POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

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

Tesi di Laurea Magistrale in
Biomedical Engineering - Ingegneria Biomedica

Marco Simoni, 251796

Advisor:
Prof. Alessandra Laura Giulia Predocchi

Co-advisors:
Prof. Vanja Miskovic
Sara Ferri

Academic year:
2025-2026

Abstract: Chimeric Antigen Receptor T-cell (CAR-T) therapy has revolutionized the treatment of relapsed or refractory B-Cell Lymphoma, but durable remission is not observed in all patients. Early identification of subjects at high risk of disease progression is crucial. This study evaluates static and dynamic Machine Learning (ML) and Deep Lerning (DL) architectures to predict Progression-Free Survival (PFS) in a real-world multicenter cohort of 1,465 patients enrolled in the CART-SIE study. I utilized clinical and laboratory variables collected at leukapheresis , infusion, and subsequent standardized follow-ups (between 30 and 180 days post-infusion). For static prediction, Cox Proportional Hazards (CPH) and ML models were compared with deep neural networks (DeepSurv). Although DeepSurv achieved superior internal discriminative performance, the standard CPH model maintained a higher C-index on the independent external validation cohort (0.73 versus 0.69), demonstrating greater generalizability. To leverage longitudinal data, dynamic predictive frameworks were developed, comparing Joint Models, Dynamic-DeepHit and superlandmarking strategies based on Recurrent Neural Networks (RNN). Results demonstrate that encoding the patient's entire trajectory through a hybrid RNN-CPH architecture offers the best predictive capabilities. Comparisons among all models were performed using time-variant AUC (tvAUC) at leukapheresis on both test and external validation sets. In these evaluations, the RNN-CPH architecture achieved achieved superior results, likely because its recurrent encoder successfully learned from the temporal evolution of patient trajectories In conclusion, while the standard CPH confirms itself as a solid model for static predictions, the integration of sequential clinical measurements through hybrid recurrent architectures within a super-landmarking framework offers a substantial advantage for continuous, dynamic risk stratification in CAR-T therapy.

Key-words: CAR-T Therapy, Dynamic Survival Analysis, Machine Learning, Super-landmarking, Recurrent Neural Networks, Large B-Cell Lymphoma

1. Introduction

1.1. Context

B-cell Non-Hodgkin Lymphomas (B-NHL) represent a heterogeneous group of malignancies characterized by diverse clinical courses, ranging from indolent to highly aggressive forms. These neoplasms account for approximately 85–90% of all Non-Hodgkin lymphomas and occur with an incidence of roughly 20 new cases per 100,000 people per year in the Western world [1].
The main categories of B-NHL include: Diffuse Large B-cell Lymphoma (DLBCL), Primary Mediastinal Large B-cell Lymphoma (PMBCL), Follicular Lymphoma (FL), and Mantle Cell Lymphoma (MCL). DLBCL is the most common aggressive B-NHL, accounting for approximately 30–40% of B-NHL cases. PMBCL is recognized as a distinct clinicopathological entity within the large B-cell lymphoma category; while it shares some features with DLBCL, it displays a unique molecular profile and typically presents as a large mediastinal mass. FL accounts for around 25% of cases and represents the most prevalent indolent lymphoma, characterized by a follicular growth pattern and a chronic, relapsing course. MCL is notable for its broad clinical spectrum, which includes both indolent leukemic non-nodal forms and highly aggressive nodal variants, with a frequency of approximately 5–7% among B-NHLs[1].

The standard treatment for these B-cell lymphomas generally involves a combination of immunotherapy (targeted antibodies) and conventional chemotherapy, with the R-CHOP regimen still considered a backbone for many patients. For DLBCL, new first-line options include combining antibodies like polatuzumab vedotin with chemotherapy, while high-risk cases may require more intensive drug protocols. FL is often managed with a "watch and wait" approach if there are no symptoms, but when treatment is needed, it typically involves antibody-chemotherapy combinations followed by two years of maintenance therapy. MCL treatment depends on the patient's age and fitness: younger patients often undergo intensive chemotherapy followed by an autologous stem cell transplant, whereas older patients receive standard drug combinations like ibrutinib [1].

