Attention-Based Transformers for Transparent Survival Analysis

Yueh-Han Chen yc7592@nyu.edu
Acey Vogelstein av3848@nyu.edu
Vaarun Muthappan vkm9614@nyu.edu
Luke Sullivan lcs9595@nyu.edu

Abstract

Predicting time-to-event outcomes is a fundamental problem in healthcare, yet survival analysis remains challenging in practice due to limited sample sizes, heterogeneous covariates, and censoring. While attention-based Transformers have shown strong generalization on small tabular datasets, their effectiveness for survival prediction without task-specific training is not well understood. We evaluate TabPFN, a general-purpose pre-trained transformer for tabular data, on two clinical benchmarks, METABRIC and SUPPORT2, comparing its zero-shot performance against classical, ensemble, deep learning, and transformer-based survival models. Zero-shot TabPFN consistently outperforms traditional and deep survival baselines and achieves performance competitive with specialized survival Transformers, particularly in low-data regimes. Fine-tuning TabPFN on a large collection of synthetic survival datasets yields modest gains in probabilistic accuracy, suggesting that survival-aware adaptation is beneficial but limited. Overall, our results highlight the promise of general-purpose transformer priors for data-efficient survival modeling and motivate future work on survival-specific pretraining.

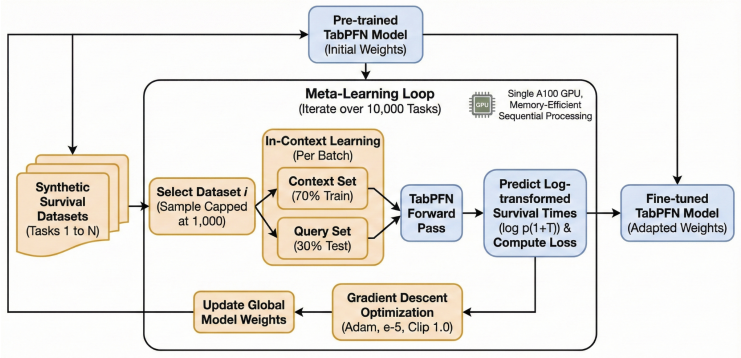

Figure 1: Meta-learning fine-tuning procedure for TabPFN on synthetic survival analysis tasks. Each dataset is treated as an independent task with context and query splits, enabling in-context adaptation and global parameter updates.

1 Introduction

In clinical settings, survival models are routinely used to estimate patient risk, inform treatment decisions, and support prognosis under uncertainty [1]. Despite extensive methodological progress, survival prediction in real-world settings remains difficult due to small sample sizes, heterogeneous covariate distributions, missing data, and complex censoring mechanisms [2, 3]. These challenges

Preprint. Under review.

limit the reliability of both classical approaches such as Cox proportional hazards models and modern deep learning methods [4, 5].

Attention-based Transformers have recently shown strong generalization ability on small tabular datasets, suggesting potential advantages for survival analysis. TabPFN [6], in particular, is a pre-trained transformer capable of performing in-context learning on tabular inputs without task-specific supervised training. However, its suitability for time-to-event prediction is not well understood. Existing work on transformer-based survival models typically relies on specialized architectures or survival-specific objectives trained on real-world datasets [3], leaving open the question of whether a general-purpose transformer prior can support accurate and transparent survival prediction—especially in data-constrained clinical environments.

This work investigates whether TabPFN can generalize effectively to survival analysis and whether its performance can be improved by adapting it to survival-specific structure. We benchmark the model on two established clinical datasets—SUPPORT2 and METABRIC [7, 8]—and compare it against widely used baselines spanning statistical, ensemble, deep learning, and transformer families. To explore whether survival-aware priors enhance in-context learning, we fine-tune TabPFN on a large collection of synthetically generated survival datasets designed to capture diverse covariate distributions, hazard structures, and censoring regimes [9].

Our contributions are threefold:

1. **Evaluation of a general-purpose Transformer for survival analysis.** We assess TabPFN's zero-shot performance against CoxPH, Random Survival Forests, DeepSurv, DeepHit, and SurvTRACE across two real-world clinical benchmarks.

