POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Longitudinal Machine Learning Models for CAR-T Therapy Outcome in Large B-Cell Lymphoma

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica

Author: Marco Simoni

Advisor: Prof. Alessandra Laura Giulia Pedrocchi

Co-advisor: Prof. Vanja Miskovic, Sara Ferri

Academic year: 2025-2026

1. Introduction

B-cell Non-Hodgkin Lymphomas (B-NHL) encompass a heterogeneous group of malignancies, with Diffuse Large B-cell Lymphoma (DLBCL) being the most common aggressive form. When standard treatments fail, Chimeric Antigen Receptor (CAR)-T cell therapy emerges as a revolutionary immunotherapeutic option for relapsed or refractory disease [1]. This highly targeted therapy engineers the patient's own T cells to selectively eliminate malignant cells, offering the potential for durable responses and long-term immune surveillance.
However, CAR-T therapy involves costly, time-consuming manufacturing and achieves complete remission in only about 50% of patients. Leveraging leukapheresis clinical characteristics to early identify non-responders is therefore critical, not only to mitigate economic and logistical burdens, but to promptly redirect these patients toward alternative, potentially more effective therapies[1].

.

Artificial Intelligence (AI) and Machine Learning (ML) approaches are increasingly utilized to learn complex patterns from historical data and predict clinical outcomes. In particular, analyzing longitudinal data, repeated measurements collected over time, allows for a more accurate modeling of disease trajectories compared to static, baseline-only approaches [2].
Conducted within the framework of the Italian multicenter CART-SIE study, the aim of this work is to develop and evaluate longitudinal ML models to predict Progression-Free Survival (PFS) in patients undergoing CAR-T therapy. By quantifying the incremental benefit of incorporating follow-up data, this study seeks to provide continuous, dynamic risk stratification to support personalized patient management.

2. Material and Methods

The methodological workflow followed in the present work is outlined in Figure 1, and distinguished with respect to the two tasks: static and dynamic prediction.

2.0.1 Dataset Collection and Preprocessing

Starting from an initial cohort of 1,465 patients diagnosed with B-NHL and candidates for CAR-T cell therapy, a rigorous selection process was applied. Patients diagnosed with Mantle Cell

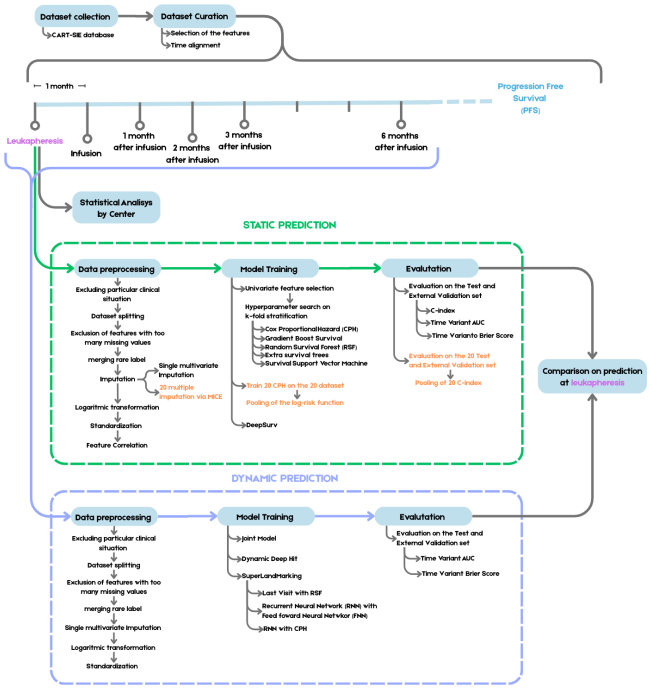

Figure 1: Methodological workflow.

Lymphoma were excluded, as they are characterized by a clinical course that is not directly comparable, alongside subjects treated with the *liso-cel* product due to immature follow-up, and records lacking the outcome event.

For spatial validation and to verify the models' generalizability, patients from the "Roma Cattolica" center were isolated to form an independent external validation set. The remaining subjects were partitioned into a training set (80%) (and further divided in training (70%) and validation (10%) if in a deep learning setting) and an internal test set (20%), ensuring stratified sampling based on the censoring indicator to preserve the same proportion of events across all partitions.

