

Deep Survival Analysis Frameworks for Personalized Prognosis Prediction



**UNIVERSITÀ
DI TORINO**

University of Turin, Italy
Department of Medical Sciences
PhD Program in Complex Systems for Quantitative Biomedicine
XXXVII cycle
Academic years: 2021/2024
Code of scientific discipline: PHYS-06/A (FIS-07)

Author: Cesare Rollo

Advisor: Prof. Piero Fariselli

Co-Advisor: Dr. Tiziana Sanavia

PhD coordinator: Prof. Francesca Cordero

Abstract

This thesis presents a series of novel computational frameworks designed to enhance survival analysis in biomedical research by integrating advanced artificial intelligence methods with privacy-preserving data sharing and mechanistic modeling.

First, the SYNDSURV framework is a synthetic data-based approach, enabling decentralized survival analysis without compromising patient sensitive data. By generating synthetic survival data at local nodes—combining Bayesian data synthesis with Accelerated Failure Time modeling—it facilitates centralized training of survival models while maintaining privacy constraints through differential privacy techniques.

Next, the VAE-Surv model introduces a deep-learning-based survival analysis approach that leverages Variational Autoencoders to extract latent representations from high-dimensional genetic and cytogenetic data. This representation, enriched with clinical features, is concatenated with a deep survival network that achieves high prognostic predictions and patient stratification, particularly in the context of Myelodysplastic Syndromes (MDS). This approach not only improves predictive accuracy but also enhances interpretability, bridging the gap between deep learning and clinical applicability.

Finally, this thesis explores a stochastic diffusion process approach, conceptualizing survival as a First Fitting Time (FHT) problem. Here, patient-specific parameters are inferred via neural networks, offering mechanistic insights into individual survival trajectories and capturing non-linear progression patterns that traditional survival models often overlook.

The experimental results across multiple datasets demonstrate that these frameworks consistently outperform traditional survival analysis methods, providing interpretable and clinically actionable insights. By addressing key challenges in federated learning, data privacy, and multimodal integration, this work contributes to the development of more effective, personalized strategies for managing rare and complex diseases, especially in the context of prognosis prediction.

During my PhD, I have collaborated with the Center for Health Data Science, University of Copenhagen, under the supervision of Professor Anders Krogh. I also thank Prof. Matteo Della Porta (Humanitas University) and Prof. Gastone Castellani (University of Bologna), my external supervisors within the GenoMed4All consortium, for their guidance and continuous support throughout my PhD. Their expertise and collaboration have significantly enriched many of my research projects.

List of publications

I, Cesare Rollo, confirm that the work presented in this thesis is my own. Where external contributions have been provided, I confirm that this has been indicated in the thesis.

Publications present in this thesis:

- Cesare Rollo et al. “SYNDSURV: A simple framework for survival analysis with data distributed across multiple institutions”. In: *Computers in Biology and Medicine* 172 (2024), p. 108288
- Cesare Rollo et al. “VAE-Surv: A novel approach for genetic-based clustering and prognosis prediction in myelodysplastic syndromes”. In: *Computer Methods and Programs in Biomedicine* (2025), p. 108605
- Saverio D’Amico et al. “MOSAIC: An Artificial Intelligence–Based Framework for Multimodal Analysis, Classification, and Personalized Prognostic Assessment in Rare Cancers”. In: *JCO Clinical Cancer Informatics* 8 (2024), e2400008 (Co-first author)
- Cesare Rollo et al. “Influence of model structures on predictors of protein stability changes from single-point mutations”. In: *Genes* 14.12 (2023), p. 2228
- Oussama Harrak et al. “Federated AdaBoost for Survival Analysis”. In: *Proceedings of the ECML-PKDD Workshops*. Springer. 2024, pp. 1–9

Publications not included in this thesis:

- Cesare Rollo et al. “Communities, gateways, and bridges: Measuring attention flow in the reddit political sphere”. In: *International Conference on Social Informatics*. Springer. 2022, pp. 3–19
- Corrado Pancotti et al. “MUSE-XAE: MUtational Signature Extraction with eXplainable AutoEncoder enhances tumour types classification”. In: *Bioinformatics* 40.5 (2024), btae320
- Francesco Casadei et al. “Survival Model Optimization via Federated Learning: A Study Combining Simulations and Experiments”. In: *2024 IEEE International Conference on Big Data (BigData)*. IEEE. 2024, pp. 7658–7667
- Corrado Pancotti et al. “Deep learning methods to predict amyotrophic lateral sclerosis disease progression”. In: *Scientific reports* 12.1 (2022), p. 13738
- Sara Ferrini et al. “A novel machine learning-based web application for field identification of infectious and inflammatory disorders of the central nervous

system in cattle”. In: *Journal of Veterinary Internal Medicine* 37.2 (2023), pp. 766–773

- Corrado Pancotti et al. “Unravelling the instability of mutational signatures extraction via archetypal analysis”. In: *Frontiers in Genetics* 13 (2023), p. 1049501
- Federico Giacobbe et al. “Impact of statin adherence and interruption within 6 months after ST-segment elevation myocardial infarction (STEMI): Results from the real-world regional registry FAST-STEMI”. in: *International Journal of Cardiology* 405 (2024), p. 131933
- Corrado Pancotti et al. “Multi-event survival prediction for amyotrophic lateral sclerosis”. In: *CEUR workshop proceedings*. Vol. 3180. Faggioli G Ferro N, Hanbury A, Potthast M. 2022, pp. 1269–1276
- Giuseppe Giannino et al. “Impact of adherence to beta-blockers in all-comers ST-segment elevation myocardial infarction (STEMI) patients and according to left ventricular ejection fraction (LVEF) at discharge: results from the real-world registry FAST-STEMI”. in: *Journal of Cardiovascular Pharmacology* (2022), pp. 10–1097
- Giovanni Birolo et al. *SurvHive: a package to consistently access multiple survival-analysis packages*. 2025. arXiv: 2502.02223 [q-bio.QM]. URL: <https://arxiv.org/abs/2502.02223>

Manuscripts accepted for publication:

- Francesco Codicè et al. “The Specification Game: Rethinking the Evaluation of Drug Response Prediction for Precision Oncology”. In: *bioRxiv* (2024), pp. 2024–10 (accepted for publication in *Journal of Cheminformatics*)

List of Tables

3.1	Summary of survival datasets.	24
4.1	Comparison of models' performance in terms of median C-Index and Confidence Intervals, evaluated in the 10-folds CV splits.	34
4.2	Performance comparison including K-modes clustering on the input space, Survival Cluster Analysis, and K-means clustering on the latent representations learned by VAE-Surv	36
4.3	Logistic Regression average performance for the cluster assignment task on the 100-times repeated internal test sets.	39
B.1	Description of GenoMed4All and IWG-PM Cohorts. Comparisons were performed using Wilcoxon Rank Sum and Chi-Square tests for continuous and categorical variables, respectively.	63
B.2	Optimal hyperparameters of the VAE-Surv Neural Network	65

List of Figures

2.1	Censoring	4
2.2	Schematic representation of a FL framework, from [<empty citation>].	16
3.1	Schematic representation of the SYNDSURV framework. In this example, there are two participating institutions that cooperate to create a global model. Each one trains a local AFT survival algorithm using only its own patients' data. Then, synthetic samples are generated through a Bayesian generative network built with patient covariates at single institution level, and local models are used to predict survival times for synthetic patients. Finally, all synthetic data are transferred from the data centres to a central aggregation server that trains the global model. Its performance is evaluated on an external test set composed of real patients from each institution.	21
3.2	Test set C-Index vs. number of generated patients per node, for the CoxPH model trained on the synthetic data. Reported values represent the mean over 10 different test sets; error bars indicate the errors of the mean.	26
3.3	Test set C-Index vs. number of generated patients per node, for the DeepHit model trained on the synthetic data. Reported values represent the mean over 10 different test sets; error bars indicate the errors of the mean.	26
3.4	Test set C-Index vs. number of generated patients per node, for the DeepHit model trained on the synthetic data, using a differential privacy setting with $\epsilon = 0.1$. Reported values represent the mean over 10 different test sets; error bars indicate the errors of the mean.	27
4.1	Schematic representation of the VAE-Surv computational framework. Genetic and cytogenetic markers are input as binary vectors into a Variational Autoencoder (VAE) for dimensionality reduction. The latent feature space, obtained by the VAE, is concatenated with clinical variables to form an enriched feature set. This combined feature set is then fed into a parallel Deep Cox Proportional Hazards Neural Network (DeepSurv) for survival risk prediction. The dimensionality of the latent space z is optimized as a hyperparameter of the model.	31
4.2	Kaplan-Meier curves stratified by risk groups based on percentiles. . . .	35
4.3	Comparison of VAE-Surv and CoxPH models using PCA-reduced genetic features and clinical covariates.	35

4.4	Left panel: t-SNE representation of the patients' genetic and cytogenetic latent space learned from VAE, coloured according to the cluster assignment. Right panel: Kaplan-Meier survival curves stratified by the identified cluster. The size of each group is specified in the legend. . . .	37
4.5	Clustermap showing the distribution of WHO 2016-defined MDS subtypes (y-axis) within each of the nine clusters identified (x-axis). Each column sums to 1. The colour intensity correlates with the percentage of patients in a given cluster belonging to a specific MDS subtype. . . .	37
4.6	Clustermaps displaying the frequency of gene mutations and cytogenetic alterations within each identified cluster. The colour intensity corresponds to the prevalence of each genetic feature in the clusters, providing insights into their mutational landscape. Panels a) and b) show the results from clustering on the VAE latent space and on the input space, respectively.	38
5.1	Mean C-Index for the 5-fold cross validation, for different loss function, model and datasets.	47
5.2	Mean integrated Brier score for the 5-fold cross validation, for different loss function, model and datasets.	48
5.3	x_0 and D learned distributions for <i>support2</i> dataset when using Brier loss and Levy model.	49
A.1	CoxPH coefficients ranked by absolute value. The weights related to the classes of the categorical features have been aggregated by summing their absolute values.	58
A.2	Synthetic times distribution compared to the real event times, for the uncensored patients.	59
A.3	Test set time-dependent AUC vs. number of generated patients per node, for the CoxPH model trained on the synthetic data.	60
A.4	Test set time-dependent AUC vs. number of generated patients per node, for the DeepHit model trained on the synthetic data.	61
A.5	Test set time-dependent auc vs. number of generated patients per node, for the DeepHit model trained on the synthetic data, using a differential privacy setting with $\epsilon = 0.1$	61
B.1	VAE's latent representation: Plot of the quality metric Q_k against the number of clusters k , determined through K-means clustering of the VAE's latent representation. The peak indicates the optimal number of clusters, $k = 9$	65
B.2	Genetic and cytogenetic input data Plot of the quality metric Q_k against the number of clusters k , determined through the Survival Cluster Analysis model. The peak indicates the optimal number of clusters, $k = 3$	66
B.3	Survival Cluster Analysis Plot of the quality metric Q_k against the number of clusters k , determined through K-modes clustering genetic and cytogenetic input data. The peak indicates the optimal number of clusters, $k = 3$	66

B.4	Left panel: t-SNE representation of the VAE-Surv latent space when using the KL-term during the training. Right panel: Kaplan-Meier survival curves stratified by the identified cluster. The size of each group is specified in the legend.	67
B.5	Left panel: t-SNE representation of the patients' genetic and cytogenetic input data, coloured according to the cluster assignment (K-modes). Right panel: Kaplan-Meier survival curves stratified by the identified cluster. The size of each group is specified in the legend.	67
B.6	Mutations count per cluster.	68
B.7	Median values of relevant clinical attributes per cluster. 95% Confidence Intervals are represented by the error bars.	68
B.8	Sankey plot showing the flow of MDS subtypes across the identified clusters. Each strand represents the proportion of a specific MDS subtype present in each cluster, providing a comprehensive view of how the MDS landscape is distributed across the clusters.	69
B.9	Comparison of the C-Index results obtained with the considered models for the Validation cohort, using the Bootstrap with resample technique.	70
B.10	t-SNE visualization of the latent space for the training and validation cohort.	70
B.11	Top-15 ranked features by shapley values in the training cohort.	71
B.12	Top-15 ranked features by shapley values in the validation cohort.	71
C.1	C-Index and Brier score distributions for datasets stratified by sample size (below vs. above the median).	72
C.2	C-Index and Brier score distributions for datasets stratified by number of features (below vs. above the median).	73

Contents

1	Introduction	2
2	Background	4
2.1	Survival Analysis	4
2.1.1	Building blocks of Survival Analysis	5
2.1.2	Likelihood-based Survival Models	6
2.1.3	Cox Proportional Hazards Model	8
2.1.4	Deep Cox Proportional Hazards Methods	10
2.1.5	Other nonlinear survival models	10
2.1.6	Metrics for evaluating survival models	11
2.2	Privacy-preserving Distributed Learning	15
2.2.1	Federated Learning	15
2.2.2	Distributed Survival Analysis	17
3	SYNDSURV: A Simple Framework for Survival Analysis with Data Distributed Across Multiple Institutions	19
3.1	Introduction	19
3.2	Methodology	20
3.2.1	Model architecture	20
3.2.2	Data Synthesizer	20
3.2.3	Synthetic labels generation	22
3.2.4	Centralised learning and validation strategy	23
3.3	Results	24
3.3.1	Survival Datasets	24
3.3.2	Performance comparison	25
3.3.3	Differential Privacy	27
3.4	Conclusion	28
3.4.1	Limitations and future works	28
4	VAE-Surv: A Novel Approach for Genetic-Based Clustering and Prognosis Prediction in Myelodysplastic Syndromes	29
4.1	Introduction	29
4.2	Materials and Methods	31
4.2.1	Datasets	31
4.2.2	Model Architecture	32
4.2.3	Post-training analysis	33
4.3	Results	34
4.3.1	VAE-Surv prediction performance	34

4.3.2	VAE-Surv latent representation and clustering performance . . .	35
4.3.3	Biological Consistency of the MDS clusters	36
4.4	Validation in the external cohort	38
4.5	Discussion	39
4.6	Conclusion and future work	40
5	Modelling Survival as a Stochastic Diffusion Process	42
5.1	Overview of Wiener processes	42
5.2	First Hitting Time for Survival	44
5.3	Parameter estimation for FHT Survival models with Neural Networks .	45
5.4	Experimental setting and models comparison	46
5.5	Discussion	49
6	Related Projects	51
6.1	MOSAIC: An Artificial Intelligence-Based Framework for Classification and Prognostic Assessment in Rare Cancers	51
6.2	Federated AdaBoost for Survival Analysis	53
6.3	Influence of Model Structures on Predictors of Protein Stability Changes from Single-Point Mutations	54
7	Conclusion	55
A	SYNDSURV Supplementary Materials	57
A.1	Datasets description	57
A.2	Features selection for SUPPORT dataset	58
A.3	Synthetic times distributions	59
A.4	Time-dependent AUC results	60
B	VAE-Surv Supplementary Materials	62
B.1	Dataset description	62
B.2	Optimal hyperparameters and VAE-Surv architecture	64
B.3	Optimal K-means clustering	65
B.4	Clustering from VAE-Surv trained with KL-term	66
B.5	Clustering on Input Data	66
B.6	Clustering on VAE latent space	67
B.6.1	Cluster-specific analysis and comparison with WHO 2016 classi- fication	67
B.7	VAE latent space: validation in the external cohort	70
B.8	VAE-Surv: Model Explanation with SHAP	71
C	Modelling Survival as a Stochastic Diffusion Process - Figures	72

Chapter 1

Introduction

Advancements in computational biology and artificial intelligence (AI) have revolutionized the field of biomedicine, enabling more precise and personalized approaches to understanding complex diseases. Within the realm of rare cancers and hematologic disorders, deep learning and survival analysis have emerged as critical tools for unraveling intricate biological mechanisms, improving prognostication, and guiding clinical decision-making. Myelodysplastic Syndromes (MDS), a clonal hematopoietic disorder characterized by substantial genomic and clinical heterogeneity, exemplifies the challenges posed by such diseases. Despite significant progress, the current classification systems for MDS, such as the WHO classification, remain limited in their ability to fully capture the genomic landscape, underscoring the need for innovative approaches to patient stratification and prognosis prediction.

This thesis is part of the GenoMed4All consortium, a European initiative focused on advancing personalized medicine in haematological diseases through the integration of artificial intelligence (AI) and the pooling of genomic and clinical data. The project aims to enhance the diagnosis, prognosis, and treatment of both common and rare blood disorders by leveraging AI algorithms and fostering cross-border data sharing. A key component of GenoMed4All is the development of a secure and privacy-respecting data sharing platform utilizing Federated Learning. This approach allows for the collaborative analysis of data across multiple institutions without the need to centralize sensitive patient information, thereby maintaining data privacy while enabling large-scale data analysis. The consortium's efforts are concentrated on three specific haematological conditions:

- Myelodysplastic Syndromes (MDS): A group of diverse bone marrow disorders characterized by ineffective hematopoiesis leading to blood cell abnormalities.
- Multiple Myeloma: A cancer of plasma cells that accumulate in the bone marrow, leading to bone damage and impaired immune function.
- Sickle Cell Disease (SCD): A genetic disorder causing red blood cells to assume a sickle shape, leading to various complications including pain and organ damage.

By applying AI-driven analysis to pooled genomic and clinical data, GenoMed4All aims to develop predictive models and personalized treatment strategies for these diseases. The project also emphasizes the importance of interoperability and standardization in data sharing to facilitate collaboration across European healthcare

institutions. The consortium comprises 23 partners, including clinical experts from the European Reference Network on Rare Hematological Diseases (ERN-EuroBloodNet), who contribute their specialized knowledge to the project’s objectives.

My research contributes to the consortium’s efforts by developing novel deep learning frameworks and multimodal survival analysis techniques, addressing key challenges in integrating high-dimensional, multimodal datasets while maintaining robust interpretability and clinical utility. In the context of MDS, this thesis investigates the application of advanced AI methods to enhance patient stratification and prognostic accuracy. Techniques such as Variational Autoencoders (VAE), federated learning, and explainable AI were employed to integrate genomic, clinical, and survival data, forming the backbone of innovative frameworks like MOSAIC [3], SYNDSURV [1], and VAE-Surv[2]. These approaches not only outperform traditional statistical models but also offer scalable and privacy-preserving solutions for multicenter collaborations, as demonstrated through GenoMed4All’s federated infrastructure.

The following chapters will delve into these contributions, outlining the methodologies, validation efforts, and their implications for the broader field of precision medicine. By bridging the gap between computational advances and clinical applications, this work seeks to establish a foundation for more effective, personalized strategies in the management of rare and complex diseases.

Chapter 2

Background

2.1 Survival Analysis

Survival analysis is a branch of statistics and machine learning designed to analyze and predict the time until the occurrence of an event of interest, such as death, disease progression, or machine failure. It is a fundamental tool for many applications in biomedicine, particularly in oncology and other fields where time-to-event data is critical for decision-making.

Given a group of individuals susceptible to experiencing an event of interest, survival analysis techniques seek to infer the probability distribution for the time of that event for each individual. Obtaining these times-to-event requires patients to be followed in long observational studies, which typically last years. Due to this long duration, patients frequently drop out from studies before any event occurs. In this case, one can only know that the event did not occur before the dropping time, which is known. This partial knowledge is known in survival analysis as right censoring.

For example, consider a study where some patients drop out before experiencing the event, as illustrated in Figure 2.1. Here for patients A, C, and E it is unknown whether they did or did not experience an event after termination of the study, so they are considered *censored*. The only available information is that they are event-free up to the last follow-up. Therefore, the exact time of an event could only be recorded for patients B and D, which are considered as *uncensored*.

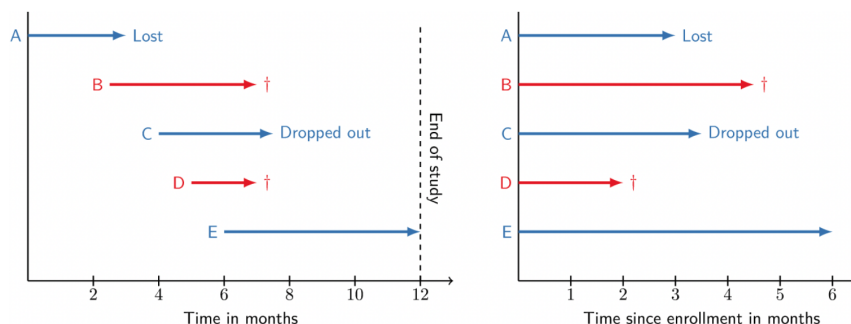


Figure 2.1: Censoring

2.1.1 Building blocks of Survival Analysis

The goal is to model the probability distribution of the actual event time $T_i \in \mathbb{R}_+$ for each individual $i \in [N] = 1, \dots, N$. The core functions in Survival Analysis are the following:

Survival Function : It represents the probability that an individual does not experience the event of interest beyond a certain time t :

$$S(t) = \Pr(T > t) \quad \text{with} \quad S(\infty) = 0 \quad \text{and} \quad S(0) = 1 \quad (2.1)$$

$S(t)$ is a monotonic decreasing function.

Cumulative Failure : It expresses the probability that the event occurs by time t and can be interpreted as the complementary of the Survival Function.

$$F(t) = 1 - S(t) \quad (2.2)$$

Failure (or density) : It is the probability density function of the event times, thus quantifying how likely the event is to occur at an exact moment in time. By definition:

$$f(t) = F'(t) \quad (2.3)$$

Hazard function : It is the approximate probability that an event occurs in the time interval $[t; t + \Delta t]$, under the condition that the individual would remain event-free up to time t :

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \quad (2.4)$$

The conditional probability in the numerator can be rewritten using the definition of conditional probability:

$$P(t \leq T < t + \Delta t \mid T \geq t) = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)}.$$

- $P(t \leq T < t + \Delta t)$ is the probability that the event occurs in $[t, t + \Delta t)$, which is given by the probability density function $f(t) \cdot \Delta t$;
- $P(T \geq t) = S(t)$ is the survival function.

Substituting these expressions into the hazard function:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\frac{f(t) \cdot \Delta t}{S(t)}}{\Delta t} = \frac{f(t)}{S(t)}.$$

This relationship shows that the hazard function $h(t)$ can be derived from the survival function $S(t)$, and vice versa, using the following relationships:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log S(t) \quad (2.5)$$

$$S(t) = e^{-\int_0^t h(\tau) d\tau} = e^{-H(t)} \quad (2.6)$$

Nonparametric Estimation

A straightforward way to estimate the survival function from observed (possibly censored) data, without assuming any specific parametric form, is via the *Kaplan–Meier* (product-limit) estimator [17]:

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{r_i}\right), \quad (2.7)$$

where $t_{(i)}$ are the distinct, ordered event times, d_i is the number of observed events at $t_{(i)}$, and r_i is the number of individuals at risk just before $t_{(i)}$. The estimator $\hat{S}(t)$ takes the form of a step function, remaining constant between observed event times and decreasing when an event occurs. It imposes only mild assumptions (primarily noninformative censoring) and is widely used for a first, model-free look at the data’s survival pattern.

- **Advantages:** Minimal assumptions and easy visual interpretation. Researchers often use Kaplan–Meier curves for comparing different groups (e.g., treatment arms) through *log-rank tests*.
- **Limitations:** Cannot directly incorporate covariates. To examine the effect of covariates, one typically resorts to parametric or semiparametric (e.g., Cox) modeling.

An alternative *nonparametric* estimator is the *Nelson–Aalen* estimator of the cumulative hazard [18]:

$$\hat{H}(t) = \sum_{i:t_{(i)} \leq t} \frac{d_i}{r_i}, \quad \text{with} \quad \hat{S}(t) = \exp(-\hat{H}(t)). \quad (2.8)$$

This approach is particularly convenient for large-sample hazard estimation.

2.1.2 Likelihood-based Survival Models

In many applications, one has *covariates* (also called *features* or *predictors*) that may influence the time to an event. We therefore seek to model the time-to-event conditional on these covariates, denoted by $\mathbf{x}$. Concretely, we replace

$$S(t) \longrightarrow S(t \mid \mathbf{x}),$$

where $S(t \mid \mathbf{x})$ is the conditional survival function for individuals whose covariates equal $\mathbf{x}$. Similarly, we define $h(t \mid \mathbf{x})$ and $f(t \mid \mathbf{x})$ for the hazard and density, respectively. When covariates are present, one typically moves beyond pure nonparametric approaches and adopts *parametric* or *semiparametric* models, defining how $\mathbf{x}$ affects the survival distribution.

Full Likelihood

While nonparametric methods such as Kaplan–Meier and Nelson–Aalen are powerful for exploring survival data, they do not incorporate covariates directly or provide a succinct functional form for $S(t)$. An alternative is to assume a *parametric* model for the time-to-event distribution. Common choices include the exponential, Weibull,

log-normal, or gamma distributions for $f(t \mid \mathbf{x})$. Under a parametric model, each individual's event time T_i (possibly conditioned on covariates $\mathbf{x}_i$) is characterized by parameters θ (e.g., scale or shape).

The full likelihood function, accounting for both uncensored and censored observations, is:

$$\begin{aligned}\mathcal{L}(D) &= \prod_{i \in \text{uncens}} P(T = T_i \mid \theta, X) \prod_{i \in \text{cens}} P(T > T_i \mid \theta, X) \\ &= \prod_{i \in \text{uncens}} f(T_i \mid \theta, X) \prod_{i \in \text{cens}} S(T_i \mid \theta, X)\end{aligned}\tag{2.9}$$

Taking the logarithm:

$$-\log \mathcal{L}(D) = - \left(\sum_{i \in \text{uncens}} \log f(T_i \mid \theta, X) + \sum_{i \in \text{cens}} \log S(T_i \mid \theta, X) \right)\tag{2.10}$$

The full likelihood can also be expressed in terms of the hazard function. From Eq. 2.9:

$$\begin{aligned}\mathcal{L}(D) &= \prod_{i \in \text{uncens}} f(T_i \mid \theta, X) \prod_{i \in \text{cens}} S(T_i \mid \theta, X) \\ &= \prod_{i \in \text{uncens}} S(t_i \mid \theta, X) h(t_i \mid \theta, X) \prod_{i \in \text{cens}} S(t_i \mid \theta, X) \\ &= \prod_{i=1}^N h(t_i \mid X)^{\delta_i} S(t_i \mid \theta, X) \\ &= \prod_{i=1}^N h(t_i \mid X)^{\delta_i} e^{-\int_0^{t_i} h(\tau \mid x_i) d\tau}\end{aligned}\tag{2.11}$$

Hence, maximizing $L(\theta)$, or its logarithm, with respect to θ yields *maximum likelihood estimates* for the model parameters.