2. **Scaling analysis.** We characterize TabPFN's performance across varying sample sizes, with particular focus on the low-data regimes that commonly arise in healthcare applications.

3. **Survival-specific fine-tuning using synthetic data.** We generate 10,000 diverse synthetic survival datasets and fine-tune TabPFN to investigate whether survival-aware adaptation improves its in-context predictions.

2 Datasets

We evaluate our fine-tuned model and all the baselines on two well-established survival analysis benchmark datasets, including METABRIC and SUPPORT2, that represent distinct clinical domains and data characteristics. We present high-level statistics in Table 1, which summarizes key characteristics of both datasets.

METABRIC The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset [8] is a comprehensive breast cancer genomics study comprising 1,904 patients with long-term follow-up data. The dataset contains 9 clinical and molecular features, including hormone receptor status, tumor size, and gene expression markers. The outcome of interest is overall survival, with 57.9% of patients experiencing the event (death) and 42.1% being right-censored. Follow-up times range from 0 to 355 months, with a median survival time of approximately 10 years. METABRIC serves as the gold standard benchmark for survival analysis methods due to its well-curated nature, balanced event rate, and clinical relevance in oncology research.

SUPPORT2 The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT2) dataset [7] is a large-scale prospective cohort study of 9,105 seriously ill hospitalized patients across multiple disease categories including acute respiratory failure, congestive heart failure, chronic obstructive pulmonary disease, cirrhosis, coma, colon cancer, lung cancer, and multiple organ system failure. After one-hot encoding of categorical variables, the dataset contains 57 features capturing demographics, physiological measurements, disease severity scores, and comorbidities. The dataset exhibits a high event rate of 68.0% mortality, with follow-up times ranging from 3 to 2,029 days (median: approximately 1.5 years). SUPPORT2 presents a challenging benchmark due to its heterogeneous patient population, real-world missing data patterns (approximately 12% missingness), and diverse disease etiologies, making it ideal for testing model robustness in complex clinical settings.

3 Related Work

Survival analysis focuses on time-to-event prediction under censoring. It has long relied on classical models such as the Kaplan–Meier estimator and the Cox proportional hazards (CoxPH) model, valued for their interpretability and efficiency despite restrictive assumptions [1]. To increase expressiveness, non-parametric ensemble methods like Random Survival Forests (RSF) extend tree-based learning to censored data [10], while deep learning approaches such as DeepSurv and DeepHit model nonlinear risk functions or discrete-time event distributions [4, 5]. More recently, transformer-based survival models, including SurvTRACE, have applied self-attention to tabular survival data using survival-specific architectures [3]. In parallel, foundation models for tabular data, notably TabPFN, show that pretrained on large collections of synthetic datasets can perform Bayesian-style inference via in-context learning and excel in low-data regimes [6], with extensions to related tasks such as causal effect estimation [9]. We bridge these lines by examining whether a general-purpose PFN prior can be effectively applied to survival analysis and how it compares to survival-specific methods.

4 Baselines

We evaluate TabPFN against five representative survival analysis baselines covering semi-parametric, ensemble-based, deep learning, and transformer-based approaches. Cox Proportional Hazards (CoxPH) serves as a standard semi-parametric reference model [1]. Random Survival Forests (RSF) represent non-parametric ensemble methods for censored data [10]. DeepSurv and DeepHit are included as established neural survival models that respectively extend CoxPH with nonlinear predictors and model discrete-time event distributions [4, 5]. Finally, SurvTRACE is included as a survival-specific transformer baseline, enabling a direct comparison between a task-specialized attention-based architecture and a general-purpose PFN prior applied to survival prediction [3].

5 Methods

5.1 Evaluation Metrics

We evaluate survival prediction using three complementary metrics. The concordance index (C-index) measures a model's ability to correctly rank survival times by computing the proportion of comparable patient pairs whose predicted risks agree with the observed ordering while accounting for right censoring [11]. The Integrated Brier Score (IBS) assesses probabilistic accuracy by integrating the squared error between predicted survival probabilities and observed outcomes over time, with lower values indicating better calibration and discrimination [12]. Finally, the integrated area under the time-dependent ROC curve (iAUC) evaluates time-dependent discrimination by averaging AUC values across the evaluation horizon. All metrics account for right censoring using standard inverse-probability-of-censoring weighting [13].