In cases where these therapeutic strategies prove ineffective, Chimeric Antigen Receptor T cell (CAR-T) therapy represent a promising and revolutionary form of immunotherapy that has emerged as a major breakthrough in cancer treatment. It is based on the genetic modification of a patient's own T cells to express synthetic (chimeric) receptors that enable specific recognition and binding to antigens expressed on the surface of malignant cells [2].
Compared with conventional cancer treatments such as chemotherapy and radiation therapy, CAR-T cell therapy offers several potential advantages, most notably its highly targeted mechanism of action. This personalized approach harnesses the patient's immune system to selectively eliminate malignant cells while largely sparing healthy tissues, thereby achieving high specificity and potency. As a consequence of this precise targeting, CAR-T cell therapy may reduce the incidence of off-target toxicities and improve overall clinical outcomes. Patients undergoing CAR-T cell therapy often experience fewer adverse effects commonly associated with traditional treatments, contributing to an improved quality of life during and after therapy [2] .

The structure of the CAR-T is composed of three main domains. The extracellular domain constitutes the outer portion of the receptor and functions as the antigen-sensing component. It contains an antibody-derived single-chain variable fragment (scFv) responsible for the specific recognition of tumor-associated antigens, as well as a hinge region that provides the structural flexibility required to facilitate effective antigen binding[2]. The transmembrane domain spans the cell membrane and serves as an anchoring element, ensuring stable integration of the receptor within the T-cell membrane[2]. The intracellular domain, located inside the cell, is responsible for initiating T-cell activation and the cytotoxic response. It includes the CD3ζ signaling chain, which delivers the primary activation signal leading to tumor cell killing, together with one or more co-stimulatory domains, such as CD28 or 4-1BB, which enhance T-cell proliferation, persistence, and long-term survival [2].

The mechanism of action of CAR-T cells can be described as a multistep process. First, recognition and binding occur when the extracellular scFv functions as a high-affinity sensor, specifically identifying and binding the target antigen (such as CD19) on the surface of the cancer cell. This interaction induces T-cell activation, triggering a conformational change that initiates intracellular signaling through the CD3ζ chain, thereby mimicking physiological T-cell receptor activation. Following activation, CAR-T cells undergo rapid expansion and cytokine release, both ex vivo and in vivo, producing potent immunomodulatory cytokines that recruit and activate additional immune effector cells. The combined effects of direct cell-mediated cytotoxicity and localized cytokine secretion result in targeted killing of malignant cells. Finally, a distinctive feature of CAR-T therapy is cellular persistence, as CAR-T cells may remain in the patient's body for extended periods, providing a durable antitumor response and acting as a living therapy capable of long-term immune surveillance against

disease recurrence [2].

The manufacturing process encompasses all steps from cell collection to final reinfusion into the patient. The process begins with leukapheresis, during which peripheral blood mononuclear cells are collected from the patient; the cellular product is then shipped to a specialized manufacturing laboratory for further processing. At the laboratory, T lymphocytes are isolated from the initial cellular product and subsequently activated, typically through stimulation of the CD3 and CD28 receptors, in order to enable genetic modification[3].
Subsequently, genetic transduction is performed, during which the gene encoding the chimeric antigen receptor is introduced into the T-cell genome. In most approved CAR-T products this step relies on viral vectors, such as lentiviruses or retroviruses, although non-viral strategies, including CRISPR/Cas9-based editing or mRNA electroporation. The genetically modified T cells then undergo ex vivo expansion to reach the therapeutic dose required for infusion [3].
While CAR-T cells are being manufactured, patients may receive bridging therapy to control disease progression. In the days immediately preceding CAR-T infusion, patients undergo lymphodepleting chemotherapy, a conditioning regimen designed to transiently reduce endogenous lymphocytes. This step creates a favorable immunological environment by decreasing competition for homeostatic cytokines and immune niches, thereby enhancing CAR-T cell engraftment, expansion, and persistence after infusion. Once the CAR-T cell product has successfully passed quality control testing, the therapy culminates in the final infusion of the engineered cells, as illustrated in Figure 1 [3].