In writing down the likelihood (2.9), two key assumptions are implicitly made:

1. **Independent Samples.** Each individual's time-to-event T_i (conditional on their covariates $\mathbf{x}_i$) is assumed independent of others.
2. **Noninformative Censoring.** The reasons for censoring are assumed unrelated to the individual's underlying risk of the event, except through known covariates. Concretely, an individual leaving the study early (or being lost to follow-up) does not depend on their unobserved future event time T_i .

Under these assumptions, maximum likelihood methods can consistently estimate how covariates $\mathbf{x}_i$ influence the parameters of the chosen parametric distribution, allowing for predictions of individual survival curves or hazard rates.

Partial Likelihood

An alternative estimation perspective uses a *partial likelihood* that focuses on the set of observed event times. In essence, for each event time t_i of an *uncensored* individual, we consider the probability that this specific individual (among those still at risk at t_i) experiences the event. Formally, if $\delta_i = 1$ indicates an observed event, then one obtains:

$$\begin{aligned}\mathcal{PL}_i &= \mathbb{P}(i \text{ has event at } t_i \mid \text{one event occurs at } t_i) \\ &= \mathbb{P}\left(\delta_i = 1, t = T_i \mid \sum_j \delta_j = 1, t = T_i\right)\end{aligned}\tag{2.12}$$

Then, if we consider:

- event A : i has event in t_i
- event B : there is one event in t_i
- event C : i is event-free up to time t_i

It leads to:

$$\begin{aligned}\mathcal{PL}_i &= \frac{\mathbb{P}(B \mid A) \mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A, C) \mathbb{P}(A \mid C)}{\mathbb{P}(B \mid C)} = \frac{\mathbb{P}(A \mid C)}{\mathbb{P}(B \mid C)} \\ &= \frac{h(t_i \mid x_i)}{\sum_{j \in \mathcal{R}_i} h(t_j \mid x_j)}\end{aligned}\tag{2.13}$$

Hence, the final Partial Likelihood for the entire dataset is:

$$\mathcal{PL}(D) = \prod_{i: \delta_i = 1} \frac{h(t_i \mid x_i)}{\sum_{j \in \mathcal{R}_i} h(t_j \mid x_j)}\tag{2.14}$$

where $\mathcal{R}(t_i)$ is the *risk set*, i.e. the set of individuals who are still at risk (have not experienced the event of interest) just before time t_i .

Comparing Eq. 2.9 with Eq. 2.14, we understand why we talk about *Partial Likelihood*:

- Eq. 2.9 includes all history of hazards up to time t_i through the integral $\int_0^{t_i}$; Eq. 2.14 only considers hazards at the event times;
- Eq. 2.9 directly uses censored samples, exploiting all the available information; Eq. 2.14 uses censored information only partially through the risk set $\mathcal{R}_i$;
- Eq. 2.14 is derived from Eq. 2.12 which is conditioned on the observed event times, treating them as fixed rather than random. Thus it is not a true probability density for the full data;
- Eq. 2.14 cannot be factorized into independent components.

This formulation leverages the *ordering* of observed events rather than the full distribution of survival times. Consequently, it is not a complete likelihood because it omits integral components of the time-to-event process (specifically, how the baseline hazard accumulates over time).

2.1.3 Cox Proportional Hazards Model

A widely adopted *semiparametric* approach for modeling how covariates affect the time-to-event distribution is the Cox Proportional Hazards (PH) model [19]. Unlike parametric methods, it does not require specifying a functional form for the baseline

hazard $h_0(t)$. Instead, it assumes that the hazard rate for an individual with covariate vector $\mathbf{x}$ is proportional to a baseline hazard function that can be factorized as:

$$h(t | \mathbf{x}) = h_0(t) \exp\{\boldsymbol{\beta}^\top \mathbf{x}\}, \quad S(t | \mathbf{x}) = [S_0(t)]^{\exp\{\boldsymbol{\beta}^\top \mathbf{x}\}}. \quad (2.15)$$

This *proportional hazards* assumption implies that the ratio of hazards for any two individuals is *constant* over time, depending only on their covariates and not on t itself.

Cox Partial Likelihood. Under the proportional hazards framework, the hazard at each observed event time t_i can be decomposed into a baseline term $h_0(t_i)$ and a covariate-dependent term $\exp\{\boldsymbol{\beta}^\top \mathbf{x}_i\}$. Substituting this into the partial likelihood (recall Eq. 2.14):

$$\mathcal{PL}(\boldsymbol{\beta}) = \prod_{i: \delta_i=1} \frac{h(t_i | \mathbf{x}_i)}{\sum_{j \in \mathcal{R}(t_i)} h(t_i | \mathbf{x}_j)} = \prod_{i: \delta_i=1} \frac{h_0(t_i) e^{\boldsymbol{\beta}^\top \mathbf{x}_i}}{\sum_{j \in \mathcal{R}(t_i)} h_0(t_i) e^{\boldsymbol{\beta}^\top \mathbf{x}_j}}. \quad (2.16)$$

Since $h_0(t_i)$ is common to both numerator and denominator, it cancels out, leaving

$$\mathcal{PL}(\boldsymbol{\beta}) = \prod_{i: \delta_i=1} \frac{e^{\boldsymbol{\beta}^\top \mathbf{x}_i}}{\sum_{j \in \mathcal{R}(t_i)} e^{\boldsymbol{\beta}^\top \mathbf{x}_j}} \quad \text{with} \quad \mathcal{R}(t_i) = \{j | t_j \geq t_i\}. \quad (2.17)$$

Maximizing $\mathcal{PL}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is typically done by taking the negative log:

$$-\log \mathcal{PL}(\boldsymbol{\beta}) = \sum_{i=1}^N \delta_i \left[\boldsymbol{\beta}^\top \mathbf{x}_i - \log \left(\sum_{j \in \mathcal{R}(t_i)} e^{\boldsymbol{\beta}^\top \mathbf{x}_j} \right) \right], \quad (2.18)$$

and applying numerical optimization. The resulting $\hat{\boldsymbol{\beta}}$ parameters capture how covariates shift the hazard rate relative to the *unspecified* baseline hazard $h_0(t)$.

Breslow Estimator for the Baseline

Although the baseline hazard $h_0(t)$ is not directly involved in the partial likelihood, it can still be estimated *a posteriori* once $\boldsymbol{\beta}$ is known. The cumulative baseline hazard $H_0(t)$ is often estimated via the *Breslow estimator* [20]:

$$H_0(t) = \sum_{i: t_i \leq t} \frac{d(t_i)}{\sum_{j \in \mathcal{R}(t_i)} e^{\hat{\boldsymbol{\beta}}^\top \mathbf{x}_j}}, \quad (2.19)$$

where $d(t_i)$ is the number of events occurring at time t_i . Integrating or differentiating $H_0(t)$ yields an estimate of $h_0(t)$. Consequently, one obtains

$$\hat{S}_0(t) = \exp\{-H_0(t)\} \quad \longrightarrow \quad \hat{S}(t | \mathbf{x}) = [\hat{S}_0(t)]^{\exp\{\hat{\boldsymbol{\beta}}^\top \mathbf{x}\}}.$$

Thus, the Cox model provides an elegant compromise between parametric and fully nonparametric methods: it avoids specifying a distribution for when events occur (the shape of h_0) but captures how covariates affect the hazard ratio.

2.1.4 Deep Cox Proportional Hazards Methods

The Cox PH model is semiparametric, as it does not specify a particular form for the baseline hazard $h_0(t)$ but assumes a log-linear link $\log h(t \mid \mathbf{x}) = \log h_0(t) + \boldsymbol{\beta}^\top \mathbf{x}$. Although widely used, this linear assumption may be too restrictive for high-dimensional or highly nonlinear data.

An early approach to relaxing the linearity assumption while retaining Cox’s partial-likelihood ideas was proposed by Faraggi and Simon [21]. They replaced the linear term $\boldsymbol{\beta}^\top \mathbf{x}$ with a multi-layer feedforward neural network $\psi(\mathbf{x}; \theta_{\text{NN}})$, obtaining

$$\log h(t \mid \mathbf{x}) = \log h_0(t) + \psi(\mathbf{x}; \theta_{\text{NN}}).$$

Maximizing the Cox partial likelihood (Eq. 2.18) now proceeds by adjusting the neural network’s parameters θ_{NN} to best separate at-risk individuals at each event time. Conceptually, this transforms the *linear predictor* of classical Cox into a flexible *nonlinear predictor*.

Building on the Faraggi–Simon framework, Katzman et al. [22] introduced *DeepSurv*, which combines modern deep-learning practices (e.g., mini-batch gradient descent, regularization, and deeper architectures) to further improve performance. In essence, these neural network implementations retain the crucial ordering-based loss from Cox and the proportional hazards assumption, while modeling the *risk score* with a deep network. Formally, if

$$r_i = \psi(\mathbf{x}_i; \theta_{\text{NN}}),$$

then the negative log partial likelihood (cf. Eq. 2.18) becomes

$$-\log \mathcal{PL}(\theta_{\text{NN}}) = \sum_{i=1}^N \delta_i \left[r_i - \log \left(\sum_{j \in \mathcal{R}(t_i)} e^{r_j} \right) \right]. \quad (2.20)$$

Eq. (2.20) is treated as a loss function and will be minimized with respect to θ_{NN} via stochastic gradient descent.

While DeepSurv and similar approaches retain Cox’s partial-likelihood framework, numerous other recent deep survival methods go beyond the proportional hazards assumption or adopt different objective functions. For example, DeepHit [23] formulates survival in discrete time and optimizes a ranking loss to simultaneously handle competing risks. Deep Survival Machines [24] embed a mixture of parametric survival distributions into a deep neural architecture, allowing for multi-modal event-time patterns and greater flexibility in high-dimensional datasets. Meanwhile, SurvTrace [25] integrates transformer-based encoding of patient covariates for enhanced feature representation. Each of these methodologies aims to capture more flexible or multi-modal event-time distributions, making them especially suitable for the complex, high-dimensional datasets frequently encountered in modern biomedical research.

2.1.5 Other nonlinear survival models

Accelerated Failure Time

The Accelerated Failure Time (AFT) model [26] is another well-known method for survival analysis. Since CoxPH returns an estimate for the risk associated with each

patient and predicts the baseline function in a separate step (usually through non-parametric approaches like the Breslow’s estimator [20]), it does not yield directly a usable prediction of the time-to-event $\hat{t}$.

With the AFT model, it is possible to predict unknown labels using only the fitted parameters and a feature vector. This is done by modeling the survival time T as:

$$\ln T = \langle \beta, \mathbf{x} \rangle + \sigma Z$$

where Z is a random variable of a known probability distribution and σ is a scale parameter. Then, nonlinear methods as XGBoost [26] can be naturally embedded in this formulation by replacing the linear term with the output from the decision tree ensemble $\mathcal{T}(\mathbf{x})$ [27]:

$$\ln T = \mathcal{T}(\mathbf{x}) + \sigma Z$$

Thus, the model is optimised by maximising a revised likelihood function $\mathcal{L}_{AFT}(t, \mathcal{T}(\mathbf{x}))$, which properly takes into consideration censored and uncensored labels. The functional form of the likelihood will depend on the chosen distribution for the random variable Z . The most commonly used are logistic, normal, and extreme distributions.

Random Survival Forest

Random Survival Forests (RSF) [28] is an ensemble learning method adapted from the traditional Random Forest algorithm for handling censored survival data. It extends the idea of decision trees to survival analysis by using a set of trees to model time-to-event data while accounting for censoring.

At its core, RSF operates as an ensemble of decision trees, each trained on a different bootstrap sample of the data. Instead of using classification or regression-based splitting criteria, RSF employs survival-specific rules to determine the best way to partition the data at each node. The most common approach relies on the log-rank test, a statistical method that maximizes the difference in survival distributions between child nodes. This ensures that each split effectively separates groups of individuals with different survival patterns.

One of the key strengths of RSF is its ability to handle complex, nonlinear relationships between variables without making strong assumptions about the underlying survival distribution. Unlike the Cox Proportional Hazards model, which assumes a log-linear relationship between covariates and hazard risk, RSF is fully nonparametric. This makes it particularly useful for biomedical applications where interactions between genomic, clinical, and demographic factors may not follow simple patterns.

After constructing a large number of survival trees, RSF aggregates the results across the ensemble to produce a cumulative hazard function for each individual. The survival probability at any given time is then obtained by exponentiating the negative cumulative hazard. This ensemble approach improves predictive accuracy and robustness by reducing the variance associated with individual trees.

2.1.6 Metrics for evaluating survival models

Evaluating the performance of survival models requires specialized metrics that account for censored data and the time-dependent nature of survival outcomes. Unlike

traditional regression or classification models, where performance can be assessed using straightforward measures like mean squared error or accuracy, survival models must be evaluated based on their ability to correctly rank survival times and provide well-calibrated probability estimates over time. Two of the most commonly used metrics in this domain are the *Concordance Index* (C-Index) and the *Brier Score*, along with its integrated version.

Concordance Index

The Concordance Index (C-Index) [29] is a widely used measure for assessing the discriminatory power of a survival model, quantifying how well the model ranks individuals in terms of their survival risk. It is conceptually similar to the area under the ROC curve (AUC) used in classification tasks but adapted for survival analysis.

Mathematically, the C-Index is defined as the proportion of all comparable pairs of individuals for which the model correctly predicts the order of survival times. A pair (i, j) is considered comparable if:

- The observed survival time T_i is less than T_j , and
- Individual i has experienced the event (i.e., is uncensored).

If the model assigns a higher risk (i.e., a lower predicted survival probability) to the individual with the shorter observed survival time, the pair is said to be concordant. The C-Index is then computed as:

$$C = \frac{\sum_{i,j} \mathbf{1}(T_i < T_j) \mathbf{1}(\hat{r}_i > \hat{r}_j)}{\sum_{i,j} \mathbf{1}(T_i < T_j)} \quad (2.21)$$

where $\hat{r}_i$ is the predicted risk score for individual i , and $\mathbf{1}()$ is an indicator function. A C-Index of 1.0 indicates perfect discrimination, meaning the model correctly ranks all individuals. A value of 0.5 corresponds to random ranking, while values below 0.5 indicate worse-than-random predictions.

Time-Dependent C-Index (Antolini et al.) The standard C-Index assumes that risk rankings remain valid across all time points, which may not always be the case in real-world survival data where the risk profile of individuals can change over time. To address this, a time-dependent C-Index extends the concept by evaluating the concordance of predictions at different time horizons. Instead of comparing survival rankings across all time points, it considers whether the predicted survival probabilities at a given time t are correctly ordering individuals who were still at risk just before t . This refinement is particularly useful in dynamic survival modeling, where a model's performance can be assessed at different stages of follow-up. According to Antolini et al. formulation [30], time-dependent C-Index is defined as:

$$C_{td} = P\left(\hat{S}(t | X_i) < \hat{S}(t | X_j) \mid T_i < T_j, \delta_i = 1\right), \quad (2.22)$$

where:

- T_i, T_j are the survival times of individuals i and j .
- δ_i is the censoring indicator ($\delta_i = 1$ if the event is observed, 0 if censored).

- $\hat{S}(t | X_i)$ is the predicted survival probability at time t for individual i .

The empirical estimation of the time-dependent C-Index is given by:

$$C_{td} = \frac{\sum_{i,j} W_{ij} \mathbf{1}(\hat{S}(t | X_i) < \hat{S}(t | X_j))}{\sum_{i,j} W_{ij}}, \quad (2.23)$$

where:

- $\mathbf{1}(\cdot)$ is the indicator function, which returns 1 if the inequality holds and 0 otherwise.
- W_{ij} is a weight function that accounts for censoring, typically defined as:

$$W_{ij} = \frac{\mathbf{1}(T_i < T_j, \delta_i = 1)}{\hat{G}(T_i)}, \quad (2.24)$$

where $\hat{G}(T_i)$ is the Kaplan-Meier estimate of the censoring distribution, ensuring proper weighting of censored observations.

Brier score

While the C-Index measures ranking performance, it does not assess the calibration of survival predictions—that is, how well the estimated survival probabilities match actual outcomes. The Brier Score provides a way to quantify this by measuring the difference between predicted and observed survival probabilities [31].

At a given time t , the Brier Score is defined as:

$$BS(t) = \frac{1}{N} \sum_{i=1}^N \left[\left(0 - \hat{S}(t | x_i)\right)^2 \mathbf{1}(T_i \leq t, \delta_i = 1) + \left(1 - \hat{S}(t | x_i)\right)^2 \mathbf{1}(T_i > t) \right] \quad (2.25)$$

where

- $\hat{S}(t | x_i)$ is the predicted survival probability for individual i at time t ,
- T_i is the observed survival time,
- δ_i is the censoring indicator (1 if the event is observed, 0 if censored),
- $\mathbf{1}(T_i \leq t, \delta_i = 1)$ represents individuals who experienced the event by time t ,
- $\mathbf{1}(T_i > t)$ represents individuals who have not yet experienced the event.

A lower Brier Score indicates better calibration, with 0 representing perfect predictions and 0.25 corresponding to a completely uninformative model in a binary setting.

Since the Brier Score is computed at a specific time t , it is often useful to integrate it over a range of time points to obtain a single, overall measure of predictive performance. The *Integrated Brier Score* (IBS) is computed as:

$$IBS = \frac{1}{t_{\max}} \int_0^{t_{\max}} BS(t) dt \quad (2.26)$$

where t_{max} is the maximum time point considered in the study. The IBS provides a summary measure of how well a survival model predicts across the entire follow-up period.

Both the C-Index and the Brier Score serve complementary roles in survival model evaluation. The C-Index focuses on discrimination—how well the model ranks individuals by survival risk—while the Brier Score assesses calibration, ensuring that predicted survival probabilities align with observed events. When combined, these metrics provide a comprehensive evaluation of survival models, ensuring they are both accurate in ranking patients and well-calibrated in their risk predictions.

2.2 Privacy-preserving Distributed Learning

Advances in machine learning and artificial intelligence have transformed healthcare by enabling precise and individualized predictions. However, their application in sensitive domains like healthcare faces significant barriers due to strict regulations on sharing patient data. These regulations are particularly challenging for rare diseases, where individual institutions often lack sufficient data to train robust predictive models. Collaborative efforts across institutions, while promising, require innovative approaches to ensure data privacy.

To address these challenges, privacy-preserving distributed learning frameworks have emerged as a powerful solution. By facilitating model training on data distributed across institutions without requiring data sharing, these frameworks provide a way to unlock the potential of collaborative learning while maintaining patient confidentiality [32, 33, 34, 35]. Among these methods, Federated Learning (FL), introduced in 2017 [36], has rapidly become a cornerstone for privacy-centric collaborative machine learning [37, 38].

2.2.1 Federated Learning

Federated Learning is a decentralized approach to machine learning where data remains localized at individual nodes (e.g., hospitals, research institutions), and only model parameters or updates are shared with a central server. This architecture ensures data privacy while enabling collaborative model training across multiple institutions.

Problem Setting

Let the dataset be distributed across K institutions (or nodes), denoted as $D = \{D_1, D_2, \dots, D_K\}$, where $D_k = \{(x_i, y_i)\}_{i=1}^{n_k}$ represents the local dataset at institution k , consisting of n_k data samples. The global dataset D can thus be expressed as $D = \bigcup_{k=1}^K D_k$, and the total number of samples is $N = \sum_{k=1}^K n_k$.

The goal of FL is to collaboratively minimize a global objective function $\mathcal{F}(w)$ over the model parameters w , without sharing raw data between nodes. The global objective function can be formulated as:

$$\mathcal{F}(w) = \frac{1}{N} \sum_{k=1}^K \sum_{i=1}^{n_k} \ell(w; x_i, y_i),$$

where $\ell(w; x_i, y_i)$ is the loss function (e.g., cross-entropy for classification or mean squared error for regression) for a single data point.

Federated Optimization

To optimize $\mathcal{F}(w)$, FL employs a decentralized iterative process:

1. **Local Model Updates:** Each node k computes updates to the global model w using its local data. Specifically, the local objective for node k is:

$$\mathcal{F}_k(w) = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(w; x_i, y_i).$$

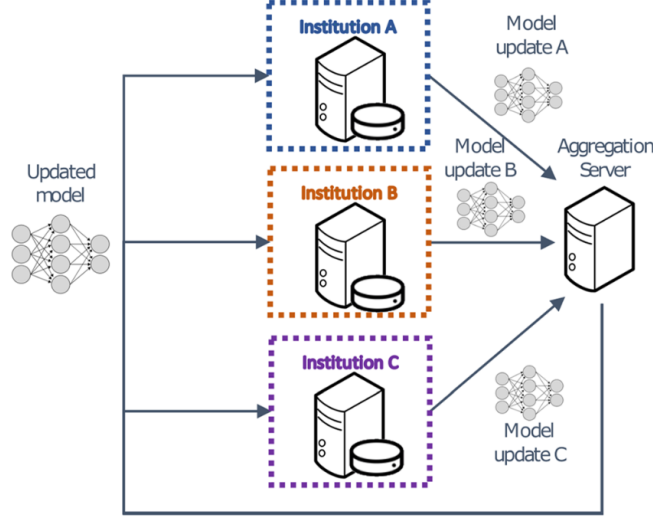


Figure 2.2: Schematic representation of a FL framework, from [empty citation].

Node k performs E steps of local optimization (e.g., stochastic gradient descent, SGD) to minimize $\mathcal{F}_k(w)$. The update rule can be written as:

$$w_k^{(t+1)} = w_k^{(t)} - \eta \nabla \mathcal{F}_k(w_k^{(t)}),$$

where η is the learning rate.

2. Parameter Aggregation: After local updates, each node sends its updated parameters $w_k^{(t+1)}$ to a central server. The server aggregates these parameters to update the global model w . A common aggregation method is Federated Averaging (FedAvg) [36]:

$$w^{(t+1)} = \frac{1}{N} \sum_{k=1}^K n_k w_k^{(t+1)}.$$

3. Broadcast Updated Model: The central server broadcasts the updated global model $w^{(t+1)}$ back to all nodes, which initialize their local models with this value.

This process is repeated for T communication rounds or until convergence.

FL significantly reduces communication overhead by performing multiple local updates (E steps) before synchronizing with the central server. The total communication cost depends on the number of rounds T , the number of nodes K , and the size of the model parameters $|w|$. The cost per round is approximately $\mathcal{O}(K \cdot |w|)$.

Privacy and Security

To ensure privacy, FL often integrates additional techniques such as:

1. Differential Privacy (DP): Adding noise to the local updates or aggregated model to obscure individual data contributions:

$$\tilde{w}^{(t+1)} = w^{(t+1)} + \mathcal{N}(0, \sigma^2),$$

where $\mathcal{N}(0, \sigma^2)$ is Gaussian noise.

2. **Secure Aggregation:** Using cryptographic methods to securely compute the aggregated model without exposing individual updates.

3. **Homomorphic Encryption (HE):** Encrypting data so that computations can be performed on encrypted values without decryption.

Convergence and Challenges

As outlined in [39] and [40], convergence in FL is influenced by factors such as:

- **Non-IID Data Distribution:** Data at different nodes may not follow the same distribution, which can slow convergence.
- **System Heterogeneity:** Nodes may have different computational capacities, leading to asynchrony.
- **Communication Constraints:** Limited bandwidth and network delays can hinder the efficiency of parameter aggregation.

Theoretical results show that under certain assumptions, the FedAvg algorithm converges to the global optimum for strongly convex and smooth loss functions:

$$\mathcal{F}(w^{(t)}) - \mathcal{F}(w^*) \leq \mathcal{O}\left(\frac{1}{\sqrt{t}}\right),$$

where w^* is the optimal model.

Finally, most FL existing strategies are optimized for standard gradient descent algorithms (like neural network-based models) for weights updating. This approach rules out the use of other algorithms that could be more suitable or perform better in some applications, thus posing a hard constraint on the development of optimal distributed solutions.

2.2.2 Distributed Survival Analysis

Among the different supervised learning tasks, Survival Analysis is intrinsically more complex to distribute compared to classification and regression models. Indeed, optimising Cox-like loss functions requires the knowledge of all individuals' sorted times, requiring a ranking evaluation (see Eq.2.20). This implies that optimising the function in *batches* (or per centre) may lead to a different solution which may not be optimal compared to the one obtained using the loss function computed on the complete and centralised data.

Recent studies faced the problem of creating distributed survival models by building *surrogate* likelihood functions [41] or aggregating different CoxPH models trained in separated nodes [42, 43]. In [44], a classical federation was built on a novel reformulation of the survival task in terms of pseudovalues, while in [45] the authors proposed a discrete-time extension of the CoxPH model that makes the loss separable and turns the survival task into a multiple classification problem. However, most of these methods (except [44]) are specifically tailored either to the linear CoxPH framework, which is additive in the input features, or more generally to the CoxPH loss function and its underlying assumptions. Thus, they cannot be easily extended to other more comprehensive and non-linear algorithms.

The field of synthetic data generation has recently gained much attention as a viable strategy for sharing information among different data holders that nevertheless avoids

the dissemination of real patient data [46, 47, 48, 49]. However, the generation of synthetic data for Survival Analysis tasks is not straightforward due to the presence of censored times, as outlined by Norcliffe et al. [50], who proposed a generative adversarial model tailored for survival predictions.