5.2 Synthetic Data Generation

To adapt TabPFN to survival analysis, we fine-tune it on a large collection of synthetic time-to-event datasets. Each dataset is sampled from a flexible prior over data-generating processes (DGPs), following the CausalPFN [9] philosophy of exposing the model to many distinct "survival worlds." This encourages the model to learn survival-specific structure and improves robustness when evaluated on real datasets.

Sampling Dataset Configurations. For each dataset, we first sample high-level configuration parameters that define the number of rows, number of features, baseline hazard family, time scale, and censoring rate. The parameter ranges used in this work are summarized in Table 2. Time scales are drawn from three intervals aligned with real data: a SUPPORT2-like range (3–2,029 days), a METABRIC-like range (0–355 months), and a medium generic range (10–500). The first two reflect the duration distributions shown in Figure 3 and Figure 4, while the medium range fills a gap between these extremes to prevent overfitting to only long or short time-scale regimes. The synthetic censoring-rate prior (30–45%) is chosen to match the real datasets' censoring levels (SUPPORT2 $\approx 32\%$, METABRIC $\approx 42\%$), ensuring that fine-tuning occurs under in-distribution conditions.

Covariate Generation. Covariates are generated according to sampled feature-type probabilities: roughly 60% continuous, 25% binary, and 15% categorical. Continuous covariates are drawn from simple families (normal, uniform, exponential), and moderate correlations are injected among 20–40% of continuous feature pairs to mimic realistic clinical structure. Between 5 and 25 features are designated as *prognostic*—that is, features that have a true, non-zero effect on the hazard function and thus influence the underlying event risk. Each prognostic feature receives a log–hazard ratio sampled from small, medium, or large effect-size ranges. The linear predictor combines these main effects with a limited set of non-linear transformations, including squared terms and pairwise interactions for continuous or binary prognostic features.

Survival Time Generation. Survival times are simulated under a proportional hazards model. Each dataset samples one of five baseline hazard families—Weibull, exponential, log-normal, log-logistic, or Gompertz—and event times are drawn using inverse transform sampling. Generated times are rescaled into the specified time range from Table 2 using either linear or log-space scaling, the latter applied when distributions are highly skewed in order to preserve shape.

Censoring Mechanisms. Right censoring is introduced through either administrative or random censoring. Administrative censoring applies a fixed study end time, while random censoring draws censoring times from an exponential distribution. In both cases, parameters are iteratively tuned until the realized censoring rate matches the sampled target. This produces observed times $T_{obs} = \min(T_{event}, T_{censor})$ and event indicators $\delta = \mathbf{1}\{T_{event} \leq T_{censor}\}$. Although the codebase includes support for MCAR missingness, we disable missing-value injection for all experiments so that all covariates remain fully observed.

Preprocessing and Target Construction. Before feeding data to TabPFN, categorical variables are one-hot encoded and continuous variables are standardized. Because TabPFN requires all training examples to share the same input dimensionality, each synthetic dataset is mapped to a fixed dimension of 70 features. If one-hot expansion produces more than 70 columns, entire non-prognostic (noise) features are removed; if fewer, zero-padding is applied. Prognostic features are never dropped. Survival labels are converted into regression targets using a log-time formulation: the model predicts $\log(T_{obs} + \varepsilon)$ with censored observations down-weighted by 0.4, a strategy well aligned with TabPFN's regression interface. In total, 10,000 datasets are generated for fine-tuning.

5.3 Fine-Tuning

We fine-tune the pre-trained TabPFN regression model using a meta-learning approach designed to improve generalization to survival analysis tasks.

Meta-Learning Framework: Unlike conventional fine-tuning approaches that combine all data into a single training set, we adopt a meta-learning paradigm where each synthetic dataset constitutes an independent learning task. This approach mirrors TabPFN's in-context learning mechanism, where the model learns to rapidly adapt to new datasets through its attention-based architecture (Figure 1).