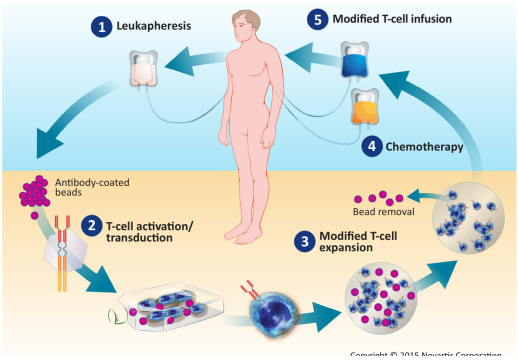

Figure 1: Overview of manufacturing process of CAR-T.[4]

In conventional manufacturing settings, the overall vein-to-vein time (i.e., the time from leukapheresis to infusion) typically ranges from 6 to 12 weeks. However, in real-world clinical practice, logistical and administrative constraints may substantially prolong this timeline, with waiting times for manufacturing slots ranging from 1 to 10 months. Additionally, approximately 4–7% of patients ultimately do not receive CAR-T therapy due to technical failures during cell manufacturing [3].
Despite the remarkable clinical efficacy and targeted mechanism of action of CAR-T cell therapy, as well as its potential for durable responses, only approximately 50% of infused patients achieve complete remission [5].
In this context, CAR-T cell therapy is a complex, time-consuming, and costly procedure. This economic and logistical burden underscores the importance of precision medicine: identifying the right patients not only improves clinical outcomes but also ensures the sustainability of healthcare systems by avoiding expensive treatments for those unlikely to respond [6].

Several clinical and laboratory parameters have been associated with outcomes following CAR-T therapy in B-Cell Lymphoma. Markers reflecting tumor burden, such as elevated lactate dehydrogenase (LDH), have consistently been linked to poorer clinical outcomes [7]. Inflammatory and general condition markers, including ferritin, albumin, and hemoglobin, also carry prognostic value: high ferritin and low albumin/hemoglobin often reflect an adverse inflammatory and have been associated with a higher risk of relapse or death. In addition, elevated inflammatory markers have been linked not only to reduced durability of response but also to a higher incidence of treatment-related toxicities, including cytokine release syndrome (CRS). Moreover, immune cell counts at leukapheresis can be informative, as low lymphocyte levels and broader signs of immune depletion may indicate reduced T-cell fitness, potentially limiting CAR-T expansion and contributing to treatment failure [8, 9].
Disease-related characteristics such as advanced Ann Arbor stage (III/IV), the presence of multiple extranodal sites (e.g., ≥ 2), bulky disease, and a high International Prognostic Index (IPI) score remain relevant prognostic factors even in the CAR-T setting. In addition, patient-specific variables including older age, poor functional status (particularly an ECOG performance status ≥ 2) and a high number of prior treatment lines have demonstrated prognostic relevance in both clinical trials and real-world studies; extensive prior therapy may also

compromise T-cell quality and increase the likelihood of manufacturing failure[8, 9].
Collectively, these findings highlight the importance of identifying and integrating reliable biomarkers to improve risk stratification and outcome prediction in CAR-T-treated patients.

Artificial Intelligence (AI) is a field of research that brings together methods and techniques aimed at reproducing, within computational systems, forms of "intelligence" applied to real-world problems across many domains (from engineering to economics and law). Within this broad framework, AI includes several sub-areas such as reasoning, expert systems, knowledge representation, neural networks, natural language understanding, data mining, and machine learning. Machine learning (ML) is described as an approach in which software acts as a "learning apprentice," meaning an agent that learns from data in order to characterize and extract structure, syntax, and meaning from information. Learning is carried out through iterative procedures and specific algorithms [10].
From a predictive standpoint, these models are valuable because they can learn patterns from historical data and use them to generate forecasts for outcomes of interest, especially in settings where relationships among variables are complex and difficult to specify through explicit rules. By transforming observed data into learned decision rules, AI-based approaches can support planning and decision-making, improved resource allocation, helping anticipate critical parameters and future trends in a wide range of applications. Among these applications, medicine occupies a central role: currently, its uses range from the virtual branch, such as ML for the identification of DNA variants and the management of electronic medical records; to the physical branch, which includes robot-assisted surgery and nanorobots for targeted drug delivery. AI is expected to become an increasingly pervasive component of the future healthcare system. This evolution is essential to enhance diagnostic accuracy through clinical decision support systems, which are key tools for achieving precision medicine and enabling early diagnosis and targeted prevention. Looking ahead, AI will not only assist physicians in making optimal decisions in real time, but will also evolve into a personal service for consumers, capable of analyzing gene–environment interactions to suggest personalized lifestyles and improve the overall design of healthcare delivery [11].
Even in the context of CAR-T for lymphomas the AI finds its application. The literature [12] report different instances of its utility, like:

