0.141 - 0.167 - 0.194)	(0.141 - 0.161 - 0.185)	**0.008**/0.088/0.247/**0.000**
	Node 3	(0.137 - 0.161 - 0.187)	(0.135 - 0.157 - 0.183)	(0.133 - 0.158 - 0.185)	(0.136 - 0.172 - 0.206)	(0.145 - 0.170 - 0.198)	1.000/1.000/1.000/1.000
Scenario 6	Node 1	(0.134 - 0.153 - 0.173)	(0.133 - 0.153 - 0.174)	(0.134 - 0.153 - 0.174)	(0.133 - 0.153 - 0.174)	(0.134 - 0.154 - 0.174)	1.000/1.000/1.000/1.000
	Node 2	(0.154 - 0.174 - 0.197)	(0.150 - 0.172 - 0.193)	(0.147 - 0.169 - 0.192)	(0.140 - 0.165 - 0.190)	(0.142 - 0.163 - 0.186)	0.088/**0.020**/0.072/**0.000**
	Node 3	(0.142 - 0.166 - 0.192)	(0.169 - 0.198 - 0.230)	(0.140 - 0.168 - 0.194)	(0.166 - 0.205 - 0.256)	(0.175 - 0.212 - 0.243)	1.000/1.000/1.000/1.000

In the GBSG dataset results from Table C.4, no significant improvements are observed, highlighting the challenging nature of achieving substantial reductions in IBS, particularly under the complexities of non-IID scenarios.

C.3. Special non-IID scenario

Table C.5 presents the results of the IBS for Scenario 7, where only FedSDS-based techniques can be applied due to the removal of the *age* column in Node 2. Significant improvements are observed in the METABRIC and GBSG datasets.

The techniques that demonstrate significant IBS improvements are those that combine column prediction with the addition of synthetic data. Among these, the most pronounced improvements are observed with the aggregation of synthetic data using the *biased* method. This approach aligns synthetic data more closely with the local distributions of Node 2, thereby enhancing the model's ability to mitigate the effects of missing critical covariates.

For METABRIC and GBSG, the *biased* technique achieves the most significant improvements, with adjusted p-values below 0.05. These results highlight the effectiveness of combining imputation and *biased* synthetic data aggregation in addressing data heterogeneity and compensating for missing information.

C.4. Discussion

Improving IBS proves more challenging than improving the C-index due to its dual focus on discrimination and calibration. While the C-index

Table C.4
IBS comparison of isolated, FedAvg, FedProx, and FedSDS (*naive* and *biased*) methods for the GBSG dataset in non-IID scenarios. Average IBS results are shown with confidence intervals. Adjusted *p*-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Scenario	Nodes	Isolated	FedAvg	FedProx	FedSDS *naive*	FedSDS *biased*	Adjusted *p*-values
Centralized	Node 1	(0.174 - 0.196 - 0.219)	–	–	–	–	–
Scenario 4	Node 1	(0.164 - 0.186 - 0.209)	(0.164 - 0.186 - 0.208)	(0.165 - 0.187 - 0.211)	(0.166 - 0.187 - 0.211)	(0.164 - 0.186 - 0.209)	1.000/1.000/1.000/1.000
	Node 2	(0.150 - 0.171 - 0.194)	(0.147 - 0.167 - 0.188)	(0.149 - 0.169 - 0.191)	(0.148 - 0.172 - 0.201)	(0.147 - 0.170 - 0.192)	0.079/1.000/1.000/1.000
	Node 3	(0.149 - 0.171 - 0.195)	(0.147 - 0.167 - 0.189)	(0.147 - 0.168 - 0.190)	(0.148 - 0.171 - 0.198)	(0.149 - 0.171 - 0.194)	0.218/1.000/1.000/1.000
Scenario 5	Node 1	(0.165 - 0.187 - 0.210)	(0.165 - 0.185 - 0.209)	(0.166 - 0.187 - 0.210)	(0.164 - 0.186 - 0.210)	(0.167 - 0.188 - 0.211)	0.722/1.000/1.000/1.000
	Node 2	(0.170 - 0.193 - 0.218)	(0.170 - 0.192 - 0.214)	(0.169 - 0.190 - 0.213)	(0.172 - 0.194 - 0.217)	(0.171 - 0.194 - 0.220)	1.000/0.636/1.000/1.000
	Node 3	(0.177 - 0.207 - 0.240)	(0.169 - 0.197 - 0.224)	(0.169 - 0.193 - 0.218)	(0.170 - 0.199 - 0.233)	(0.167 - 0.192 - 0.224)	0.642/0.207/1.000/0.180
Scenario 6	Node 1	(0.165 - 0.188 - 0.212)	(0.165 - 0.186 - 0.208)	(0.166 - 0.187 - 0.210)	(0.166 - 0.187 - 0.209)	(0.165 - 0.188 - 0.212)	1.000/1.000/1.000/1.00
	Node 2	(0.171 - 0.194 - 0.217)	(0.170 - 0.191 - 0.214)	(0.170 - 0.191 - 0.213)	(0.174 - 0.198 - 0.225)	(0.167 - 0.193 - 0.218)	0.057/0.076/1.000/1.000
	Node 3	(0.177 - 0.204 - 0.236)	(0.170 - 0.193 - 0.217)	(0.182 - 0.205 - 0.233)	(0.178 - 0.205 - 0.234)	(0.177 - 0.201 - 0.229)	0.218/1.000/1.000/1.000

