FSA-Bench: Benchmarking Federated Survival Models

Sultan Ahmed SULTAN.AHMED@UMBC.EDU
University of Maryland, Baltimore County
Baltimore, Maryland, USA.

Sanjay Purushotham PSANJAY@UMBC.EDU
University of Maryland, Baltimore County
Baltimore, Maryland, USA

Clinical time-to-event data are routinely collected across multiple healthcare institutions but are rarely pooled due to privacy regulations such as GDPR and HIPAA. This fragmentation limits the development of robust and generalizable survival models for clinical decision-making. Federated Survival Analysis (FSA) enables collaborative modeling without sharing patient-level data, but the reliability, calibration, and robustness of survival models in federated settings remain insufficiently understood. We present **FSA-Bench**, a comprehensive benchmark for evaluating survival models under realistic federated and multi-institutional healthcare conditions. FSA-Bench includes classical statistical methods (Kaplan Meier, Weibull Accelerated Failure Time, and Cox Proportional Hazards), machine learning models, and modern deep learning approaches, including attention-based federated architectures such as FedPAttn. The benchmark evaluates these models across diverse clinical datasets under realistic challenges, including heterogeneous patient populations, varying censoring rates, and non-IID data distributed across multiple institutions. Across diverse clinical benchmark datasets, we evaluate model performance using clinically meaningful metrics, including discrimination (time-dependent C-index), calibration (Integrated Brier Score and negative log-likelihood), and ranking consistency using Spearman's correlation, Kendall's rank correlation, stability, and Jaccard similarity. An aggregated cross-dataset analysis further provides a comprehensive assessment of ranking robustness, revealing that Fed CoxPH offers the strongest overall ranking consistency, while FedPLSTM and FedPAttn achieve the highest stability and top-model agreement across heterogeneous federated settings. Collectively, FSA-Bench establishes a standardized and reproducible evaluation framework for benchmarking federated survival models and offers practical guidance for selecting reliable survival modeling approaches in privacy-preserving, multi-institutional healthcare environments.

1. Introduction

Survival (time-to-event) analysis plays a central role in clinical research and decision-making, supporting applications such as prognosis, disease progression modeling, and treatment evaluation. In practice, survival data are distributed across multiple healthcare institutions and cannot be readily pooled because of privacy regulations such as HIPAA and GDPR. This fragmentation limits the development of robust and generalizable survival models that reflect diverse patient populations. **Federated Survival Analysis (FSA)** addresses this challenge by enabling collaborative model training without sharing patient-level data. Instead, institutions train models locally and exchange only model parameters or summary statistics, preserving privacy while leveraging distributed clinical data. However,

federated learning introduces additional challenges, including heterogeneous patient populations, institutional variability, non-IID data distributions, and varying censoring patterns, all of which can substantially affect model reliability and generalizability.

Several survival models has been adapted to federated settings, including classical statistical approaches such as Kaplan–Meier estimators (von Maltitz et al., 2021), Accelerated Failure Time (AFT) models (Saikia and Barman, 2017), and Cox proportional hazards (CoxPH) models (Seidi et al., 2024), as well as machine learning and deep learning methods such as DeepHit (Lee et al., 2018), DeepPseudo (Rahman et al., 2021), and attention-based federated architectures including FedPAttn (Rahman and Purushotham, 2023). While these approaches differ in their modeling assumptions and complexity, there remains no systematic benchmark that evaluates their performance under realistic federated, multi-institutional healthcare conditions. Existing benchmark studies have focused primarily on centralized settings. For example, SurvBenchmark (Zhang et al., 2022b) provides a comprehensive comparison of survival models using pooled datasets, while subsequent studies (Burk et al., 2024; Oexner et al., 2025) evaluate statistical and machine learning methods under centralized learning. Although these benchmarks provide valuable insights into predictive performance, they do not address the unique challenges of federated healthcare, including heterogeneous client distributions, varying censoring rates, institutional variability, or the reproducibility of model explanations. Consequently, they provide limited guidance for selecting reliable survival models for privacy-preserving, multi-institutional clinical applications. Table 1 summarizes the key differences between existing benchmark studies and FSA-Bench.

In this work, we introduce **FSA-Bench**, the first comprehensive benchmark for federated survival analysis. We systematically evaluate statistical, machine learning, and deep learning survival models across ten real-world clinical datasets under realistic federated learning settings. Beyond conventional predictive performance, FSA-Bench evaluates discrimination, calibration, robustness, and explanation stability, providing a comprehensive assessment of model reliability. In particular, we introduce an explanation-stability evaluation that measures the consistency of feature importance rankings across cross-validation folds, federated clients, and centralized versus federated training using rank-based agreement metrics. Our main contributions are summarized as follows:

- We introduce **FSA-Bench**, the first comprehensive benchmark for federated survival analysis, comprising ten diverse multi-institutional clinical datasets with realistic heterogeneity, censoring patterns, and federated data distributions.
- We develop a **unified evaluation framework** that systematically evaluates federated survival models using complementary metrics for discrimination, calibration, probabilistic accuracy, robustness, and explanation stability.
- We present the **first benchmark of explanation stability** for federated survival analysis by quantifying the consistency and reproducibility of feature importance rankings across folds, sites, and learning paradigms.
- We provide the **first comprehensive comparison** of statistical, machine learning, and attention-based federated survival models under realistic heterogeneous, multi-institutional healthcare settings.
- We provide **practical guidelines for trustworthy federated survival model selection** through the joint evaluation of discrimination, calibration, robustness, and explanation stability.

Table 1: Comparison of existing benchmark studies for survival analysis. Prior work primarily evaluates centralized settings, whereas our work provides the first comprehensive benchmark for federated survival analysis.

Aspect	**Property**	Zhang et al. (2022b)	Dirick et al. (2017)	Burk et al. (2024)	Oexner et al. (2025)	**This Work**
Data	Centralized	✓	✓	✓	✓	✓
	Federated	–	–	–	–	✓
Statistical	Non-param.	–	–	✓	–	✓
	Semi-param.	✓	✓	✓	✓	✓
	Parametric	–	✓	–	–	✓
ML	Tree-based	✓	–	✓	✓	✓
	Deep	✓	–	✓	✓	✓
	Attention	–	–	–	–	✓
Evaluation	Ranking	✓	✓	✓	✓	✓
	Discrimination	✓	–	✓	–	✓
	Censoring	–	–	–	–	✓
	Stability	–	–	–	–	✓
Events	Cancer	✓	–	✓	✓	✓
	Infectious	✓	–	✓	✓	✓
	Organ failure	✓	–	✓	✓	✓
	Other	✓	✓	–	✓	✓

Generalizable Insights about Machine Learning in the Context of Healthcare

Survival (time-to-event) analysis is fundamental to clinical decision-making, supporting prognosis, risk stratification, and treatment evaluation. In real-world healthcare settings, these data are inherently distributed across institutions and cannot be centrally aggregated because of privacy regulations, motivating Federated Survival Analysis (FSA) as a framework for privacy-preserving, multi-institutional modeling. Our results show that model performance in federated settings is driven primarily by data characteristics, including heterogeneity, censoring, and institutional variability, rather than model class or training paradigm alone. While federated models achieve discrimination comparable to centralized approaches, they exhibit important differences in calibration, robustness, and explanation stability. Attention-based federated models (e.g., FedPAttn) provide more stable and better-calibrated survival predictions, whereas federated Cox proportional hazards models demonstrate the strongest overall consistency in feature importance rankings across sites and data partitions. Our explanation-stability analysis further reveals that models differ substantially in the reproducibility of their feature importance rankings under heterogeneous federated settings. In particular, classical statistical models generally produce more consistent explanations, while some deep learning models exhibit greater variability, indicating that strong predictive performance does not necessarily imply reliable or reproducible explanations. Overall, our benchmark demonstrates that trustworthy evaluation of federated survival models should extend beyond discrimination to jointly assess calibration, robustness, and explanation stability. These findings provide practical guidance for selecting survival models that are not only accurate but also reliable, interpretable, and reproducible for privacy-preserving, multi-institutional healthcare applications.

2. Related Work

Survival analysis has long been used for clinical time-to-event prediction, with classical statistical methods such as the Kaplan–Meier estimator (Kaplan and Meier, 1958), Cox proportional hazards (CoxPH) (Cox, 1972), and Accelerated Failure Time (AFT) models (Kalbfleisch and Prentice, 2002) serving as the foundation of clinical risk modeling. More recently, these approaches have been extended to federated learning to enable privacy-preserving model development across multiple institutions. To improve modeling flexibility, deep learning methods such as DeepHit (Lee et al., 2018), DeepPseudo (Rahman et al., 2021), and the FedPseudo framework (Rahman and Purushotham, 2023) have been proposed, allowing nonlinear survival modeling under heterogeneous federated settings. However, existing studies typically focus on individual models or a single model family rather than providing a systematic comparison across diverse survival approaches.

Several benchmark studies have evaluated survival models in centralized settings. For example, SurvBenchmark (Zhang et al., 2022b) compares statistical and deep learning survival models using pooled datasets, while large-scale evaluations (Burk et al., 2024; Oexner et al., 2025) assess classical and machine learning approaches across multiple datasets. Other studies, such as (Dirick et al., 2017), focus on specific application domains. Although these benchmarks provide valuable insights into predictive performance, they assume centralized data and do not evaluate the challenges introduced by federated healthcare, including heterogeneous client distributions, varying censoring rates, and institutional variability.

Table 1 summarizes the key differences between existing benchmarks and FSA-Bench. In contrast to prior work, FSA-Bench provides the first unified benchmark for federated survival analysis, evaluating statistical, machine learning, and attention-based models under realistic multi-institutional settings. Beyond conventional predictive performance, our benchmark jointly evaluates discrimination, calibration, robustness, and explanation stability, enabling a comprehensive assessment of both predictive reliability and the reproducibility of model explanations across heterogeneous federated healthcare environments.

3. FSA-Bench

In this section, we describe the design of FSA-Bench, including the benchmark datasets, survival models, and evaluation framework used throughout the study. An overview of the proposed framework is shown in Fig. 1. The pipeline begins with data extraction, preprocessing, and feature selection to construct standardized time-to-event datasets distributed across multiple federated clients. We evaluate a diverse set of federated survival models spanning three categories: classical statistical models (Kaplan–Meier, AFT Weibull, and Cox proportional hazards), statistical neural models (NNPH and NNnPH), and deep learning approaches (DeepPseudo, DeepHit, FedPDNN, FedPLSTM, and FedPAttn).

To ensure a fair and reproducible comparison, all models are evaluated using a consistent cross-validation protocol across all datasets. Performance is assessed using clinically relevant survival analysis metrics, including the time-dependent C-index, Integrated Brier Score (IBS), and negative log-likelihood (NLL) to measure discrimination, calibration, and goodness-of-fit. In addition, we evaluate explanation stability by quantifying the consistency of feature importance rankings across folds, sites, and data partitions using rank-based agreement metrics. This comprehensive evaluation framework enables a systematic comparison of predictive performance, robustness, and interpretability, providing a rigorous

assessment of federated survival models under realistic heterogeneous, multi-institutional healthcare settings.