- ML-models to quantify CAR-T cell immunological synapse quality, demonstrating a correlation with antitumor activities [13].
- explore the spectrum of cellular responses achievable through the CAR-T and enhanced their design with a more quantitative and predictive approach [14]
- detect the onset of Cytokine release syndrome (CRS) [15] or predicts the risk of progression to high-grade CRS after its initial onset in patients receiving CAR-T therapy [16].

In healthcare, due to the nature of clinical events evolving over time, being able to collect and analyze repeated measurements is essential to properly model disease trajectories and make accurate predictions. Longitudinal data allow researchers to capture temporal correlations, dynamic changes in biomarkers, and complex nonlinear relationships that cannot be fully represented using cross-sectional information . Compared with traditional statistical approaches, machine learning methods applied to longitudinal data are particularly well suited for handling high-dimensional variables and modeling nonlinear effects that are not known a priori . Moreover, prediction accuracy has been shown to improve when longitudinal information is adequately incorporated, enabling earlier identification of risk and better outcome forecasting. However, several challenges limit their application. Longitudinal datasets frequently contain missing measurements, patient dropouts, and uneven time intervals between observations, which for standard ML time-series algorithms is a problem since are typically built on the assumption of complete and regular samples, thus requiring additional preprocessing strategies. Repeated measures for an individual are inherently correlated, which tend to break the standard "independent and identically distributed" (i.i.d.) assumption of many ML algorithms, leading to biased results. Again longitudinal data trajectories may be highly complex and non-linear (e.g., large variations between individuals), breaking again the i.i.d. assumption. More complex models, such as recurrent architectures, typically require large sample sizes and adapted interpretation tools , while many machine learning approaches remain difficult to interpret due to their "black-box" nature. Consequently, although longitudinal models offer substantial predictive advantages, their implementation demands careful methodological choices and robust validation [17].

1.2. Aim of the study

This work is conducted within the framework of the Italian Observational Study on CAR-T Therapy for Lymphoma (CART-SIE), a prospective, multicenter, observational real-world study aimed at evaluating the feasibility and effectiveness of chimeric antigen receptor T-cell therapy in patients with relapsed or refractory B-cell lymphomas. The study consecutively enrolls patients with DLBCL, HGBCL, PMBCL, MCL, and FL referred to Italian hematologic centers qualified for CAR-T treatment and assessed for eligibility according to current

regulatory criteria [18].
CART-SIE is designed to systematically collect clinical data across the CAR-T treatment pathway, including patient eligibility status, leukapheresis, manufacturing outcomes, and treatment administration, together with outcome-related information such as disease response and patient status over time. This enables the assessment of treatment effectiveness in a real-life clinical setting and the characterization of patient trajectories after CAR-T evaluation and treatment [18].

The aim of this thesis is to develop and evaluate longitudinal machine learning models for outcome prediction in patients with B-Cell Lymphoma undergoing CAR-T cell therapy. Specifically, this work evaluate how the inclusion of follow-up data impacts predictive performance relative to static models, and quantify the incremental benefit of longitudinal information beyond what can be achieved using leukapheresis data alone.
In a prospective clinical perspective, such time-updated risk estimates could be used to support longitudinal patient monitoring and to inform risk-adapted management during follow-up, i.e., to anticipate unfavorable evolution and consider timely adjustments of supportive care or additional therapeutic interventions in light of the expected patient trajectory.

2. Materials and Methods

The methodological workflow followed in the present work is outlined in Figure 2, and distinguished with respect to the two tasks: static and dynamic prediction.