Table C.5
IBS comparison in Scenario 7 of the three different FedSDS settings. Average IBS results are shown with confidence intervals. Adjusted *p*-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Dataset	Imputation	Imputation	Imputation + Synthetic Data *naive*	Imputation + Synthetic Data *biased*	Adjusted *p*-values
METABRIC	(0.154 - 0.176 - 0.199)	(0.156 - 0.178 - 0.201)	(0.139 - 0.168 - 0.196)	(0.136 - 0.160 - 0.183)	0.992 / **0.044** / **0.000**
GBSG	(0.175 - 0.197 - 0.220)	(0.180 - 0.216 - 0.255)	(0.157 - 0.190 - 0.223)	(0.169 - 0.192 - 0.218)	0.992 / 0.122 / **0.009**

solely evaluates a model's ability to rank survival probabilities correctly, the IBS additionally penalizes errors in the predicted probabilities' calibration relative to actual event outcomes. This makes IBS particularly sensitive to misalignments in survival predictions, especially under conditions of heterogeneity, censoring, or limited data. Consequently, achieving significant improvements in IBS requires methods capable of maintaining ranking accuracy and producing well-calibrated survival probabilities, such as FedSDS *biased*, which leverages its alignment of synthetic data to local distributions to achieve superior performance.

Appendix D. Additional validation on complex real-world data

To further evaluate the proposed FedSDS framework, we performed an additional experiment using real-world clinical data from the publicly available TCGA Pan-Cancer clinical resource [36]. This dataset contains information from multiple cancer types alongside metadata such as Tissue Source Site (TSS), which enables the simulation of an FL scenario by partitioning the data across different nodes according to TSS identifiers.

For this supplementary analysis, we focused on breast cancer (BRCA) cases, which represent the largest cohort within the Pan-Cancer dataset, ensuring sufficient sample size for meaningful evaluation. However, as is commonly the case with public clinical datasets, the number of samples per institution remains limited. Therefore, we constructed a two-node federation based on TSS, where Node 1 includes 250 patients and Node 2 includes 134 patients for training. This configuration simulates a highly constrained real-world situation with both data scarcity and imbalance across participating institutions.

In this configuration, Node 1 contains the majority of the samples, while Node 2 has a smaller cohort, introducing both data scarcity and distribution imbalance. This setup mimics a realistic real-world scenario in which institutions contribute heterogeneous amounts of data, reflecting their clinical capacity or data accessibility. As reflected in the confidence intervals of Tables D.1 and D.2, Node 1 benefits from a more stable estimation, while Node 2 represents a low-resource node that can benefit most from collaborative learning.