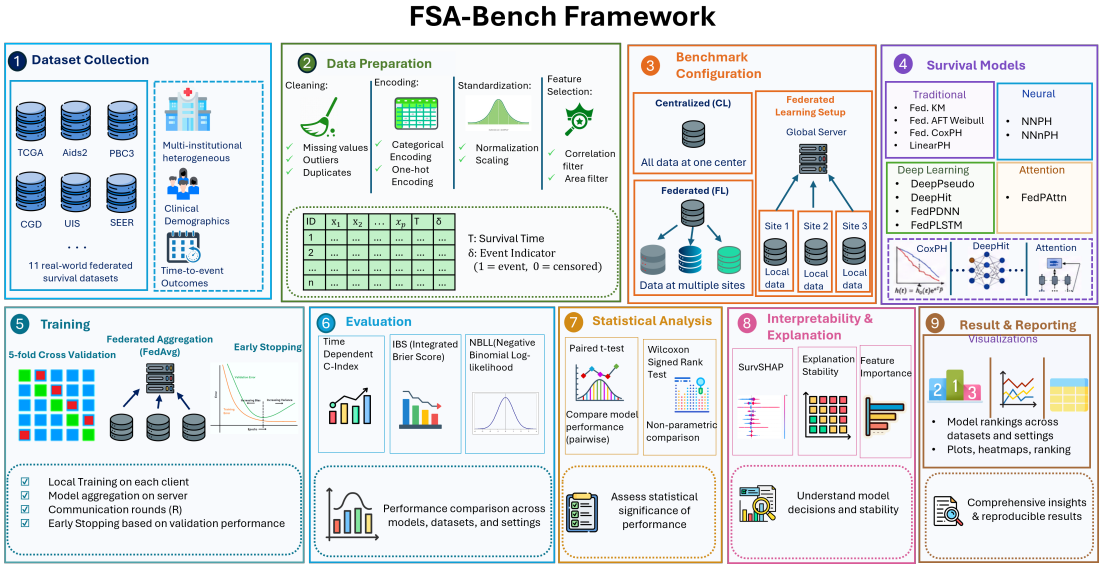

Figure 1: Overview of FSA-Bench Framework

3.1. Real-world Distributed Survival Datasets

We evaluate FSA-Bench on multiple real-world distributed survival datasets with diverse characteristics in terms of sample size, feature dimensionality, event rates, and censoring patterns. Key dataset characteristics are summarized in Table 2, while their distributions are illustrated in Appendix Fig. 7. Collectively, these datasets represent a broad spectrum of healthcare applications, encompassing diverse patient populations, clinical characteristics, disease domains, and geographical regions. TCGA (National Cancer Institute, 2026a) is a large multi-center cancer dataset comprising 17 cancer types collected across institutions in North America and Europe. Aids2 (Kemp, 2003) contains survival data for AIDS patients in Australia, whereas LeukSurv provides survival outcomes for patients with acute myeloid leukemia in the United Kingdom. Pbc3 (Lombard et al., 1993; de Boer et al., 2024) includes patients with primary biliary cirrhosis treated at multiple European hospitals and contains extensive clinical and biochemical measurements. Dialysis (SÁ CARVALHO et al., 2003) provides survival data for hemodialysis patients from 67 clinics across Brazil. The d.oropha.rec (Drysdale, 2022) and CGD (Fleming and Harrington, 2013) datasets originate from multicenter clinical trials on oropharyngeal carcinoma and chronic granulomatous disease conducted in the United States and Europe, respectively. UIS (Auton, 2000) includes survival data for drug-dependent individuals in the United States, while Scania (Dribe and Quaranta, 2020) contains historical survival records of adults aged 50 years and older from nineteenth-century rural Sweden. Finally, SEER (Rahman and Purushotham, 2022) is a large U.S. population-based cancer registry containing detailed demographic, tumor, and treatment information.

To emulate realistic federated healthcare environments, we preserve naturally occurring institutional or geographically defined partitions whenever available. For example, Pbc3

Table 2: Data sets summary

Dataset	Samples	Variables	Event(%)	Censoring(%)	# clients	Data link
TCGA (National Cancer Institute, 2026a) (Rahman and Purushotham, 2023)	6084	283	30.50%	69.50%	7	https://www.cancer.gov/ccg/research/genome-sequencing/tcga
Aids2 (Kemp, 2003; Drysdale, 2022)	2839	3	62.03%	37.97%	4	https://rdrr.io/cran/MASS/man/Aids2.html
LeukSurv (Drysdale, 2022) (Henderson et al., 2002)	1043	4	84.28%	15.72%	24	https://rdrr.io/cran/spBayesSurv/man/LeukSurv.html
Pbc3 (Lombard et al., 1993) (de Boer et al., 2024)	349	12	17.48%	82.52%	6	https://www.rdocumentation.org/packages/pec/versions/2025.06.24/topics/Pbc3
Dialysis (Drysdale, 2022) (SÁ CARVALHO et al., 2003)	6805	2	23.56%	76.44%	67	https://rdrr.io/cran/RcmdrPlugin.survival/man/Dialysis.html
Oropharyngeal Carcinoma (Aalen et al., 2008) (Drysdale, 2022)	192	12	72.40%	27.60%	6	https://rdrr.io/cran/invGauss/man/d.oropha.rec.html
cgd (Drysdale, 2022) (Fleming and Harrington, 2013)	128	8	34.38%	65.62%	4	https://rdrr.io/cran/survival/man/cgd.html
uis (Drysdale, 2022) (Auton, 2000)	628	7	80.89%	19.11%	2	https://github.com/graemeleehickey/hosmer-lemeshow/blob/master/edition2/uis.txt
scania (Dribe and Quaranta, 2020) (Drysdale, 2022)	1040	4	52.60%	47.40%	5	https://cran.r-project.org/web/packages/eha/refman/eha.html#scania
SEER (Rahman and Purushotham, 2022) (National Cancer Institute, 2026b)	25749	11	25.36%	74.64%	3	https://seer.cancer.gov/

consists of 349 patients distributed across six European hospitals with substantial variation in client size, event prevalence, and censoring rates (Table 5), while Dialysis comprises 6,805 patients from 67 Brazilian clinics with clinic-level censoring rates ranging from 37.5% to nearly 100%. Preserving these natural client partitions captures realistic heterogeneity in patient populations, sample sizes, event distributions, and censoring patterns, providing a more representative and clinically relevant evaluation setting for federated survival analysis.

3.2. Federated Survival Models

Federated survival models enable collaborative learning across distributed healthcare institutions without sharing raw patient data. To provide a comprehensive benchmark, we evaluate a diverse set of survival models spanning classical statistical methods, neural survival models, and deep learning approaches under a common federated learning framework.

3.2.1. Federated Statistical Survival Models

Federated Kaplan–Meier (Fed. KM) The Federated Kaplan–Meier estimator (Hansen et al., 2022) is a non-parametric method that estimates the global survival function by aggregating client-level summary statistics, including the number at risk and the number of observed events, without sharing individual patient records.

Federated Accelerated Failure Time (Fed. AFT–Weibull) The Federated Accelerated Failure Time (AFT) model (Hu et al., 2025) employs a Weibull distribution to model

the relationship between covariates and survival time. Each client estimates local model parameters, which are aggregated to construct a global parametric survival model.

Federated Cox Proportional Hazards (Fed. CoxPH) The Federated Cox proportional hazards model (Seidi et al., 2024) estimates a global hazard function by collaboratively learning the effects of covariates on survival risk through the exchange of intermediate statistics. As a semi-parametric approach, it remains one of the most widely adopted methods for clinical survival analysis.

3.2.2. Federated Machine Learning Survival Models

To overcome the modeling assumptions of classical survival methods, we also evaluate several machine learning and deep learning approaches capable of learning complex nonlinear relationships from heterogeneous federated data.

DeepPseudo DeepPseudo (Rahman et al., 2021) uses pseudo-values derived from survival estimates as training targets. A feed-forward neural network is trained locally at each client, and model parameters are periodically aggregated to obtain a global survival model.

DeepHit DeepHit (Lee et al., 2018) discretizes survival time into intervals and directly models the joint distribution of event times and outcomes using a deep neural network. The model jointly optimizes likelihood and ranking objectives and is trained collaboratively through federated parameter aggregation.

Federated Neural Cox Models (NNPH and NNnPH) We evaluate two neural extensions of the Cox PH model (Zhang et al., 2022a): NNPH, which preserves the proportional hazards assumption, and NNnPH, which relaxes this assumption by incorporating time-varying effects. Both models employ neural networks to learn nonlinear risk representations while optimizing survival-specific loss functions under federated training.

FedPseudo Framework FedPseudo (Rahman and Purushotham, 2023) employs Federated Pseudo Values (FPVs) as supervision instead of raw survival times and event indicators. Clients compute local FPVs using globally aggregated summary statistics and train one of three neural architectures:

- **FedPDNN:** A fully connected feed-forward neural network for modeling nonlinear survival relationships.
- **FedPLSTM:** An LSTM-based architecture (Hochreiter and Schmidhuber, 1997) that captures temporal dependencies in pseudo-values.
- **FedPAttn:** A Bi-LSTM with a global attention mechanism (Sun et al., 2021) that models complex temporal patterns while emphasizing informative time-dependent representations.

Collectively, these models represent a broad spectrum of statistical and learning-based approaches for FSA. Their diverse modeling assumptions and architectural designs enable a comprehensive evaluation of predictive performance, calibration, robustness, and explanation stability under realistic heterogeneous, multi-institutional healthcare settings.

3.3. Predictive Performance Evaluation

We evaluate predictive performance using three complementary survival analysis metrics that assess discrimination, calibration, and probabilistic accuracy: the time-dependent C-index, Integrated Brier Score (IBS), and negative log-likelihood (NLL). Together, these

metrics provide a comprehensive evaluation of survival prediction quality under heterogeneous federated settings.

Time-dependent C-index. We adopt the time-dependent C-index proposed by Antolini *et al.* (Antolini et al., 2005), based on the concordance framework of Harrell *et al.* (Harrell et al., 1982), to evaluate a model's ability to correctly rank patients according to their risk of experiencing an event. Higher values indicate better discrimination.

Integrated Brier Score (IBS). IBS (Graf et al., 1999) evaluates calibration by measuring the discrepancy between predicted survival probabilities and observed outcomes while accounting for censoring through inverse probability of censoring weighting (IPCW). Lower values indicate better calibration.

Negative Log-Likelihood (NLL). Negative log-likelihood (NLL, also referred to as NBLL) (Yang and Berdine, 2015) measures the agreement between the predicted survival distribution and observed event times. Lower values indicate better probabilistic accuracy and complement the C-index and IBS by evaluating the overall quality of predicted survival distributions.