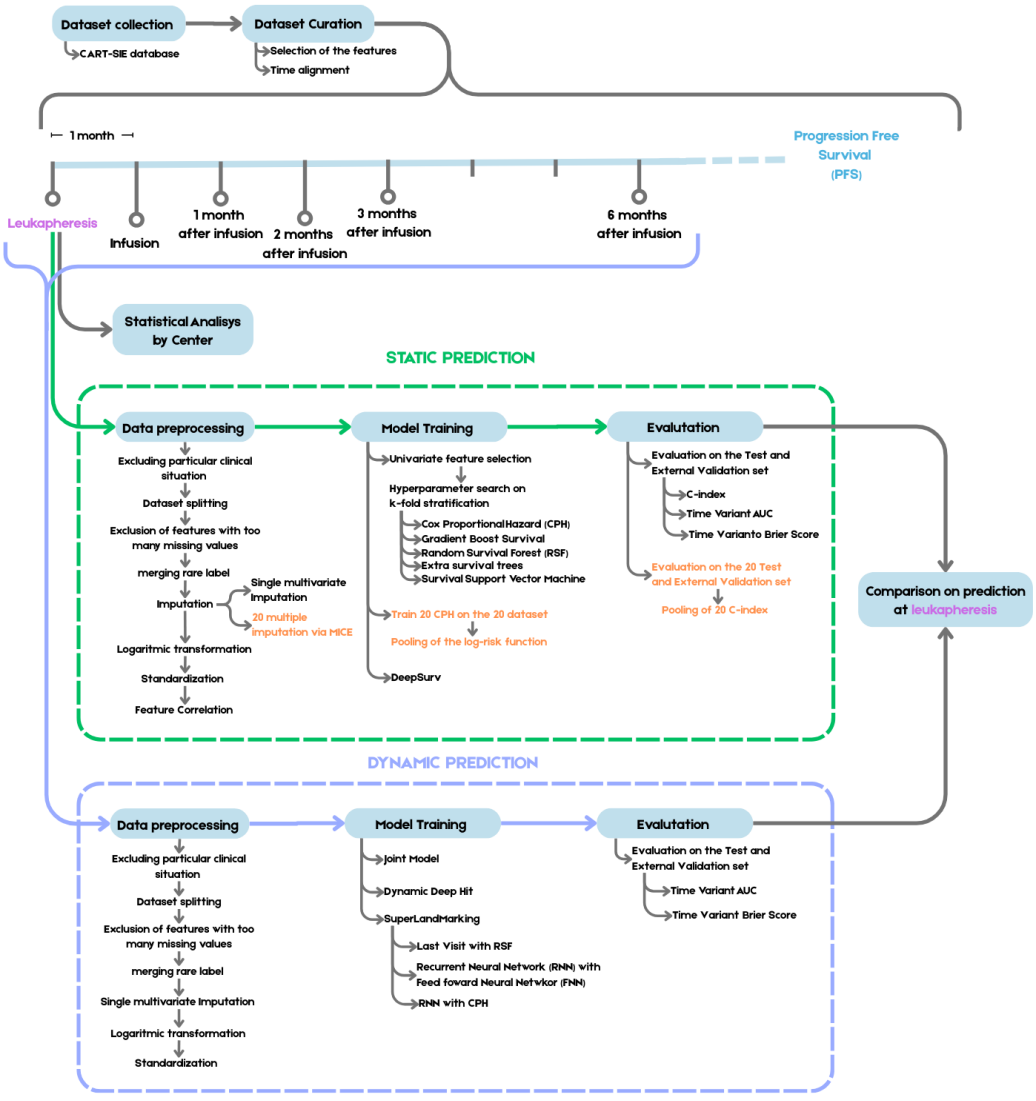

Figure 2: Methodological workflow

2.1. Dataset Collection

As part CART-SIE, a prospective, multicenter, observational cohort study, clinical data were collected on patients with a diagnosis of DLBCL, PMBCL, MCL, or FL, who were eligible for CAR-T treatment with commercially available products in Italy: tisagenlecleucel (tisa-cel), axicabtagene ciloleucel (axi-cel), brexucabtagene autoleucel (brexu-cel), lisocabtagene maraleucel (liso-cel). The dataset was built to represent the real-world CAR-T pathway, capturing patients across key stages of the process: screening/eligibility, leukapheresis, manufacturing, and infusion

Different categories of variables were collected throughout the study trajectory. The dataset includes clinical characteristics and disease-related information, treatment-pathway variables, and longitudinal outcome data. Outcomes include response assessment in routine clinical practice according to Lugano criteria (PET/CT-based), as well as major time-to-event endpoints such as overall survival (OS), progression-free survival (PFS), and duration of response (DoR). In the present work, PFS is adopted as the prediction target for the developed models. Safety and tolerability were documented through the monitoring of early and late adverse events (including, among others, CRS and neurotoxicity, infections, cytopenias, and second malignancies) throughout follow-up. When available, additional biologically oriented variables and biomarker-related information were collected within the optional biological component of the study [18].