Chapter 3

SYNDSURV: A Simple Framework for Survival Analysis with Data Distributed Across Multiple Institutions

Data sharing among different institutions represents one of the major challenges in developing distributed machine learning approaches, especially when data is sensitive, such as in medical applications. Federated learning is a possible solution, but requires fast communications and flawless security. SYNDSURV (SYNthetic Distributed SURVival) is an alternative approach that simplifies the current state-of-the-art paradigm by allowing different centres to generate local simulated instances from real data and then gather them into a centralized hub, where an Artificial Intelligence (AI) model can learn in a standard way. The main advantage of this procedure is that it is model-agnostic; therefore, prediction models can be directly applied in distributed applications without requiring particular adaptations as the current federated approaches do. To show the validity of this approach for medical applications, we tested it on a survival analysis task, offering a viable alternative to train AI models on distributed data. While federated learning has been mainly optimized for gradient-based approaches so far, this framework works with any predictive method, proving to be a comparable way of performing distributed learning without being too demanding towards each participating institute in terms of infrastructural requirements.

3.1 Introduction

The past few decades have witnessed a massive and growing increase in well-structured clinical data on a global scale [51, 52]. The development of increasingly sophisticated Artificial Intelligence (AI) techniques capable of fully exploiting large, heterogeneous, and complex data will be of crucial importance for healthcare systems in the coming years [53, 54], and Federated Learning (FL) initiatives are being developed in healthcare to collaboratively train predictive models without the need to centralise sensitive personal data [55]. However, a main constraint in using clinical data for predictive purposes is that, despite the relative global abundance, they are limited in individual centres, especially for rare diseases. Furthermore, sharing and aggregating local

datasets is often impossible due to their strict privacy regulations. Therefore, it is essential to analyse the available clinical data in a manner that adheres to privacy regulations while still preserving high accuracy in the predictions.

One of the main applications in healthcare is to predict when specific adverse events of interest to the patient occur. Survival analysis models address this task and several machine learning-based techniques have been proposed for prognosis predictions. However, making distributed versions of survival models is more complex than others for classification and regression models, since it requires ranking evaluations and dealing with censored instances. These instances are partially missing temporal data, which are common in real-world clinical scenarios, where patients can only sometimes be subject to an arbitrarily long follow-up period.

This work addresses data-sharing challenges in critical contexts by proposing a distributed model tailored for survival analysis. The proposed framework generates simulated survival data based on real local datasets and aggregates them into a centralized hub for training. The primary aims for this work are: 1) to develop accurate predictive models that can improve patient outcomes and advance medical research without compromising data security; 2) to overcome current limitations of federated approaches by proposing a simple model-agnostic data sharing framework which can be easily implemented by the participating institutes.

Relevant related works are outlined and discussed in Section 2.2.2.

3.2 Methodology

3.2.1 Model architecture

With the proposed framework, named *SYNDSURV* (SYNthetic Distributed SURVival), we address the task of developing a common survival model in a decentralised scenario, where data are stored in separated centres (e.g. healthcare institutions) and cannot be shared due to privacy limitations. The core idea behind our contribution is to build a unique survival model using an arbitrarily large set of aggregated synthetic data, that are independently generated at each node level. The pseudo-code of the whole procedure is shown in Algorithm 1.

The global model $\mathcal{M}$ has been tested twice, using two different time-to-event predictive methods (CoxPH and DeepHit). The performance was assessed in terms of two survival-specific metrics (C-Index and time-dependent AUC), which were compared to the upper bound (centralised scenario) and lower bound (isolated nodes) scores. A schematic representation of the framework is presented in Figure 3.1.

3.2.2 Data Synthesizer

To generate the synthetic data, we have employed the *DataSynthesizer* tool [56], which is a Python implementation of the *PrivBayes* algorithm [57]. It is a generative model that can be represented as a directed acyclic graph. In the graph, each node X_i represents a feature, and each edge from X_j to X_i corresponds to $\Pr(X_i|X_j)$, the probability of X_i given X_j .

In summary, the algorithm operates in two steps. First, it creates a k -degree Bayesian

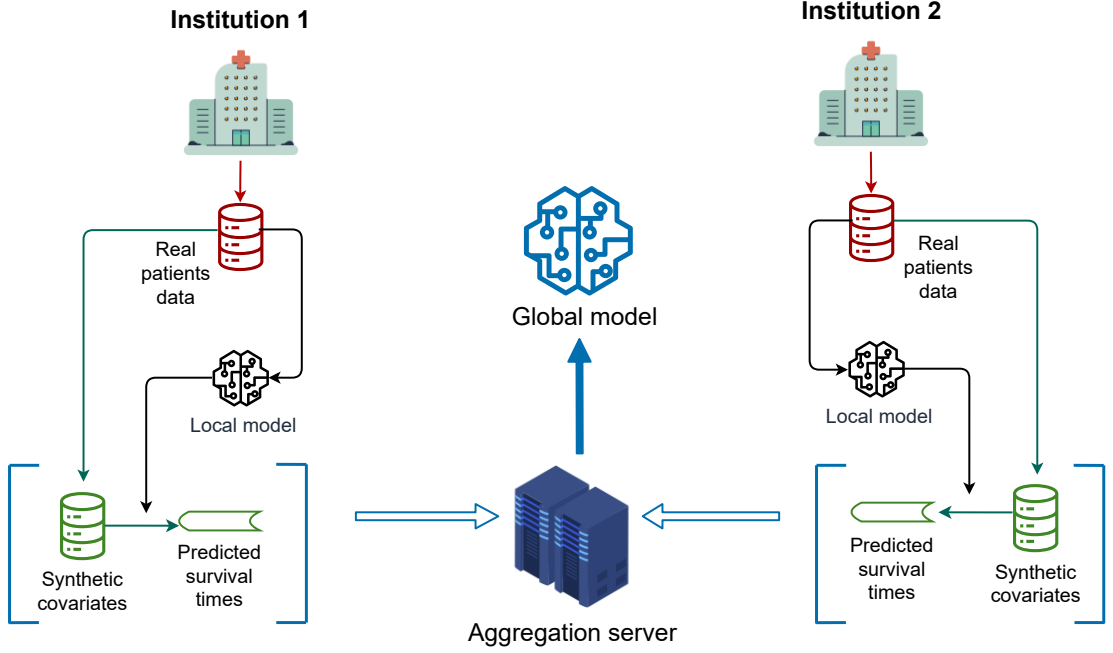


Figure 3.1: Schematic representation of the SYNDSURV framework. In this example, there are two participating institutions that cooperate to create a global model. Each one trains a local AFT survival algorithm using only its own patients' data. Then, synthetic samples are generated through a Bayesian generative network built with patient covariates at single institution level, and local models are used to predict survival times for synthetic patients. Finally, all synthetic data are transferred from the data centres to a central aggregation server that trains the global model. Its performance is evaluated on an external test set composed of real patients from each institution.

Network $\mathcal{N}$ (where k defines the maximum size of any parent node set Π_i in $\mathcal{N}$) that approximates the real distribution of the features set $\mathcal{A}$, such that $\Pr_{\mathcal{N}}[\mathcal{A}] \approx \Pr[\mathcal{A}]$. In particular, under the assumption that any X_i and any $X_j \notin \Pi_i$ are conditionally independent given Π_i , we have:

$$\Pr[\mathcal{A}] = \Pr[X_1, X_2, \dots, X_d] = \prod_{i=1}^d \Pr[X_i | \Pi_i] \quad (3.1)$$

Thus, the construction of $\Pr_{\mathcal{N}}[\mathcal{A}]$ can be viewed as an optimization problem, where the objective is to minimize the *KL-divergence* $D_{KL}(\Pr[\mathcal{A}] \parallel \Pr_{\mathcal{N}}[\mathcal{A}])$, by choosing the optimal parent node set Π_i related to each feature X_i . If k is a small value, this procedure becomes computationally feasible.

The second step is to generate synthetic data from the network $\mathcal{N}$. This is done by sampling each X_i from its conditional distribution $\Pr[X_i | \Pi_i]$ independently, without considering any attribute not in $\Pi_i \cup \{X_i\}$, which is possible thanks to Eq. 3.1.

Moreover, the generative process can also be combined with privacy-preserving approaches like Differential Privacy, which guarantees that the output of an algorithm

Algorithm 1: Pseudo-code of the proposed framework for distributed survival analysis.

Input: Datasets $(\mathbf{X}_1, t_1, \delta_1), (\mathbf{X}_2, t_2, \delta_2), \dots, (\mathbf{X}_C, t_C, \delta_C)$ from C nodes

Output: Trained global survival model $\mathcal{M}$

```

1 for each node  $c$  do
2    $((\mathbf{X}_c^{\text{tr}}, t_c^{\text{tr}}, \delta_c^{\text{tr}}), (\mathbf{X}_c^{\text{tt}}, t_c^{\text{tt}}, \delta_c^{\text{tt}})) = \text{train\_test\_split}(\mathbf{X}_c, t_c, \delta_c)$ 
3    $\tilde{\mathbf{X}}_c = \text{Synthetic\_generator}(\mathbf{X}_c^{\text{tr}})$ 
4    $\mathcal{M}_c = \text{FitAFTModel}(\mathbf{X}_c^{\text{tr}})$  # fit AFT-survival model to original data
5    $\tilde{t}_c = \mathcal{M}_c(\tilde{\mathbf{X}}_c^{\text{tr}})$  # predict event times for the synthetic data
6    $\tilde{\delta}_c = \begin{cases} 0 & \text{if } \tilde{t}_c > \max(t_c^{\text{tr}}) \\ 1 & \text{otherwise} \end{cases}$  # set synthetic event indicators
7    $\text{Send\_To\_Aggregator}(\tilde{\mathbf{X}}_c, \tilde{t}_c, \tilde{\delta}_c)$ 

8  $(\tilde{\mathbf{X}}, \tilde{t}, \tilde{\delta}) = \text{AggregateData}()$ 
9  $\mathcal{M} = \text{FitModel}(\tilde{\mathbf{X}}, \tilde{t}, \tilde{\delta})$  # train global survival model Evaluate  $(\mathcal{M}, (\mathbf{X}^{\text{tt}}, t^{\text{tt}}, \delta^{\text{tt}}))$ 
   # evaluate M on aggregated test data

```

is statistically indistinguishable on a pair of *neighbouring* databases (i.e. a pair of databases that differ by only one tuple). This concept is formalized with the following definition [58]:

Definition 1 (ϵ -Differential privacy). *A randomised algorithm G satisfies ϵ -differential privacy if, for any two datasets D_1 and D_2 that differ only in one tuple x , and for any possible output O of G , we have:*

$$\Pr[G(D_1) = O] \leq e^\epsilon \cdot \Pr[G(D_2) = O].$$

The parameter ϵ is called privacy budget, and it controls how much the distribution of outputs can depend on data from an individual x . In practice, differential privacy is obtained by introducing a Laplacian noise $\text{Lap}(\frac{2(d-k)}{n\epsilon})$ into each joint distribution $\Pr[X_i, \Pi_i]$, from which the differential private conditional distributions $\Pr^*[X_i | \Pi_i]$ are then derived. Here, d is the number of features and n is the number of samples in the dataset. It is worth highlighting that higher values of ϵ correspond to a lower amount of noise injected into the synthetic data, and hence to a lower level of privacy. Usually ϵ is set in a range of values that goes from 0.1 to 10 ([59], [57]).

Once completed, this whole procedure leads to the generation of a new synthetic dataset $\tilde{\mathbf{X}}_c$ for each node c , with the desired samples size, which depends on how far the sampling process is carried out.

3.2.3 Synthetic labels generation

The generated patients, in order to be used for training a common survival model, need to be paired with an estimated time-to-event label. Indeed, this information could

not fit properly into the Bayesian Network due to the presence of the censored data. To overcome this issue, the proposed solution is to define, for each node, a survival regression with an accelerated failure time (AFT) model using XGBoost (see Section 2.1.5).

An advantage of applying an AFT model is that, unlike traditional Cox-PH, it yields directly a usable prediction of the time to the event $\hat{t}$, thus avoiding its estimation from the survival function. In addition, the AFT framework provides a better fit when the proportional hazard assumption does not hold and, combined with a tree-based Gradient Boosting method like XGBoost, it also considers nonlinear relationships among covariates, enriching the model’s capabilities.

For each node, a separated AFT model $\mathcal{M}_c$ is trained by using only the real data of the node c and then applied to predict the time-to-event $\tilde{t}_c$ related to the synthetic samples $\widetilde{\mathbf{X}}_c$. In addition, in order to generate realistic survival data, the synthetic patients with a predicted time greater than the maximum observed event time of that node ($\tilde{t}_{c,i} > t_{max,c}$) were considered right-censored at the time $t_{max,c}$, by setting $\widetilde{\delta}_{c,i} = 0$. This truncation avoids having long-tailed distributions for the synthetic times.

3.2.4 Centralised learning and validation strategy

The synthetic tuples $\{(\widetilde{\mathbf{X}}_c, \tilde{t}_c, \widetilde{\delta}_c)\}_{c=1,\dots,C}$ are then shared with an aggregation server in order to be jointly used to build a common survival model $\mathcal{M}$. This allows the exploration of any desired survival algorithm in a centralised-like scenario, thus avoiding all the possible limitations related to model-specific federation processes.

In this work, to prove the effectiveness of the architecture, we applied both CoxPH model and DeepHit, which is a more complex neural network-based model that avoids the proportional hazard and linearity assumptions. In the CoxPH model we included an ElasticNet penalty to prevent overfitting [60], treating the penalty strength α and the l_1 ratio as tunable hyperparameters.

DeepHit is a recently developed survival model, which has been shown to outperform standard CoxPH in both single-event and competing-risk survival predictions [23]. DeepHit discretizes the survival times and trains a neural network to learn the estimated joint distribution of survival time and event, while capturing the right-censored nature inherent in survival data. The neural network returns a softmax layer that outputs an estimated probability for each discrete time interval, and it is trained by using as loss function a convex combination of a negative log-likelihood and a ranking loss.

Both ElasticNet-CoxPH and DeepHit hyperparameters were optimized using an internal validation set, using the tree-structured Parzen estimator algorithm [61] (implemented within the *Optuna* optimization framework [62]).

The final model, despite being trained on the synthetic dataset, needs to be reliable also when applied to real patients. Therefore, for all simulations, we tested the performance with an external test set of the real data, which had never been used by the models $\mathcal{M}_c$ for the synthetic labels’ prediction, nor by the centralised synthetic model $\mathcal{M}$. Moreover, in order to analyse the effect of the synthetic dataset’s size on the final performance, we evaluated different sets of generated patients, ranging from 10^1 to 3×10^3 samples per node.

As in the case of FL strategies, the performance naturally has an upper bound and a lower bound. The *upper bound* represents the centralised scenario on the real data, i.e. when patients’ data can be directly shared to a central node without restrictions. In this case, it is the performance of the model trained on the aggregated real data from all nodes. In contrast, the *lower bound* indicates the worst-case scenario, i.e. when the model is trained for each isolated node using only its own data. Therefore, it is the average performance across all nodes. The expected behavior of our method in order to demonstrate its effectiveness is that its performance on the test set has to lie somewhere within these two limits.

The whole procedure, including the synthetic patients’ generation from the training data, has been repeated 10 times by changing the original random train-test split.

3.3 Results

3.3.1 Survival Datasets

To validate our framework, we considered four publicly available survival datasets, commonly used to assess the performance of the survival models: the Assay of Serum Free Light Chain [63] (FLCHAIN), the Molecular Taxonomy of Breast Cancer International Consortium [64] (METABRIC), the Study to Understand Prognoses Preferences Outcomes and Risks of Treatment [65] (SUPPORT) and the Rotterdam tumour bank and German Breast Cancer Study Group [66], [67] (GBSG+Rotterdam). FLChain and METABRIC datasets were preprocessed according to [68], while for GBSG+Rotterdam the same preprocessing schema of [69] was applied, with the only difference that the data from the two studies were finally merged to create a single dataset. As regards SUPPORT, we used the same procedure of [25] for data cleaning and imputation, followed by a final feature selection to decrease the data dimensionality, as it is summarised in Figure A.2, Appendix A.

Each dataset was randomly split into training and test sets (75% – 25%), repeating the partitioning 10 times, by changing the random seeds. Then, the samples of the training sets were further distributed among C fictitious centres to simulate a decentralised scenario. The number of nodes C for each dataset was selected so that the performance gap between upper and lower bounds (in at least one evaluation metric) was large enough to justify an advantage in applying a decentralised approach. All the created fictitious centres included $\approx 2 \times 10^2$ samples, which is also a reasonable number of patients for real-world clinical trials of medium size. Summary statistics for the preprocessed survival datasets are reported in Table 3.1.

	Size	# Features	% Censored	C	Nodes size
FLChain	6524	7	70%	32	153
METABRIC	1904	9	42%	8	178
SUPPORT	9105	20	32%	32	213
GBSG+Rotterdam	3668	8	50%	16	172

Table 3.1: Summary of survival datasets.

3.3.2 Performance comparison

As mentioned above, to properly evaluate the performance of the synthetic model $\mathcal{M}$ in the test data, it must be compared to the upper bound (centralised scenario) and the lower bound (isolated nodes) boundary situations. From a preliminary comparison of the performance of CoxPH and DeepHit in the centralised scenario, as shown in Tables 3.2a and 3.2b, it clearly results that DeepHit outperforms or at least equals the standard CoxPH method for both metrics, which is expected due to the greater flexibility and capacity of the model. Conversely, in the isolated nodes the same consideration does not hold and CoxPH performs better than DeepHit in 2 out of 4 datasets, for both metrics. This is probably due to the reduced size of the fictitious data, which mostly affects DeepHit.

	<i>Lower bound</i>		<i>Upper bound</i>	
	COXPH	DEEPHIT	COXPH	DEEPHIT
FLChain	0.766(0.003)	0.779 (0.003)	0.789(0.003)	0.789(0.003)
METABRIC	0.612 (0.004)	0.604(0.005)	0.630(0.005)	0.674 (0.005)
SUPPORT	0.776(0.001)	0.785 (0.001)	0.802(0.001)	0.843 (0.001)
GBSG+Rotter.	0.641 (0.003)	0.604(0.005)	0.671(0.003)	0.680 (0.003)
(a)				
	<i>Lower bound</i>		<i>Upper bound</i>	
	COXPH	DEEPHIT	COXPH	DEEPHIT
FLChain	0.790 (0.004)	0.743(0.004)	0.812(0.004)	0.813 (0.004)
METABRIC	0.659 (0.006)	0.631(0.004)	0.688(0.006)	0.695 (0.005)
SUPPORT	0.866(0.001)	0.867 (0.001)	0.894(0.002)	0.920 (0.001)
GBSG+Rott.	0.678(0.005)	0.678(0.005)	0.716(0.005)	0.727 (0.004)
(b)				

Table 2: C-Index (a) and Time-dependent AUC (b) metrics on the test set in the centralised scenario (upper bound) and in the case of isolated nodes (lower bound). The reported scores represent the mean values over the 10 different test sets; error of the means are written in brackets. Best scores are highlighted in bold.

It is also worth noticing that the performance gap between upper and lower bounds is not negligible for each dataset and for both metrics. This confirms the necessity of the proposed approach. Hence, we present the results of the global models, trained only with the synthetic data and evaluated on the external test sets of real patients. C-index and time-dependent AUC were considered for a variable number of synthetic patients generated per node, and, for each, the mean and standard deviation values over 10-times repeated simulations were computed. The C-Indices for the CoxPH model are reported in Figure 3.2, while those related to DeepHit are shown in Figure 3.3. In both figures, each panel corresponds to a separate survival dataset; red and black lines represent the lower and upper bounds, respectively. The time-dependent AUCs are shown in Figures A.3 to A.5 in Appendix A.

For CoxPH, the synthetic model reaches the upper limit of the C-Index in all four datasets, even with less than 10^3 generated patients per node; when considering the time-dependent AUCs, this happens in 3 out of 4 datasets, even though in the remaining one (METABRIC) the average scores fall clearly within the desired range. Considering DeepHit, in all simulations the average performance of the synthetic model exceeds the lower bound as the number of patients generated per node is above 10^2 ,

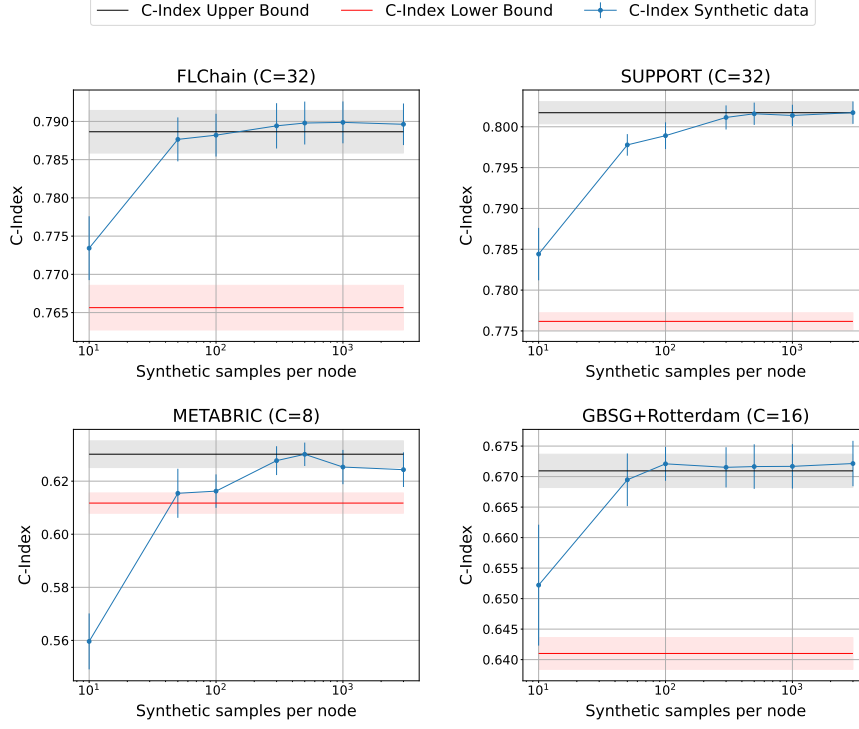


Figure 3.2: Test set C-Index vs. number of generated patients per node, for the CoxPH model trained on the synthetic data. Reported values represent the mean over 10 different test sets; error bars indicate the errors of the mean.

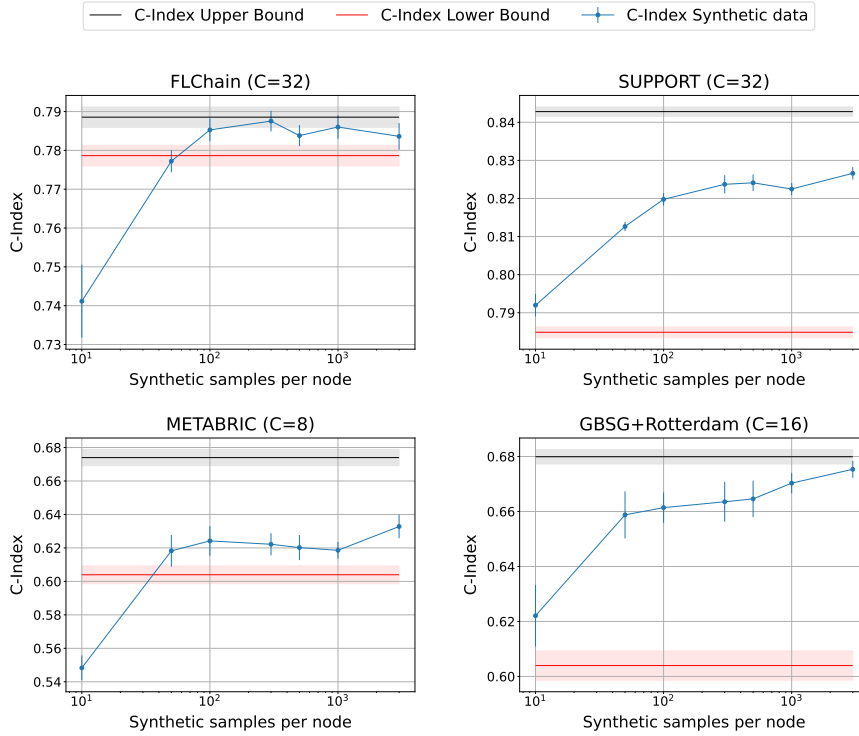


Figure 3.3: Test set C-Index vs. number of generated patients per node, for the DeepHit model trained on the synthetic data. Reported values represent the mean over 10 different test sets; error bars indicate the errors of the mean.

and most of the time it is closer to the upper limit when more than 10^3 patients are

generated.

These results show that the survival models trained on the synthetic data generated through the proposed method are capable of generalizing well to real patients' data, achieving a performance always comparable to that of the centralized scenario or considerably higher than the lower bound. Moreover, such achievements are not model-dependent, since the same considerations apply for both CoxPH and DeepHit models, which are different in terms of baseline assumptions, optimization function, and complexity.

3.3.3 Differential Privacy

Finally, we also tested the goodness of the proposed framework in a Differential Privacy setting. This is done by adding to the Bayesian Network (from which synthetic covariates are sampled) a certain amount of noise parameterized by the privacy budget ϵ , as explained in 3.2.2. In our experiments, we set $\epsilon = 0.1$, which produces a high level of noise at the expense of the synthetic data quality and hence of the model's performance. The results for DeepHit on the datasets considered in this scenario are presented in Figure 3.4.