Following the approach proposed in FedSDS, synthetic data generation was performed exclusively at Node 1, which possesses more data to ensure robust generative modeling. This decision aligns with our methodology rationale, as nodes with very limited or biased data may lead to poorly trained generative models that could propagate noise or artifacts to other nodes. This parallels the weighting schemes of classical

Table D.1
C-index comparison of isolated, FedAvg, FedProx, and FedSDS (*naive* and *biased*) methods for the BRCA dataset. Average C-index results are shown with confidence intervals. Adjusted *p*-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Nodes	Isolated	FedAvg	FedProx	FedSDS *naive*	FedSDS *biased*	Adjusted *p*-values
Node 1	(0.531 - 0.605 - 0.688)	(0.534 - 0.594 - 0.649)	(0.506 - 0.567 - 0.621)	(0.532 - 0.604 - 0.681)	(0.560 - 0.621 - 0.677)	0.821/0.990/0.520/0.111
Node 2	(0.481 - 0.552 - 0.654)	(0.509 - 0.577 - 0.660)	(0.529 - 0.595 - 0.653)	(0.513 - 0.607 - 0.700)	(0.555 - 0.617 - 0.698)	0.101/**0.020**/**0.014**/**0.004**

Table D.2
IBS comparison of isolated, FedAvg, FedProx, and FedSDS (*naive* and *biased*) methods for the BRCA dataset. Average IBS results are shown with confidence intervals. Adjusted *p*-values below 0.05, indicating significant differences compared to the isolated case, are highlighted in bold.

Nodes	Isolated	FedAvg	FedProx	FedSDS *naive*	FedSDS *biased*	Adjusted *p*-values
Node 1	(0.130 - 0.167 - 0.213)	(0.135 - 0.174 - 0.216)	(0.137 - 0.174 - 0.216)	(0.129 - 0.167 - 0.210)	(0.130 - 0.166 - 0.207)	0.974/0.973/0.514/0.287
Node 2	(0.159 - 0.200 - 0.261)	(0.181 - 0.226 - 0.287)	(0.168 - 0.214 - 0.264)	(0.141 - 0.192 - 0.249)	(0.154 - 0.193 - 0.240)	0.997/0.980/0.153/0.092

FL methods (e.g., FedAvg), where nodes with more samples have a larger contribution to the global update.

The experimental results show that Node 2, which suffers from limited data availability, exhibits statistically significant improvements when applying both FedProx and FedSDS. However, the benefits of FedSDS, particularly with the *biased* aggregation strategy, are more pronounced, as evidenced by stronger gains in C-index (Table D.1). In IBS (Table D.2), while improvements are less statistically significant, FedSDS *biased* consistently achieves lower average values, further supporting its superior performance. These results highlight the ability of FedSDS to provide clinically meaningful improvements, even in highly constrained settings with very limited sample sizes.

Overall, this additional evaluation supports the applicability and robustness of FedSDS even under highly challenging real-world conditions with limited open-access clinical data. Despite the data scarcity, FedSDS demonstrates its ability to effectively leverage limited cross-institutional data distributions to improve model performance while preserving local data privacy.

Appendix E. Empirical analysis for synthetic data privacy

To empirically assess the privacy-preserving capabilities of the FedSDS framework, we evaluated whether the synthetic data generated by the VAE-BGM model exhibited signs of overfitting or memorization of the training data. Specifically, we compared the distributions of minimum pairwise distances between (i) real–real sample pairs and (ii) synthetic–real sample pairs. The rationale for this analysis is that if synthetic samples replicate or closely mimic real observations, the distances between synthetic and real samples would be comparable to (or smaller than) those among real samples. In contrast, if the synthetic data are sufficiently different, we would expect to observe larger minimum distances in the synthetic–real case.