A crucial methodological challenge in the analysis of real-world CAR-T therapy data is the temporal heterogeneity of the logistical pathway: the time elapsed between leukapheresis and the actual infusion of the modified cells varies considerably from patient to patient. In most cases, infusion occurs within two months of leukapheresis; however, this interval can exceed one year in the event of complications related to prior treatments or disease progression. To ensure the comparability of longitudinal trajectories across subjects, an artificial temporal alignment was imposed, assuming that infusion occurs exactly one month after leukapheresis for all patients. This standardization of time windows was essential for the evaluation of dynamic models.

Missing data management involved the removal of variables with over 30% missing values. For the remaining variables, which included clinical and laboratory measurements at leukapheresis, infusion and subsequent follow-ups (30, 60, 90, and 180 days post-infusion), a single multivariate imputation procedure (IterativeImputer) was applied. Rare labels in the categorical features were merged, highly skewed continuous variables were logarithmically transformed and all numerical feature finally standardized.

2.0.2 Outcome definition and Survival Analysis

The primary endpoint of the study is PFS, defined as the time elapsed from infusion to disease progression or death from any cause. As is typical in observational studies, PFS is subject to right-censoring for patients who do not experience the event by the end of the observation period. To properly handle this incomplete event-time data, the methodological framework of survival analysis is utilized. This approach accommodates right-censored observations by assuming non-informative censoring, meaning the censoring mechanism itself does not convey prognostic information. Within this framework, time-to-event data are modeled using two fundamental concepts: the survival function, which describes the cumulative probability of a patient remaining event-free up to a specific time, and the hazard function, which characterizes the instantaneous risk of experiencing the event conditional on having survived up to that moment.

2.0.3 Static Prediction

To establish a solid baseline, a suite of static prediction models trained exclusively on variables available at the time of leukapheresis was initially evaluated. This framework included traditional statistical algorithms (the semi-parametric Cox Proportional Hazards - CPH model), ML approaches based on trees and support vectors (Gradient Boost Survival, Random Survival Forests, Extra Survival Trees, Survival Support Vector Machine), and deep learning architectures dedicated to survival analysis

(DeepSurv). The overall discriminative accuracy of these models was primarily measured using Harrell's C-index [2].

2.0.4 Dynamic Predictions

The methodological core of this study lies in the development and comparison of dynamic predictive architectures capable of overcoming the intrinsic limitations of static models. Static models, by ignoring data collected during subsequent clinical follow-ups, are unable to reflect the evolution of the patient's health status. Dynamic prediction, instead, aims to continuously update the risk estimate as new longitudinal measurements (e.g., biomarkers, blood tests) become available at various landmark times (infusion, and 30, 60, 90, 180 days after infusion). To accurately assess how well the models perform as these predictions are continuously updated, traditional static metrics are replaced by the time-varying AUC (tvAUC), which effectively evaluates the dynamic nature of discrimination over time.

To this end, several algorithmic families were explored, ranging from statistical approaches to complex neural networks. Joint Models (JM) simultaneously analyze longitudinal data and time-to-event outcomes. Linear mixed-effects models (fixed effects plus patient-specific random effects) capture the temporal evolution of a patient's time-varying features; the same patient-specific random effects are then carried into the survival part, linking the longitudinal and time-to-event processes so that the CPH model is driven by the patient-specific biomarker trajectory. Despite their theoretical rigor, these models require highly stringent distributional assumptions and become computationally prohibitive as the number of longitudinal variables increases, often proving unstable in software implementations [3]. Alternatively, Dynamic-DeepHit (DDH) is a purely deep learning approach to handle multiple events over discrete time by trying to estimate the joint distribution of event time and event type over discrete time intervals; for our setting only one event type is considered, so it reduces to estimating the discrete-time distribution of the event time [4].

the patient's features history with a mixed-effects model The longitudinal model is then linked to the survival model though survival model is then linked to this longitudinal process through the patient-specific *latent* biomarker level implied by the mixed-effects model, so the event risk is informed by the full biomarker trajectory rather than by a single observed value.