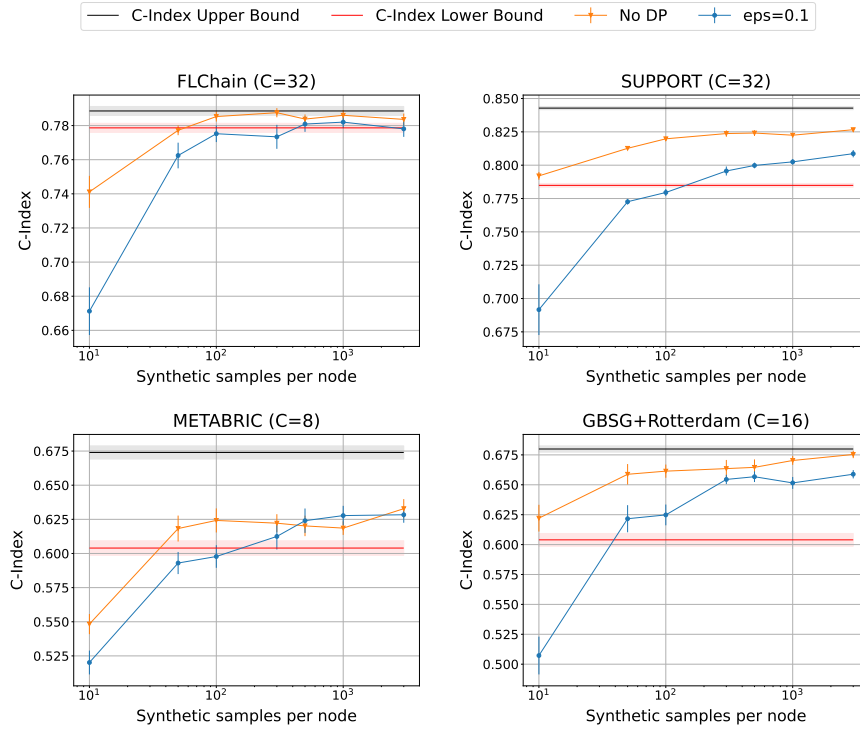


Figure 3.4: Test set C-Index vs. number of generated patients per node, for the DeepHit model trained on the synthetic data, using a differential privacy setting with $\epsilon = 0.1$. Reported values represent the mean over 10 different test sets; error bars indicate the errors of the mean.

Due to the added noise, both C-Indices and time-dependent AUCs are slightly deteriorated on the test set with respect to the scenario without differential privacy, but they are higher than the lower bound in all four datasets when more than 5×10^2 patients per node are generated. This achievement shows that the *SYNDSURV* framework is robust to the presence of high levels of noise in the synthetic data.

Moreover, by controlling the privacy budget ϵ , it is possible to choose the desired trade-off between model performance and privacy requirements.

3.4 Conclusion

We proposed a distributed survival analysis framework to address the task of creating a common survival model that indirectly uses time-to-event data from multiple participating institutions, without sharing patients’ sensitive data. This is achieved by generating synthetic patients at the level of the single data centres through a two-steps procedure that properly handles survival labels: 1) the synthetic patient’s covariates are sampled from a Bayesian generative network; 2) a survival AFT model trained on the original data is used to predict the survival times of the synthetic patients. Thus, the synthetic samples from each institution are shared with a common aggregation server to train the desired global model.

As demonstrated by simulations performed over four distinct survival datasets (split among separated fictitious nodes), the global model’s performance on a test set composed of real patients is often comparable to the centralised scenario, where data are shared without limitations, or at least significantly greater than the lower bound (isolated nodes scenario). These results prove the applicability of the method as a valid alternative to the current FL approaches used to learn common predictive models. A main advantage of this approach is that it is flexible to the choice of any desired algorithm for designing the global model, since it does not depend on the algorithm structure as it happens for the FL methods. In addition, it is also suitable to properly handle time-to-event data in presence of censoring. Finally, we also demonstrated that our approach is robust to the presence of noise into the synthetic data, hence it can be naturally embedded in differential privacy settings, which can be necessary when controlling the privacy budget.

3.4.1 Limitations and future works

Despite its advantages, the proposed framework may have some limitations. A possible limitation is that the performance of the global model depends on the quality of the synthetic samples generated by the local data centres. If the synthetic samples are not representative of the true data, the global model may not perform well.

Another limitation is that the proposed method assumes that the censoring mechanism is the same across all local data centres, which may only sometimes be the case in practice.

Finally, generating synthetic samples at local data centres might be challenging, especially for datasets with many features. However, current federated approaches may suffer from the same problem.

Future developments include further experiments to test if our framework can address more troublesome real-world scenarios, where nodes are highly unbalanced or data are non-IID across the participating centres. Likewise, different strategies for synthetic data generation should be investigated, including Autoencoders (Decoders), Generative Adversarial Networks (GANs) for both tabular and imaging data, or other generative Bayesian network techniques.

Chapter 4

VAE-Surv: A Novel Approach for Genetic-Based Clustering and Prognosis Prediction in Myelodysplastic Syndromes

4.1 Introduction

Survival analysis is one of the most used approaches in clinical prognostic investigations, aiming to predict the time to an event, such as death or the progression of a disease. A key aspect of this analysis is the presence of censored data, indicating that the event of interest did not occur during the study, therefore requiring the use of specialized statistical methods. Traditionally, the Cox proportional hazards model (CoxPH) [19] has been the most widely used technique to analyze censored data, but was designed for small data sets and does not scale well to high dimensions. In addition, most of the state-of-the-art methods assume only linear relationships among covariates. Recently, several machine learning (ML) algorithms have been adapted to work with censored data and can give more accurate results than traditional statistical methods since they are able to consider nonlinear relationships among the features [70].

Another main goal in clinics when censored data are analyzed is disease subtyping through risk stratification of patients. To address this task, cluster analysis of survival data can identify similar groups of patients in time-to-event distributions, in order to enhance a precision medicine approach for clinical decision-making. So far, only a few approaches have been suggested to cluster patients according to their survival function, yet this topic is beginning to gain increased attention. Liverani et al. proposed a Dirichlet process mixture model with cluster-specific parameters for the Weibull distribution, addressing cases where it is difficult to apply usual survival models due to multicollinearity [71]. Chapfuwa et al. proposed a Bayesian non-parametric time-to-event approach with structured latent representations that can be clustered through a prior for infinite mixture of distributions [72]. However, it is difficult to transfer these approaches to medical studies when multiple types of data are available (omics, clinical, imaging, etc.), which are often high-dimensional, heterogeneous and with missing information, representing challenges to current statistical methods.

Recent research has shown that the integration of multiple sources of data (e.g. clinical, genomic data) leads to better prediction of the prognostic risk than the use of a single source [73]. In this context, some studies proved the effectiveness of Autoencoder architectures in integrating clinical and multiomics data, exploiting the learned latent representations to predict the outcomes of interest [74, 75, 76]. Kim et al. [77] applied a Variational Autoencoder (VAE) model to pan-cancer RNA-seq gene expression data from TCGA, combined through a transfer learning setting with a survival analysis neural network. However, the authors focused the approach in the risk prediction, without exploiting the latent features obtained by the autoencoder and without integrating the clinical variables. In the work by Hira et al. [78] instead, multiple VAEs were used to integrate RNA-seq, CNVs and DNA-methylation data in the context of ovarian cancer, identifying disease subtypes through a classification schema applied to the latent features. This approach optimizes a combined loss that accounts for both the VAE reconstruction and the classification task. The latent features were also used to provide a stratification of the patients. However, the authors focused their approach on the classification task and performed linear CoxPH survival predictions to each classified subgroup only *a posteriori*.

In this study, we propose VAE-Surv, a computational framework based on a VAE with the aim to both stratify the patients and assess their prognostic risk by optimizing both data reconstruction and the survival prediction task. Specifically, the framework integrates genetic, cytogenetic and clinical features to uncover novel insights into genetic-based prognostic risk predictions. To show the robustness of our approach, we focused our application on the prognosis prediction of Myelodysplastic Syndromes (MDS), which represent a heterogeneous group of clonal hematopoietic stem cell disorders characterized by ineffective hematopoiesis. These syndromes predominantly affect elderly populations and manifest as bone marrow failure, leading to varying degrees of cytopenia (low blood cell counts) in one or more myeloid lineages. Additionally, MDS has a substantial risk of progression to Acute Myeloid Leukemia (AML). From a genetic standpoint, MDS showcase an intricate mutational landscape with numerous genetic abnormalities, including point mutations, chromosomal deletions, and translocations [79, 80, 81]. Although these mutations contribute to the onset and progression of the disease, their predictive utility for treatment response remains a topic of ongoing research. The current WHO 2016 classification of MDS patients [82], which is based on morphological, clinical, and genetic characteristics, cannot fully encapsulate the complex genetic and cytogenetic landscape of this disorder, limiting the granularity of the patients' stratification [83, 84] and the therapeutic options, which require personalized approaches [85].

Here, we will apply VAE-surv to two of the largest cohorts of MDS patients currently available (2,043 patients from Genomed4all MDS cohort [84] and 2,384 patients from the International Working Group for the study of Prognosis in MDS cohort [86]) to demonstrate that this computational model both provides clinically valid stratification of different MDS subtypes and achieves high accuracy in predicting patients' survival. The underlying code for this study is available at the link <https://github.com/comphiomed-unito/VAE-Surv>.

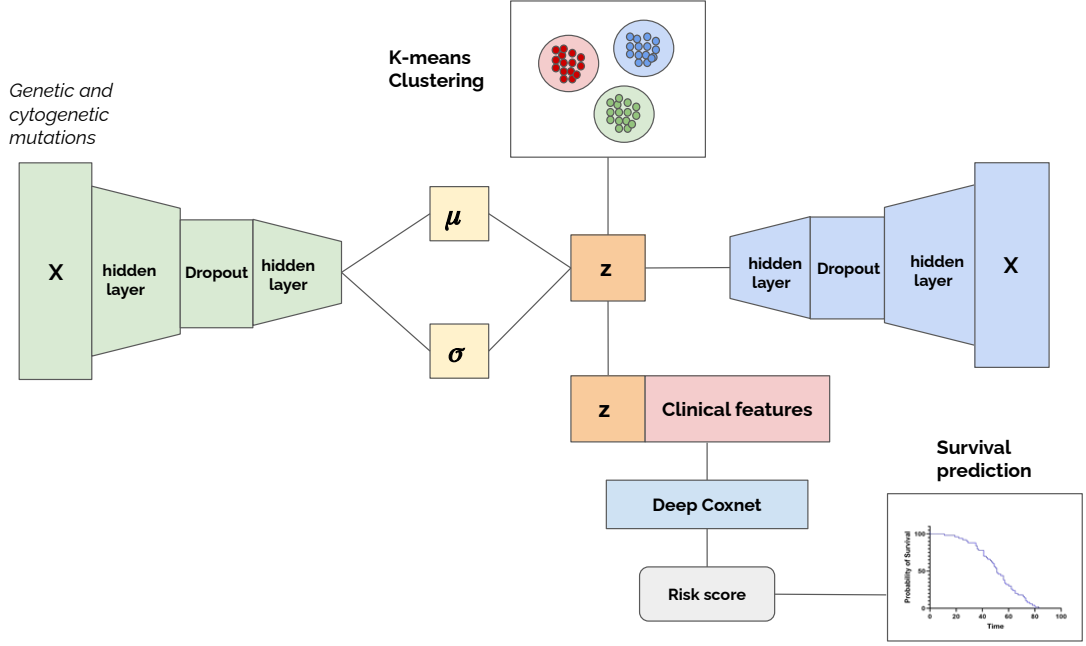


Figure 4.1: Schematic representation of the VAE-Surv computational framework. Genetic and cytogenetic markers are input as binary vectors into a Variational Autoencoder (VAE) for dimensionality reduction. The latent feature space, obtained by the VAE, is concatenated with clinical variables to form an enriched feature set. This combined feature set is then fed into a parallel Deep Cox Proportional Hazards Neural Network (DeepSurv) for survival risk prediction. The dimensionality of the latent space z is optimized as a hyperparameter of the model.

4.2 Materials and Methods

4.2.1 Datasets

The proposed framework was built considering an international retrospective cohort of 2,043 MDS patients, available through the GenoMed4All consortium [84, 87, 3]. The study included patients diagnosed with MDS based on the 2016 WHO classification criteria. Laboratory and clinical data were collected at diagnosis or within six months of diagnosis for consistency. DNA sequencing was performed in bone marrow mononuclear cells or peripheral blood granulocytes, ensuring robust genomic profiling. Patients with therapy-related myeloid neoplasms, paroxysmal nocturnal hemoglobinuria, aplastic anemia, or MDS/myeloproliferative neoplasm with ring sideroblasts and thrombocytosis were excluded, as described in Supplementary File 2 of the Bersanelli et al. study [83]. The dataset collects clinical, demographic and molecular features (a selected panel of 58 genetic and cytogenetic mutations deemed relevant for MDS). Demographic and clinical covariates include age of diagnosis (AOD), gender, neutrophils, hemoglobin, platelets, and bone marrow blasts (BMB). At diagnosis, cytogenetic analysis was performed using standard G-banding and karyotypes were classified using the International System for Cytogenetic Nomenclature Criteria. Mutation screening of 46 genes related to myeloid neoplasms was performed on DNA from peripheral blood granulocytes or bone marrow mononuclear cells. Overall survival was considered as the outcome of interest for our analyses. Both survival time and censoring status were available for each patient (censoring rate of 71%). Further details are reported in Bersanelli et al. [83].

In order to assess the generalisability of the results, the model was also validated on an external cohort of 2,384 MDS patients, provided by the International Working Group for the study of Prognosis in MDS (IWG-PM) [86]. The inclusion criteria for this cohort were: age of diagnosis ≥ 18 years, a diagnosis of MDS according to WHO 2016 criteria and information available on demographics, clinical features, mutational screening/chromosomal abnormalities, treatment and overall survival. A comparison between IWG-PM and Genomed4all cohorts is provided in Appendix B, Section 1.

4.2.2 Model Architecture

A schematic representation of the framework is presented in Fig. 4.1. The core of our computational approach is a Variational Autoencoder (VAE) [88], designed to intake genetic and cytogenetic markers as binary vectors (presence/absence of mutations and abnormalities). The goal of the VAE is to reduce the sparse and high-dimensional genetic-based input space into a latent feature space of lower dimensionality. This latent representation is concatenated with the patients' clinical variables to generate an enriched feature set. A parallel Deep Cox Proportional Hazards Neural Network (DeepSurv [22]) is applied to this enriched feature set for a survival analysis task.

In conventional VAE architectures, a loss function often includes both a reconstruction term $\mathcal{L}_{recon}$ and a Kullback-Leibler divergence term:

$$\begin{aligned}\mathcal{L}_{VAE}(\theta, \phi; x) &= \mathcal{L}_{recon} + \text{KL-term} \\ &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x | z)] - D_{KL}(q_{\phi}(z | x) || p(z)),\end{aligned}$$

where $p_{\theta}(x | z)$ is the decoder parametrized by the weights of the neural network θ , $q_{\phi}(z | x)$ approximates the posterior distribution of the latent variables (encoder parametrized by ϕ) and $p(z)$ is the prior distribution over the latent variables assumed to be Gaussian ($z \sim p(z) = \mathcal{N}(0, I)$). Typically, the KL-term regularizes the latent space by encouraging $q_{\phi}(z | x)$ to adhere closely to the Gaussian prior. This allows the model to learn a smooth and continuous latent space, which is particularly beneficial for generative tasks. However, in our approach, we experiment a removal of the KL-term from the loss function. This decision is motivated by the fact that we are not primarily interested in the generative capabilities of the model or in ensuring that the latent space adheres to a predefined Gaussian prior. Instead, our focus is on learning compact representations of the genetic input, with a particular emphasis on enabling the latent space to exhibit clustered structures. Moreover, by removing the KL divergence, we prevent the phenomenon of posterior collapse [89], where the learned posterior distribution $q(z | x)$ becomes overly similar to the prior distribution $p(z)$, effectively ignoring the input x . Therefore, the risk is that the latent variables z become uninformative, and the decoder almost entirely relies on its learned capacity to reconstruct x without using the latent representation.

The VAE was initially pretrained separately from the DeepSurv model. Its objective was to minimize the reconstruction error through a *logcosh* loss function over a predefined number of epochs. Upon achieving a stable latent feature representation, the VAE's parameters were frozen. After VAE pretraining, the DeepSurv model was fine-tuned. The input feature set of the DeepSurv model includes the frozen latent representation from the VAE concatenated with the clinical variables, allowing for a

comprehensive survival risk assessment. Hyperparameters optimization was applied, employing a 10-fold cross-validation scheme, stratified by censoring indicator, using the Concordance Index (C-Index), as the validation metric [90]. C-Index is a widely used ranking metric, quantifying the model’s ability to correctly order survival times, ranging from zero to one, with a value of 0.5 corresponding to the performance of a random ranking and 1 to perfect discrimination. We compared the performance of our VAE-Surv framework against baseline linear Cox Proportional Hazards (CoxPH) regression models, which do not exploit latent feature extraction, and against the Survival Cluster Analysis (SCA) nonlinear method [72], using the median C-Index obtained across the 10 cross-validation folds as the principal performance metric. Specifically, CoxPH models were implemented with various regularization strategies, including no regularization, Lasso (L1), Ridge (L2), and ElasticNet penalties [91]. Regularization is a method for preventing overfitting by controlling the model complexity. It accomplishes this by penalizing the coefficients of predictors that do not provide meaningful information to the model. Lasso penalization allows for the generation of a sparse model (setting unimportant covariates to 0), while Ridge penalization avoids the feature selection set by Lasso by encouraging a grouping effect of the covariates used in the model. Finally, Cox regression combined with ElasticNet penalization combines both Lasso and Ridge penalty terms to identify the more representative covariates in each group that contribute most to modelling the outcome. On the other hand, SCA is a Bayesian non-parametric approach that represents the patients in a clustered latent space and encourages latent representations to behave as a mixture of distributions, following a Dirichlet Process structure via a distribution matching approach, to provide clusters (subpopulations of patients) with distinct risk profiles.

4.2.3 Post-training analysis

Following the training phase, once the best set of hyperparameters had been selected (see Section 2 of Appendix B for further details), the model was retrained on the whole dataset, and a K-means clustering on the latent representations generated by the VAE was performed using the Euclidean distance metric. Importantly, these latent variables were devoid of the clinical features and the survival outcome, preserving their nature as a distilled representation of genetic and cytogenetic markers.

To determine the optimal number of clusters k , we selected the one maximizing the quantity:

$$Q_k = Silh_{mean}(k) * \frac{N_{min}(k)}{N_{max}(k)}$$

where $Silh_{mean}$ is the mean Silhouette score [92], N_{max} and N_{min} are the sizes of the largest and the smallest clusters, respectively. The ratio behind this optimization was to maximize the cluster quality, measured by the Silhouette score, while simultaneously avoiding the presence of excessively small or large clusters.

Once the optimal clustering was determined, survival curves were then generated to evaluate the clinical validity of these clusters, and gene mutation frequencies were examined to understand the underlying genetic and cytogenetic landscape.

4.3 Results

The analysis of VAE-Surv’s performance is two-fold, focusing on both the accuracy of survival predictions and the quality of the genetic-based clusters.

4.3.1 VAE-Surv prediction performance

The results of the survival predictions on the cross-validation sets are summarized in Table 4.1. The performance of the VAE-Surv model was systematically compared against widely adopted state-of-the-art survival analysis methods. Specifically, we benchmarked it against Cox Proportional Hazards regression models, implemented with various regularization strategies, including no regularization, Lasso (L1), Ridge (L2), and ElasticNet penalties [91]. Finally, the Survival Cluster Analysis (SCA) [72] framework was also employed as a baseline for comparative evaluation.

It is important to highlight that none of these methods provides the capability to treat genetic and cytogenetic data independently of clinical and demographic information. Among all comparisons, VAE-Surv consistently achieved either superior ($p=0.01$ vs. SCA, Wilcoxon Rank Sum Test) or comparable ($p \geq 0.16$ vs. all Cox models, Wilcoxon Rank Sum Test) median C-Index scores, highlighting its robust predictive accuracy for survival outcomes.

Notably, excluding the KL-term from the VAE loss function improves the performance. The effect of the KL-term on the learned latent space is clearly visible from Figure B.4 in the Supplementary Materials. Furthermore, in order to assess the reliability of the survival predictions obtained from VAE-Surv, we divided the predicted risk scores into 3 risk groups based on percentiles (Low, Medium and High risk) and plotted the corresponding Kaplan-Meier curves (see Figure 4.2). A pairwise log-rank test revealed that the survival distributions of the 3 groups are statistically different, with all the p -values < 0.005 .

	Median C-Index	95% CI
CoxPH	0.754	(0.726, 0.784)
CoxPH - L2	0.754	(0.729, 0.785)
CoxPH - L1	0.775	(0.745, 0.796)
Elastic-Net	0.775	(0.745, 0.796)
SCA	0.744	(0.725, 0.758)
VAE-Surv w/ KL	0.751	(0.719, 0.767)
VAE-Surv w/o KL	0.780	(0.747, 0.790)

Table 4.1: Comparison of models’ performance in terms of median C-Index and Confidence Intervals, evaluated in the 10-folds CV splits.

Additionally, we aimed to assess whether the dimensionality reduction of genetic features applied by the VAE module provides a significant advantage in the prediction performance over other linear encoding representations of the input data. To this end, we conducted a Principal Component Analysis (PCA) on the raw input genetic data. We then concatenated the first two principal components, which accounted for an

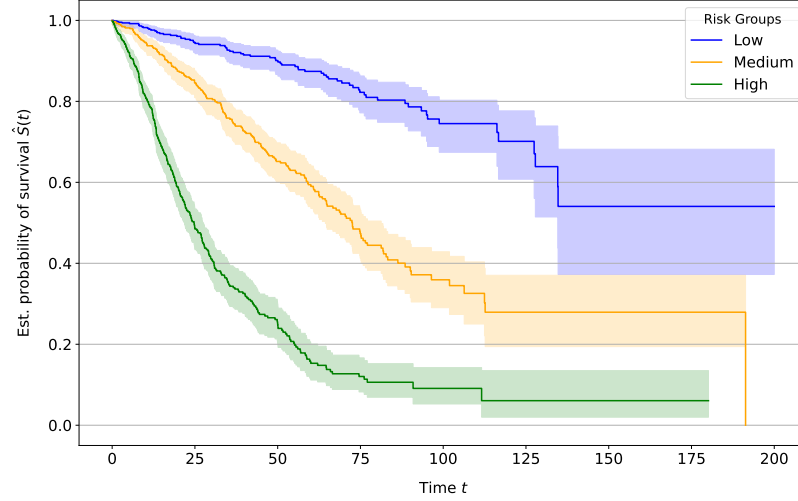


Figure 4.2: Kaplan-Meier curves stratified by risk groups based on percentiles.

explained variance of 21% and 15%, respectively, with the clinical covariates. This concatenated vector was subsequently used to train the different CoxPH models.

The performance outcomes of these models are illustrated in Figure 4.3. VAE-Surv consistently exhibits a higher C-Index. This comparative analysis supports the efficacy of our VAE-Surv framework in providing more accurate and consistent survival risk assessments, outperforming the linear Cox models, especially when high-dimensional genetic and cytogenetic data are considered.

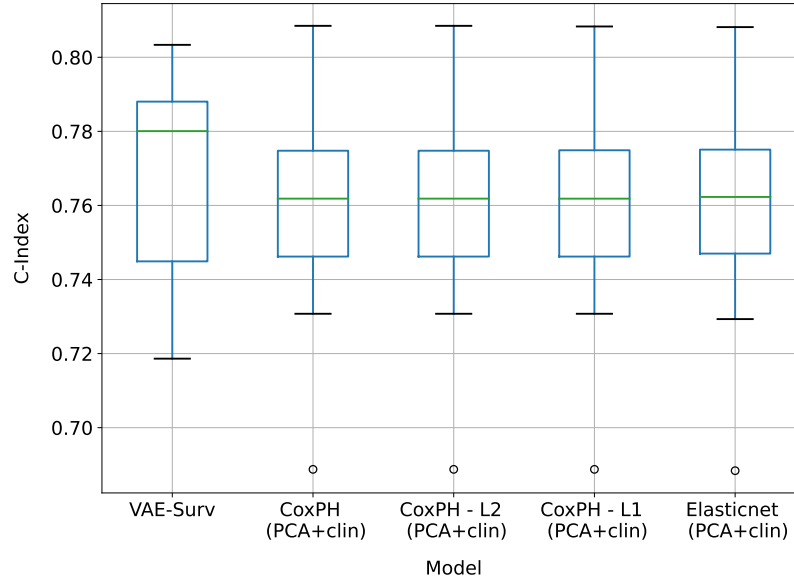


Figure 4.3: Comparison of VAE-Surv and CoxPH models using PCA-reduced genetic features and clinical covariates.

4.3.2 VAE-Surv latent representation and clustering performance

The dimensionality of z was optimized as a hyperparameter of the model and fixed to $z_{dim} = 7$. We employed K-means clustering on the z space to explore the utility of

the latent representation created by the VAE. The optimal number of clusters was determined according to the quality metric Q_k . As shown in Supplementary Figure B.1, the optimal configuration that ensures the maximum efficiency is when $k = 9$.

A t-SNE visualization of the latent representation points, coloured by the assigned cluster, is displayed in Fig. 4.4, left panel. The Kaplan-Meier survival curves, displayed in the right panel of Figure 4.4, reveal distinct survival profiles, thereby validating that the clusters obtained from the latent space (whose creation was not survival-informed) provide meaningful stratification in the context of patients’ survival.

To evaluate the effectiveness of the clustering on the VAE-Surv latent space, we compared the K-means approach to two alternative methods: (1) K-modes clustering, applied directly to the raw genetic and cytogenetic data, and (2) the Survival Cluster Analysis (SCA) framework [72], which integrates survival data into its clustering process. For both K-modes on the input space and SCA, the optimal number of clusters was determined as $k = 3$ based on the predefined quality metrics Q_k (Supplementary Figures B.3, B.2). As summarized in Table 4.2, K-means clustering on the VAE-Surv latent space outperformed both K-modes and SCA in terms of the Silhouette score and the Q_k metric. Importantly, while SCA integrates survival data into the clustering process, it does not leverage genetic and cytogenetic data as effectively as the latent representations created by VAE-Surv, as evident from its lower performance metrics.

	k	Q_k	Silhouette
K-modes on input space	3	0.041	0.15
Survival Cluster Analysis	3	0.035	0.26
K-means on VAE latent space	9	0.046	0.32

Table 4.2: Performance comparison including K-modes clustering on the input space, Survival Cluster Analysis, and K-means clustering on the latent representations learned by VAE-Surv

4.3.3 Biological Consistency of the MDS clusters

To investigate the biological consistency of the clusters defined by the genetic and cytogenetic features, we focused both on the distribution of MDS subtypes within the clusters and on the mutational profiling of each group. Supplementary Figures B.6 and B.7 show the number of mutations observed and the median values of the most relevant clinical attributes per cluster, respectively.