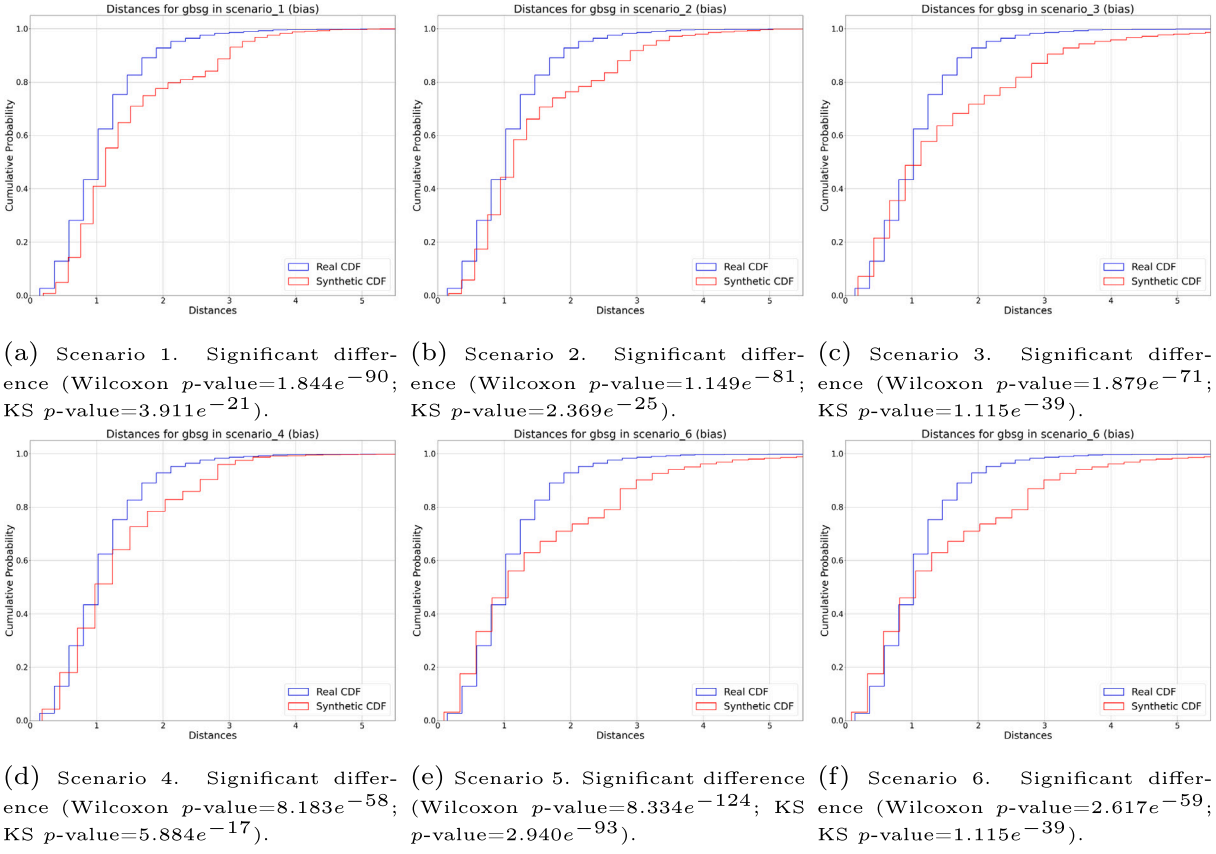
Fig. E.1 presents the CDFs of these minimum distances for the five representative scenarios using the GBSG dataset. In all scenarios, the minimum distances between synthetic and real samples were consistently larger than those among real samples and strictly positive. To statistically confirm these observations, we performed one-sided Wilcoxon signed-rank and Kolmogorov-Smirnov (KS) tests, evaluating whether the distribution of distances in the real–synthetic case was significantly shifted to the right. The resulting p-values, reported in each panel of Fig. E.1, were below the 0.05 significance threshold, indicating strong evidence that the distributions differ.

These results demonstrate that the generative process implemented in FedSDS does not replicate individual data points and avoids memorization, even under scarce data conditions or in the presence of model biases. Maintaining a meaningful margin between real and synthetic data distributions ensures a low privacy leakage risk while preserving statistical resemblance and utility.

Appendix F. Empirical convergence analysis

To provide additional insights into the convergence properties of the methodology, we tracked the C-index during training across all nodes and experimental scenarios. Figs. F.1 and F.2 present the evolution of the validation C-index throughout the training epochs at each node for the different scenarios using METABRIC and GBSG datasets, respectively.

Across all scenarios, we observe that the C-index curves exhibit a consistent convergence trend. After an initial increase during the early training stages, the metric stabilizes for all nodes, indicating that the models reach a steady performance state. Importantly, no evidence of divergence or instability is observed, even under the challenging non-IID



(a) Scenario 1. Significant difference (Wilcoxon p-value=$1.844e^{-90}$; KS p-value=$3.911e^{-21}$).

(b) Scenario 2. Significant difference (Wilcoxon p-value=$1.149e^{-81}$; KS p-value=$2.369e^{-25}$).