2.2. Dataset Curation

Across all modeling approaches, leukapheresis data provide the earliest standardized information available for all subjects and are essential for early response prediction. Accordingly, leukapheresis variables were included in every model, and the candidate features are summarized in Table 10.
For clarity and downstream modeling purposes, variables were organized by data type into two categories: Categorical data (further divided into Binary, Nominal, and Ordinal) and Numerical data.
Given the longitudinal design of CART-SIE, it is important to note that not all features listed in Table 10 are systematically collected at every assessment time. While the complete feature set is available at leukapheresis, only a subset of features is repeatedly recorded at infusion and during follow-up visits at approximately 1, 2, 3, and 6 months after infusion.
Finally, because the time interval between leukapheresis and infusion varies across patients, for modeling purposes, infusion was assumed to occur one month after leukapheresis for all subjects, thereby defining a common temporal reference across patients.

2.3. Outcome

Among the time-to-event outcomes collected from CART-SIE, progression-free survival (PFS) was selected as the primary endpoint. PFS is defined as the time from treatment infusion to disease progression or death from any cause, whichever occurs first [19].
However, in a real-world observational setting, the exact time to progression cannot always be observed for all patients. In longitudinal observational studies such as CART-SIE, follow-up may end before the event of interest occurs for several reasons, including: administrative end of study, transfer of care to other institutions, withdrawal of consent, or loss to follow-up due to clinical or logistical factors. As a consequence, for a subset of patients the exact time to progression or death is not observed. This results in right censoring, whereby the true event time is unknown but is known to occur after the last recorded follow-up time. Right-censored observations therefore contribute only partial information on the underlying time-to-event process and require analytical approaches that properly accommodate incomplete event-time data, as illustrated in Figure 3 [18] [20]. To account for this, the PFS outcome is represented by a pair (t, δ), where t denotes the observed follow-up time and δ is an event indicator: $\delta = 1$ if disease progression or death is observed at time t, and $\delta = 0$ if the patient is right-censored, meaning that the event has not been observed by the end of follow-up and is only known to occur after t.

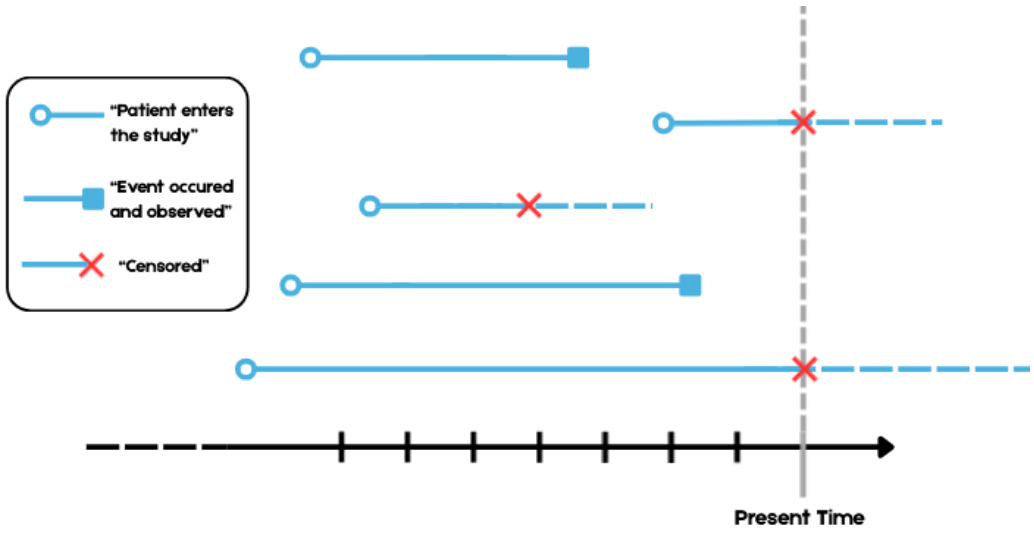

Figure 3: Visual representation of right censoring in time-to-event data. The blue lines denote individual PFS trajectories. If the actual event (disease progression or death) is not observed for any reason, the patient is right-censored at the time of their last available follow-up.