For each synthetic dataset $\mathcal{D}_i = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n_i}$, we perform the following:

1. Split $\mathcal{D}_i$ into context set $\mathcal{D}_i^{ctx}$ (70%) and query set $\mathcal{D}_i^{qry}$ (30%)
2. The model conditions on $\mathcal{D}_i^{ctx}$ via in-context learning
3. Predictions are generated for $\mathcal{D}_i^{qry}$
4. Loss is computed and gradients are backpropagated to update model parameters

This procedure enables the model to learn generalizable representations across 10,000 diverse survival tasks rather than overfitting to any single data distribution.

Training Procedure: We fine-tune TabPFN v2.5 using the official fine-tuning API, which provides utilities for proper data preprocessing and batch collation. The training configuration is shown in Table 3.

The training procedure utilizes TabPFN's `get_preprocessed_datasets()` function to prepare each synthetic dataset and `meta_dataset_collator()` for proper batching. For each batch, we:

1. Preprocess features and targets using TabPFN's internal normalization
2. Fit the model on the context set using `fit_from_preprocessed()`

3. Generate predictions on the query set via the forward pass
4. Compute the loss using TabPFN's z-normalized loss function
5. Backpropagate gradients and update parameters with gradient clipping

To manage GPU memory constraints on a single NVIDIA T4 GPU (16GB), we process datasets sequentially and clear the CUDA cache every 100 iterations. The total training comprises $10,000 \times 10 = 100,000$ gradient updates across all epochs.

5.4 Inference for Survival Analysis

At inference time, the fine-tuned TabPFN model is applied to real survival datasets using a wrapper that converts regression outputs to survival predictions:

Risk prediction: The model predicts log-transformed survival times $\hat{y} = \log(1 + \hat{T})$. Risk scores are computed as $r = -\hat{y}$, where higher values indicate higher risk (shorter predicted survival).

Survival function estimation: We estimate survival probabilities by combining Kaplan-Meier baseline estimates with individual risk adjustments. For a patient with predicted risk percentile p, the survival function is computed as:

$$\hat{S}(t) = S_0(t)^{\gamma(p)} \tag{1}$$

where $S_0(t)$ is the Kaplan-Meier baseline survival and $\gamma(p) = 2.0 - 1.5p$ maps the risk percentile to a hazard multiplier.

This approach leverages TabPFN's strength in relative risk ranking while providing calibrated survival probability estimates for metrics such as the Integrated Brier Score.

6 Results

We evaluate the pre-trained TabPFN (Baseline) and our fine-tuned variant (TabPFN-FT) against the five baselines described in Section 4, using the metrics in Section 5.1. We conduct a scaling analysis by training all models on increasing fractions of the training data (10%, 25%, 50%, 75%, and 100%) for both METABRIC and SUPPORT2; results are shown in Figure 2.