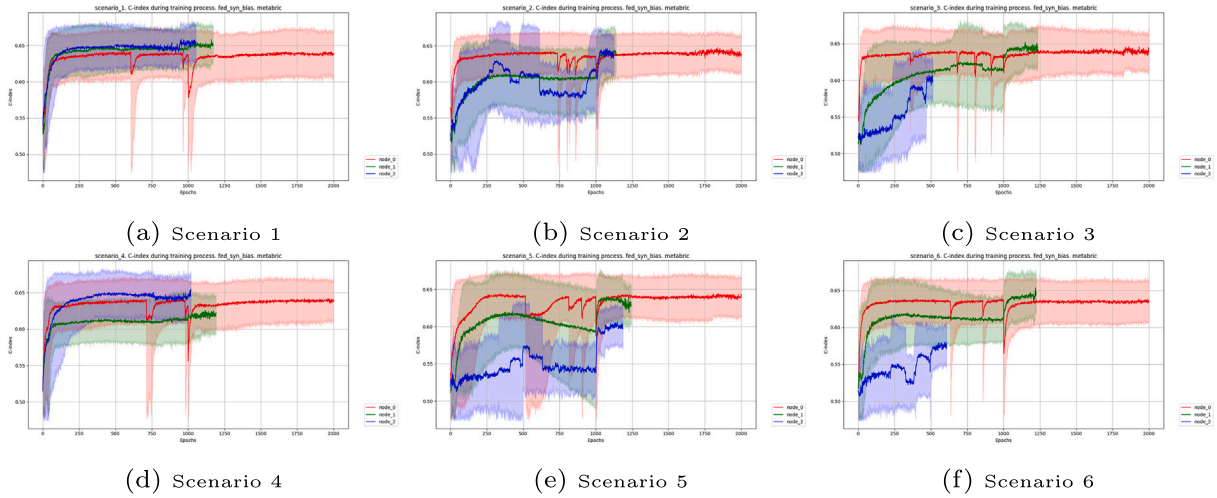
(c) Scenario 3. Significant difference (Wilcoxon p-value=$1.879e^{-71}$; KS p-value=$1.115e^{-39}$).

(d) Scenario 4. Significant difference (Wilcoxon p-value=$8.183e^{-58}$; KS p-value=$5.884e^{-17}$).

(e) Scenario 5. Significant difference (Wilcoxon p-value=$8.334e^{-124}$; KS p-value=$2.940e^{-93}$).

(f) Scenario 6. Significant difference (Wilcoxon p-value=$2.617e^{-59}$; KS p-value=$1.115e^{-39}$).

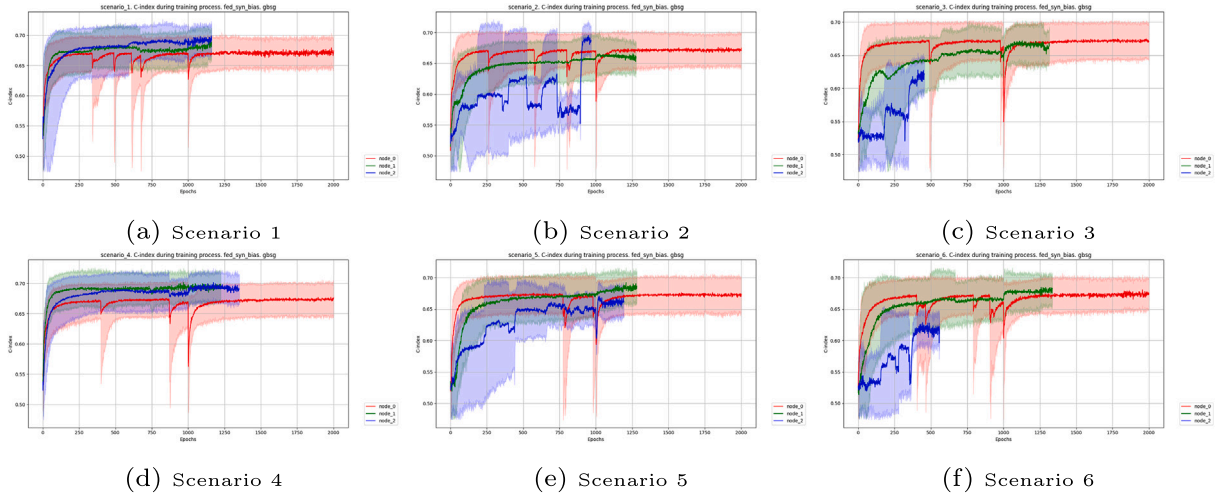
Fig. E.1. Comparison of minimum distance distributions between real and synthetic samples across different scenarios. Significant differences are observed in all cases, confirming the absence of direct duplication or data leakage.