3.4. Interpretability and Explanation-Stability Analysis

To assess the robustness and reproducibility of model explanations, we introduce an explanation-stability analysis based on SurvSHAP (Krzyziński et al., 2023). SurvSHAP computes time-dependent feature attribution values for survival predictions, enabling quantitative comparisons of feature importance across different training settings. Let $\phi_{i,j}(t)$ denote the SurvSHAP value of feature j for test instance i at evaluation time t. The overall importance of feature j is obtained by averaging the absolute attribution values across all test instances and evaluation time points: $I_j = \frac{1}{N|\mathcal{T}|}\sum_{i=1}^{N}\sum_{t\in\mathcal{T}}|\phi_{i,j}(t)|$, where N denotes the number of test instances and $\mathcal{T}$ represents the evaluation time grid. Features are ranked according to I_j, and ranking consistency is evaluated across three complementary settings (Appendix Fig. 8):

- **Fold-level consistency:** compares feature rankings obtained from globally trained federated models across different cross-validation folds.
- **Site-level consistency:** evaluates whether feature importance rankings remain consistent across heterogeneous federated clients using local training data as the reference dataset.
- **Partition-level consistency:** compares feature rankings generated by centralized and federated models using identical reference and test datasets, isolating the effect of the learning paradigm on model explanations.

For all analyses, explanations are computed over a common evaluation grid defined by the 10th, 20th, ..., 90th, and 99th percentiles of the observed survival-time distribution. Time-dependent SurvSHAP values are aggregated across all evaluation time points before constructing the final feature rankings.

Explanation stability is quantified using three complementary metrics. Spearman's rank correlation and Kendall's rank correlation coefficient measure agreement between feature rankings, while Jaccard similarity measures the overlap among the highest-ranked features. Together, these metrics provide a comprehensive assessment of the consistency, robustness,

and reproducibility of model explanations across folds, clinical sites, and federated data partitions.

4. Experiments

We conduct experiments on ten real-world survival datasets to evaluate federated survival models under realistic multi-institutional healthcare settings. Our experiments are designed to answer the following research questions:

(Q1) How does predictive performance differ between centralized and federated survival models?
(Q2) How do federated survival models perform under varying censoring rates?
(Q3) How does the number of participating clients affect model performance?
(Q4) What insights can be drawn regarding predictive reliability, calibration, and explanation stability in federated survival models?

Feature Selection. To ensure a fair and consistent comparison across datasets, we adopt a filter-based feature selection strategy (Bagherzadeh-Khiabani et al., 2016). Location-specific variables (e.g., spatial coordinates and institutional identifiers) are removed to prevent information leakage and artificially inflated performance. This ensures that models are evaluated using clinically meaningful predictors under realistic heterogeneous and non-IID federated settings.

Implementation Details. All statistical and machine learning survival models are implemented in Python, while the FedPDNN, FedPLSTM, and FedPAttn models are implemented using NVIDIA FLARE (Roth et al., 2022). All experiments employ stratified 5-fold cross-validation, with three folds used for training, one for validation, and one for testing. Reported results correspond to the mean and standard deviation across the test folds. To ensure a fair comparison, we use standard model configurations without extensive dataset-specific hyperparameter tuning, thereby avoiding bias toward individual methods. Deep learning models are trained using the Adam optimizer with a learning rate of 10^{-3}, dropout of 0.05, and a maximum of 250 training epochs. Early stopping with a patience of 25 epochs is applied based on validation performance. Federated models are trained for up to 50 communication rounds with early stopping. Evaluation metrics are computed over a common time grid spanning the 10th to the 99th percentile of the observed survival times at 10-percentile intervals. Continuous features are standardized using statistics computed from the training data within each cross-validation fold to prevent data leakage. To facilitate reproducibility, the source code, benchmark configuration, and processed datasets are publicly available on GitHub. Unless otherwise stated, all experiments use identical data partitions, evaluation protocols, and preprocessing pipelines across models to ensure a fair and reproducible benchmark.

4.1. Centralized versus Federated Predictive Performance

Observation: Figure 2 compares the discrimination performance of centralized (CL) and federated learning (FL) models using the time-dependent C-index. Overall, federated learning achieves discrimination comparable to centralized learning across most datasets, particularly for larger cohorts such as TCGA, LeukSurv, SEER, and Dialysis. In several cases, FL slightly outperforms CL, especially for attention-based and neural survival models, suggesting that collaborative learning can effectively leverage distributed clinical data without

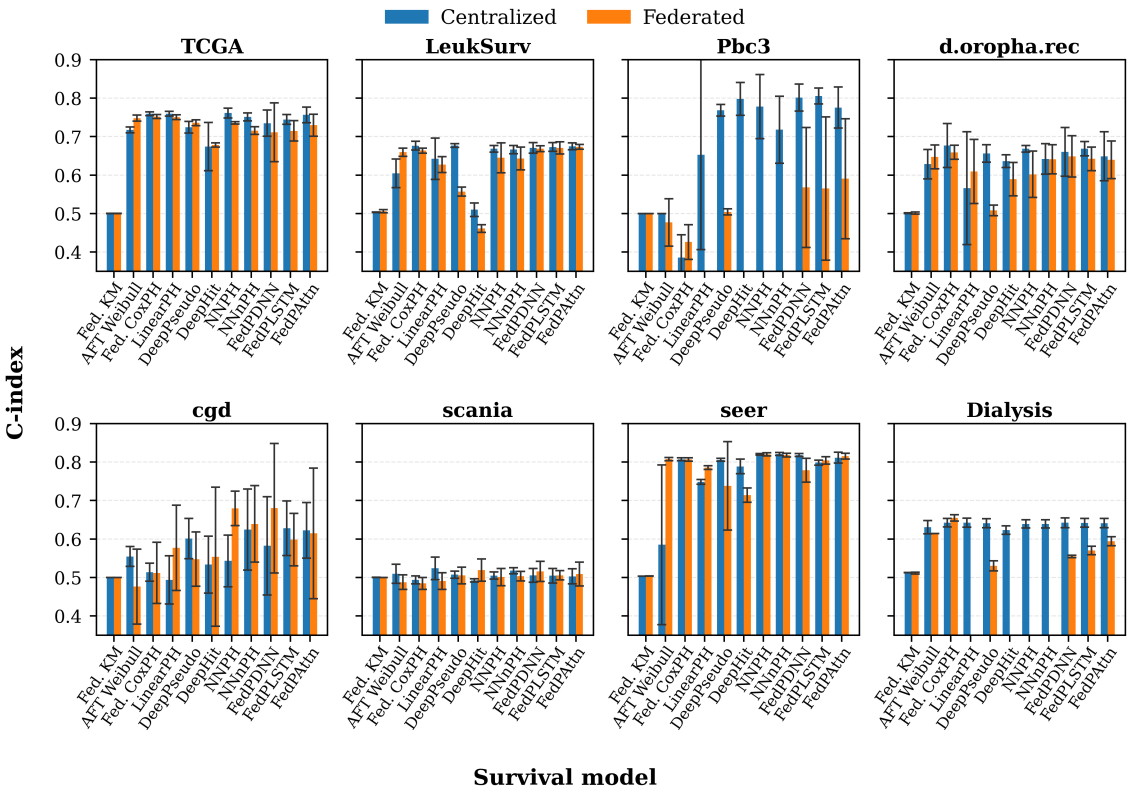

Figure 2: Comparison of time-dependent C-index (mean ± standard deviation) for centralized (CL) and federated learning (FL) models across the ten benchmark datasets. Error bars denote the standard deviation across the cross-validation folds.

requiring centralized data sharing. In contrast, performance differences become more pronounced for smaller and more heterogeneous datasets, including Pbc3, UIS, and Scania, where federated models exhibit greater variability. These results indicate that predictive performance is influenced more by dataset characteristics, such as sample size, censoring, and client heterogeneity, than by the learning paradigm itself.

Discussion: The results demonstrate that federated learning can closely match centralized discrimination performance when sufficient data are available across participating institutions. However, discrimination alone does not fully characterize model reliability. As demonstrated in the following sections, calibration, robustness, and explanation stability provide complementary insights that are essential for selecting reliable survival models in federated healthcare environments.

Clinical Implications: Comparable discrimination between centralized and federated learning suggests that privacy-preserving collaborative modeling can be achieved without substantial loss in predictive accuracy. This finding supports the adoption of federated survival analysis in multi-institutional healthcare systems where patient data cannot be centrally pooled. Nevertheless, institutions with limited sample sizes or highly heterogeneous patient populations should carefully evaluate calibration, robustness, and explanation stability before clinical deployment.

Limitations: Centralized and federated learning operate under fundamentally different

assumptions, as centralized models have access to pooled data whereas federated models are trained on distributed, potentially non-IID datasets. Although this analysis focuses on discrimination, calibration results show only modest differences between the two paradigms. Averaged across the ten benchmark datasets, the IBS increased by only 0.0163, 0.0159, and 0.0091 for FedPDNN, FedPLSTM, and FedPAttn, respectively, under federated training, indicating only a small reduction in calibration relative to centralized learning. Finally, differences in client distributions, sample sizes, and censoring patterns may also contribute to the observed variability across datasets.

4.2. Performance under Varying Censoring Rates

Observation: Figure 3 summarizes predictive performance across datasets with varying censoring rates using discrimination (time-dependent C-index), probabilistic accuracy (NLL), and calibration (IBS). Across all three metrics, predictive performance generally degrades as censoring increases, with a more pronounced decline beyond approximately 50% censoring. Nevertheless, the federated deep learning models (FedPDNN, FedPLSTM, and FedPAttn) consistently maintain competitive performance across a wide range of censoring levels. In particular, FedPAttn frequently achieves the highest C-index, while FedPDNN and FedPAttn generally obtain lower NLL and IBS values, indicating improved calibration and probabilistic accuracy. Statistical models, particularly Fed CoxPH, remain competitive on several datasets but exhibit greater variability under highly censored conditions.

Discussion: These results demonstrate that censoring is a major factor influencing predictive performance in federated survival analysis. Although all methods experience performance degradation as censoring increases, federated deep learning models exhibit smaller relative declines in discrimination and calibration, suggesting greater robustness to incomplete follow-up. Furthermore, the complementary behavior of the C-index, IBS, and NLL highlights that strong discrimination does not necessarily imply well-calibrated survival probabilities, reinforcing the importance of evaluating multiple performance dimensions.

Clinical Implications: High censoring rates are common in longitudinal clinical studies because of patient dropout, loss to follow-up, and limited observation periods. Models that maintain robust discrimination and calibration under these conditions are therefore more suitable for clinical risk stratification, prognosis estimation, and treatment planning. The consistently strong performance of the federated deep learning models suggests that they provide reliable survival predictions even in challenging, highly censored federated healthcare environments.

Limitations: This analysis evaluates predictive performance across a representative collection of benchmark datasets but does not include formal statistical significance testing. In addition, differences in sample size, feature distributions, client heterogeneity, and censoring patterns may contribute to the observed variability and should be considered when interpreting model performance across datasets.