The Super-Landmarking Framework Training predictive models on complete longitudinal trajectories introduces a significant bias: models tend to over-rely on the most recent measurements, which are often collected very close to the event. Consequently, they struggle to make accurate predictions at early horizons when only limited history is available. To address this challenge, the study implements a landmarking strategy. This approach evaluates patients at pre-defined time points (landmark times), truncating their longitudinal history at that specific moment and including only individuals who are still at risk. To further optimize this solution and maximize data utility, the study extends this concept to super-landmarking. By stacking multiple landmarked versions of the original dataset, it becomes possible to train a single, robust global model that rigorously preserves the essential "at-risk" conditioning. This framework is highly flexible, allowing for various modeling approaches. For instance, the Last-visit RSF utilizes only the most recent measurements available up to each landmark time to train a RSF on the population still at risk. Alternatively, to better capture temporal evolution, the longitudinal history can be encoded using a Recurrent Neural Network (RNN). The RNN extracts latent representations of the patient's trajectory, which are subsequently processed by a Feed-Forward Neural Network (FNN) for survival prediction.

The RNN-CPH is a hybrid architecture trained in two distinct stages. In the first stage the history of the patient is compressed into a latent summary (Z_i), optimized through longitudinal loss to capture biological trends. In the second stage, this representation is "frozen" and feed a CPH module. The hazard function for the patient is modeled linearly with respect to the extracted representations, assuming the form

$$h_i(t) = h_0(t) \exp\Big(Z_i^\top \beta\Big).$$

During this second stage, the CPH regression coefficients (β) are optimized by minimizing the negative partial log-likelihood on the now-fixed latent representations [5].

3. Results

3.1. Dataset Characteristics and Splitting

Following the application of the predefined exclusion criteria, the final analytical cohort comprised 1,071 patients. The external validation set consisted of 67 patients (6% of the overall dataset), the remaining subjects were randomly split into a test set of 201 patients (20%) and a training set of 803 patients (80%). When required by the neural network training pipelines, the 80% training set was further partitioned into 702 training and 101 validation subjects. To preserve a comparable event-to-censoring composition across all splits, stratified sampling based on the PFS event indicator was systematically employed. The resulting distribution of patients across the datasets is illustrated in Figure 2.

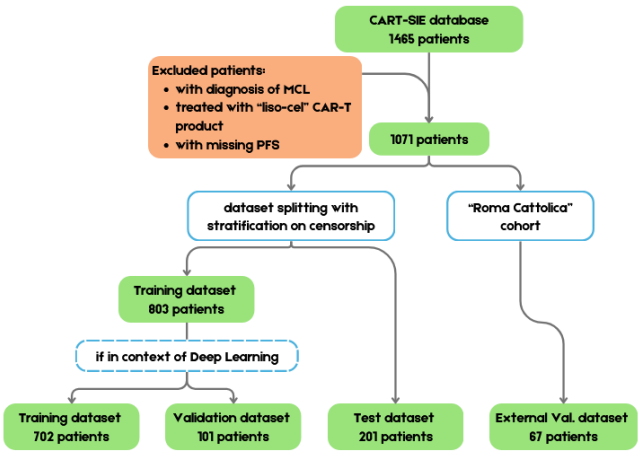

Figure 2: Distribution of patients across training, validation, test, and external validation datasets.

3.2. Static Prediction

Static models, trained exclusively on clinical data available at leukapheresis, were evaluated using Harrell's Concordance Index (C-index) to measure global discriminative performance. Table 1 reports the results of the CPH baseline alongside DeepSurv, which emerged as the best-performing model among the advanced machine learning and deep learning approaches evaluated. A part from the perfomances (0.64 on the Test set and 0.73 on the Ext. Val.), the CPH is also a simpler and more interpretable framework, and so represents the most logical choice among static models for its role as reference static baseline to benchmark our dynamic architectures across all evaluated time horizons using tvAUC.

	Test	Ext Val
CPH	0.64	0.73
DeepSurv	0.67	0.69

Table 1: C-index values for baseline models across test, and external validation sets.

While the DeepSurv architecture demonstrated superior discriminative capabilities on the internal test set(0.67), this dynamic was reversed during external evaluation (0.69).