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

(d) Scenario 4 (e) Scenario 5 (f) Scenario 6

Fig. F.1. Validation C-index evolution during training for all nodes with the METABRIC. Each curve represents the mean C-index across three runs per node, with shaded areas denoting the confidence interval.



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

(d) Scenario 4 (e) Scenario 5 (f) Scenario 6

Fig. F.2. Validation C-index evolution during training for all nodes with the GBSG. Each curve represents the mean C-index across three runs per node, with shaded areas denoting the confidence interval.

and data-scarce conditions simulated in several experimental settings. It is worth noting that some fluctuations are visible in the curves, particularly around certain epochs. These correspond to the moments where synthetic data is exchanged and integrated into the training process. In FL, such transitions are common and expected, especially when introducing external samples into local training sets. These fluctuations reflect the adaptation of the local model to newly received information. Crucially, after these transitions, the metric returns to a stable trend, confirming that convergence is ultimately preserved. The shaded regions represent confidence intervals across different random seeds, and their progressive narrowing also reflects increased stability and reduced variance over the epochs. While occasional fluctuations occur, especially in low-data nodes, these are typical in federated setups and do not compromise the overall convergence behavior.

These empirical findings support the convergence of the proposed FedSDS *biased* framework when applied to real-world medical datasets, even in decentralized and heterogeneous environments. A similar convergence dynamics is observed across both datasets, reinforcing the robustness of the approach.

Appendix G. Computational resources and training time analysis

To complement our evaluation of communication efficiency and predictive performance, we compare the computational overhead introduced by each FL strategy. In particular, we measure the total training time (in seconds) required to complete each experimental scenario for METABRIC and GBSG. The reported times include the full training loop for all clients (including model updates and synthetic data generation where applicable). Experiments were conducted using the same hardware for all configurations. Table G.1 presents the results.

These results demonstrate that FedSDS achieves competitive or even lower training times than traditional FL approaches such as FedAvg and FedProx. This efficiency stems from two design factors: first, FedSDS requires only a single communication round for synthetic data sharing, eliminating the iterative parameter exchange that characterizes standard FL methods; second, the underlying VAE architecture used for synthetic data generation is deliberately lightweight. While the biased variant of FedSDS incurs slightly more computation than its naive counterpart due to the additional latent-space filtering step, both configurations remain substantially more efficient than FedProx and faster than FedAvg in

Table G.1
Total training time (seconds) for each FL strategy across scenarios and datasets.

Dataset	Scenario	Isolated	FedAvg	FedProx	FedSDS Naive	FedSDS Biased
METABRIC	centralized	417.75	N/A	N/A	N/A	N/A
	scenario_1	348.59	768.80	2905.24	477.43	596.55
	scenario_2	174.03	464.26	3019.70	324.82	341.99
	scenario_3	161.15	481.75	2074.23	316.49	341.85
	scenario_4	290.57	655.58	1109.71	438.97	557.68
	scenario_5	171.80	514.06	512.07	348.41	332.25
	scenario_6	171.34	458.53	505.16	325.26	347.04
GBSG	centralized	401.22	N/A	N/A	N/A	N/A
	scenario_1	343.54	717.11	576.73	525.45	588.21
	scenario_2	178.92	478.56	386.51	317.12	338.05
	scenario_3	176.92	488.55	301.31	328.04	337.42
	scenario_4	339.25	711.45	591.35	468.33	574.31
	scenario_5	169.27	507.60	373.19	326.21	322.84
	scenario_6	172.06	513.93	346.13	338.76	317.77
Average	–	248.26	529.78	733.18	368.46	404.39

most scenarios. These characteristics reinforce the practical viability of FedSDS for real-world deployment, particularly in resource-constrained environments where both communication and computational costs must be minimized.

Data availability statement

All datasets used in this research are publicly available and can be found in the repository https://github.com/Patricia-A-Apellaniz/fed_savae.