2.4. Statistical Analysis by center

Since the data come from different centers, it was first investigate whether patients from different centers show significant differences across features [20]. For each feature I start with an omnibus (global) hypothesis test that evaluates the null hypothesis that all centers share the same underlying distribution for that feature [21].

For Numerical and Ordinal Features, I first check whether the within-center samples are compatible with a normal model. If normality is plausible for all groups, I use a parametric global test one-way ANOVA, where the null hypothesis is that the mean value of the feature is equal across all centers; rejecting it indicates that at least one center differs in average level [22–24]. If normality is not plausible, I adopt the Kruskal–Wallis test, a non-parametric omnibus procedure that serves as an analog to one-way ANOVA. This method converts raw data into ranks to compare the centrality of distributions, posing the null hypothesis that all independent groups originate from populations with the same median. Rejecting this hypothesis suggests a systematic shift in the feature's distribution across centers, allowing for the detection of differences without relying on Gaussian assumptions [21, 22, 25]. When the global test is significant, I perform post-hoc pairwise comparisons to identify which centers differ from which others, Tukey's procedure after ANOVA, or Dunn's test after Kruskal–Wallis [21, 23]. Because many pairwise tests are run, p-values are adjusted for multiple comparisons to control false positives (type 1 error) with the Bonferroni correction [21–23, 25]. The resulting pairwise significance pattern is then used to highlight potential "outlier" centers, defined as centers that differ significantly from a large fraction of the others, and visual diagnostics (density curves for numerical features, proportion plots for categorical features) are used to interpret the direction and magnitude of these differences.

For binary and multi-class features, I aim to check whether different centers show similar or different proportions across the categories. I first run a global chi-square test of independence using a contingency table where rows represent centers and columns represent the category levels. The null hypothesis is that the categorical outcome is independent of the center, meaning that all centers have the same distribution of categories (e.g., similar percentages of 0/1 in a binary variable, or similar percentages across all levels in a nominal variable) . If this global test is significant, it suggests that at least one center differs [26, 27]. To understand where the differences come from, I then perform pairwise comparisons between centers by repeating the chi-square test on two centers at a time. Because many pairwise tests are carried out, I still apply a multiple-comparison correction, such as Bonferroni [26, 27].

2.5. Survival Analysis

Survival analysis provides a methodological framework for the analysis of time-to-event outcomes (like PFS) in the presence of censoring. A central assumption is that censoring is non-informative, meaning that the

censoring mechanism does not convey prognostic information about the subsequent event process. Moreover, an adequate duration of follow-up is needed to observe a sufficient number of events, and, in multicenter settings, methodological consistency across participating centers is essential to ensure the comparability and validity of survival estimates [20].
Two fundamental functions are used to describe time-to-event data in survival analysis: the survival function and the hazard function.

The survival function, denoted as $S(t)$, represents the probability that an individual survives from the time origin (for example, diagnosis or treatment initiation) up to a specified future time t. From a clinical perspective, the survival function provides a concise and interpretable summary of the survival experience of a cohort, directly quantifying the proportion of individuals who remain free from the event of interest over time. By its nature, $S(t)$ is a cumulative quantity, as it reflects the cumulative non-occurrence of the event, that is, the probability that the event of interest has not yet occurred up to time t. As such, it provides a global description of the survival experience of the study population and is central to the interpretation of time-to-event outcomes [20].

In practice, the survival function is commonly estimated using the non-parametric *Kaplan-Meier estimator*, which allows both observed event times and right-censored observations to be appropriately incorporated into the analysis. When survival curves are estimated for different groups, it is often necessary to formally assess whether observed differences in survival experience are statistically meaningful. The *log-rank test* is a widely used non-parametric hypothesis test for comparing the survival distributions of two or more groups. Its null hypothesis states that the survival functions of the groups being compared are identical over time. A statistically significant result indicates that the survival experiences of the groups differ more than would be expected by chance [20].