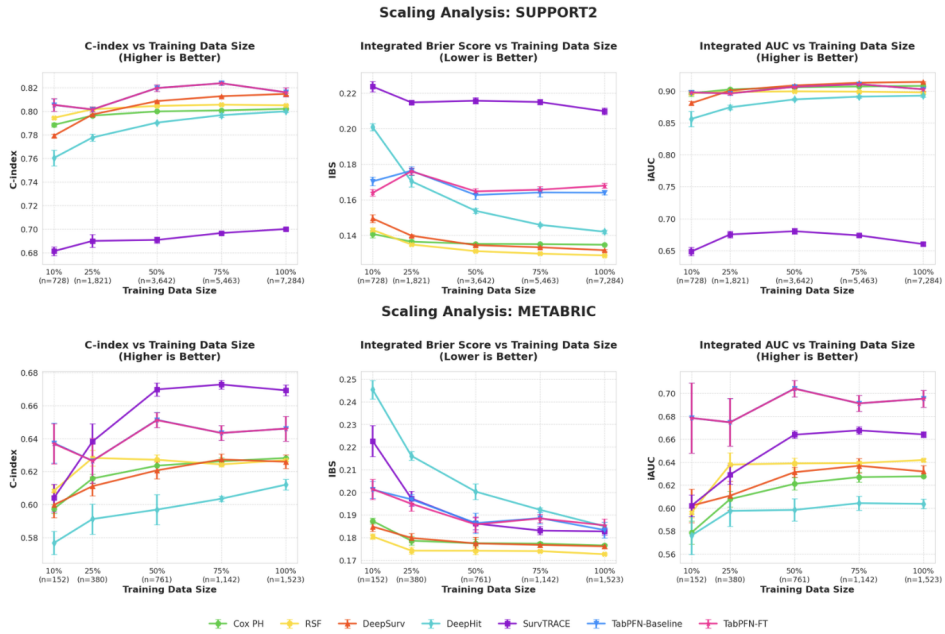

Figure 2: Scaling analysis on SUPPORT2 (top) and METABRIC (bottom). Performance versus training set size for the C-index (left), Integrated Brier Score (middle; lower is better), and integrated AUC (right).

Performance on METABRIC: The baseline TabPFN exhibits strong in-context learning, achieving discriminative performance competitive with specialized deep survival models without task-specific training. Across all data splits, the C-index and iAUC curves for the baseline (blue) and TabPFN-FT (pink) largely overlap, indicating similar ranking performance and only marginal discriminative gains from fine-tuning. In contrast, fine-tuning improves probabilistic calibration in low-data settings: at the 25% split ($n \approx 380$), TabPFN-FT yields a noticeable reduction in Integrated Brier Score, suggesting more accurate probability estimates when data are scarce.

Performance on SUPPORT2: On SUPPORT2, TabPFN-FT performs similarly to the baseline across most metrics and training sizes, underscoring the strength and robustness of the pre-trained priors and suggesting that synthetic fine-tuning does not consistently yield additional gains. Both TabPFN variants remain competitive with established baselines such as DeepSurv and Random Survival Forests throughout the scaling range. Notably, fine-tuning provides a clear benefit in specific low-data conditions: at the 10% split ($n = 728$), TabPFN-FT improves probabilistic accuracy, reducing the Integrated Brier Score from ≈ 0.171 (Baseline) to ≈ 0.165.

Main takeaway: Overall, these results indicate that survival-aware fine-tuning can outperform the baseline TabPFN in certain regimes—particularly for probabilistic calibration in low-data settings—but that the performance boost is not consistent across datasets or sample sizes, with both variants generally exhibiting similar performance.

7 Limitations

The observed gains are constrained by several aspects of our experimental design.

Fine-Tuning vs. Pre-Training: We fine-tune an existing TabPFN model rather than pre-training a survival Transformer from scratch. The mismatch between the original priors and survival-specific objectives likely limits improvements and contributes to the largely similar performance of the baseline and fine-tuned models. Consistent gains may require full survival-specific pre-training.

Inconsistent Fine-Tuning Gains: Fine-tuning improves performance only in specific low-data settings (e.g., SUPPORT2 at $n = 728$) and does not yield consistent benefits across datasets or metrics.

Synthetic Data Diversity: Despite their scale ($N = 10{,}000$), the synthetic datasets rely on simplified parametric assumptions that may not capture the heterogeneity and noise of real clinical data, limiting generalization.

Compute-Constrained Fine-Tuning Setup: Due to compute constraints, we fine-tune with only two ensemble members and cap the maximum number of samples per synthetic dataset at 1,000.

8 Conclusion

In this work, we investigate the applicability of general-purpose tabular Transformers to survival analysis. Our evaluation demonstrates that TabPFN possesses strong capabilities for time-to-event prediction, performing competitively with specialized deep learning models without any task-specific training. By fine-tuning the model on a large number of synthetic survival datasets, we successfully adapted its in-context learning mechanism to better handle censored data and hazard estimation.

Our results indicate that while the fine-tuned model maintained discriminative performance (C-index) comparable to the baseline, it achieved meaningful gains in probabilistic calibration (IBS) within low-data regimes. Specifically, the fine-tuned model reduced calibration error at the 25% scale on METABRIC and the 10% scale on SUPPORT2. These findings imply that while general-purpose models are robust starting points, fully bridging the gap to state-of-the-art survival performance requires addressing specific methodological and data-centric constraints.