References

[1] E.L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, J. Am. Stat. Assoc. 53 (282) (1958) 457–481.
[2] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc. Ser. B 34 (2) (1972) 187–202.
[3] H. Ishwaran, U.B. Kogalur, Random survival forests for r, R news 7 (2) (2007) 25–31.
[4] H. Binder, M. Schumacher, Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models, BMC Bioinformatics 9 (2008) 1–10.
[5] P. Wang, Y. Li, C.K. Reddy, Machine learning for survival analysis: A survey, ACM Comput. Surv. 51 (6) (2019) 1–36.
[6] S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, A. Bender, Deep learning for survival analysis: a review, Artif. Intell. Rev. 57 (3) (2024) 65.
[7] J.L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Y. Kluger, Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network, BMC Med. Res. Methodol. 18 (2018) 1–12.
[8] H. Kvamme, Ø. Borgan, I. Scheel, Time-to-event prediction with neural networks and cox regression, J. Mach. Learn. Res. 20 (129) (2019) 1–30.
[9] C. Lee, W. Zame, J. Yoon, M. Van Der Schaar, Deephit: A deep learning approach to survival analysis with competing risks, in: Proceedings of the AAAI conference on artificial intelligence, vol. 32, 2018.
[10] C. Lee, J. Yoon, M. Van Der Schaar, Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data, IEEE Trans. Biomed. Eng. 67 (1) (2019) 122–133.
[11] P.A. Apellániz, J. Parras, S. Zazo, Leveraging the variational bayes autoencoder for survival analysis, Sci. Rep. 14 (1) (2024) 24567.
[12] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional gan, Adv. Neural Inf. Process. Syst. 32 (2019).
[13] P.A. Apellaniz, J. Parras, S. Zazo, An improved tabular data generator with vae-gmm integration, in: 2024 32nd European Signal Processing Conference (EUSIPCO), 2024, pp. 1886–1890.
[14] P. A. Apellániz, A. Jiménez, B. A. Galende, J. Parras, S. Zazo, Artificial inductive bias for synthetic tabular data generation in data-scarce scenarios, arXiv:2407.03080, 2024.
[15] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International conference on machine learning, PMLR, 2017, pp. 1126–1135.
[16] K. Gao, O. Sener, Modeling and optimization trade-off in meta-learning, Adv. Neural Inf. Process. Syst. 33 (2020) 11154–11165.
[17] K.A. Henry, X. Niu, F.P. Boscoe, Geographic disparities in colorectal cancer survival, Int. J. Health Geogr. 8 (2009) 1–13.
[18] M.A. Mullee, B. De Stavola, M. Romanengo, M.P. Coleman, Geographical variation in breast cancer survival rates for women diagnosed in england between 1992 and 1994, Br. J. Cancer 90 (11) (2004) 2153–2156.
[19] M.-Y. Tseng, J.-H. Tseng, E. Merchant, Comparison of effects of socioeconomic and geographic variations on survival for adults and children with glioma, J. Neurosurg. Pediatr. 105 (4) (2006) 297–305.
[20] C.-L. Lu, S. Wang, Z. Ji, Y. Wu, L. Xiong, X. Jiang, et al., Webdisco: a web service for distributed cox model learning without patient-level data sharing, J. Am. Med. Inform. Assoc. 22 (6) (2015) 1212–1219.
[21] R. Duan, C. Luo, M.J. Schuemie, J. Tong, C.J. Liang, H.H. Chang, et al., Learning from local to global: An efficient distributed algorithm for modeling time-to-event data, J. Am. Med. Inform. Assoc. 27 (7) (2020) 1028–1036.
[22] D.K. Zhang, F. Toni, M. Williams, A federated cox model with non-proportional hazards, in: Multimodal AI in healthcare: A paradigm shift in health intelligence, Springer, 2022, pp. 171–185.
[23] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282.
[24] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, in: Proceedings of Machine Learning and Systems, vol. 2, 2020, pp. 429–450.
[25] J. Wang, Q. Liu, H. Liang, G. Joshi, H.V. Poor, Tackling the objective inconsistency problem in heterogeneous federated optimization, in: Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 7611–7623.