The clustermap reported in Figure 4.5 and the Sankey plot displayed in Supplementary Figure B.8 compare the identified clusters with the state-of-the-art MDS subtypes from WHO 2016 classification, based on both genetic and clinical features (including morphological criteria and haematologic parameters). The colour intensity in each cell of the clustermap correlates with the percentage of patients in a given cluster belonging to a specific MDS subtype. While the clustermap explains the composition of each cluster according to the WHO MDS classification, the Sankey plot highlights how the MDS subtypes are distributed across all the new groups.

Focusing on each cluster, we found that cluster 6 collects patients with both subtypes of MDS with Ring Sideroblasts (RS-MLD and RS-SLD), which are not significantly present in the other clusters, as can also be noticed in the Sankey plot. Cluster 2

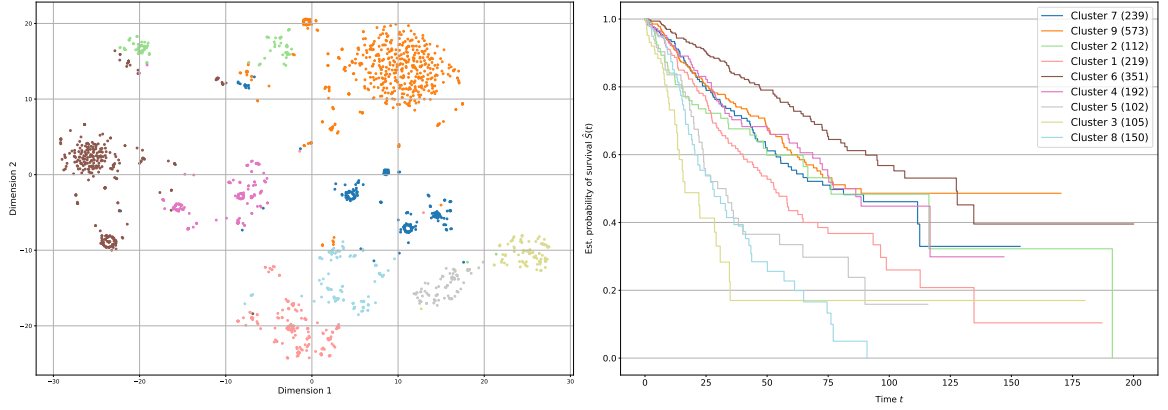


Figure 4.4: Left panel: t-SNE representation of the patients' genetic and cytogenetic latent space learned from VAE, coloured according to the cluster assignment. Right panel: Kaplan-Meier survival curves stratified by the identified cluster. The size of each group is specified in the legend.

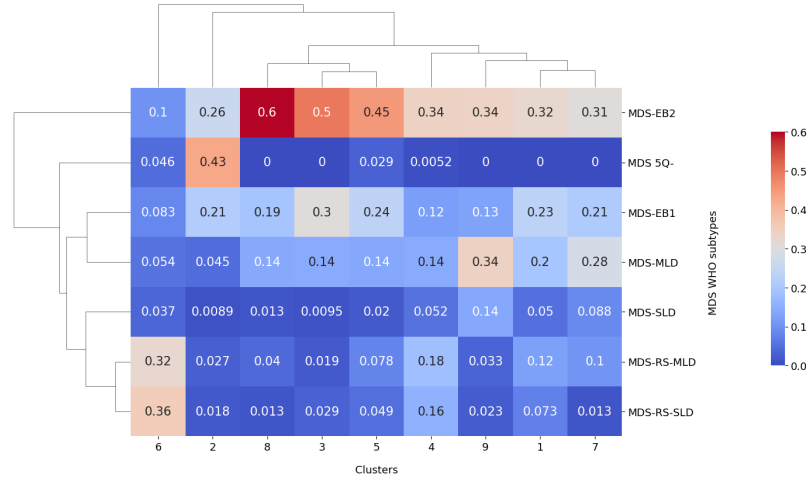


Figure 4.5: Clustermap showing the distribution of WHO 2016-defined MDS subtypes (y-axis) within each of the nine clusters identified (x-axis). Each column sums to 1. The colour intensity correlates with the percentage of patients in a given cluster belonging to a specific MDS subtype.

has a composite characterization, including Excess Blasts-1 (EB1), Excess Blasts-2 (EB2) and MDS with deletions in the long arm of chromosome 5 (5Q-) subtypes. It is worth highlighting that the MDS subtypes EB2, EB1 and the multilineage dysplasia (MLD) are spread across the clusters, with clusters 8, 3 and 9 represented mainly by these classes, respectively (Fig. 4.5). Therefore, it seems that some subtypes from WHO-based classification can be explained by potential subgroups identified by VAE-Surv. Indeed, focusing on the mutational landscape, the VAE-Surv clusters are represented by different underlying genetic features. Clusters 1, 6, 4, 8, 5 and 3 are strongly characterized by ASXL1, SF3B1, DNMT3A, RUNX1, Gainofchr8 and Lossofchr7ordel7q alterations, respectively, as displayed in Fig. 4.6a. This clustermap associates each identified cluster with the frequency of mutated genes or cytogenetic alterations in that cluster. Cluster 7 shows higher frequencies of SRSF2 and TET2 gene mutation with respect to the other groups, while cluster 2 is enriched with del5q, TP53 and Lossofchr5ordel5qPLUSother alterations. Notably, TP53 mutation is also frequent in cluster 3 (which is strongly defined by the Lossofchr7ordel7q alteration). In

perform a fair comparison of the results, the final model was tested using a 1000 times repeated bootstrap with resampling technique, ensuring that, for each repetition, the censoring rate matched exactly the one of the training cohort (71.1%). The resulting median C-Index was equal to 0.74 (Standard Deviation = 0.01). Further details of the performance on the validation cohort are reported in Appendix B (Sections 7,8).

Then, a Logistic Regression classifier was trained on the latent space representation of the GenoMed4all patients and applied to the IWG-PW cohort to assign its patients to the previously defined genetic clusters. The performance of the classifier were assessed by creating random internal test sets within the training cohort and repeating the random split 100 times for statistical consistency. The results in terms of accuracy and Matthew’s Correlation Coefficients are reported in Table 4.3. The latent representation of the training and validation cohorts are shown in Supplementary Figure B.10.

	Mean	Standard Deviation
Accuracy	0.995	0.004
MCC	0.994	0.005

Table 4.3: Logistic Regression average performance for the cluster assignment task on the 100-times repeated internal test sets.

Finally, we compared the interpretability of the results obtained by the survival models of VAE-Surv applied on both the Genomed4all and the IWG-PM cohorts. To this aim, we estimated the SHAP values, which quantify the contribution of each feature to the model’s prediction for an individual patient’s overall survival outcome, which in this case is the relative risk. Originating from cooperative game theory, SHAP values are based on the Shapley values, which assigns a fair distribution of the ‘payout’ (prediction effect) among the ‘players’ (features) [95]. By integrating SHAP with our VAE-Surv model, we can dissect the nonlinear, complex interactions captured by the deep learning architecture, providing insights into how specific genetic and clinical variables influence survival predictions.

The shapley values have been calculated for the model retrained on the entire GenoMed4all cohort. Summary plots for the top-15 ranked features according to SHAP are shown in Supplementary Figures B.11 and B.12, for training and validation cohorts respectively. It is possible to observe a consistency of the main clinical/genetic features involved in both models, confirming the reproducibility and generalizability of the results in different patients’ cohorts.

4.5 Discussion

This study provides a robust computational framework for patient stratification in Myelodysplastic Syndromes (MDS) by integrating a Variational Autoencoder (VAE) with a DeepSurv model. The framework outperformed traditional CoxPH models in survival prediction tasks, reinforcing that deep learning can add value to understanding and managing complex diseases like MDS.

The proposed model effectively exploits the rich genetic and cytogenetic landscape of MDS for feature extraction and subsequent risk stratification. The latent representa-

tions obtained by the VAE served as an informative summary of the genetic landscape, increasing the granularity of our survival model. Removing the KL-term in the VAE loss function allowed a local interpretation of the feature space, helping to capture the inherent heterogeneity of the disease.

The application of K-means clustering on the latent space identified distinct patient groups, each with unique genetic and clinical characteristics. The visualization of these clusters in relation to WHO 2016 MDS subtypes indicates that our framework is capable of capturing existing clinical categorizations while also suggesting novel, data-driven patient groups. In fact, according to the modern oncological view, the WHO 2016 MDS classification is considered too coarse as it groups genetically different cancer subtypes into the same class [83, 96, 97]. Our method, VAE-Surv, refines the WHO classification and the non-uniform distribution of MDS subtypes across the new clusters implies that the model recognizes both well-defined and less-recognized patients' subgroups, potentially facilitating individualized treatment strategies. In contrast, we found that input data clustering reduces the number of cancer type classes to only 3, thus highlighting its inability to capture subtle distinctions in the input space. Furthermore, we also characterized the mutational landscape within each cluster, providing a more specific understanding of genetic factors driving different MDS subtypes.

Regarding the limitations of our study, its retrospective nature represents the primary limitation, including the potential for selection bias inherent in a single cohort and the high heterogeneity of the disease under investigation. To address these limitations, the model was validated in a second large cohort of patients. When comparing the two cohorts, the presence of significant differences, particularly in the WHO categories and risk classes, further highlights the heterogeneity of the disease under investigation. At the same time, these differences suggest the potential to generalize the findings to patients with characteristics distinct from those of the training cohort.

On the external validation cohort consisting of 2,384 patients [86], VAE-Surv model performance resulted in an average C-Index equal to 0.74 and a cluster assignment accuracy of 0.995. These scores display the robustness and consistency of the model. Moreover, as shown by the SHAP analysis, the features that contribute most to the model's prediction are almost the same on both training and validation cohorts.

Finally, our findings align well with the previous results [83]. In particular, our analysis confirmed the pivotal role of gene SF3B1 mutation in improving the outcome and markedly characterizing a specific group of patients. In both works, patients with a flat mutational pattern fall into a well-defined cluster whose survival rate is higher than the average. The association of TP53 mutation with complex karyotypes, as highlighted by Bersanelli et al., is confirmed. However, our analysis delineates that this gene is related to extremely poor survival outcomes only when combined with Lossofchr7ordel7q (cluster 3), and not in cluster 2, where it co-occurs with del5q and Lossofchr5ordel5qPLUSother alterations.

4.6 Conclusion and future work

In summary, our VAE-Surv framework demonstrates the power of deep learning in handling the intricate genetic landscape of MDS, offering a novel, robust methodology

for patient stratification and survival prediction. It stands to contribute significantly to personalized medicine in the context of haematologic disorders, facilitating more accurate diagnosis and tailored treatment plans for MDS patients.

In future works, the integration with additional omics data like transcriptomics or epigenetics could be easily embodied in the model, hopefully providing a more comprehensive view of the disease. Moreover, the VAE-Surv architecture is not limited to the MDS case study. The tool can be extended to the study of other haematologic or even solid tumour malignancies where both clinical and genetic mutations data are available. By retraining the model on different disease-specific datasets, clinicians and researchers could gain similarly actionable insights for patient stratification and survival prediction across a wider array of conditions.

Chapter 5

Modelling Survival as a Stochastic Diffusion Process

Survival analysis, a fundamental statistical approach in biomedical research, is traditionally based on models such as the Cox Proportional Hazards (CoxPH) model or parametric survival distributions. However, in certain contexts, particularly in biological and medical applications where disease progression or system failure can be modeled as a stochastic process, First Hitting Time (FHT) models offer a more mechanistic and flexible alternative [98]. These models describe the time at which a stochastic process, often modeled as a Brownian motion, first reaches a predefined threshold, which can represent events such as disease onset, relapse, or death [99, 100].

5.1 Overview of Wiener processes

Brownian motion, also known as the Wiener process, is a fundamental stochastic process widely used in physics, finance, and biology to model systems subject to random fluctuations. A *Wiener process* is a continuous-time stochastic process $\{W_t : t \geq 0\}$ that satisfies the following properties:

1. **Initial Value:**

$$W_0 = 0 \quad \text{almost surely.}$$

2. **Independent Increments:** For any times $0 \leq t_1 < t_2 < \dots < t_n$, the increments

$$W_{t_2} - W_{t_1}, \quad W_{t_3} - W_{t_2}, \quad \dots, \quad W_{t_n} - W_{t_{n-1}}$$

are independent random variables.

3. **Stationary Increments:** The distribution of each increment depends only on the length of the time interval. In particular,

$$W_t - W_s \sim \mathcal{N}(0, t - s) \quad \text{for } 0 \leq s < t.$$

This means the increment over $[s, t]$ is normally distributed with mean 0 and variance $t - s$.

4. **Almost Sure Continuity:** With probability 1, the sample paths $t \mapsto W_t$ are continuous functions of t .

The Wiener process W_t serves as the foundation for modeling stochastic dynamics. A key extension of the Wiener process is the general stochastic process $X(t)$, that can be expressed as:

$$\mathbf{X}(t) = \mathbf{x}_0 + \mu t + \sigma \mathbf{W}(t)$$

where $\mathbf{x}_0$ is the initial state, μ is the drift term and σ is a volatility coefficient determining the amplitude of the stochastic fluctuations. This process evolves according to a stochastic differential equation (SDE):

$$d\mathbf{X}_t = \mu dt + \sigma d\mathbf{W}_t \quad (5.1)$$

If $\mu = 0$ the process is a pure isotropic diffusion (unbiased random motion); instead, if $\mu \neq 0$ the process exhibits a drift, leading to biased diffusion where $X(t)$ systematically trends toward a specific direction.

In general, the drift term μ and the diffusion coefficient σ may be functions of time t and/or the state $X(t)$, leading to a more general formulation of the stochastic differential equation:

$$d\mathbf{X}_t = \mu(\mathbf{X}_t, t) dt + \sigma(\mathbf{X}_t, t) d\mathbf{W}_t$$

In this case, the evolution of $X(t)$ depends not only on an external constant force and a fixed level of randomness, but also on its own state or time, allowing for complex behaviors such as state-dependent diffusion (where fluctuations are stronger or weaker depending on the value of X_t) and time-dependent drifts (where the tendency to increase or decrease varies over time). These generalizations are widely used in modeling various natural and biological processes, including gene regulation, neural activity, and financial markets.

However, in this work, we focus on the simplified case where both the drift and diffusion coefficients are constant and the process is one-dimensional. This assumption allows for analytical tractability while still capturing the essential stochastic nature of survival phenomena modeled through First Hitting Time (FHT) processes. A possible generalization to multi-dimensional processes will be proposed for a special class of survival analysis tasks (competing events) in the last section of this chapter.

The Fokker-Planck Equation: Evolution of Probability Distributions

A fundamental aspect of stochastic processes is the evolution of their probability distribution over time. Instead of describing individual sample paths of $X(t)$, one can study the behavior of the probability density function $p(x, t)$, which describes the likelihood of finding the process at position x at time t . The time evolution of $p(x, t)$ is governed by the Fokker-Planck equation (also known as the Kolmogorov forward equation), which is derived from the underlying stochastic differential equation. For the process described by Equation 5.1, the associated one-dimensional Fokker-Planck equation is:

$$\frac{\partial p(x, t)}{\partial t} = -\mu \frac{\partial p(x, t)}{\partial x} + D \frac{\partial^2 p(x, t)}{\partial x^2}, \quad (5.2)$$

where D is the diffusion coefficient and is defined as:

$$D = \frac{\sigma^2}{2}$$

Equation 5.2 is a partial differential equation (PDE) that can be solved using different techniques, depending on the boundary conditions and initial conditions.

For a free diffusion process (without boundaries), starting from an initial condition $p(x, 0) = \delta(x - x_0)$ (i.e., the process is initially located at x_0), the solution of the Fokker-Planck equation is:

$$p(x, t | x_0) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(-\frac{(x - \mu t - x_0)^2}{4Dt}\right) \quad (5.3)$$

This represents a Gaussian distribution with:

- Mean $\mathbb{E}[X_t] = x_0 + \mu t$ (drift shifts the probability density),
- Variance $\text{Var}[X_t] = \sigma^2 t$ (diffusion spreads the probability density over time).

5.2 First Hitting Time for Survival

In many survival scenarios, particularly in biomedical applications, the event of interest occurs when a stochastic process $X(t)$ first reaches a predefined threshold. This threshold may represent a physiological limit (e.g., a critical biomarker level) or a system failure (e.g., an organ reaching a critical state). The time at which this happens is called the First Hitting Time (FHT).

Formally, given a process evolving according to the stochastic differential equation 5.1, the first hitting time T for a fixed absorbing boundary at $x = 0$ is defined as:

$$T = \inf\{t > 0 \mid X(t) = 0\}$$

This formulation captures a wide range of survival phenomena, including disease progression, failure of biological systems, and crossing of regulatory thresholds in genetic processes.

To determine the probability of survival over time, we analyze the Fokker-Planck equation with an absorbing boundary condition at $x = 0$. That is, we impose:

$$p(0, t) = 0, \text{ for all } t > 0,$$

which reflects the fact that once the process reaches $x = 0$ it is absorbed (i.e., the event has occurred). To solve the Fokker-Planck equation 5.2 with absorbing boundary conditions we use the method of images, obtaining:

- if $\mu = 0$:

$$p(x, t | x_0) = \frac{1}{\sqrt{4\pi Dt}} \left[\exp\left(-\frac{(x - x_0)^2}{4Dt}\right) - \exp\left(-\frac{(x + x_0)^2}{4Dt}\right) \right] \quad (5.4)$$

- if $\mu \neq 0$:

$$p(x, t | x_0) = \frac{1}{\sqrt{4\pi Dt}} \left[\exp\left(-\frac{(x - x_0 - \mu t)^2}{4Dt}\right) - \exp\left(-\frac{\mu x_0}{D}\right) \exp\left(-\frac{(x + x_0 - \mu t)^2}{4Dt}\right) \right] \quad (5.5)$$

The Survival function, representing the probability of remaining event-free (i.e., within the absorbing boundaries) in time, can then be computed by integrating the solution of the Fokker-Planck equation:

$$S(t) = \int_0^\infty dx p(x, t | x_0) \quad (5.6)$$

Then, the First Hitting Time distribution is the density defined in 2.3. For the one-dimensional Brownian process with constant diffusion, we obtain:

- if $\mu = 0$:

$$S(t) = \operatorname{erf} \left(\frac{x_0}{\sqrt{4Dt}} \right) \quad (5.7)$$

$$f(t | x_0) = \frac{x_0 D}{2\sqrt{\pi}} \frac{e^{-\frac{x_0^2}{4Dt}}}{(Dt)^{3/2}} \quad (5.8)$$

- if $\mu \neq 0$:

$$S(t) = \Phi \left(\frac{x_0 + \mu t}{\sqrt{2Dt}} \right) - \exp \left(-\frac{\mu x_0}{D} \right) \Phi \left(\frac{\mu t - x_0}{\sqrt{2Dt}} \right) \quad (5.9)$$

$$f(t | x_0) = \frac{x_0}{\sqrt{4\pi Dt^3}} \exp \left(-\frac{(x_0 + \mu t)^2}{4Dt} \right) \quad (5.10)$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. The FHT density obtained in the absence of drift 5.8 is a *Lévy* distribution; when drift is present we get an *Inverse Gaussian* distribution (5.10).

The Lévy distribution for purely diffusive processes implies that first passage times have no characteristic scale—events can happen very quickly, but a long delay is also possible. The Inverse Gaussian distribution for drifted processes reflects asymmetry in first passage times due to a deterministic tendency to reach the absorbing boundary. These distributions may potentially provide mechanistic interpretations of survival processes in cancer progression (e.g., time to tumor metastasis), neuroscience (e.g., neuronal firing thresholds), molecular biology (e.g., gene expression reaching activation thresholds), medical diagnostics (e.g., biomarkers reaching critical levels).

5.3 Parameter estimation for FHT Survival models with Neural Networks

Accurately modeling survival times in biomedical applications requires estimating the underlying parameters that govern the stochastic dynamics of the event of interest. In the case of First Hitting Time (FHT) models, the survival process is assumed to follow a stochastic diffusion equation, whose first hitting time distribution $f(t)$ is either a Lévy distribution (for pure diffusion) or an Inverse Gaussian distribution (for drifted diffusion) [101, 102, 103]. In this section, we propose a parametric approach where the key parameters of the stochastic process (initial position x_0 , diffusion coefficient D , and drift μ if present) are learned directly from data using a neural network. Given a dataset with time-to-event information and corresponding covariates, we leverage a deep learning model to predict the parameters that best describe the observed survival

times by maximizing the likelihood (or some other loss function) of the assumed FHT distribution.

Instead of estimating the parameters explicitly from data using traditional statistical methods, we propose using a neural network to learn the mapping from input covariates $\mathbf{z}$ (e.g., patient features, genomic information, clinical variables) to the parameters of the chosen FHT distribution. The neural network takes as input a set of covariates $\mathbf{z}$ and outputs the estimated parameters:

$$(x_0, D, \mu) = \mathcal{F}_\theta(\mathbf{z}),$$

where $\mathcal{F}_\theta$ is the neural network with trainable parameters θ . A major advantage of using a neural network-based approach over traditional survival models is that we estimate individualized values for x_0 , D and μ for each patient. That is, instead of assuming fixed parameters for the entire population, the network predicts a unique set of survival parameters for each individual based on their covariates. This enables the model to capture patient-specific variations, leading to a more flexible and realistic survival prediction.

5.4 Experimental setting and models comparison

To systematically evaluate the performance of the proposed First Hitting Time (FHT) survival models, we implemented a robust experimental pipeline using a diverse set of survival datasets. The goal was to assess whether Brownian motion-based FHT models—yielding Lévy and Inverse Gaussian first hitting time distributions—provide superior or comparable predictive performance compared to traditional fully-parametric survival models, namely: Exponential, Weibull and LogNormal distribution.

To perform this analysis we used SurvSet [104], an open-source time-to-event dataset repository. The data in SurvSet are consistently formatted so that a single preprocessing method will work for all datasets. SurvSet currently has 76 datasets which vary in dimensionality, time dependency, and background (the majority of which come from biomedicine). Among all the datasets in the repository, we selected those with at least 500 samples, ensuring a sufficiently large dataset to enable effective training of a nonlinear neural network while maintaining statistical reliability in performance evaluation. This resulted in a pool of 31 survival datasets in total. Further details can be found in [104].

For each dataset and for each model we applied a standardized pipeline for validation. Specifically, we applied a 5-fold cross-validation scheme (stratified on censoring information) and hyperparameter optimization using the mean Concordance Index as evaluation metrics. The optimized hyperparameters were: number of epochs, learning rate, dropout level, activation function, use of batch normalization and weight decay. Moreover, we tested 3 different loss functions: full log-likelihood 2.10, partial log-likelihood 2.14 and Brier loss 2.26. For the Inverse Gaussian model, we assumed a positive drift ($\mu > 0$), meaning that the stochastic process exhibits a systematic tendency to move toward the absorbing boundary at $x = 0$. This assumption implies that, on average, each sample follows a trajectory that gradually progresses toward the event occurrence, rather than diffusing away indefinitely. The comparison results are shown in Figures 5.1 and 5.2 in terms of best C-Index and Brier score, obtained for each loss. The performance distributions with datasets stratified by sample size and feature size are shown in Figures C.1 and C.2 of Appendix C, respectively.



Figure 5.1: Mean C-Index for the 5-fold cross validation, for different loss function, model and datasets.

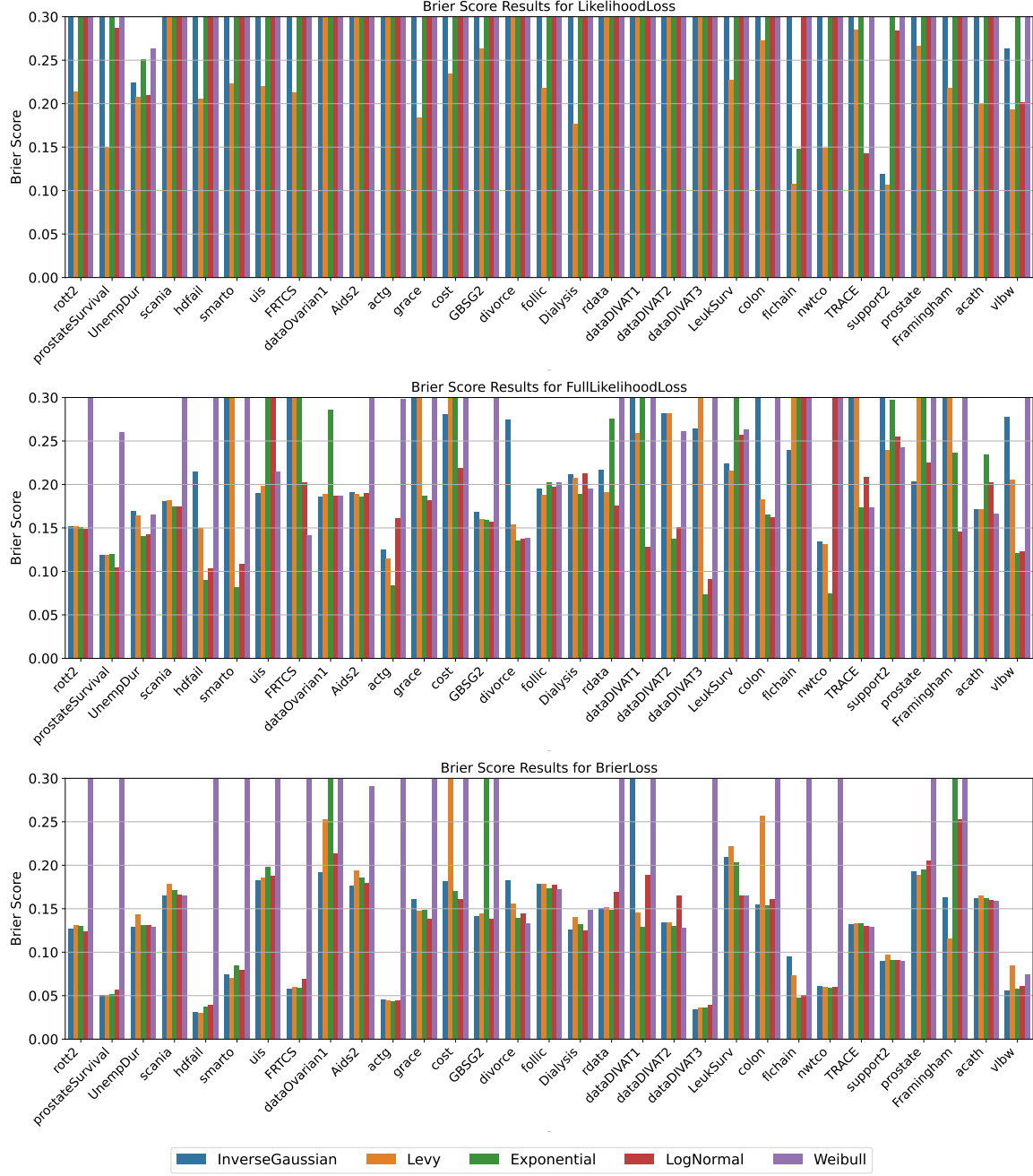


Figure 5.2: Mean integrated Brier score for the 5-fold cross validation, for different loss function, model and datasets.