4.3. Effect of the Number of Participating Clients

Observation: Figure 4 illustrates the relationship between calibration performance (IBS) and the number of participating clients, representing different levels of federation. Overall, the federated deep learning models (FedPDNN, FedPLSTM, and FedPAttn) maintain competitive calibration across both small- and large-scale federated settings. For datasets with relatively few participating clients (e.g., UIS and SEER), differences among models are

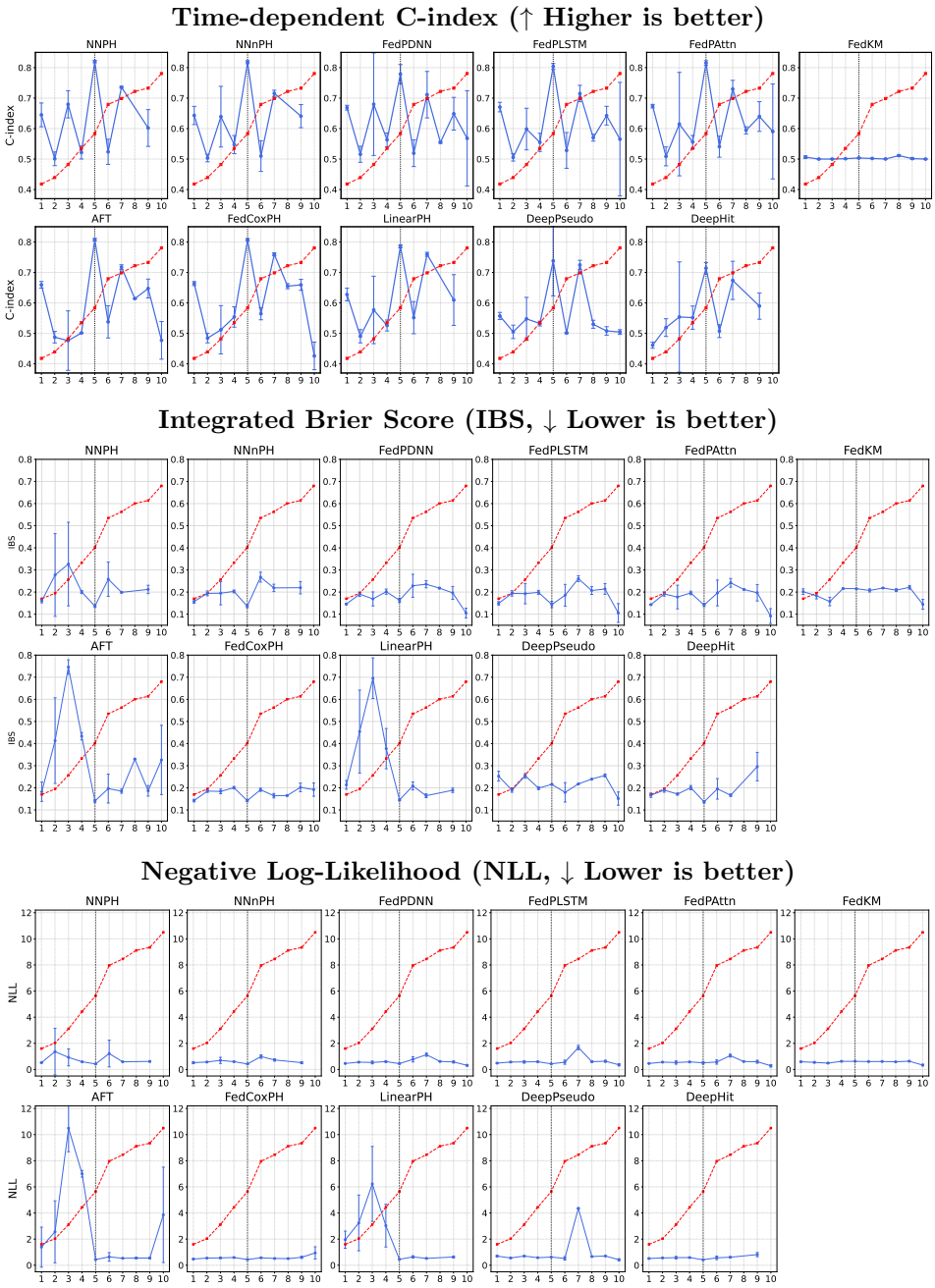

Figure 3: Predictive performance under varying censoring rates across all evaluated survival models. The three panels report (top) time-dependent C-index (↑, higher is better), (middle) Integrated Brier Score (IBS, ↓, lower is better), and (bottom) Negative Log-Likelihood (NLL, ↓, lower is better). Results are shown as the mean ± standard deviation across benchmark datasets. In the top panel, the red dashed curve denotes the censoring rate (right axis), and the vertical dotted line indicates the approximate 50% censoring threshold.
Dataset indices: 1–LeukSurv, 2–UIS, 3–d.oropha.rec, 4–Aids2, 5–Scania, 6–cgd, 7–TCGA, 8–SEER, 9–Dialysis, 10–PBC3.

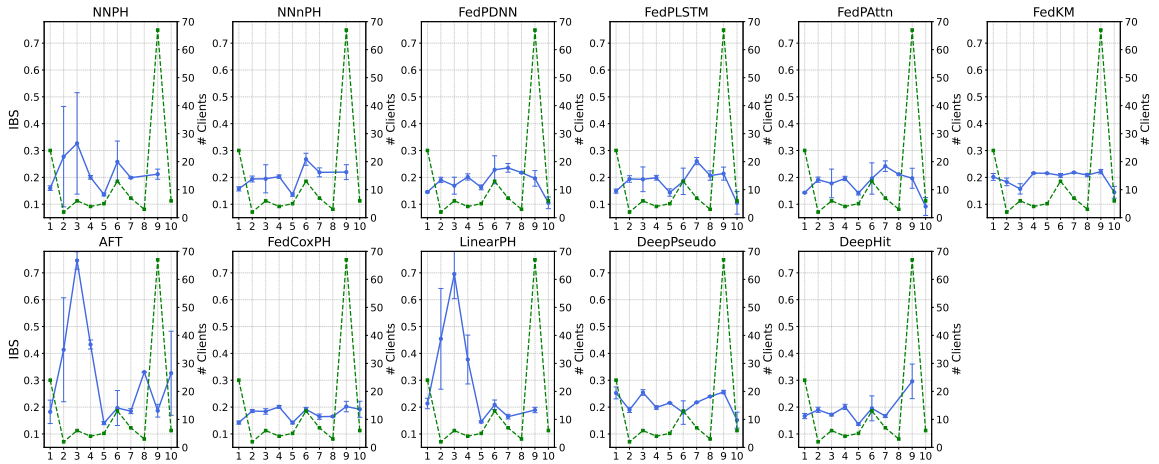

Figure 4: Integrated Brier Score (IBS; lower is better) as a function of the number of participating clients. Each subplot corresponds to a survival model and reports the mean ± standard deviation across the benchmark datasets. The green dashed line denotes the number of participating clients (right axis).

generally small, suggesting limited sensitivity to federation scale when client heterogeneity is modest. In contrast, datasets with a larger number of participating clients (e.g., Dialysis) exhibit greater variability, where FedPDNN and FedPAttn consistently achieve lower IBS values, indicating more robust calibration under increased data decentralization.

Discussion: These results suggest that predictive performance is influenced more by client heterogeneity than by the number of participating clients alone. Increasing the number of clients does not consistently improve or degrade model performance; rather, calibration depends on the underlying distribution of data across institutions. The ability of the federated deep learning models to maintain stable calibration across diverse federation scales indicates their robustness to varying degrees of data decentralization.

Clinical Implications: Healthcare federations may range from a small number of hospitals to large multi-institutional networks. The ability of federated deep models to maintain consistent calibration across different federation scales supports their applicability in privacy-preserving collaborations without requiring centralized data sharing.

Limitations: The number of participating clients is only one aspect of federation complexity. Client heterogeneity, non-IID data and feature distributions, and censoring patterns may have a greater influence on model performance than federation scale alone.

4.4. Statistical Significance and Evaluation Rigor

To assess whether FedPAttn provides statistically significant improvements over other federated survival models, pairwise comparisons are performed using the two-sided Wilcoxon signed-rank test. Datasets are treated as independent experimental units, using the mean C-index from cross-validation for each model and dataset. FedPAttn is compared with Fed. KM, AFT Weibull, Fed. CoxPH, LinearPH, DeepPseudo, DeepHit, NNPH, and NNnPH. Comparisons with missing results are restricted to datasets with valid paired observations. To account for multiple comparisons, p-values are adjusted using Holm's procedure, with $p_{\text{Holm}} < 0.05$ considered statistically significant. We additionally report wins-losses, median paired C-index differences (ΔC), and rank-biserial correlations (r_{rb}).

Table 3: Pairwise Wilcoxon signed-rank comparisons of FedPAttn with federated survival baselines. Holm correction was applied across eight comparisons.

Comparator	n	**W–L**	W	p	p_{Holm}	**Med.** ΔC	r_{rb}
Fed. KM	10	10–0	0	**0.00195**	**0.0156**	+0.1026	+1.000
AFT Weibull	10	7–3	14	0.1934	0.7734	+0.0107	+0.491
Fed. CoxPH	10	6–4	23	0.6953	0.7734	+0.0051	+0.164
LinearPH	8	6–2	4	0.0547	0.2734	+0.0298	+0.778
DeepPseudo	10	9–1	2	**0.00586**	**0.0410**	+0.0657	+0.927
DeepHit	8	7–1	2	0.0234	0.1406	+0.0508	+0.889
NNPH	8	5–3	11	0.3828	0.7734	+0.0124	+0.389
NNnPH	8	5–3	9	0.2500	0.7734	+0.0070	+0.500

Notes: n: paired datasets; W–L: FedPAttn wins–losses; W: Wilcoxon statistic; ΔC: median C-index difference (FedPAttn − comparator); r_{rb}: rank-biserial correlation. Positive values favor FedPAttn. Bold indicates $p_{\text{Holm}} < 0.05$.

Observation: FedPAttn significantly outperformed Fed. KM ($p_{\text{Holm}} = 0.0156$) and DeepPseudo ($p_{\text{Holm}} = 0.0410$), with 10-0 and 9-1 wins-losses, respectively. Both comparisons showed large effect sizes ($r_{\text{rb}} = 1.000$ and 0.927). FedPAttn also showed favorable performance against DeepHit & LinearPH, with large positive effect sizes, although these comparisons did not remain statistically significant after Holm correction.

Discussion: The median paired C-index difference was positive across eight baseline comparisons, indicating that the overall dataset-level performance consistently favored FedPAttn. These results provide statistically supported evidence of improvement over Fed. KM & DeepPseudo, while the large effects against DeepHit & LinearPH suggest additional performance advantages may require more datasets to establish statistical significance.

Limitations: Statistical power is limited by the small number of datasets and, for some baselines, missing paired results. Consequently, potentially meaningful differences may not reach significance after multiple-comparison correction. Evaluation on additional datasets would provide a more comprehensive assessment of statistical robustness.

4.5. Explanation Stability Analysis

Observation: Table 4 summarizes explanation stability across the four benchmark datasets (LeukSurv, SEER, PBC, and CGD). Fed CoxPH achieves the strongest overall ranking consistency, obtaining the best average performance in four of the eight evaluation metrics, particularly at the site and partition levels. Among the federated deep learning models, FedPLSTM achieves the highest overall stability, while FedPAttn obtains the highest Jaccard similarity, indicating the most consistent preservation of the highest-ranked features across different data partitions. Overall, classical statistical survival models, particularly Fed CoxPH, remain highly competitive for producing consistent feature rankings, whereas FedPLSTM and FedPAttn demonstrate the most stable explanations among the deep learning approaches. In contrast, DeepHit exhibits the greatest variability across datasets, indicating lower explanation robustness under heterogeneous federated settings.