[26] S.P. Karimireddy, S. Kale, M. Mohri, S.J. Reddi, S.U. Stich, A.T. Suresh, Scaffold: Stochastic controlled averaging for federated learning, in: International Conference on Machine Learning, PMLR, 2020, pp. 5132–5143.
[27] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, Y. Khazaeni, Federated learning with matched averaging, arXiv preprint arXiv:2002.06440, 2020.
[28] J. Goetz, A. Tewari, Federated learning via synthetic data, arXiv preprint arXiv:2008.04489, 2020.
[29] N. Guha, A. Talwalkar, V. Smith, One-shot federated learning, arXiv preprint arXiv:1902.11175, 2019.
[30] R. Ormándi, I. Hegedűs, M. Jelasity, Gossip learning with linear models on fully distributed data, Concurr. Comput. Pract. Exp. 25 (4) (2013) 556–571.
[31] A. Lalitha, O.C. Kilinc, T. Javidi, F. Koushanfar, Peer-to-peer federated learning on graphs, arXiv preprint arXiv:1901.11173, 2019.
[32] H. Xing, O. Simeone, S. Bi, Federated learning over wireless device-to-device networks: Algorithms and convergence analysis, IEEE J. Sel. Areas Commun. 39 (12) (2021) 3723–3741.
[33] B. Pereira, S.-F. Chin, O.M. Rueda, H.-K.M. Vollan, E. Provenzano, H.A. Bardwell, et al., The somatic mutation profiles of 2, 433 breast cancers refines their genomic and transcriptomic landscapes, Nat. Commun. 7 (2016) 11479, https://doi.org/10.1038/ncomms11479, https://europepmc.org/articles/PMC4866047.
[34] M. Schumacher, G. Bastert, H. Bojar, K. Hübner, M. Olschewski, W. Sauerbrei, et al., Randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group, J. Clin. Oncol. 12 (10) (1994) 2086–2093.
[35] J.A. Foekens, H.A. Peters, M.P. Look, H. Portengen, M. Schmitt, M.D. Kramer, et al., The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients, Cancer Research 60 (3) (2000) 636–643.
[36] J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Shaw, B.A. Ozenberger, K. Ellrott, et al., The cancer genome atlas pan-cancer analysis project, Nat. Genet. 45 (10) (2013) 1113–1120.
[37] L. Antolini, P. Boracchi, E. Biganzoli, A time-dependent discrimination index for survival data, Stat. Med. 24 (24) (2005) 3927–3944.
[38] F.E. Harrell, R.M. Califf, D.B. Pryor, K.L. Lee, R.A. Rosati, Evaluating the yield of medical tests, Jama 247 (18) (1982) 2543–2546.
[39] J.W. Tukey, The philosophy of multiple comparisons, Statistical Science (1991) 100–116.
[40] E.L. Lehmann, J.P. Romano, Generalizations of the familywise error rate, Springer, 2012.
[41] M.J. Van der Laan, S. Dudoit, K.S. Pollard, Multiple testing. part ii. step-down procedures for control of the family-wise error rate, Stat. Appl. Genet. Mol. Biol. 3 (1) (2004).
[42] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. (1979) 65–70.
[43] A. Garrido, A. Almodóvar, P.A. Apellániz, J. Parras, S. Zazo, Deep survival analysis in multimodal medical data: A parametric and probabilistic approach with competing risks, arXiv preprint arXiv:2507.07804, 2025.
[44] A.E.W. Johnson, T.J. Pollard, L. Shen, L.-W.H. Lehman, M. Feng, M. Ghassemi, et al., Mimic-iii, a freely accessible critical care database, Sci. Data. 3 (1) (2016) 1–9.
[45] S. Liverani, L. Leigh, I.L. Hudson, J.E. Byles, Clustering method for censored and collinear survival data, Comput. Stat. 36 (2021) 35–60.
[46] R. Ranganath, D. Tran, J. Altosaar, D. Blei, Operator variational inference, Adv. Neural Inf. Process. Syst. 29 (2016).
[47] G.W. Brier, Verification of forecasts expressed in terms of probability, Mon. Weather Rev. 78 (1) (1950) 1–3.
[48] E. Graf, C. Schmoor, W. Sauerbrei, M. Schumacher, Assessment and comparison of prognostic classification schemes for survival data, Stat. Med. 18 (17–18) (1999) 2529–2545.
[49] J.M. Robins, et al., Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers, in: Proceedings of the biopharmaceutical section, American statistical association, vol. 24, San Francisco CA, 1993, pp. 3.