3.3. Prediction at Leukapheresis

Figure 3 presents the comparison between static and longitudinal models evaluated on the test set. By relying exclusively on information available at leukapheresis, this analysis highlights distinct differences in discriminative performance over time. While most static models demonstrate strong initial performance, their accuracy steadily degrades as the prediction horizon extends. Conversely, several dynamic models, particularly Dynamic-DeepHit (mean tvAUC of 0.71) and Last-visit RSF (mean tvAUC of 0.68), demonstrate the ability to maintain or increase their discriminative power at extended future horizons. However, not all dynamic approaches systematically improve discrimination, as the Joint Model and RNN-FNN often perform comparably to, or worse than, static models (DeepSurv has a mean tvAUC of 0.69). Overall, the hybrid RNN-CPH model achieved the most consistent and superior performance across most landmark times. (This result is reconfirmed on the external validation set, where the RNN-CPH maintained its position as the top-performing model, see Section 3.2.3 of the Thesis).

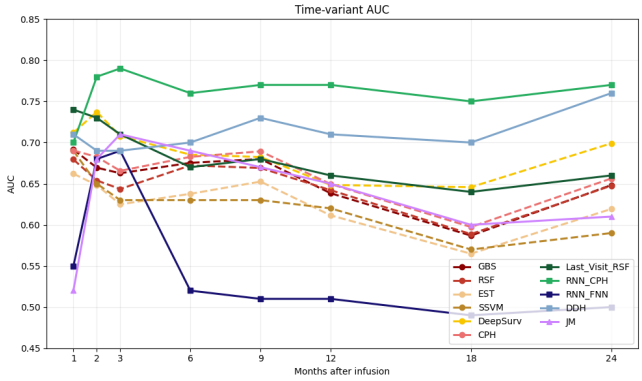

Figure 3: tvAUC on test set over months after infusion. Comparison between longitudinal models (solid lines) and static benchmarks (dashed lines) using exclusively data at leukapheresis.

3.4. Dynamic Prediction

Dynamic architectures were additionally evaluated based on their ability to update risk predictions as new longitudinal measurements became available. Performance was still quantified using the tvAUC evaluated at multiple future horizons.

To show how performance varies as data updates, we report the mean tvAUC at different landmarks in Table 3. Joint Models demonstrated low discriminative performance. Conversely,Dynamic-DeepHit exhibited a strong temporal imbalance, showing poor performance at short prediction horizons and a tendency to overfit based on an over-reliance on late-stage clinical history. Within the superlandmarked framework, naive approaches such as the Last visit Random Survival Forest (Last-visit RSF) showed clear limitations, losing information on the temporal trend of the disease. Similarly, pairing a sequential extractor with a standard neural network for survival (RNN-FNN) caused catastrophic overfitting, as the intra-subject correlation introduced by super-landmarking destabilized the network's gradients.

Among all evaluated dynamic strategies, the hybrid RNN-CPHmodel, trained within the superlandmarking framework, emerged as a highly promising approach, generally yielding more favorable metrics than the other alternatives tested. Table 2 summarizes its robust discriminative performance across key prediction horizons.

Table 2: tvAUC values for the RNN-CPH model on the Test set.

	1 month after infusion	2 months after infusion	3 months after infusion	6 months after infusion	9 months after infusion	12 months after infusion	18 months after infusion	24 months after infusion
Leukaph.	0.70	0.78	0.79	0.76	0.77	0.77	0.75	0.78
Infusion	0.71	0.78	0.79	0.76	0.77	0.77	0.75	0.78
Day 30	X	0.80	0.81	0.76	0.77	0.77	0.75	0.78
Day 60	X	X	0.92	0.71	0.74	0.74	0.72	0.75
Day 90	X	X	X	0.65	0.70	0.71	0.68	0.73
Day 180	X	X	X	X	0.75	0.76	0.69	0.75

mean tvAUC of Longitudinal Models

Landmark	JM	DDH	RSF	RNN-FNN	RNN-CPH
Leukapheresis	0.65	0.72	0.69	0.57	0.77
Infusion	0.63	0.70	0.69	0.54	0.77
Day 30	0.58	0.76	0.65	0.55	0.78
Day 60	0.53	0.77	0.58	0.47	0.74
Day 90	0.54	0.79	0.56	0.47	0.70
Day 180	0.45	0.64	0.53	0.66	0.74

Table 3: mean tvAUC values for dynamic models across all prediction horizons (1 to 24 months) for each landmark, calculated from the test dataset.