Discussion: These results demonstrate that predictive accuracy alone is insufficient for selecting federated survival models. Models with comparable predictive performance may differ substantially in the consistency and reproducibility of their explanations. Evaluating explanation stability therefore provides complementary information that is particularly important for trustworthy clinical decision support.

Clinical Implications: In clinical practice, consistent feature importance rankings im-

Table 4: Aggregate ranking consistency and robustness across the four benchmark datasets (LeukSurv, SEER, PBC, cgd). Results are reported as the mean ± standard deviation computed across datasets. Fold-, site-, and partition-level consistency are evaluated using Spearman's rank correlation (ρ) and Kendall's rank correlation coefficient (τ). Stability measures the consistency of model rankings across federated partitions, while Jaccard similarity quantifies the overlap among the top-ranked models. Higher values indicate better ranking consistency and robustness. The best average performance in each metric is highlighted in **bold**.

Model	Fold ρ	Fold τ	Stability	Site ρ	Site τ	Jaccard	Partition ρ	Partition τ
FedPDNN	0.673 ± 0.292	0.562 ± 0.267	0.609 ± 0.187	0.801 ± 0.123	0.709 ± 0.137	0.732 ± 0.119	0.726 ± 0.209	0.616 ± 0.230
FedPLSTM	$\mathbf{0.741 \pm 0.218}$	$\mathbf{0.632 \pm 0.238}$	$\mathbf{0.779 \pm 0.175}$	0.882 ± 0.056	0.790 ± 0.072	0.768 ± 0.105	0.664 ± 0.232	0.570 ± 0.259
FedPAttn	0.732 ± 0.221	0.630 ± 0.261	0.703 ± 0.196	0.870 ± 0.066	0.784 ± 0.093	$\mathbf{0.802 \pm 0.134}$	0.611 ± 0.281	0.515 ± 0.312
AFT Weibull	0.620 ± 0.264	0.516 ± 0.289	0.564 ± 0.240	0.890 ± 0.046	0.814 ± 0.057	0.762 ± 0.172	0.624 ± 0.259	0.517 ± 0.292
DeepHit	0.539 ± 0.341	0.454 ± 0.290	0.608 ± 0.055	0.523 ± 0.222	0.418 ± 0.174	0.570 ± 0.276	0.579 ± 0.269	0.413 ± 0.229
LinearPH	0.762 ± 0.300	0.682 ± 0.298	0.703 ± 0.252	0.857 ± 0.095	0.768 ± 0.097	0.603 ± 0.368	0.618 ± 0.304	0.515 ± 0.261
NNPH	0.673 ± 0.215	0.543 ± 0.221	0.656 ± 0.172	0.708 ± 0.247	0.597 ± 0.223	0.653 ± 0.161	0.772 ± 0.190	0.659 ± 0.243
NNnPH	0.490 ± 0.372	0.340 ± 0.357	0.705 ± 0.108	0.601 ± 0.212	0.498 ± 0.181	0.461 ± 0.187	0.534 ± 0.333	0.431 ± 0.300
Fed CoxPH	0.727 ± 0.302	0.633 ± 0.302	0.684 ± 0.220	$\mathbf{0.938 \pm 0.051}$	$\mathbf{0.872 \pm 0.094}$	0.791 ± 0.207	$\mathbf{0.828 \pm 0.130}$	$\mathbf{0.706 \pm 0.189}$
DeepPseudo	0.601 ± 0.239	0.443 ± 0.160	0.601 ± 0.221	0.827 ± 0.099	0.742 ± 0.111	0.777 ± 0.112	0.574 ± 0.286	0.492 ± 0.219

prove the transparency and reproducibility of survival models across healthcare institutions. The strong explanation stability of Fed CoxPH, together with the robust performance of FedPLSTM and FedPAttn, suggests that these models provide reliable explanations suitable for multi-institutional clinical deployment.

Limitations: Explanation stability was evaluated using SurvSHAP and may vary across explanation methods. Future work should assess robustness using alternative explainability techniques and larger clinical datasets.

4.6. Overall Insights, Clinical Relevance, and Limitations

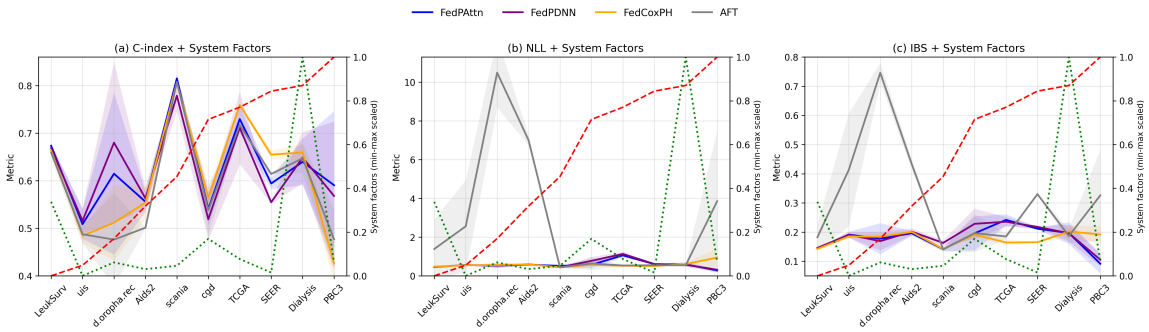

Figure 5: Comparison of representative survival models across datasets. Lines denote mean performance and shaded regions represent ± standard deviation. Top row: (a) C-index, (b) NLL, and (c) IBS. Bottom row: Corresponding trends with censoring rate (red dashed) and number of clients (green dotted, scaled). Federated deep models demonstrate improved stability across varying censoring levels and distributed settings.

Observation: Figure 5 summarizes the overall benchmark results across discrimination, calibration, and probabilistic accuracy. The federated deep learning models, particularly FedPAttn and FedPDNN, consistently achieve strong predictive performance across multiple datasets while exhibiting relatively small variability. Performance generally decreases as censoring increases, whereas the number of participating clients has a comparatively smaller influence. Together with the explanation-stability analysis, these results indicate that data characteristics, including censoring and client heterogeneity, are more influential than the learning paradigm or federation scale alone.

Discussion: Collectively, the experiments demonstrate that no single model is uniformly optimal across all evaluation dimensions. Federated deep learning models provide a favorable balance between discrimination, calibration, and robustness, whereas classical statistical models, particularly Fed CoxPH, remain highly competitive in terms of explanation stability and ranking consistency. These findings highlight the importance of evaluating predictive performance and interpretability jointly when benchmarking FSA models.

Clinical Implications: The results support the use of federated survival analysis for privacy-preserving, multi-institutional healthcare applications. Evaluating discrimination, calibration, and explanation stability together enables the selection of models that are not only accurate but also reliable and interpretable, facilitating trustworthy clinical decision support for prognosis estimation, patient stratification, and treatment planning.

Limitations: Although FSA-Bench provides a comprehensive evaluation across ten clinical datasets, several limitations remain. The benchmark focuses on commonly used federated survival models and evaluation metrics, and additional model architectures, client configurations, and explanation methods warrant further investigation. Moreover, real-world federated deployments may involve more complex communication constraints, client dynamics, and privacy mechanisms than those considered in this study.

5. Conclusion

This paper introduced **FSA-Bench**, the first comprehensive benchmark for federated survival analysis, providing a systematic evaluation of statistical, machine learning, and deep learning survival models under realistic multi-institutional healthcare settings. Beyond predictive performance, FSA-Bench jointly evaluates discrimination, calibration, robustness, and explanation stability, enabling a more comprehensive assessment of model reliability. Our results show that federated survival models achieve predictive performance comparable to centralized approaches while preserving patient privacy. Federated deep learning models, particularly FedPDNN, FedPLSTM, and FedPAttn, consistently demonstrate strong discrimination, calibration, and robustness under heterogeneous and highly censored settings, whereas classical statistical models, especially Fed CoxPH, provide the most consistent and reproducible feature importance rankings. Overall, model performance is driven more by data characteristics, including censoring and client heterogeneity, than by the learning paradigm itself. These findings demonstrate that evaluating discrimination alone is insufficient for selecting federated survival models. Instead, jointly assessing predictive performance and explanation stability provides a more reliable basis for deploying trustworthy models for prognosis estimation, risk stratification, and treatment planning in privacy-preserving, multi-institutional healthcare systems. Future work will extend FSA-Bench to additional clinical datasets, federated optimization strategies, and explainability methods while incorporating advanced privacy-preserving techniques and addressing practical deployment challenges, including heterogeneous computational resources, communication constraints, and dynamic client participation. Overall, FSA-Bench establishes a standardized, reproducible evaluation framework and provides practical guidance for selecting accurate, robust, and interpretable survival models for distributed healthcare applications.

Acknowledgments

This work was partially supported by NSF grant# 2238743

References

Odd O Aalen, Ørnulf Borgan, and Håkon K Gjessing. *Survival and event history analysis: a process point of view*. Springer, 2008.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

Tim Auton. Applied survival analysis: regression modelling of time to event data. *Journal of Applied Statistics*, 27(1):132, 2000.

Farideh Bagherzadeh-Khiabani, Azra Ramezankhani, Fereidoun Azizi, Farzad Hadaegh, Ewout W Steyerberg, and Davood Khalili. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *Journal of clinical epidemiology*, 71:76–85, 2016.

Lukas Burk, John Zobolas, Bernd Bischl, Andreas Bender, Marvin N Wright, and Raphael Sonabend. A large-scale neutral comparison study of survival models on low-dimensional data. *arXiv preprint arXiv:2406.04098*, 2024.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Jasper de Boer, Klest Dedja, and Celine Vens. Survivallvq: Interpretable supervised clustering and prediction in survival analysis via learning vector quantization. *Pattern Recognition*, 153:110497, 2024.

Lore Dirick, Gerda Claeskens, and Bart Baesens. Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, 68 (6):652–665, 2017.

Martin Dribe and Luciana Quaranta. The scanian economic-demographic database (sedd). 2020.

Erik Drysdale. Survset: An open-source time-to-event dataset repository. *arXiv preprint arXiv:2203.03094*, 2022.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*. John Wiley & Sons, 2013.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Christian Rønn Hansen, Gareth Price, Matthew Field, Nis Sarup, Ruta Zukauskaite, Jørgen Johansen, Jesper Grau Eriksen, Farhannah Aly, Andrew McPartlin, Lois Holloway, et al. Larynx cancer survival model developed through open-source federated learning. *Radiotherapy and Oncology*, 176:179–186, 2022.

Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

Robin Henderson, Silvia Shimakura, and David Gorst. Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association*, 97(460):965–972, 2002.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Mengtong Hu, Xu Shi, Ziyang Gong, and Peter X-K Song. Collaborative inference for accelerated failure time model using clinical center-level summary statistics. *Statistics in Medicine*, 44(23-24):e70279, 2025.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2002.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Freda Kemp. Modern applied statistics with s, 2003.