5.5 Discussion

The comparison of different survival models across 31 datasets reveals an overall heterogeneous performance, highlighting how different datasets exhibit distinct survival dynamics that are better captured by different modeling approaches. The results presented in Figures 5.1 and 5.2 provide insights into the predictive capability of each model, evaluated through C-Index (ranking accuracy) and Brier Score (calibration).

From the C-Index and Brier Score results, it is evident that no single model dominates across all datasets. Some datasets favor traditional parametric models (Exponential, Weibull, LogNormal), while others perform equally well or better under the First Hitting Time (FHT) models (Lévy and Inverse Gaussian survival models). This suggests that different underlying survival mechanisms are at play across datasets, making it crucial to test multiple models rather than assuming a single best-fit distribution. The performance variability may stem from dataset-specific characteristics such as censoring rates, feature distributions, and survival curve shapes. In addition, when analyzing performance stratified by the number of samples and dataset features (as shown in Figures C.1 and C.2), no clear trend emerges. This suggests that dataset dimensionality does not significantly impact either the diffusion-based models or the other parametric models.

Moreover, while models such as Weibull and LogNormal often provide a reasonable fit for survival data, they rely on predefined functional forms that may not be well-suited for processes governed by stochastic variability. In contrast, FHT models naturally incorporate random fluctuations in survival time, making them particularly useful for applications where biological systems evolve via random accumulation of damage (e.g., neurodegenerative diseases, molecular degradation) or event occurrence follows an activation threshold (e.g., genetic mutations triggering cancer progression).

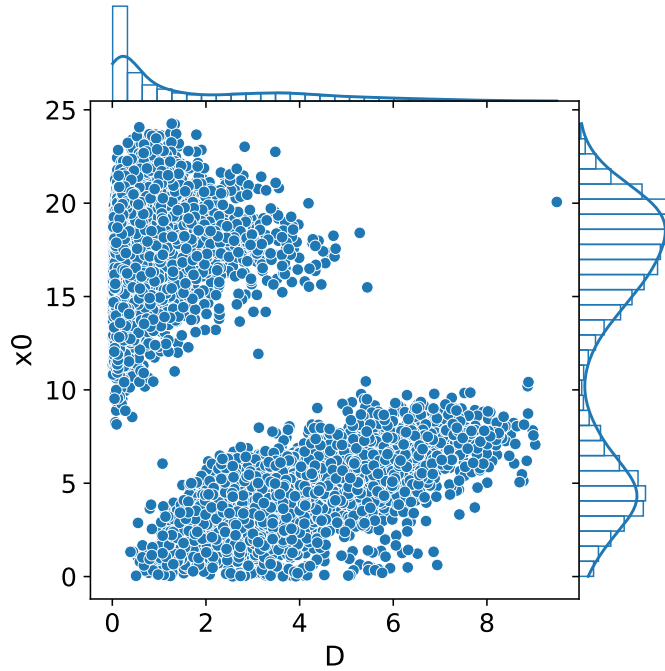


Figure 5.3: x_0 and D learned distributions for *support2* dataset when using Brier loss and Levy model.

Unlike traditional parametric models, FHT models offer a direct interpretation in terms of stochastic processes, allowing us to analyze learned diffusion dynamics for each patient. Specifically, our neural network-based approach estimates patient-specific parameters x_0 , D , μ which describe, respectively, the initial state of the survival process (e.g., baseline health, initial risk factor levels), the diffusion coefficient (reflecting the random variability in survival time) and the drift, representing a systematic trend toward event occurrence (e.g., disease progression). To further explore the interpretability of FHT models, it is also possible to analyze the distribution of the learned parameters for a specific dataset. This potentially allows to identify patient subgroups with different survival dynamics. For example, patients with higher diffusion D tend to have more unpredictable survival times, whereas those with lower D exhibit more deterministic survival trajectories. Higher x_0 values suggest a longer time to reach the event, while lower values indicate patients closer to event realization. As an example, Figure 5.3 shows the learned distribution of D and x_0 for the *support2* dataset. It is possible to clearly distinguish 2 groups of patients based on the initial "health state" x_0 , with a long-tailed distribution of the Diffusion parameter. By analyzing the learned parameters at the patient level, we can move beyond mere risk ranking (as done in Cox models) and toward a mechanistic understanding of individual survival processes.

Chapter 6

Related Projects

6.1 MOSAIC: An Artificial Intelligence–Based Framework for Classification and Prognostic Assessment in Rare Cancers

Background Myeloid neoplasms (MN) present clinical and molecular heterogeneity and therefore a risk-adapted treatment strategy is mandatory. In MN, classification and prognostic tools based on clinical and morphologic criteria are being complemented by introducing genomic features. The clinical implementation of next-generation classifications and prognostic systems requires the availability of a robust methodological framework together with a solution to provide access to these technologies for clinicians.

Aims Machine learning (ML) and Deep Learning (DL) approaches produce powerful predictive models and offer explainable solutions to assure full interpretability of a model when applied in clinical settings. Here we provide a comprehensive assessment of explainable ML/DL-based methods for classification and prognostic assessment of MN, and we developed a solution to apply these methods across different clinical Centres through a Federated Learning (FL) approach.

Methods We analysed two cohorts of patients from GenoMed4All consortium with myelodysplastic syndrome (MDS), $n=2,043$ and $n=2,384$, with available clinical and molecular features to train and validate the models. Methods were then applied to other MN, i.e. acute myeloid leukemia (AML, $n=1154$) and chronic myelomonocytic leukemia (CMML, $n=1037$). We stratified patients by two clustering approaches based on Hierarchical Dirichlet Process (HDP) and HDBSCAN combined with UMAP data reduction. We trained a Random Forest (RF) classifier to assign new patients to the existing clusters, considering Balanced Accuracy (BA) and Cohen’s K (CK) as performance metrics. We then compared different survival prediction methods: CoxPH model (and its penalized version), Random Survival Forests, DeepCox, Gradient Boosting and XGboost survival methods. Models’ explainability was performed through SHapley Additive exPlanations approach (SHAP). C-index was used to evaluate the models performance. Finally, we developed a Federated Learning (FL) environment together with an imputation approach to handle missing values by a deep decoder model.

Results In MDS training cohort, we identified 18 and 8 clusters by using HDBSCAN and HDP, respectively (Figure 1). We measured the average Silhouette Coefficient on the data space obtaining the following performance in terms of classification task: HDBSCAN (BA:92.7±1.3%, CK:92.1±1.4%) and HDP (BA:85.8±0.8%, CK:83.3±0.9%). Similar distributions were observed when focusing on the validation cohort. Model explainability analysis (SHAP) showed that in both populations similar features drive patients’ classification.

Comparison of survival prediction for MDS is displayed in Figure 2, showing the models’ performance in the two cohorts considering demographics, clinical, cytogenetics and genomic features. Non-linear ML/DL-based methods outperformed classical CoxPH-based approaches without requiring huge data pre-processing. Moreover, all the models showed higher C-indices with respect to that of conventional IPSS-R score. SHAP analysis showed similar feature importance ranking for both training and validation cohorts. Models were then applied to AML and CMML cohorts, showing consistent results across different types of MN.

Finally, we aimed to develop a federated learning (FL) solution (FedAvg, with a deep decoder model for missing data imputation) to favour a wide clinical implementation of the models. Data were collected to a single server and used to build and train a centralized model. Using global data training was expected to improve the model efficiency. This approach also ensured that the data in each node adhere to data privacy policies. We implemented CoxPH model in a setting of 3 nodes (Centers) respectively contributing to 60%, 30% and 10% of the training data. We observed that the poor node (i.e., node contributing to 10% of data) benefits from FedAvg with respect to working on an isolated setting (C-index 0.63 vs. 0.54). The centralized model trained on the whole dataset presented the highest efficiency (C-index 0.74).

Conclusion Machine Learning/Deep Learning approaches produce explainable and robust solutions to optimize classification and prognostic assessment in MN, as a basis for personalized medicine programs in these disorders. Federated learning algorithms allow a wide clinical implementation of the models by ensuring high performance and data protection.

6.2 Federated AdaBoost for Survival Analysis

FedSurvBoost is a novel federated learning approach designed for survival analysis, leveraging an adaptation of the AdaBoost.F algorithm [105]. The primary motivation behind this work is to enhance predictive performance in survival analysis while preserving data privacy, making it particularly relevant for sensitive domains such as healthcare. Traditional survival analysis techniques often require centralized data collection, which poses significant privacy risks, especially under regulatory constraints like GDPR. Federated learning addresses this issue by enabling collaborative model training across multiple institutions without the need to share raw patient data. Instead, only model parameters are exchanged, ensuring privacy and compliance with data protection regulations.

The core innovation of FedSurvBoost lies in its modification of AdaBoost.F to support survival analysis, which is inherently a regression problem rather than a classification task. By eliminating the dependence on class coefficients and redefining how errors and weights are computed, the algorithm becomes well-suited for handling censored survival data. Unlike existing federated survival analysis models such as FedSurF [106], which primarily rely on decision trees, FedSurvBoost is model-agnostic and can integrate various weak learners, including Cox Proportional Hazards models, Random Survival Trees, and DeepSurv neural networks.

The effectiveness of FedSurvBoost is demonstrated through extensive experimentation on two widely used survival datasets, METABRIC and SUPPORT. The results show that FedSurvBoost consistently outperforms state-of-the-art federated survival analysis methods, achieving higher concordance index (C-index) scores and lower Integrated Brier Scores (IBS), which measure predictive accuracy. The model scales efficiently as the number of federated clients increases, maintaining or even improving performance despite the distributed nature of the data. Notably, DeepSurv (PyCox) emerges as the best-performing weak learner within the FedSurvBoost framework, highlighting the adaptability of the approach to neural network-based survival models.

This work establishes FedSurvBoost as a powerful alternative to existing federated survival analysis methods, particularly in applications requiring strong predictive capabilities alongside privacy preservation. Future research directions include extending the approach to other boosting algorithms such as PreWeak.F and DistBoost.F and exploring its performance on non-IID (non-independent and identically distributed) datasets. By bridging the gap between privacy-preserving machine learning and survival analysis, FedSurvBoost paves the way for more robust and scalable solutions in time-to-event prediction tasks across multiple domains.

6.3 Influence of Model Structures on Predictors of Protein Stability Changes from Single-Point Mutations

In this study we explored the influence of using computationally modeled protein structures on the prediction of stability changes caused by single-point mutations. Understanding these stability changes is critical in molecular biology and medicine, as they can affect protein function and contribute to disease development. Our research compares experimental protein structures with those generated using two widely used computational modeling approaches, Modeller [107] and Rosetta [108], to evaluate their impact on predictive accuracy.

A key challenge in protein stability prediction is the availability of high-resolution experimental structures for both wild-type and mutant proteins. Computational modeling fills this gap, but the fidelity of these structures can significantly affect predictive performance. The study extends previous work by systematically analyzing how minor structural variations, resulting from a single amino acid substitution, impact different $\Delta\Delta G$ (stability change) predictors. Seven widely used predictors, including ACDC-NN [109], MAESTRO [110], DDGun [111], FoldX [112], Korpm [113], DDMut [114], and PoPMuSiC [115], were assessed using both experimental and modeled structures.

The findings reveal that predictors utilizing coarse-grained representations, such as ACDC-NN and DDGun, are relatively insensitive to the choice of structure, performing well regardless of whether experimental or computational models were used. In contrast, methods relying on detailed molecular interactions, like FoldX and PoPMuSiC, show significant sensitivity to the accuracy of the input structure. This suggests that for certain applications, selecting the right predictive tool is crucial when only computationally modeled structures are available.

The study also highlights differences between Modeller and Rosetta in terms of their impact on prediction accuracy. While both tools generate structures that are geometrically close to experimental ones, their influence on prediction outcomes varies. Rosetta-derived structures generally produce more reliable predictions for most models, whereas Modeller is faster and computationally more efficient. The results suggest that, depending on the choice of predictor, Modeller can serve as a viable alternative to Rosetta, especially in large-scale studies where computational resources are a concern.

Overall, the study underscores the importance of carefully selecting both protein modeling methods and stability prediction tools, particularly in scenarios where experimental structures are unavailable. The findings provide valuable insights for researchers seeking to optimize predictive performance in protein stability analysis while balancing computational efficiency and accuracy.

Chapter 7

Conclusion

In this thesis, we have developed and evaluated several novel computational frameworks that push forward the state-of-the-art in survival analysis by integrating advanced artificial intelligence techniques with privacy-preserving data sharing, multimodal data integration, and stochastic process modeling. Our work spans multiple complementary directions:

- **SYNDSURV: A Model-Agnostic Framework for Distributed Survival Analysis.** We introduced SYNDSURV, a simple yet effective approach that circumvents the infrastructural and methodological constraints of traditional federated learning. By generating synthetic survival data—where local nodes first simulate covariates via a Bayesian generative network and then predict time-to-event labels through an Accelerated Failure Time model—the framework allows for centralized training of survival models while preserving data privacy. Experiments on several publicly available survival datasets demonstrated that the performance of global models trained on synthetic data approaches that of centralized models and significantly outperforms models built on isolated datasets. Moreover, the method proved robust in a differential privacy setting, reinforcing its applicability in real-world, privacy-sensitive environments.
- **VAE-Surv: Integrating Genetic Data for Enhanced Prognosis Prediction in Myelodysplastic Syndromes.** Recognizing the limitations of conventional survival models in capturing the heterogeneity of complex diseases, we developed VAE-Surv, a framework that leverages a Variational Autoencoder to extract latent representations from high-dimensional genetic and cytogenetic data. By concatenating these latent features with clinical variables and feeding the enriched representation into a Deep Cox Proportional Hazards network, VAE-Surv achieves superior survival prediction performance and meaningful patient stratification. The framework was validated on large cohorts of MDS patients and externally on an independent dataset, consistently yielding higher C-Indices compared to baseline Cox models. Additionally, clustering on the latent space uncovered distinct subgroups with characteristic mutational profiles and divergent survival outcomes, thus providing new insights into disease biology and potential avenues for personalized treatment strategies.
- **Modelling Survival as a Stochastic Diffusion Process.** To further bridge the gap between mechanistic modeling and data-driven survival analysis, we

explored a framework that models survival as a First Hitting Time (FHT) problem derived from stochastic diffusion processes. By assuming that the time to an event corresponds to the first passage of a Brownian (or drifted Brownian) motion to an absorbing boundary, we derived analytical expressions for the survival function and density in terms of Lévy and Inverse Gaussian distributions. We then proposed a neural network-based approach to estimate patient-specific parameters (initial state, diffusion coefficient, and drift), thereby enabling a mechanistic interpretation of individual survival trajectories. Comparative experiments over a diverse set of survival datasets revealed that FHT models are competitive with—and in some cases outperform—traditional parametric survival models, particularly when the underlying survival mechanisms are driven by stochastic dynamics.

Collectively, the contributions of this thesis demonstrate that the integration of synthetic data generation for distributed scenarios, deep representation learning, and mechanistic stochastic modeling can substantially enhance survival analysis, even in the presence of censored and high-dimensional data. Our frameworks offer several practical advantages: they are model-agnostic, privacy preserving, and capable of integrating heterogeneous data sources. Moreover, by providing interpretable models—through techniques such as SHAP analysis and latent-space clustering—they support both robust prognostic prediction and biologically meaningful patient stratification.

Future Directions. Despite these advances, several promising directions remain for future research. Extending the proposed frameworks to handle more complex, multi-dimensional stochastic processes (e.g., for competing risks) and integrating additional data modalities (such as imaging or epigenetic data) could further enhance predictive performance and clinical utility. In addition, further exploration of advanced federated learning strategies and differential privacy techniques will be critical to ensure scalability and robustness in real-world, multi-institutional settings. In conclusion, this work lays a strong foundation for next-generation survival analysis methods that are both scientifically rigorous and clinically actionable, supporting the development of more effective, personalized strategies in the management of rare and complex diseases.

Appendix A

SYNDSURV Supplementary Materials

A.1 Datasets description

Here we provide a short description of the four publicly available survival dataset used for our study:

- **SUPPORT**: Study to Understand Prognoses Preferences Outcomes and Risks of Treatment [25]. The SUPPORT dataset resulted from a study conducted to describe a prognostic model to estimate survival over a 180-day period for 9,105 seriously ill hospitalized patients. Of these 9,105 patients, 6,201 patients (68.1%) were followed to death, with a median survival time of 58 days. We considered the following patient’s covariates: age, sex, years of education, income, whether the patient has cancer, diabetes, dementia, mean arterial blood pressure, counts of white blood cells, heart rate, respiration rate, temperature, the ratio of arterial oxygen partial pressure to fractional inspired oxygen (indicator of hypoxaemia), bilirubin levels, serum sodium concentration, arterial blood pH, glucose levels, urine, index of Activities of Daily Living (ADL) of the patient;
- **FLCHAIN**: a clinical dataset to study the relationship between serum free light chain (FLChain) and mortality [63]. The dataset includes demographic and laboratory variables alongside with the recruitment year and collects 6,524 patients aged 50 or greater; 7 covariates were considered: age, sex, the calendar year in which a blood sample was obtained, serum free light chain (kappa and lambda portions), the FLC group for the subject, serum creatinine, diagnosis with monoclonal gammopathy;
- **GBSG & Rotterdam**: a dataset that combines breast cancer data from the Rotterdam tumor bank [66] and patients enrolled by GBSG (German Breast Cancer Study Group) in a randomized clinical trial that studied the effects of chemotherapy and hormone treatment on survival rate [67]. The final dataset includes 3,668 patients. We considered 8 demographic and clinical features: age, menopausal status (0= premenopausal, 1= postmenopausal), tumor size, differentiation grade, number of positive lymph nodes, progesterone receptors (fmol/l), estrogen receptors (fmol/l), hormonal treatment;

- **METABRIC**: a dataset composed of 1,904 patients with breast cancer at different stages [64]. Clinical covariates considered include age at diagnosis, tumour size, number of positive lymph node, tumour grade, ER status, PR status, HER2 status, radio therapy received and hormone therapy received.

A.2 Features selection for SUPPORT dataset

SUPPORT study investigates the survival time of seriously-ill hospitalized adults. The preprocessed dataset (which results from the same procedure applied in [25]) was composed of 9,105 patients and 35 covariates, including 5 non-binary categorical features. In order to reduce the dimensionality of the dataset, we performed a preliminary feature selection. Specifically, we trained a CoxPH regression model on the whole dataset and ranked the covariates according to the absolute value of the related regression coefficients (Figure A.1). Then, we selected the top-20 ranked covariates. The intent of this procedure was to reduce the complexity of the dataset, speeding up the synthetic data generation for each fictitious data centre without affecting the validity of the overall results.

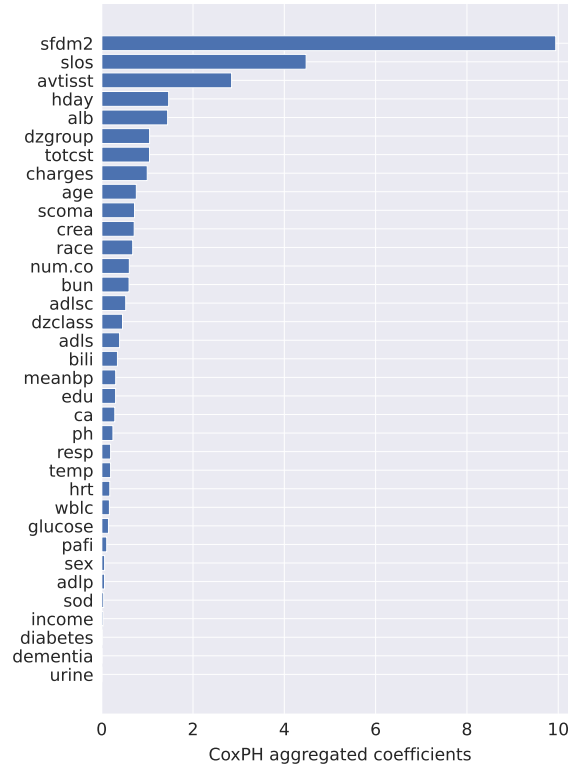


Figure A.1: CoxPH coefficients ranked by absolute value. The weights related to the classes of the categorical features have been aggregated by summing their absolute values.

A.3 Synthetic times distributions

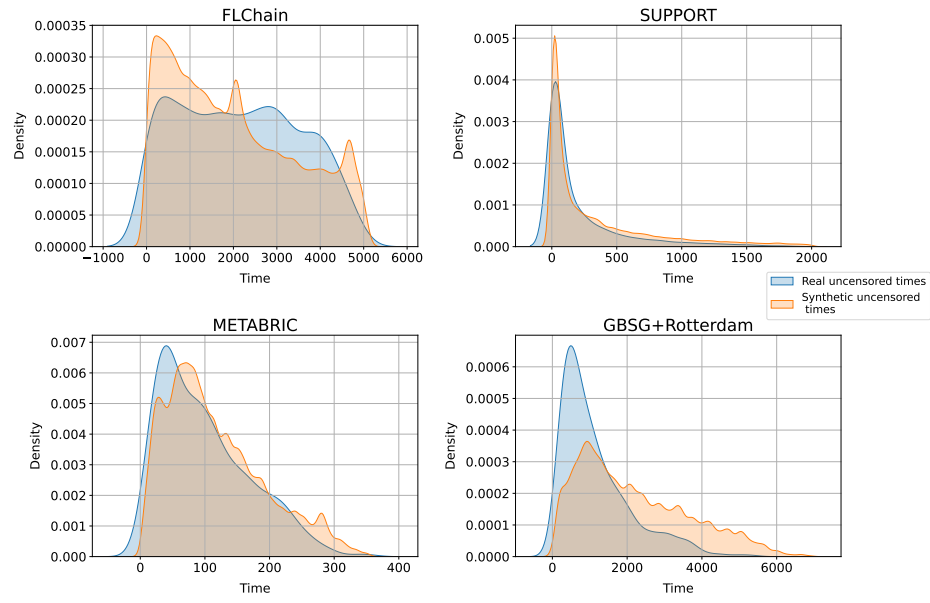


Figure A.2: Synthetic times distribution compared to the real event times, for the uncensored patients.

A.4 Time-dependent AUC results

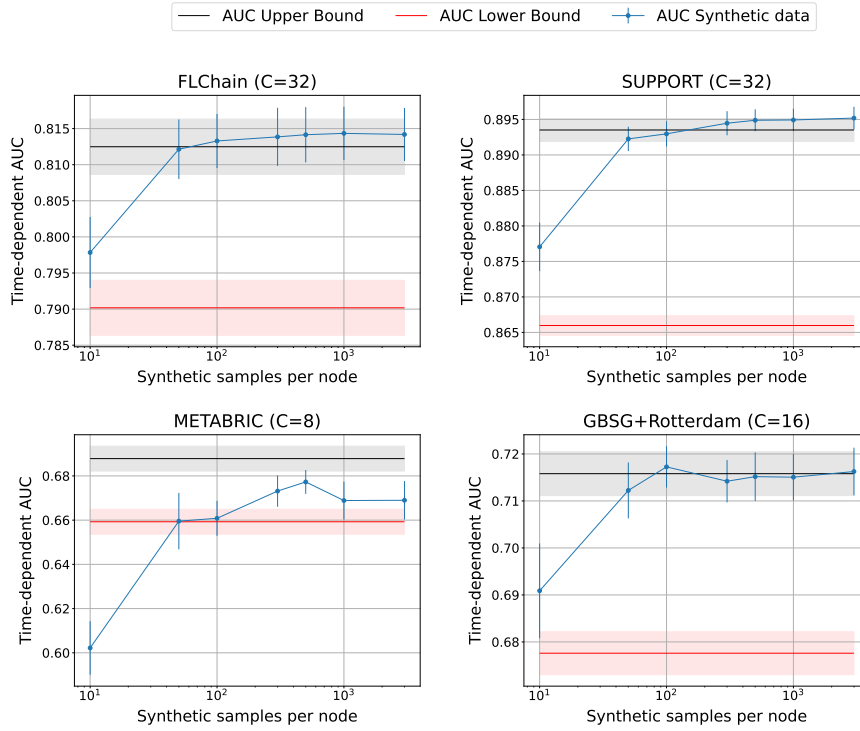


Figure A.3: Test set time-dependent AUC vs. number of generated patients per node, for the CoxPH model trained on the synthetic data.