4. Discussion

The analysis of clinical outcomes in CAR-T therapy highlights the critical role of temporal alignment and model architecture in predicting PFS. By standardizing the timeline with infusion as a fixed reference anchor, the study successfully harmonized heterogeneous real-world data, enabling a consistent comparison of patient trajectories regardless of manufacturing delays or logistical variations.

In static predictions, the CPH model proved to be the most pragmatic choice. Although deep learning models like DeepSurv showed superior internal performance, the CPH model maintained higher generalizability on the independent external validation cohort. The study suggest that for a focused set of clinical variables, simpler models are often sufficient to capture the underlying prognostic signal without the risks of overfitting associated with more complex architectures.

Transitioning to dynamic prediction, Joint Models faced severe computational limitations and required strict distributional assumptions. Conversely, while Dynamic-DeepHit circumvented these constraints, they exhibited an unbalanced performance, struggling with early-stage predictions due to a potential over-reliance on late-stage history. More broadly, training models on complete longitudinal trajectories seems to overrely on recent measurements collected close to an event, leading to unpredictable behavior in clinical practice when a physician requests an early risk estimate based on limited history. Additionally, early-failing patients influence the training phase but are absent during later landmark evaluations (which consider only event-free subjects), creating a severe train-evaluation population mismatch. To resolve this, the study implemented a super-landmarking strategy.

The study identifies the hybrid RNN-CPH architecture, implemented within a super-landmarking framework, as the most effective predictive tool examined. The RNN-CPH achieves superior performance because its recurrent encoder seems to learn the entire biological trajectory and clinical history of the patient. If it captures these complex temporal patterns during training, the model can effectively anticipate clinical evolution even when predictions are based solely on the baseline leukapheresis stage. This allows the framework to maintain robust, stable discriminative performance (tvAUC around 0.75) across all time horizons, effectively mitigating conditioning bias through the super-landmarking approach.

This framework could allow to advance personalized medicine by enabling optimal patient selection before therapy and dynamic risk adaptation during follow-up. Clinicians can leverage these periodically updated risk estimates to anticipate unfavorable clinical evolutions and adjust supportive care or therapeutic interventions in real-time.

5. Conclusions

This study aimed to employ longitudinal models for the prediction of CAR-T therapy efficacy in patients with B-NHL. The findings suggest that integrating longitudinal clinical measure-

ments through a hybrid RNN-CPH architecture, utilizing a super-landmarking framework, could enhance risk prediction in CAR-T therapy. This model appeared to maintain robust performance across time horizons, indicating a potential to anticipate disease progression even when relying solely on baseline leukapheresis data.
However, several limitations should be acknowledged. These include the need for broader external validation to evaluate true generalizability, potential intra-subject correlation introduced by the super-landmarking strategy, the artificial temporal alignment of follow-up windows, and the challenges posed by data missingness and limited model interpretability.
Despite these constraints, this framework demonstrates the potential to support personalized medicine. By enabling optimal patient selection and dynamic treatment adaptation, it could help maximize the efficacy of CAR-T therapy in real-world clinical settings.

References

[1] Na Wang, Yunchong Meng, Yaohui Wu, Jing He, and Fang Liu. Efficacy and safety of chimeric antigen receptor t cell immunotherapy in b-cell non-hodgkin lymphoma: a systematic review and meta-analysis. *Immunotherapy*, 2021. doi: 10.2217/imt-2020-0221.

[2] Ping Wang, Yan Li, and Chandan K. Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51(6):110, 2019. doi: 10.1145/3214306.

[3] Eleni-Rosalina Andrinopoulou, Michael O Harhay, Sarah J Ratcliffe, and Dimitris Rizopoulos. Reflection on modern methods: dynamic prediction using joint models of longitudinal and time-to-event data. *International Journal of Epidemiology*, 50(5):1731–1743, 2021.

[4] Changhee Lee, Jinsung Yoon, and Mihaela Van Der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2019.

[5] Wieske K de Swart, Marco Loog, and Jesse H Krijthe. A comparative study of methods for dynamic survival analysis. *Frontiers in Neurology*, 16:1504535, 2025.