Mateusz Krzyziński, Mikołaj Spytek, Hubert Baniecki, and Przemysław Biecek. Survshap (t): time-dependent explanations of machine learning survival models. *Knowledge-Based Systems*, 262:110234, 2023.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Martin Lombard, Bernard Portmann, James Neuberger, Roger Williams, Niels Tygstrup, Leo Ranek, Helmer Ring-Larsen, Juan Rodes, Miguel Navasa, Christian Trepo, et al. Cyclosporin a treatment in primary biliary cirrhosis: results of a long-term placebo controlled trial. *Gastroenterology*, 104(2):519–526, 1993.

National Cancer Institute. The Cancer Genome Atlas (TCGA) Dataset. https://portal.gdc.cancer.gov/, 2026a. Accessed: January 6, 2026.

National Cancer Institute. The Surveillance, Epidemiology, and End Results (SEER) Dataset. https://seer.cancer.gov/, 2026b. Accessed: January 6, 2026.

Rafael R Oexner, Robin Schmitt, Hyunchan Ahn, Ravi A Shah, Anna Zoccarato, Konstantinos Theofilatos, and Ajay M Shah. Comprehensive benchmarking of machine learning methods for risk prediction modelling from large-scale survival data: A uk biobank study. *arXiv preprint arXiv:2503.08870*, 2025.

Md Mahmudur Rahman and Sanjay Purushotham. Fair and interpretable models for survival analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1452–1462, 2022.

Md Mahmudur Rahman and Sanjay Purushotham. Fedpseudo: Privacy-preserving pseudo value-based deep learning models for federated survival analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1999–2009, 2023.

Md Mahmudur Rahman, Koji Matsuo, Shinya Matsuzaki, and Sanjay Purushotham. Deeppseudo: Pseudo value based deep learning models for competing risk analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 479–487, 2021.

Holger R Roth, Yan Cheng, Yuhong Wen, Isaac Yang, Ziyue Xu, Yuan-Ting Hsieh, Kristopher Kersten, Ahmed Harouni, Can Zhao, Kevin Lu, et al. Nvidia flare: Federated learning from simulation to real-world. *arXiv preprint arXiv:2210.13291*, 2022.

MARILIA SÁ CARVALHO, Robin Henderson, Silvia Shimakura, and Ines Pereira Silva Cunha Sousa. Survival of hemodialysis patients: modeling differences in risk of dialysis centers. *International Journal for Quality in Health Care*, 15(3):189–196, 2003.

Rinku Saikia and Manash Pratim Barman. A review on accelerated failure time models. *International Journal of Statistics and Systems*, 12(2):311–322, 2017.

Navid Seidi, Satyaki Roy, Sajal K Das, and Ardhendu Tripathy. Addressing data heterogeneity in federated learning of cox proportional hazards models. In *2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, pages 1–7. IEEE, 2024.

Zhaohong Sun, Wei Dong, Jinlong Shi, Kunlun He, and Zhengxing Huang. Attention-based deep recurrent model for survival prediction. *ACM Transactions on Computing for Healthcare*, 2(4):1–18, 2021.

Marcel von Maltitz, Hendrik Ballhausen, David Kaul, Daniel F Fleischmann, Maximilian Niyazi, Claus Belka, and Georg Carle. A privacy-preserving log-rank test for the kaplan-meier estimator with secure multiparty computation: algorithm development and validation. *JMIR medical informatics*, 9(1):e22158, 2021.

Shengping Yang and Gilbert Berdine. The negative binomial regression. *The Southwest respiratory and critical care chronicles*, 3(10):50–54, 2015.

D Kai Zhang, Francesca Toni, and Matthew Williams. A federated cox model with non-proportional hazards. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, pages 171–185. Springer, 2022a.

Yunwei Zhang, Germaine Wong, Graham Mann, Samuel Muller, and Jean YH Yang. Survbenchmark: comprehensive benchmarking study of survival analysis methods using both omics data and clinical data. *GigaScience*, 11:giac071, 2022b.

Appendix A. Federated Survival Analysis Overview

Figure 6 illustrates the overall workflow of the federated survival analysis framework used throughout this study. Multiple healthcare institutions collaboratively train a shared survival model while keeping patient-level data local. Model parameters are periodically aggregated by a central server to produce a global model, enabling privacy-preserving learning across distributed clinical datasets.

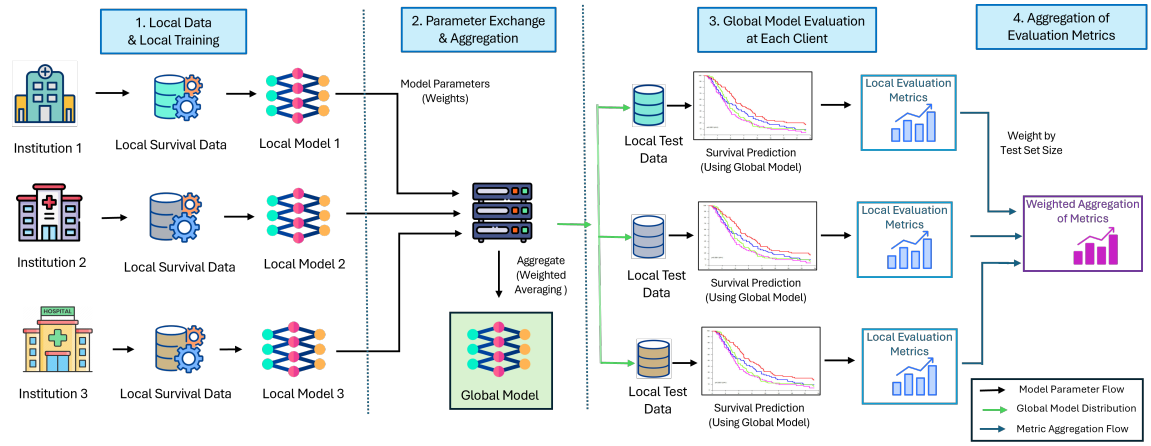

Figure 6: Overview of the federated survival modeling framework.

Appendix B. Federated Survival Dataset Statistics

Figure 7 summarizes the characteristics of the survival datasets included in FSA-Bench. The visualization highlights the diversity of the benchmark in terms of sample size, feature dimensionality, censoring rate, and the number of federated clients. Together, these datasets span a wide range of clinical scenarios and data distributions, providing a comprehensive evaluation setting for federated survival analysis methods.

Appendix C. Model-interpretation and explanation-stability Analysis

Figure 8 presents the workflow for the model interpretation and explanation-stability analysis used in FSA-Bench. The framework evaluates the consistency of feature importance and model explanations across federated clients and the aggregated global model. By comparing local and global interpretations, it enables assessment of whether federated training preserves clinically meaningful explanations while maintaining stable and reliable predictive behavior across heterogeneous institutions.

Appendix D. Client-level characteristics of the FSA datasets

The following tables(Tables 5, 6, and 7), provide detailed client-level statistics for representative federated survival datasets used in this study. For each institution (client), we report the number of samples, observed events, and censored cases. These statistics illustrate the heterogeneity in client sizes and event distributions across datasets, reflecting realistic multi-institutional federated learning scenarios and providing context for the experimental evaluation.

Table 5: Client-level characteristics of the PBC3 dataset.

Hospital/client	**Samples,** n **(%)**	**Events,** n **(%)**	**Censored,** n **(%)**
London	150 (42.98)	29 (19.33)	121 (80.67)
Copenhagen	46 (13.18)	8 (17.39)	38 (82.61)
Hvidovre	23 (6.59)	4 (17.39)	19 (82.61)
Barcelona	79 (22.64)	16 (20.25)	63 (79.75)
Munich	23 (6.59)	2 (8.70)	21 (91.30)
Lyon	28 (8.02)	2 (7.14)	26 (92.86)
Total	**349 (100)**	**61 (17.48)**	**288 (82.52)**

Table 6: Client-level characteristics of the TCGA dataset.

Region/client	**Samples,** n **(%)**	**Events,** n **(%)**	**Censored,** n **(%)**
Europe	901 (14.81)	260 (28.86)	641 (71.14)
Canada	461 (7.58)	116 (25.16)	345 (74.84)
Northeast	1210 (19.89)	361 (29.83)	849 (70.17)
Midwest	1174 (19.30)	383 (32.62)	791 (67.38)
South	1326 (21.80)	400 (30.17)	926 (69.83)
West	697 (11.46)	169 (24.25)	528 (75.75)
Others	315 (5.18)	165 (52.38)	150 (47.62)
Total	**6084 (100)**	**1854 (30.47)**	**4230 (69.53)**

Table 7: Client-level characteristics of the Aids2 dataset.

Region/client	**Samples,** n **(%)**	**Events,** n **(%)**	**Censored,** n **(%)**
NSW	1780 (62.61)	1116 (62.70)	664 (37.30)
Other	249 (8.76)	142 (57.03)	107 (42.97)
QLD	226 (7.95)	148 (65.49)	78 (34.51)
VIC	588 (20.68)	355 (60.37)	233 (39.63)
Total	**2843 (100)**	**1761 (61.94)**	**1082 (38.06)**

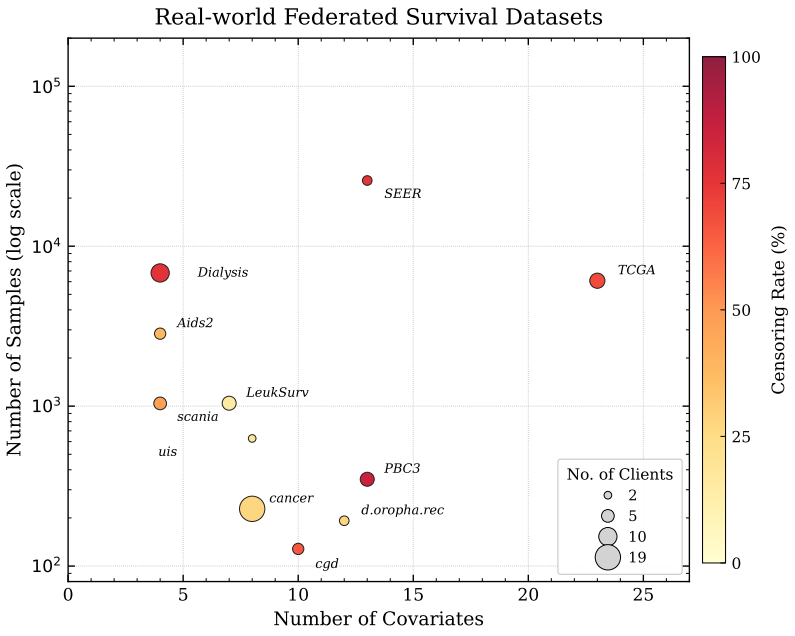

Figure 7: Visualization of 11 real-world survival datasets used in this study. Each bubble represents one dataset; its **vertical position** encodes the number of samples (log scale), its **horizontal position** the number of covariates, its **colour** the censoring rate (%), and its **size** the number of federated clients. Larger datasets appear at the top; high-dimensional ones on the right.