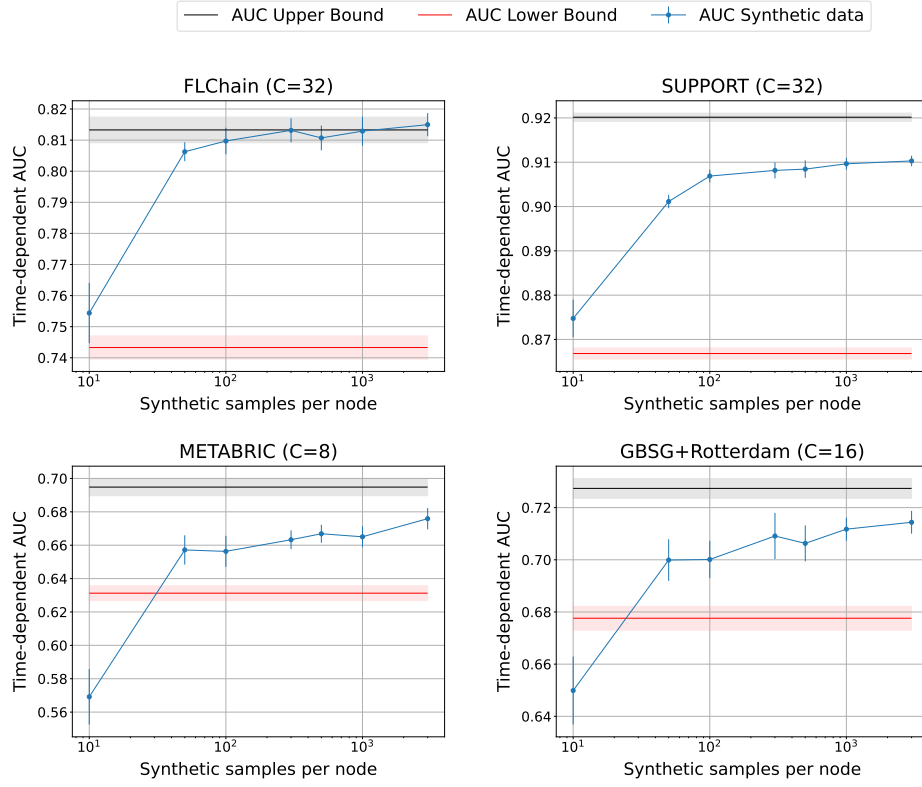


Figure A.4: Test set time-dependent AUC vs. number of generated patients per node, for the DeepHit model trained on the synthetic data.

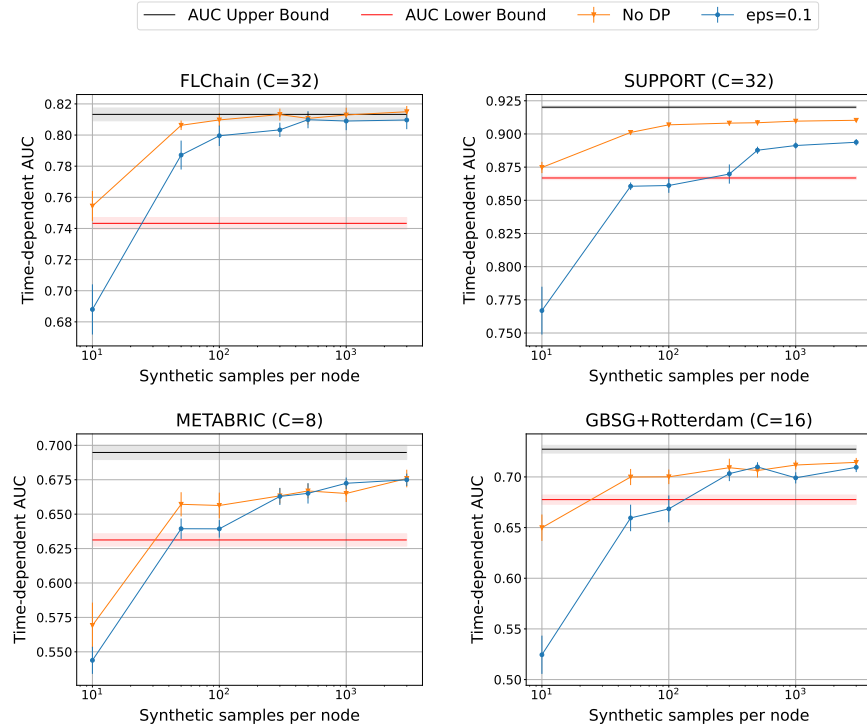


Figure A.5: Test set time-dependent auc vs. number of generated patients per node, for the DeepHit model trained on the synthetic data, using a differential privacy setting with $\epsilon = 0.1$.

Appendix B

VAE-Surv Supplementary Materials

B.1 Dataset description

Table B.1: Description of GenoMed4All and IWG-PM Cohorts. Comparisons were performed using Wilcoxon Rank Sum and Chi-Square tests for continuous and categorical variables, respectively.

Cohort	GenoMed4All	IWG-PM	p
Patient n.	2043	2384	
Female/male	816/1221 (39.9%, 60.1%)	937/1447 (39.3%, 60.7%)	0.528
Age median	68 (19–94)	72 (12–100)	< 0.0001
Clinical and hematological features			
WHO category (2016 criteria):			< 0.0001
MDS-5q-	75 (3.7%)	110 (4.61%)	
MDS-SLD	139 (6.8%)	189 (7.92%)	
MDS-MLD	405 (19.8%)	537 (22.53%)	
MDS-RS-SLD	198 (9.7%)	219 (9.18%)	
MDS-RS-MLD	235 (11.5%)	188 (7.89%)	
MDS-EB1	335 (16.4%)	391 (16.4%)	
MDS-EB2	656 (32.1%)	376 (15.77%)	
MDS-MPN	0	307 (12.87%)	
MDS-U	0	67 (2.81%)	
Hemoglobin median (g/l)	96 (21–196)	97 (40–166)	0.181
Neutrophils median ($10^9/l$)	1.89 (0–24)	1.92 (0–11.9)	0.786
Platelets median ($10^9/l$)	122 (2–1491)	129 (2–1451)	0.153
Bone marrow blast median (%)	4 (0–19)	3 (0–19.5)	< 0.0001
Chromosomal alterations:			
			< 0.0001
Normal karyotype	1195 (58.5%)	1409 (59.2%)	
Abnormal karyotype	651 (31.9%)	975 (40.8%)	
Failed	197 (9.6%)	0	
Isolated del(5q)	90	200	
Loss of chr 5 OR double clone including del(5q)*	83	159	
Loss of chr 7 or del(7q)	105	82	
Gain of chr 8	111	178	
Loss of chr 9 or del(9q)	18	22	
Loss of chr 11 or del(11q)	29	50	
Loss of chr 12 or del(12p) or t(12p)	42	67	
Loss of chr 13 or del(13q)	33	51	
Isochr 17q or t(17p)	60	126	
Loss of chr 20 or del(20q)	77	125	
Loss of chr Y (males only)	1	18	
Idic(X)(q13)	1	0	
Complex karyotype with > 3 abnormalities	142	233	
Cytogenetic risk (IPSS-R):			
			0.019
Very good	64 (3.5%)	103 (4.32%)	
Good	1339 (72.5%)	1669 (70.01%)	
Intermediate	222 (12%)	313 (13.13%)	
Poor	109 (5.9%)	111 (4.66%)	
Very poor	112 (6.1%)	188 (7.89%)	
Age-adjusted IPSS-R risk group:			
			< 0.0001
Very low	199 (12.3%)	409 (17.16%)	
Low	591 (36.5%)	910 (38.17%)	
Intermediate	332 (20.5%)	497 (20.85%)	
High	283 (17.5%)	322 (13.51%)	
Very high	216 (13.3%)	246 (10.32%)	

The GenoMed4All cohort included 2,043 patients (median age 68 years, range 19–94; 60.1% male), while the IWG-PM cohort comprised 2,384 patients (median age 72 years, range 12–100; 60.7% male). The distribution of IPSS-R risk groups differed between the cohorts: in GenoMed4All, 12.3% of patients were classified as very low risk, 36.5% as low, 20.5% as intermediate, 17.4% as high, and 13.3% as very high risk. In the IWG-PM cohort, the respective percentages were 17.2%, 38.2%, 20.8%, 13.5%, and 10.3%. Median values of basic blood exams were comparable. The median hemoglobin level was 96 g/L (range 21–196) in the GenoMed4All cohort and 97 g/L (range 40–166) in the IWG-PM cohort. The median platelet counts were $122 \times 10^9/\text{L}$ (range 2–1,491) and $122 \times 10^9/\text{L}$ (range 6–1,149), respectively. Median neutrophil counts were $1.89 \times 10^9/\text{L}$ (range 0–24) in GenoMed4All and $1.92 \times 10^9/\text{L}$ (range 0–11.9) in IWG-PM. Cytogenetic risk distribution was also slightly distinct. In GenoMed4All, 76% of patients had very good/good cytogenetic risk, 12% had intermediate risk, and 12% had poor/very poor risk. In the IWG-PM cohort, these proportions were 74.3%, 13.2%, and 12.5%, respectively. These differences highlight the heterogeneity between the two cohorts, which serve as complementary datasets for model development and validation.

B.2 Optimal hyperparameters and VAE-Surv architecture

The VAE-surv model architecture follows a structure of **Linear** $\rightarrow$ **ReLU** $\rightarrow$ **Dropout** layers in each component:

- **Encoder:** [58, 64, 32] (Dropout $p = 0.05$)
- **Latent Space:** μ , $\log \sigma^2$ layers: [32, z_{dim}]
- **Decoder:** [z_{dim} , 32, 64, 58] (Dropout $p = 0.05$)
- **Survival Layers (CoxNet):** [z_{dim} +clinical, 32, 16, 1] (Dropout $p = 0.25$)

All the final hyperparameters of the model are reported in the following Table:

Hyperparameter	Value
General Settings	
Batch size	128
Batch norm	None
Epochs	160
Pretrain epochs (VAE only)	500
Optimization Parameters	
Learning rate (VAE pretrain)	0.00056
Learning rate (Survival Net)	0.00045
L2 regularization	2.86e-05
Dropout Rates	
Dropout (VAE)	0.05
Dropout (CoxNet)	0.25
Architecture-Specific Parameters	
z dimension	7
Activation	ReLU

Table B.2: Optimal hyperparameters of the VAE-Surv Neural Network

B.3 Optimal K-means clustering

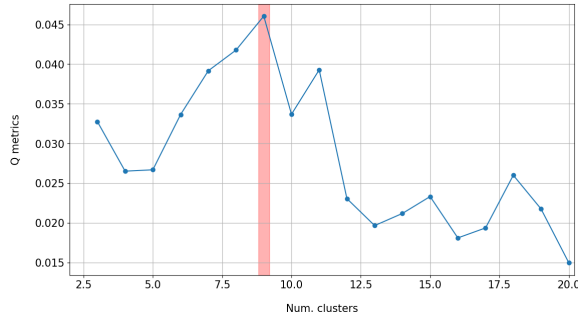


Figure B.1: **VAE’s latent representation:** Plot of the quality metric Q_k against the number of clusters k , determined through K-means clustering of the VAE’s latent representation. The peak indicates the optimal number of clusters, $k = 9$.

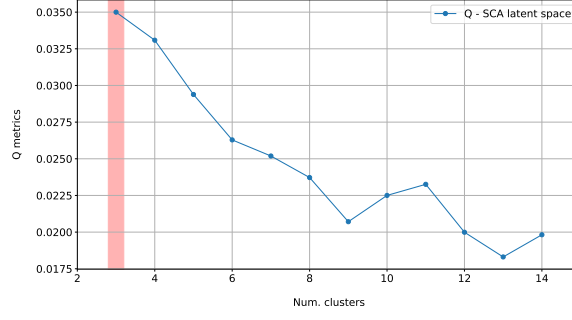


Figure B.2: **Genetic and cytogenetic input data** Plot of the quality metric Q_k against the number of clusters k , determined through the Survival Cluster Analysis model. The peak indicates the optimal number of clusters, $k = 3$.

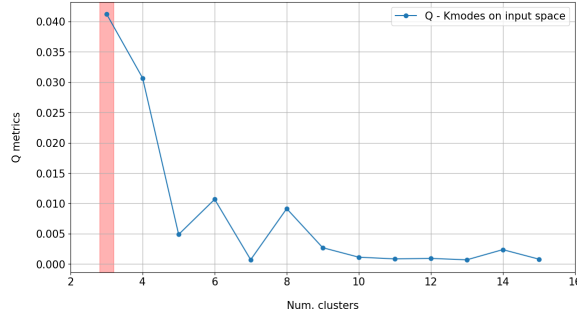


Figure B.3: **Survival Cluster Analysis** Plot of the quality metric Q_k against the number of clusters k , determined through K-modes clustering genetic and cytogenetic input data. The peak indicates the optimal number of clusters, $k = 3$.

B.4 Clustering from VAE-Surv trained with KL-term

B.5 Clustering on Input Data

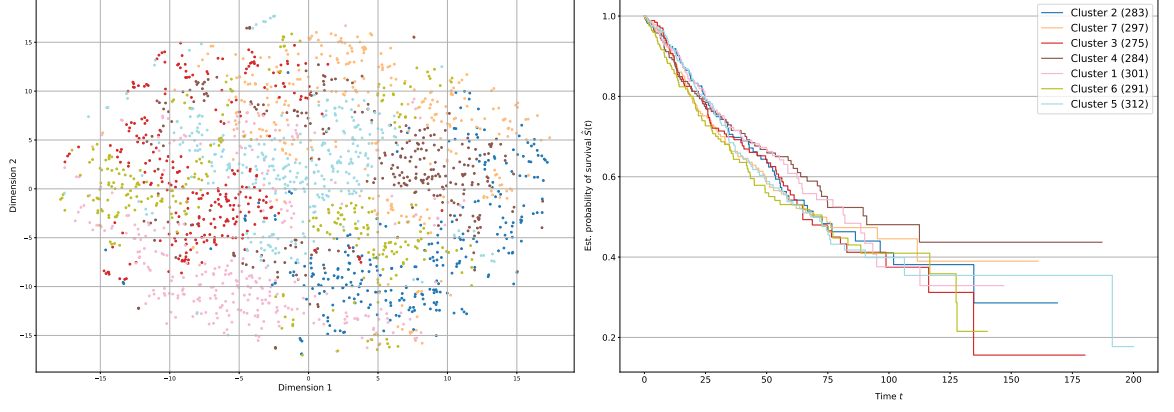


Figure B.4: Left panel: t-SNE representation of the VAE-Surv latent space when using the KL-term during the training. Right panel: Kaplan-Meier survival curves stratified by the identified cluster. The size of each group is specified in the legend.

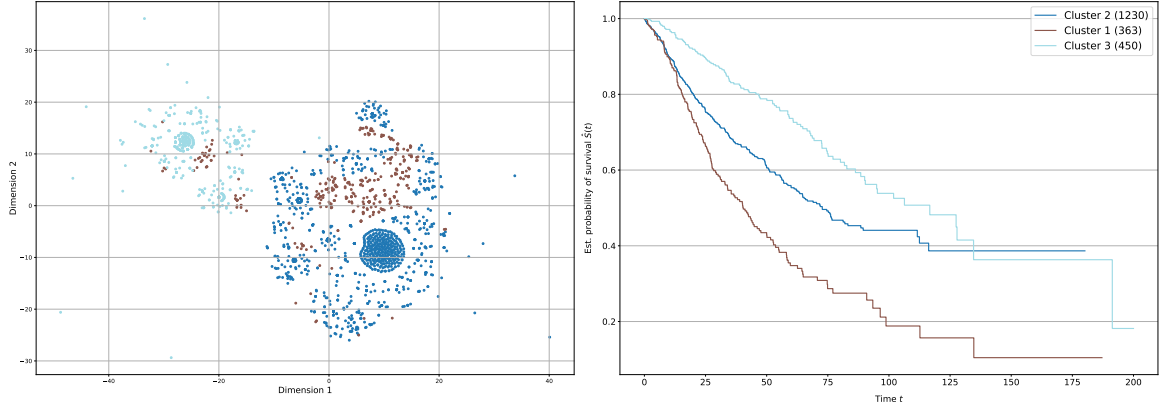


Figure B.5: Left panel: t-SNE representation of the patients' genetic and cytogenetic input data, coloured according to the cluster assignment (K-modes). Right panel: Kaplan-Meier survival curves stratified by the identified cluster. The size of each group is specified in the legend.

B.6 Clustering on VAE latent space

B.6.1 Cluster-specific analysis and comparison with WHO 2016 classification

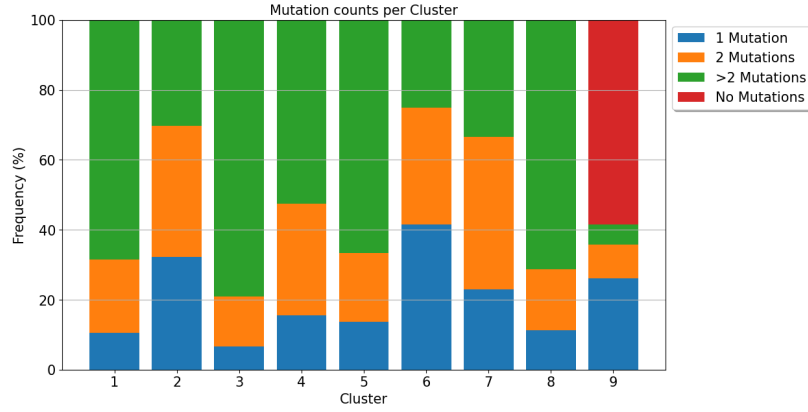


Figure B.6: Mutations count per cluster.

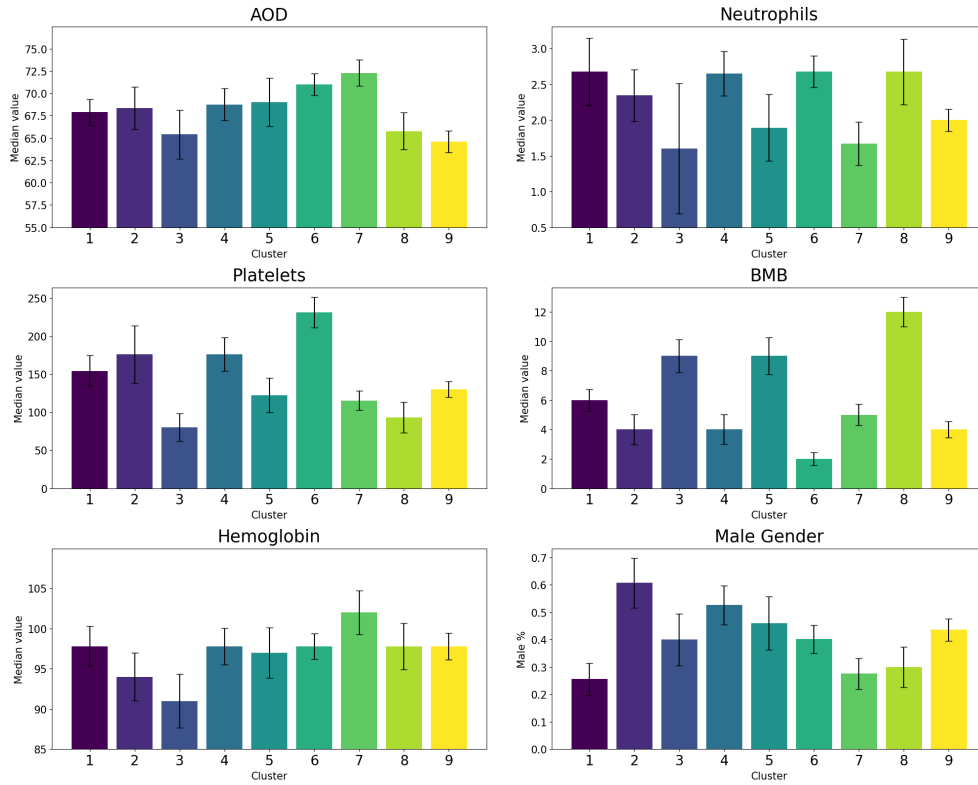


Figure B.7: Median values of relevant clinical attributes per cluster. 95% Confidence Intervals are represented by the error bars.

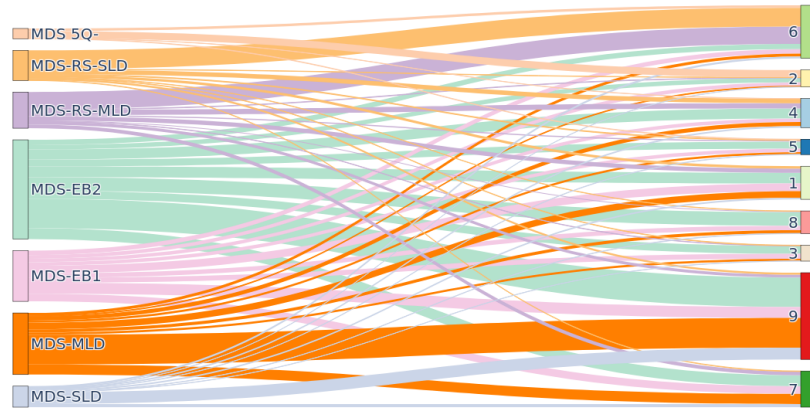


Figure B.8: Sankey plot showing the flow of MDS subtypes across the identified clusters. Each strand represents the proportion of a specific MDS subtype present in each cluster, providing a comprehensive view of how the MDS landscape is distributed across the clusters.

B.7 VAE latent space: validation in the external cohort

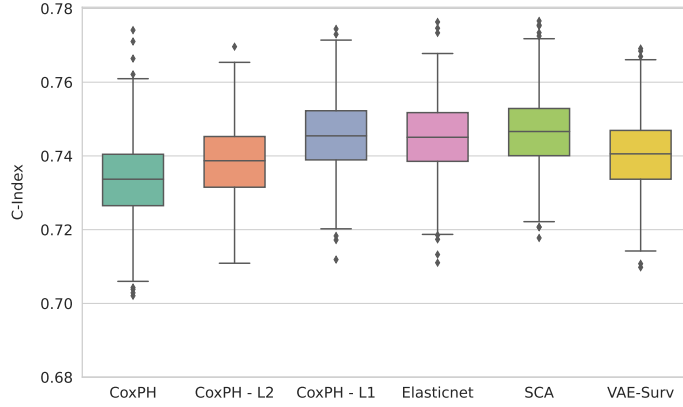


Figure B.9: Comparison of the C-Index results obtained with the considered models for the Validation cohort, using the Bootstrap with resample technique.

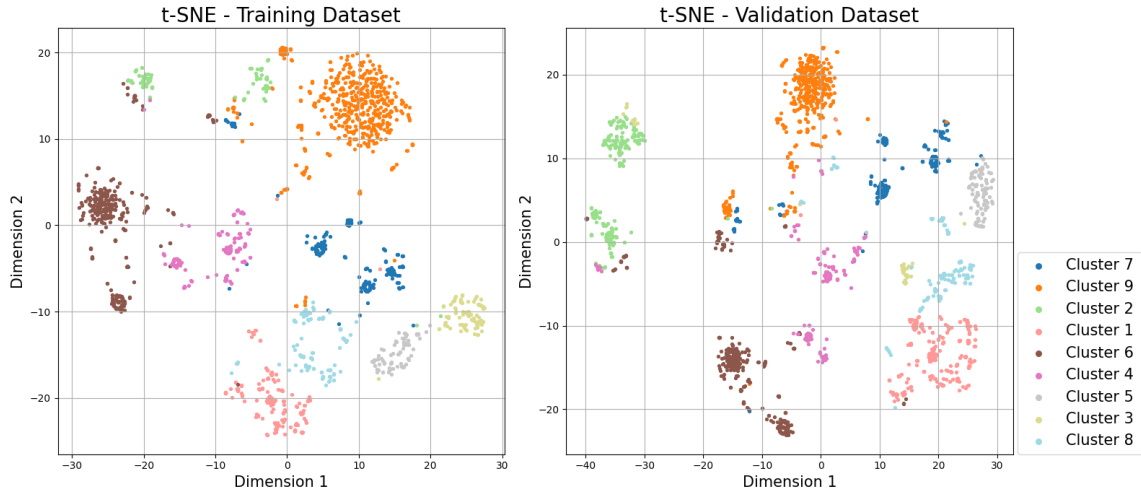


Figure B.10: t-SNE visualization of the latent space for the training and validation cohort.

B.8 VAE-Surv: Model Explanation with SHAP

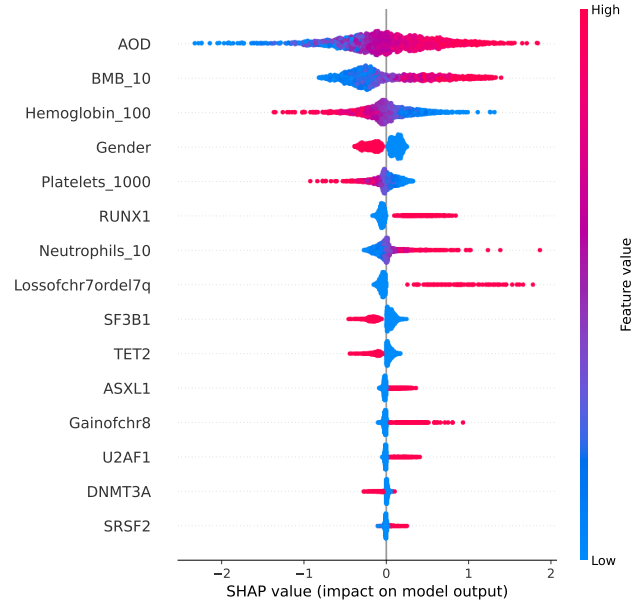


Figure B.11: Top-15 ranked features by shapley values in the training cohort.

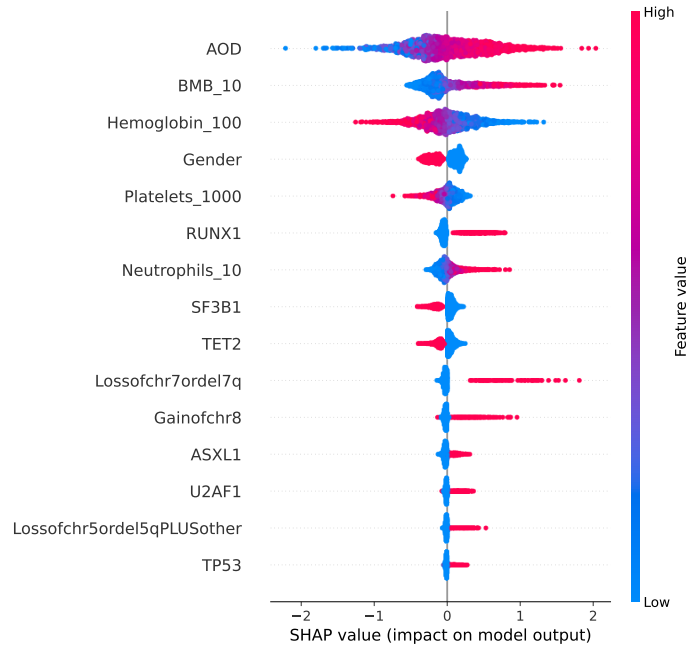


Figure B.12: Top-15 ranked features by shapley values in the validation cohort.