Acknowledgments

We gratefully acknowledge Jiayi Cheng, Michal Mankowski, and Robert Steele of NYU Langone for their mentorship and valuable feedback during the development of this work. We also thank them for sharing preliminary work on in-context survival modeling.

References

[1] David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–220, 1972.

[2] Hemant Ishwaran and Udaya B. Kogalur. Consistency of random survival forests. *Statistics & Probability Letters*, 80(13–14):1056–1064, 2010. doi: 10.1016/j.spl.2010.02.020.

[3] Zifeng Wang. Survtrace: Transformers for survival analysis with competing events. *arXiv preprint*, 2021.

[4] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 2018.

[5] 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, pages 2314–2321, 2018.

[6] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint*, 2022.

[7] William A. Knaus, Frank E. Harrell, Joanne Lynn, Lee Goldman, Russell S. Phillips, and Alfred F. Connors. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122(3):191–203, 1995. doi: 10.7326/0003-4819-122-3-199502010-00007.

[8] Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012. doi: 10.1038/nature10983.

[9] Samuel Müller et al. Prior-data fitted networks for causal effect inference. *arXiv preprint*, 2023.

[10] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008. doi: 10.1214/08-AOAS169.

[11] Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

[12] 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. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5.

[13] Thomas A. Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

A Appendix

Table 1: Summary of evaluation datasets.

Characteristic	**METABRIC**	**SUPPORT2**
Samples	1,904	9,105
Features	9	57
Event rate	57.9%	68.0%
Censoring rate	42.1%	32.0%
Follow-up range	0–355 months	3–2,029 days
Median follow-up	119 months	546 days
Domain	Breast cancer	Critical care

Table 2: Key configuration ranges sampled for synthetic dataset generation

Config	**Min**	**Max**
# Rows	100	10,000
# Features	15	50
Survival time (METABRIC-aligned scale in months)	0	355
Survival time (SUPPORT2-aligned scale in days)	3	2,029
Survival time (medium scale)	10	500
Censoring rate	30%	45%

Table 3: Fine-tuning hyperparameters.

Hyperparameter	**Value**
Number of datasets (tasks)	10,000
Maximum samples per dataset	1,000
Epochs	10
Optimizer	Adam
Learning rate	5×10^{-5}
Gradient clipping	max_norm = 1.0
Ensemble members	2

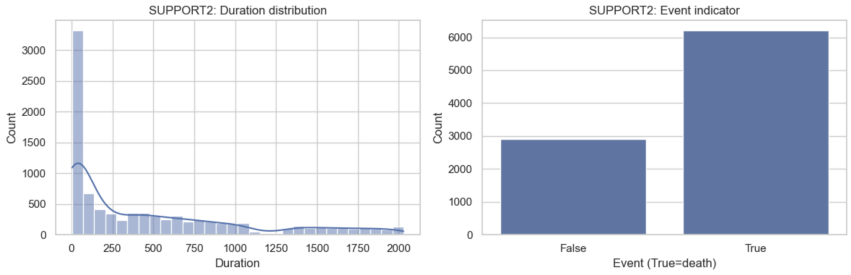

Figure 3: SUPPORT2 duration distribution (left) and event rate (right).

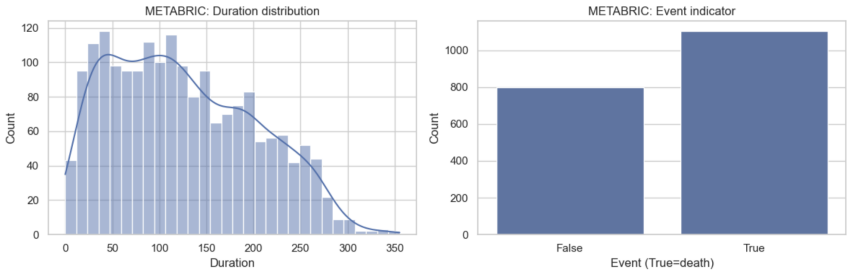

Figure 4: METABRIC duration distribution (left) and event rate (right).