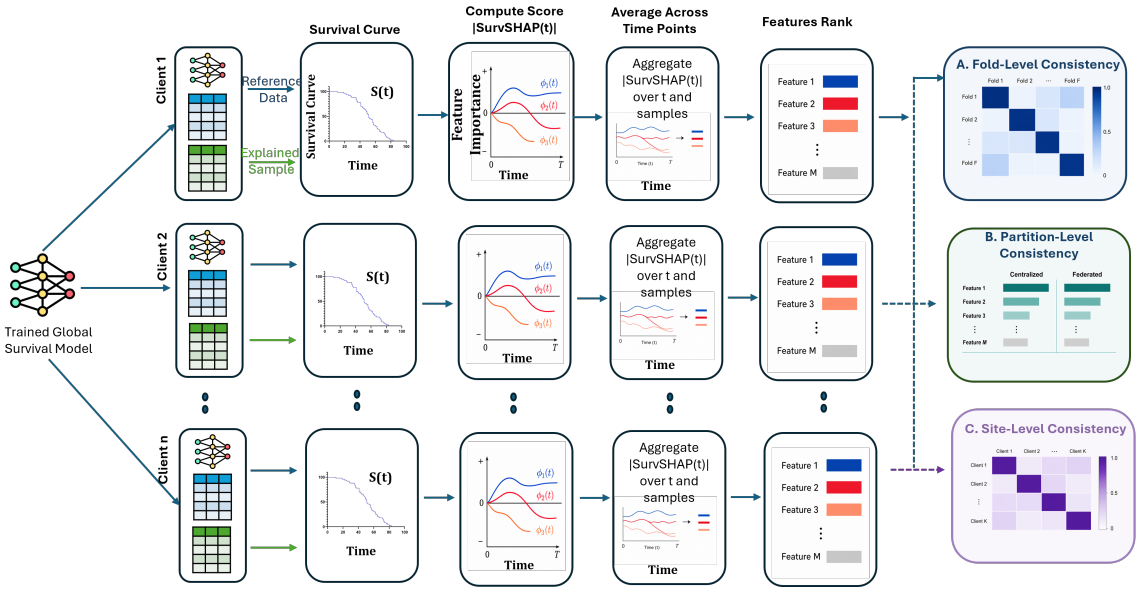

Figure 8: Overview of the model-interpretation and explanation-stability framework.

Appendix E. Model Performance on CL vs FL Partition

Figure 9 compares the predictive performance of survival models under centralized learning (CL) and federated learning (FL) on the Aids2 and UIS datasets. The results illustrate how federated training affects discrimination performance across statistical, machine learning,

and deep learning approaches, highlighting the relative robustness of different models when learning from distributed clinical data.

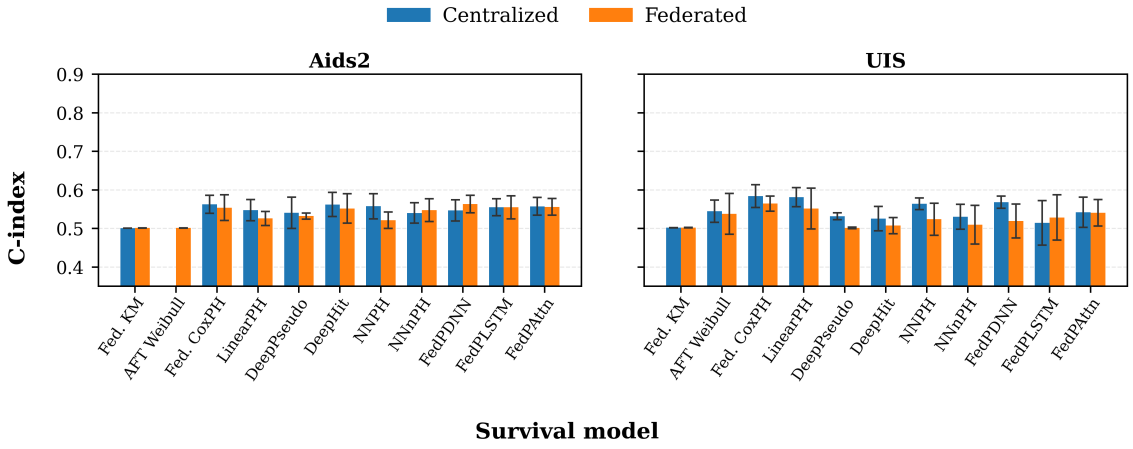

Figure 9: Comparison of predictive performance (C-index, mean ± standard deviation) across Aids2 and UIS datasets for centralized (CL) and federated learning (FL) settings. Each subplot corresponds to a dataset and reports results for classical survival models (Fed. KM, AFT Weibull, CoxPH, LinearPH) and neural approaches (DeepPseudo, DeepHit, NNPH, NNnPH, FedPDNN, FedPLSTM, FedPAttn). Bars denote mean C-index and error bars indicate standard deviation across runs.

Appendix F. Model Performance

The following tables [8,9] report the Negative Log-Likelihood (NLL) results for all evaluated models under centralized learning (CL) and federated learning (FL) across the benchmark datasets. Overall, federated learning generally results in modest increases in NLL compared with centralized training, reflecting the challenges introduced by distributed and heterogeneous data. Classical statistical models exhibit relatively stable performance across learning settings, whereas deep learning models show greater variability on several datasets. Among the federated neural approaches, attention-based and fully connected architectures (e.g., FedPAttn and FedPDNN) frequently achieve competitive performance and remain comparatively robust across datasets, while recurrent models (FedPLSTM) occasionally exhibit higher variance. These results complement the main paper by providing complete NLL scores for all datasets and model families.

The following tables [10, 11] present the Integrated Brier Score (IBS) results for all evaluated models under centralized learning (CL) and federated learning (FL). Lower IBS values indicate better calibration of predicted survival probabilities. Consistent with the observations in the main paper, centralized training generally provides slightly better calibration than federated learning, although the performance gap varies across datasets and model families. Classical survival models maintain relatively stable performance under federation, whereas deep neural models exhibit greater sensitivity to heterogeneous data distributions. Among the federated neural architectures, FedPDNN and FedPAttn consistently demonstrate competitive calibration performance across multiple datasets, while recurrent models (FedPLSTM) show larger variability on some benchmarks. Overall, the detailed IBS results indicate that model robustness depends on both the learning paradigm and dataset charac-

Table 8: Model performance comparison across datasets and settings. CL = Centralized Learning, FL = Federated Learning. Metrics reported is Negative Log-Likelihood (NLL); lower values indicate better performance.

		Non-param.	Semi-param.		Parametric (DL)		
Dataset	Setting	Fed. KM	AFT Weibull	Fed. CoxPH	LinearPH	DeepPseudo	DeepHit
TCGA	CL	0.6246 ± 0.0026	0.5497 ± 0.0186	0.4932 ± 0.0240	0.4933 ± 0.0210	0.6234 ± 0.0020	0.4996 ± 0.0117
	FL	0.6086 ± 0.0044	0.5358 ± 0.0193	0.5201 ± 0.0157	0.5262 ± 0.0158	4.3553 ± 0.0186	0.6128 ± 0.0435
Aids2	CL	0.5902 ± 0.0084	N/A	0.5867 ± 0.0140	3.3794 ± 1.9554	0.5849 ± 0.0077	0.5865 ± 0.0112
	FL	0.6216 ± 0.0048	7.0024 ± 0.2530	0.5912 ± 0.0149	3.0322 ± 1.6366	0.5777 ± 0.0207	0.5890 ± 0.0225
LeukSurv	CL	0.4872 ± 0.0157	1.6352 ± 1.3433	0.4383 ± 0.0141	0.9052 ± 0.8109	0.4833 ± 0.0155	0.4772 ± 0.0524
	FL	0.5914 ± 0.0289	1.3881 ± 1.5279	0.4764 ± 0.0317	1.9563 ± 0.6547	0.6989 ± 0.0445	0.5165 ± 0.0208
Pbc3	CL	0.4218 ± 0.0339	13.5256 ± 0.2975	0.4709 ± 0.0493	2.9803 ± 4.5280	0.4006 ± 0.0374	0.3753 ± 0.0322
	FL	0.3341 ± 0.0611	3.8583 ± 3.6518	0.9443 ± 0.4686	N/A	0.4209 ± 0.0779	N/A
d.oropha.rec	CL	0.6135 ± 0.0075	0.7549 ± 0.2938	0.5838 ± 0.0139	2.7385 ± 3.7524	0.5957 ± 0.0203	0.5867 ± 0.0444
	FL	0.6328 ± 0.0193	0.5505 ± 0.0580	0.6086 ± 0.0833	0.6316 ± 0.0741	0.7053 ± 0.0131	0.8157 ± 0.1615
cgd	CL	0.5168 ± 0.0167	12.3243 ± 0.2504	0.5142 ± 0.0286	4.3683 ± 1.5024	0.5570 ± 0.0868	0.5629 ± 0.0818
	FL	0.4865 ± 0.0431	10.4917 ± 1.8050	0.5619 ± 0.0442	6.2246 ± 2.8699	0.7020 ± 0.0211	0.5805 ± 0.0936
uis	CL	0.5821 ± 0.0075	2.8405 ± 1.9493	0.5686 ± 0.0160	0.5696 ± 0.0184	0.5854 ± 0.0052	0.5870 ± 0.0131
	FL	0.6034 ± 0.0153	0.6394 ± 0.3282	0.5691 ± 0.0182	0.6372 ± 0.0875	0.5279 ± 0.1296	0.5681 ± 0.1339
scania	CL	0.5406 ± 0.0142	2.6519 ± 1.7709	0.5441 ± 0.0131	3.7802 ± 1.3249	0.5407 ± 0.0113	0.5427 ± 0.0140
	FL	0.5426 ± 0.0338	2.5571 ± 2.3722	0.5474 ± 0.0131	3.2377 ± 2.1356	0.5572 ± 0.0242	0.5592 ± 0.0218
seer	CL	0.6208 ± 0.0020	6.4997 ± 5.0275	0.4426 ± 0.0096	0.4789 ± 0.0078	0.6195 ± 0.0016	0.4190 ± 0.0107
	FL	0.6217 ± 0.0022	0.4389 ± 0.0140	0.4417 ± 0.0123	0.4560 ± 0.0094	0.6253 ± 0.0024	0.4229 ± 0.0122
Dialysis	CL	0.5834 ± 0.0037	3.3051 ± 4.7757	0.5438 ± 0.0091	0.5447 ± 0.0108	0.5805 ± 0.0062	0.5468 ± 0.0139
	FL	0.5855 ± 0.0132	0.5505 ± 0.0580	0.5077 ± 0.0102	N/A	0.6653 ± 0.0015	N/A

Table 9: Model performance comparison across datasets and settings. CL = Centralized Learning, FL = Federated Learning. Metrics reported is Negative Log-Likelihood (NLL); lower values indicate better performance.