Appendix C

Modelling Survival as a Stochastic Diffusion Process - Figures

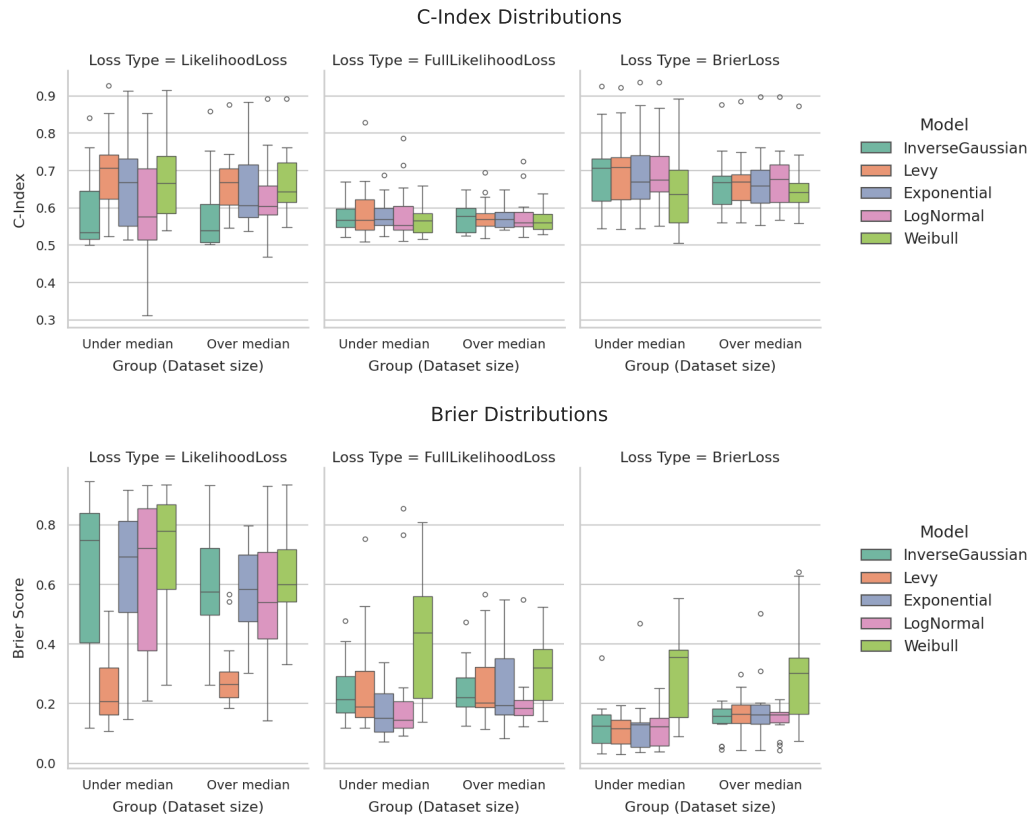


Figure C.1: C-Index and Brier score distributions for datasets stratified by sample size (below vs. above the median).

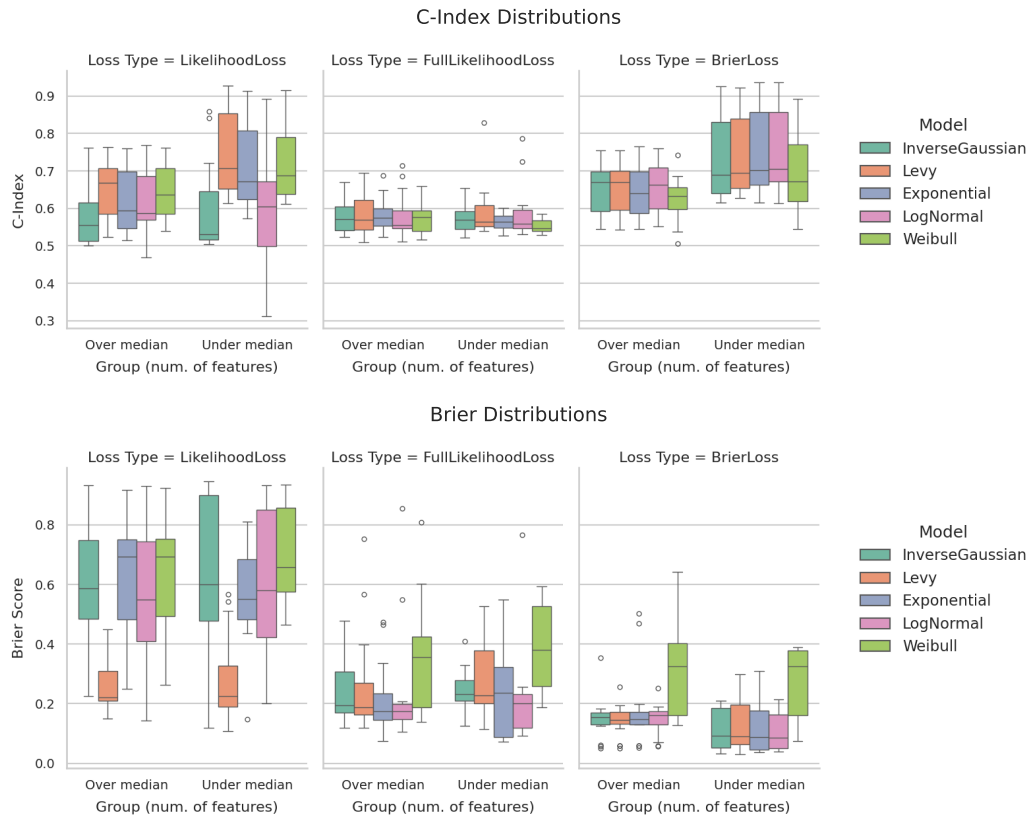


Figure C.2: C-Index and Brier score distributions for datasets stratified by number of features (below vs. above the median).

Bibliography

- [1] Cesare Rollo et al. “SYNDSURV: A simple framework for survival analysis with data distributed across multiple institutions”. In: *Computers in Biology and Medicine* 172 (2024), p. 108288.
- [2] Cesare Rollo et al. “VAE-Surv: A novel approach for genetic-based clustering and prognosis prediction in myelodysplastic syndromes”. In: *Computer Methods and Programs in Biomedicine* (2025), p. 108605.
- [3] Saverio D’Amico et al. “MOSAIC: An Artificial Intelligence–Based Framework for Multimodal Analysis, Classification, and Personalized Prognostic Assessment in Rare Cancers”. In: *JCO Clinical Cancer Informatics* 8 (2024), e2400008.
- [4] Cesare Rollo et al. “Influence of model structures on predictors of protein stability changes from single-point mutations”. In: *Genes* 14.12 (2023), p. 2228.
- [5] Oussama Harrak et al. “Federated AdaBoost for Survival Analysis”. In: *Proceedings of the ECML-PKDD Workshops*. Springer. 2024, pp. 1–9.
- [6] Cesare Rollo et al. “Communities, gateways, and bridges: Measuring attention flow in the reddit political sphere”. In: *International Conference on Social Informatics*. Springer. 2022, pp. 3–19.
- [7] Corrado Pancotti et al. “MUSE-XAE: MUtational Signature Extraction with eXplainable AutoEncoder enhances tumour types classification”. In: *Bioinformatics* 40.5 (2024), btae320.
- [8] Francesco Casadei et al. “Survival Model Optimization via Federated Learning: A Study Combining Simulations and Experiments”. In: *2024 IEEE International Conference on Big Data (BigData)*. IEEE. 2024, pp. 7658–7667.
- [9] Corrado Pancotti et al. “Deep learning methods to predict amyotrophic lateral sclerosis disease progression”. In: *Scientific reports* 12.1 (2022), p. 13738.
- [10] Sara Ferrini et al. “A novel machine learning-based web application for field identification of infectious and inflammatory disorders of the central nervous system in cattle”. In: *Journal of Veterinary Internal Medicine* 37.2 (2023), pp. 766–773.
- [11] Corrado Pancotti et al. “Unravelling the instability of mutational signatures extraction via archetypal analysis”. In: *Frontiers in Genetics* 13 (2023), p. 1049501.
- [12] Federico Giacobbe et al. “Impact of statin adherence and interruption within 6 months after ST-segment elevation myocardial infarction (STEMI): Results from the real-world regional registry FAST-STEMI”. In: *International Journal of Cardiology* 405 (2024), p. 131933.
- [13] Corrado Pancotti et al. “Multi-event survival prediction for amyotrophic lateral sclerosis”. In: *CEUR workshop proceedings*. Vol. 3180. Faggioli G Ferro N, Hanbury A, Potthast M. 2022, pp. 1269–1276.
- [14] Giuseppe Giannino et al. “Impact of adherence to beta-blockers in all-comers ST-segment elevation myocardial infarction (STEMI) patients and according to

- left ventricular ejection fraction (LVEF) at discharge: results from the real-world registry FAST-STEMI”. In: *Journal of Cardiovascular Pharmacology* (2022), pp. 10–1097.
- [15] Giovanni Birolo et al. *SurvHive: a package to consistently access multiple survival-analysis packages*. 2025. arXiv: 2502.02223 [q-bio.QM]. URL: <https://arxiv.org/abs/2502.02223>.
- [16] Francesco Codicè et al. “The Specification Game: Rethinking the Evaluation of Drug Response Prediction for Precision Oncology”. In: *bioRxiv* (2024), pp. 2024–10.
- [17] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. “Understanding survival analysis: Kaplan-Meier estimate”. In: *International journal of Ayurveda research* 1.4 (2010), p. 274.
- [18] Enrico Colosimo et al. “Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators”. In: *Journal of Statistical Computation and Simulation* 72.4 (2002), pp. 299–308.
- [19] D.R. Cox. “Regression Models and Life-Tables”. In: *J. R. Stat. Soc.* 34 (1972), pp. 187–220.
- [20] DY Lin. “On the Breslow estimator”. In: *Lifetime data analysis* 13.4 (2007), pp. 471–480.
- [21] David Faraggi and Richard Simon. “A neural network model for survival data”. In: *Statistics in medicine* 14.1 (1995), pp. 73–82.
- [22] J.L. Katzman et al. “DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network”. In: *BMC medical research methodology* 18.1 (2018), pp. 1–12.
- [23] Changhee Lee et al. “Deephit: A deep learning approach to survival analysis with competing risks”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.
- [24] Chirag Nagpal, Xinyu Li, and Artur Dubrawski. “Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks”. In: *IEEE Journal of Biomedical and Health Informatics* 25.8 (2021), pp. 3163–3175.
- [25] Paidamoyo Chapfuwa et al. “Survival cluster analysis”. In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. 2020, pp. 60–68.
- [26] Lee-Jen Wei. “The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis”. In: *Statistics in medicine* 11.14-15 (1992), pp. 1871–1879.
- [27] Avinash Barnwal, Hyunsu Cho, and Toby Hocking. “Survival regression with accelerated failure time model in XGBoost”. In: *Journal of Computational and Graphical Statistics* just-accepted (2022), pp. 1–25.
- [28] Hemant Ishwaran et al. “Random survival forests”. In: (2008).
- [29] Enrico Longato, Martina Vettoretti, and Barbara Di Camillo. “A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models”. In: *Journal of biomedical informatics* 108 (2020), p. 103496.
- [30] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. “A time-dependent discrimination index for survival data”. In: *Statistics in medicine* 24.24 (2005), pp. 3927–3944.

- [31] Thomas A Gerds and Martin Schumacher. “Consistent estimation of the expected Brier score in general survival models with right-censored event times”. In: *Biometrical Journal* 48.6 (2006), pp. 1029–1040.
- [32] Manish Gawali et al. “Comparison of privacy-preserving distributed deep learning methods in healthcare”. In: *Annual Conference on Medical Image Understanding and Analysis*. Springer. 2021, pp. 457–471.
- [33] Fadila Zerka et al. “Systematic review of privacy-preserving distributed machine learning from federated databases in health care”. In: *JCO clinical cancer informatics* 4 (2020), pp. 184–200.
- [34] Tania Pereira et al. “Sharing biomedical data: Strengthening ai development in healthcare”. In: *Healthcare*. Vol. 9. 7. MDPI. 2021, p. 827.
- [35] Margarita Kirienko et al. “Distributed learning: a reliable privacy-preserving strategy to change multicenter collaborations using AI”. In: *European Journal of Nuclear Medicine and Molecular Imaging* 48.12 (2021), pp. 3791–3804.
- [36] Brendan McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.
- [37] Micah J Sheller et al. “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data”. In: *Scientific reports* 10.1 (2020), pp. 1–12.
- [38] Nicola Rieke et al. “The future of digital health with federated learning”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–7.
- [39] Priyanka Mary Mammen. “Federated learning: opportunities and challenges”. In: *arXiv preprint arXiv:2101.05428* (2021).
- [40] Peter Kairouz et al. “Advances and open problems in federated learning”. In: *Foundations and Trends® in Machine Learning* 14.1–2 (2021), pp. 1–210.
- [41] Rui Duan et al. “Learning from local to global: An efficient distributed algorithm for modeling time-to-event data”. In: *Journal of the American Medical Informatics Association* 27.7 (2020), pp. 1028–1036.
- [42] Chia-Lun Lu et al. “WebDISCO: a web service for distributed cox model learning without patient-level data sharing”. In: *Journal of the American Medical Informatics Association* 22.6 (2015), pp. 1212–1219.
- [43] Xuan Wang et al. “SurvMaximin: Robust federated approach to transporting survival risk prediction models”. In: *Journal of biomedical informatics* 134 (2022), p. 104176.
- [44] Md Mahmudur Rahman and Sanjay Purushotham. “FedPseudo: Pseudo value-based Deep Learning Models for Federated Survival Analysis”. In: *arXiv preprint arXiv:2207.05247* (2022).
- [45] Mathieu Andreux et al. “Federated survival analysis with discrete-time Cox models”. In: *arXiv preprint arXiv:2006.08997* (2020).
- [46] Jinsung Yoon, Lydia N Drumright, and Mihaela Van Der Schaar. “Anonymization through data synthesis using generative adversarial networks (ads-gan)”. In: *IEEE journal of biomedical and health informatics* 24.8 (2020), pp. 2378–2388.
- [47] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. “PATE-GAN: Generating synthetic data with differential privacy guarantees”. In: *International conference on learning representations*. 2019.
- [48] Juan de Benedetti et al. “Practical Lessons from Generating Synthetic Healthcare Data with Bayesian Networks”. In: *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discov-*

- ery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings.* Springer. 2020, pp. 38–47.
- [49] Zhenchen Wang, Puja Myles, and Allan Tucker. “Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy”. In: *Computational Intelligence* 37.2 (2021), pp. 819–851.
 - [50] Alexander Norcliffe et al. “SurvivalGAN: Generating Time-to-Event Data for Survival Analysis”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 10279–10304.
 - [51] Sabyasachi Dash et al. “Big data in healthcare: management, analysis and future prospects”. In: *Journal of Big Data* 6.1 (2019), pp. 1–25.
 - [52] Sonja Zillner and Sabrina Neururer. “Big data in the health sector”. In: *New Horizons for a Data-Driven Economy*. Springer, Cham, 2016, pp. 179–194.
 - [53] Arjun Panesar. *Machine learning and AI for healthcare*. Springer, 2019.
 - [54] Fei Jiang et al. “Artificial intelligence in healthcare: past, present and future”. In: *Stroke and vascular neurology* 2.4 (2017).
 - [55] Nicola Rieke et al. “The future of digital health with federated learning”. In: *npj Digit. Med.* 3 (2020), p. 119.
 - [56] Haoyue Ping, Julia Stoyanovich, and Bill Howe. “Datasyntesizer: Privacy-preserving synthetic datasets”. In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 2017, pp. 1–5.
 - [57] Jun Zhang et al. “Privbayes: Private data release via bayesian networks”. In: *ACM Transactions on Database Systems (TODS)* 42.4 (2017), pp. 1–41.
 - [58] Cynthia Dwork et al. “Calibrating noise to sensitivity in private data analysis”. In: *Theory of cryptography conference*. Springer. 2006, pp. 265–284.
 - [59] Jaewoo Lee and Chris Clifton. “How much is enough? choosing ϵ for differential privacy”. In: *International Conference on Information Security*. Springer. 2011, pp. 325–340.
 - [60] Noah Simon et al. “Regularization paths for Cox’s proportional hazards model via coordinate descent”. In: *Journal of statistical software* 39.5 (2011), p. 1.
 - [61] James Bergstra et al. “Algorithms for hyper-parameter optimization”. In: *Advances in neural information processing systems* 24 (2011).
 - [62] Takuya Akiba et al. “Optuna: A next-generation hyperparameter optimization framework”. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2623–2631.
 - [63] Terry Therneau et al. “A package for survival analysis in S”. In: *R package version 2.7* (2015).
 - [64] Christina Curtis et al. “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups”. In: *Nature* 486.7403 (2012), pp. 346–352.
 - [65] William A Knaus et al. “The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults”. In: *Annals of internal medicine* 122.3 (1995), pp. 191–203.
 - [66] John A Foekens et al. “The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients”. In: *Cancer research* 60.3 (2000), pp. 636–643.
 - [67] Martin Schumacher et al. “Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients.

- German Breast Cancer Study Group.” In: *Journal of Clinical Oncology* 12.10 (1994), pp. 2086–2093.
- [68] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. “Time-to-event prediction with neural networks and Cox regression”. In: *arXiv preprint arXiv:1907.00825* (2019).
- [69] Patrick Royston and Douglas G Altman. “External validation of a Cox prognostic model: principles and methods”. In: *BMC medical research methodology* 13.1 (2013), pp. 1–15.
- [70] Ping Wang, Y. Li, and C.K. Reddy. “Machine Learning for Survival Analysis: A Survey”. In: *ACM Comput. Surv.* 51.6 (2019), pp. 1–36.
- [71] Silvia Liverani et al. “Clustering method for censored and collinear survival data”. In: *Comput Stat.* 36 (2021), pp. 35–60.
- [72] Paidamoyo Chapfuwa et al. “Survival cluster analysis”. In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. CHIL ’20. Toronto, Ontario, Canada: Association for Computing Machinery, 2020, 60–68. ISBN: 9781450370462. DOI: 10.1145/3368555.3384465. URL: <https://doi.org/10.1145/3368555.3384465>.
- [73] Kevin. Boehm et al. “Harnessing multimodal data integration to advance precision oncology”. In: *Nat Rev Cancer.* 22.2 (2022), pp. 114–126.
- [74] Ruoqi Wei and A. Mahmood. “Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey”. In: *Ieee Access* 9 (2020), pp. 4939–4956.
- [75] David Wissel, D. Rowson, and V. Boeva. “Hierarchical autoencoder-based integration improves performance in multi-omics cancer survival models through soft modality selection”. In: *bioRxiv* (2021), pp. 2021–09.
- [76] Lindong Jiang et al. “Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data”. In: *NPJ Precision Oncology* 8.1 (2024), p. 4.
- [77] Sunkyu Kim et al. “Improved survival analysis by learning shared genomic information from pan-cancer data”. In: *Bioinformatics* 36.Supplement_1 (2020), pp. i389–i398.
- [78] Muta Tah Hira et al. “Integrated multi-omics analysis of ovarian cancer using variational autoencoders”. In: *Scientific reports* 11.1 (2021), p. 6265.
- [79] Seishi Ogawa. “Genetics of MDS”. In: *Blood, The Journal of the American Society of Hematology* 133.10 (2019), pp. 1049–1059.
- [80] Guillermo Garcia-Manero. “Myelodysplastic syndromes: 2023 update on diagnosis, risk-stratification, and management”. In: *American Journal of Hematology* 98.8 (2023), pp. 1307–1325.
- [81] Rory Shallis, R. Ahmad, and A.M. Zeidan. “The genetic and molecular pathogenesis of myelodysplastic syndromes”. In: *European journal of haematology* 101.3 (2018), pp. 260–271.
- [82] Daniel Arber et al. “The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia”. In: *Blood, The Journal of the American Society of Hematology* 127.20 (2016), pp. 2391–2405.
- [83] Matteo Bersanelli et al. “Classification and personalized prognostic assessment on the basis of clinical and genomic features in myelodysplastic syndromes”. In: *Journal of Clinical Oncology* 39.11 (2021), p. 1223.

- [84] Saverio D’Amico et al. “Multi-Modal Analysis and Federated Learning Approach for Classification and Personalized Prognostic Assessment in Myeloid Neoplasms”. In: *Blood* 140.Supplement 1 (2022), pp. 9828–9830.
- [85] Mikkael Sekeres and J. Taylor. “Diagnosis and treatment of myelodysplastic syndromes: a review”. In: *Jama* 328.9 (2022), pp. 872–880.
- [86] Elsa Bernard et al. “Molecular international prognostic scoring system for myelodysplastic syndromes”. In: *NEJM evidence* 1.7 (2022), EVIDoa2200008.
- [87] Francesco Cremonesi et al. “The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform”. In: *Journal of Biomedical Informatics* 141 (2023), p. 104338.
- [88] Diederik Kingma and M. Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).
- [89] James Lucas et al. “Understanding posterior collapse in generative latent variable models”. In: (2019).
- [90] Frank Jr. Harrell, K.L. Lee, and D.B. Mark. “Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors”. In: *Statistics in medicine* 15.4 (1996), pp. 361–387.
- [91] Noah Simon et al. “Regularization paths for Cox’s proportional hazards model via coordinate descent.” In: *J Stat Softw.* 39 (2011), pp. 1–13.
- [92] Peter Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.
- [93] Luca Malcovati et al. “SF3B1-mutant MDS as a distinct disease subtype: a proposal from the International Working Group for the Prognosis of MDS”. In: *Blood, The Journal of the American Society of Hematology* 136.2 (2020), pp. 157–170.
- [94] Luca Malcovati et al. “SF3B1 mutation identifies a distinct subset of myelodysplastic syndrome with ring sideroblasts”. In: *Blood, The Journal of the American Society of Hematology* 126.2 (2015), pp. 233–241.
- [95] Andreas Messalas, Y. Kanellopoulos, and C. Makris. “Model-agnostic interpretability with shapley values”. In: *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE. 2019, pp. 1–7.
- [96] Yudi Zhang et al. “Comparison of the revised 4th (2016) and 5th (2022) editions of the World Health Organization classification of myelodysplastic neoplasms”. In: *Leukemia* 36.12 (2022), pp. 2875–2882.
- [97] Andrea Kündgen et al. “Therapy-related myelodysplastic syndromes deserve specific diagnostic sub-classification and risk-stratification—an approach to classification of patients with t-MDS”. In: *Leukemia* 35.3 (2021), pp. 835–849.
- [98] Dimitrios Stogiannis et al. “Comparing first hitting time and proportional hazards regression models”. In: *Journal of Applied Statistics* 38.7 (2011), pp. 1483–1492.
- [99] Mei-Ling Ting Lee and George A Whitmore. “Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary”. In: (2006).
- [100] Odd O Aalen. *Survival and Event History Analysis: A Process Point of View*. Springer-Verlag, 2008.

- [101] Sangbum Choi et al. “A semiparametric inverse-Gaussian model and inference for survival data with a cured proportion”. In: *Canadian Journal of Statistics* 42.4 (2014), pp. 635–649.
- [102] Jonathan A Race and Michael L Pennell. “Semi-parametric survival analysis via Dirichlet process mixtures of the First Hitting Time model”. In: *Lifetime Data Analysis* 27.1 (2021), pp. 177–194.
- [103] Riccardo De Bin and Vegard Grødem Stikbakke. “A boosting first-hitting-time model for survival analysis in high-dimensional settings”. In: *Lifetime Data Analysis* 29.2 (2023), pp. 420–440.
- [104] Erik Drysdale. “SurvSet: An open-source time-to-event dataset repository”. In: *arXiv preprint arXiv:2203.03094* (2022).
- [105] Mirko Polato, Roberto Esposito, and Marco Aldinucci. “Boosting the federation: Cross-silo federated learning without gradient descent”. In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2022, pp. 1–10.
- [106] Alberto Archetti and Matteo Matteucci. “Federated survival forests”. In: *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2023, pp. 1–9.
- [107] Benjamin Webb and Andrej Sali. “Protein structure modeling with MODELLER”. In: *Functional genomics: Methods and protocols* (2017), pp. 39–54.
- [108] Elizabeth H Kellogg, Andrew Leaver-Fay, and David Baker. “Role of conformational sampling in computing mutation-induced changes in protein structure and stability”. In: *Proteins: Structure, Function, and Bioinformatics* 79.3 (2011), pp. 830–838.
- [109] Silvia Benevenuta et al. “An antisymmetric neural network to predict free energy changes in protein variants”. In: *Journal of Physics D: Applied Physics* 54.24 (2021), p. 245403.
- [110] Josef Laimer et al. “MAESTRO-multi agent stability prediction upon point mutations”. In: *BMC bioinformatics* 16 (2015), pp. 1–13.
- [111] Ludovica Montanucci et al. “DDGun: an untrained predictor of protein stability changes upon amino acid variants”. In: *Nucleic Acids Research* 50.W1 (2022), W222–W227.
- [112] Oliver Buß, Jens Rudat, and Katrin Ochsenreither. “FoldX as protein engineering tool: better than random based approaches?” In: *Computational and structural biotechnology journal* 16 (2018), pp. 25–33.
- [113] Iván Martín Hernández et al. “Predicting protein stability changes upon mutation using a simple orientational potential”. In: *Bioinformatics* 39.1 (2023), btad011.
- [114] Yunzhuo Zhou et al. “DDMut: predicting effects of mutations on protein stability using deep learning”. In: *Nucleic Acids Research* 51.W1 (2023), W122–W128.
- [115] Yves Dehouck et al. “PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality”. In: *BMC bioinformatics* 12 (2011), pp. 1–12.