		Federated Survival Models				
Dataset	Setting	NNPH	NNnPH	FedPDNN	FedPLSTM	FedPAttn
TCGA	CL	0.5043 ± 0.0214	0.5185 ± 0.0228	0.5144 ± 0.0274	0.6207 ± 0.1686	0.5744 ± 0.0830
	FL	0.5872 ± 0.0090	0.7284 ± 0.0889	1.1323 ± 0.1248	1.6785 ± 0.1692	1.0607 ± 0.1103
Aids2	CL	1.2222 ± 1.3714	0.9072 ± 0.7117	0.5837 ± 0.0132	0.5873 ± 0.0089	0.5834 ± 0.0125
	FL	0.5879 ± 0.0170	0.5972 ± 0.0160	0.6012 ± 0.0429	0.5916 ± 0.0296	0.5786 ± 0.0190
LeukSurv	CL	0.4417 ± 0.0141	0.4347 ± 0.0129	0.4425 ± 0.0158	0.4436 ± 0.0168	0.4392 ± 0.0096
	FL	0.5243 ± 0.0285	0.5206 ± 0.0799	0.4582 ± 0.0051	0.4851 ± 0.0212	0.4574 ± 0.0019
Pbc3	CL	0.3452 ± 0.0270	2.1609 ± 3.0607	0.3394 ± 0.0029	0.3420 ± 0.0196	0.3580 ± 0.0311
	FL	N/A	N/A	0.3153 ± 0.0574	0.3567 ± 0.0893	0.2789 ± 0.0935
d.oropha.rec	CL	0.5567 ± 0.0242	0.4347 ± 0.0128	0.5697 ± 0.0470	0.5560 ± 0.0209	0.5606 ± 0.0385
	FL	0.6159 ± 0.0451	0.5206 ± 0.0799	0.5745 ± 0.0760	0.6291 ± 0.0720	0.5932 ± 0.1133
cgd	CL	1.3707 ± 0.8378	3.8726 ± 1.5038	0.5381 ± 0.0449	0.5396 ± 0.0471	0.5489 ± 0.0775
	FL	0.9251 ± 0.6476	0.6951 ± 0.2340	0.5370 ± 0.1065	0.5774 ± 0.0937	0.5300 ± 0.1279
uis	CL	0.5842 ± 0.0108	1.4118 ± 1.7874	0.5833 ± 0.0241	0.5876 ± 0.0252	0.5728 ± 0.0135
	FL	1.2184 ± 1.0224	0.9867 ± 0.1334	0.7766 ± 0.1968	0.5578 ± 0.1588	0.5686 ± 0.1637
scania	CL	0.5563 ± 0.0216	0.5812 ± 0.0876	0.5418 ± 0.0141	0.5427 ± 0.0131	0.5438 ± 0.0143
	FL	1.3669 ± 1.7857	0.5706 ± 0.0239	0.5602 ± 0.0228	0.5691 ± 0.0300	0.5629 ± 0.0211
seer	CL	0.4214 ± 0.0140	0.4143 ± 0.0150	0.4301 ± 0.0122	0.4541 ± 0.0057	0.4386 ± 0.0204
	FL	0.4219 ± 0.0148	0.4224 ± 0.0164	0.4454 ± 0.0240	0.4304 ± 0.0135	0.5121 ± 0.0979
Dialysis	CL	0.5448 ± 0.0101	0.5466 ± 0.0112	0.5461 ± 0.0138	0.5462 ± 0.0124	0.5442 ± 0.0093
	FL	N/A	N/A	0.6135 ± 0.0042	0.5977 ± 0.0499	0.6058 ± 0.0093

teristics, with larger heterogeneous datasets generally exhibiting more stable behavior than smaller cohorts.

Table 10: Model performance comparison across datasets and settings. CL = Centralized Learning, FL = Federated Learning. Metric reported is Integrated Brier Score (IBS); lower values indicate better performance.

		Non-param.	Semi-param.		Parametric (DL)		
Dataset	Setting	Fed. KM	AFT Weibull	Fed. CoxPH	LinearPH	DeepPseudo	DeepHit
TCGA	CL	0.2183 ± 0.0013	0.1850 ± 0.0093	0.1644 ± 0.0099	0.1644 ± 0.0087	0.2177 ± 0.0010	0.1666 ± 0.0054
	FL	0.2109 ± 0.0021	0.1808 ± 0.0071	0.1755 ± 0.0066	0.1783 ± 0.0069	0.2174 ± 0.0012	0.2049 ± 0.0124
Aids2	CL	0.2015 ± 0.0037	-	0.1994 ± 0.0054	0.3801 ± 0.0953	0.1992 ± 0.0034	0.1998 ± 0.0050
	FL	0.2159 ± 0.0023	0.4332 ± 0.0168	0.2008 ± 0.0054	0.3772 ± 0.0910	0.1985 ± 0.0066	0.2009 ± 0.0096
LeukSurv	CL	0.1562 ± 0.0065	0.1931 ± 0.0283	0.1372 ± 0.0052	0.1653 ± 0.0540	0.1453 ± 0.0063	0.1402 ± 0.0051
	FL	0.2017 ± 0.0128	0.1824 ± 0.0436	0.1424 ± 0.0058	0.2135 ± 0.0196	0.2528 ± 0.0221	0.1667 ± 0.0097
Pbc3	CL	0.1301 ± 0.0130	0.8391 ± 0.0182	0.1445 ± 0.0197	0.3407 ± 0.4040	0.1180 ± 0.0160	0.1104 ± 0.0080
	FL	0.1445 ± 0.0221	0.3261 ± 0.1566	0.1919 ± 0.0298	N/A	0.1514 ± 0.0298	N/A
d.oropha.rec	CL	0.2120 ± 0.0037	0.2321 ± 0.0662	0.1876 ± 0.0155	0.2810 ± 0.1552	0.2040 ± 0.0087	0.2009 ± 0.0187
	FL	0.2214 ± 0.0082	0.1866 ± 0.0238	0.2021 ± 0.0194	0.1891 ± 0.0100	0.2560 ± 0.0065	0.2959 ± 0.0644
cgd	CL	0.1699 ± 0.0069	0.7649 ± 0.0152	0.1762 ± 0.0117	0.7150 ± 0.0546	0.1655 ± 0.0145	0.1890 ± 0.0352
	FL	0.1572 ± 0.0192	0.7460 ± 0.0326	0.1838 ± 0.0103	0.6956 ± 0.0916	0.2544 ± 0.0105	0.1719 ± 0.0047
uis	CL	0.1976 ± 0.0033	0.3516 ± 0.0211	0.1923 ± 0.0073	0.1922 ± 0.0064	0.1986 ± 0.0026	0.1998 ± 0.0058
	FL	0.2074 ± 0.0070	0.1966 ± 0.0652	0.1914 ± 0.0074	0.2086 ± 0.0178	0.1796 ± 0.0437	0.1954 ± 0.0464
scania	CL	0.1827 ± 0.0061	0.5167 ± 0.1122	0.1834 ± 0.0050	0.5699 ± 0.0921	0.1823 ± 0.0049	0.1835 ± 0.0057
	FL	0.1833 ± 0.0142	0.4136 ± 0.1935	0.1854 ± 0.0058	0.4543 ± 0.1876	0.1895 ± 0.0096	0.1898 ± 0.0096
seer	CL	0.2148 ± 0.0009	0.4676 ± 0.2723	0.1425 ± 0.0039	0.1531 ± 0.0038	0.2139 ± 0.0010	0.1349 ± 0.0045
	FL	0.2153 ± 0.0010	0.1404 ± 0.0057	0.1421 ± 0.0048	0.1452 ± 0.0038	0.2157 ± 0.0009	0.1366 ± 0.0047
Dialysis	CL	0.1982 ± 0.0017	0.3604 ± 0.3000	0.1816 ± 0.0030	0.1821 ± 0.0046	0.1969 ± 0.0027	0.1833 ± 0.0062
	FL	0.2087 ± 0.0062	0.3302	0.1652 ± 0.0022	N/A	0.2395 ± 0.0016	N/A

Table 11: Model performance comparison across datasets and settings. CL = Centralized Learning, FL = Federated Learning. Metric reported is Integrated Brier Score (IBS); lower values indicate better performance.

		Federated Survival Models				
Dataset	Setting	NNPH	NNnPH	FedPDNN	FedPLSTM	FedPAttn
TCGA	CL	0.1681 ± 0.0101	0.1705 ± 0.0096	0.1722 ± 0.0111	0.1699 ± 0.0085	0.1739 ± 0.0145
	FL	0.1988 ± 0.0032	0.2188 ± 0.0166	0.2355 ± 0.0162	0.2606 ± 0.0133	0.2418 ± 0.0197
Aids2	CL	0.2461 ± 0.0919	0.2394 ± 0.0868	0.1985 ± 0.0058	0.2002 ± 0.0040	0.1984 ± 0.0055
	FL	0.2003 ± 0.0069	0.2029 ± 0.0063	0.2017 ± 0.0116	0.1987 ± 0.0089	0.1961 ± 0.0080
LeukSurv	CL	0.1395 ± 0.0049	0.1375 ± 0.0060	0.1399 ± 0.0052	0.1392 ± 0.0066	0.1387 ± 0.0043
	FL	0.1600 ± 0.0086	0.1574 ± 0.0079	0.1458 ± 0.0033	0.1489 ± 0.0077	0.1431 ± 0.0022
Pbc3	CL	0.1017 ± 0.0016	0.3478 ± 0.4195	0.1039 ± 0.0016	0.1058 ± 0.0063	0.1091 ± 0.0081
	FL	N/A	N/A	0.1053 ± 0.0217	0.1054 ± 0.0423	0.0916 ± 0.0333
d.oropha.rec	CL	0.1891 ± 0.0100	0.2006 ± 0.0087	0.1937 ± 0.0197	0.1882 ± 0.0069	0.1901 ± 0.0158
	FL	0.2120 ± 0.0188	0.2197 ± 0.0277	0.1963 ± 0.0290	0.2141 ± 0.0242	0.1968 ± 0.0373
cgd	CL	0.4310 ± 0.2171	0.7075 ± 0.0450	0.1771 ± 0.0187	0.1760 ± 0.0174	0.1841 ± 0.0350
	FL	0.3267 ± 0.1889	0.1948 ± 0.0524	0.1691 ± 0.0316	0.1929 ± 0.0460	0.1775 ± 0.0534
uis	CL	0.1980 ± 0.0042	0.2380 ± 0.0753	0.1976 ± 0.0080	0.2003 ± 0.0116	0.1941 ± 0.0052
	FL	0.2581 ± 0.0773	0.2675 ± 0.0224	0.2284 ± 0.0520	0.1854 ± 0.0491	0.1959 ± 0.0583
scania	CL	0.1891 ± 0.0090	0.1992 ± 0.0340	0.1831 ± 0.0059	0.1832 ± 0.0056	0.1839 ± 0.0062
	FL	0.2772 ± 0.1869	0.1941 ± 0.0109	0.1905 ± 0.0099	0.1941 ± 0.0122	0.1914 ± 0.0098
seer	CL	0.1363 ± 0.0054	0.1333 ± 0.0058	0.1411 ± 0.0092	0.1465 ± 0.0054	0.1416 ± 0.0046
	FL	0.1364 ± 0.0054	0.1358 ± 0.0062	0.1629 ± 0.0090	0.1437 ± 0.0138	0.1413 ± 0.0068
Dialysis	CL	0.1825 ± 0.0042	0.1830 ± 0.0047	0.1829 ± 0.0059	0.1826 ± 0.0048	0.1820 ± 0.0040
	FL	N/A	N/A	0.2178 ± 0.0034	0.2068 ± 0.0175	0.2116 ± 0.0025