Complementarily to the Survival function, the hazard function $h(t)$ characterizes the instantaneous risk of experiencing the event at time t, conditional on having remained event-free up to that time point. Unlike the survival function, which focuses on the cumulative absence of the event, the hazard function captures the local intensity of event occurrence and provides insight into how risk evolves over time. It represents the instantaneous event rate for an individual who has survived up to time t and therefore offers a conditional measure of failure risk. This function plays a central role in survival modeling and forms the basis for regression-based approaches in time-to-event analysis [20, 28].

The Cox Proportional Hazards (CPH) model, is the most widely used multivariable technique in medical research for the analysis of time-to-event data. Unlike univariate approaches such as the log-rank test, it enables the simultaneous assessment of the effects of multiple covariates on the risk of experiencing the event of interest, providing clinically interpretable estimates in terms of hazard ratios (HRs). The CPH model is defined as semi-parametric, as it does not require specification of the underlying distribution of survival times and instead directly models the hazard function, representing the instantaneous event risk at time t conditional on survival up to that time. The model is formally expressed as:

$$h(t) = h_0(t)\exp(h_\beta(x)) = h_0(t)\ \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right), \tag{1}$$

where $h_0(t)$ denotes the baseline hazard, representing the underlying risk when all covariates are equal to zero, and $\beta_1, \ldots, \beta_p$ are regression coefficients quantifying the effect of each covariate on the hazard. In this formulation, the quantity $h_\beta(x) = (\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$ will be referred to as the "log-partial hazard" or "log-risk function", while its exponentiated counterpart $\exp\left(h_\beta(x)\right)$ will be referred to as the partial hazard [28, 29]. Exponentiating the regression coefficients yields the HR, defined as $\mathrm{HR}_i = \exp(\beta_i)$ which measure the relative change in the instantaneous risk associated with a one-unit increase in the covariate x_i, holding all other variables constant. An HR > 1 indicates an increased hazard of the event, whereas an HR < 1 denotes a reduced hazard [28, 30].

The baseline hazard function is estimated nonparametrically; therefore, unlike fully parametric survival models, the CPH model does not require assuming a specific distribution for survival times. The regression coefficients β are learned from the training data by maximizing the CPH partial likelihood, which can be interpreted as the product, over all observed events, of the conditional probability that the individual who fails at time T_i is the one who experiences the event, given the set of individuals still at risk at that time. The CPH partial likelihood is parameterized by β and is defined as

$$L_c(\beta) = \prod_{i:E_i=1} \frac{\exp\left(\beta^\top x_i\right)}{\sum_{j\in\mathcal{R}(T_i)} \exp\left(\beta^\top x_j\right)}, \tag{2}$$

where the values T_i, E_i, and x_i are the respective event time, event indicator, and data for the i^{th} observation. The product is defined over the set of patients with an observable event $E_i = 1$. The risk set $\mathcal{R}(t) = \{\, i : T_i \geq t \,\}$ is the set of patients still at risk of failure at time t [28, 31].

For the CPH model to yield valid and interpretable results, several key assumptions must be satisfied. The central assumption is that of proportional hazards, which requires that the ratio of hazards between any two individuals or groups remains constant over time. Additionally, the model assumes non-informative censoring, independence of observations, and a linear and multiplicative relationship between the logarithm of the hazard and the covariates. Violation of these assumptions may lead to biased estimates and misinterpretation of covariate effects [28].

2.6. Data preprocessing

As a first preprocessing step, I excluded patients with missing outcome data (PFS), as these observations cannot be used for model training and evaluation.

Following clinical advice, I excluded patients with a diagnosis of mantle cell lymphoma (MCL) from the analysis. This choice was motivated by the fact that MCL shows a clinical course and treatment pathway that are often not directly comparable with those of DLBCL, PMBCL, and FL. Predictive modeling in oncology achieves optimal performance when tailored to specific patient populations, as prognostic drivers often vary significantly across different malignancies. By excluding MCL patients from the study, I mitigate the risk of risk heterogeneity, preventing these aggressive and unique biological patterns from distorting the model's variable weights and compromising overall calibration. This methodological choice facilitates the development of a more robust and specialized model for the remaining diagnoses, which share more consistent prognostic indicators, thereby ensuring higher predictive accuracy within a more homogeneous population [32, 33] .
In addition, still following clinical advice, I excluded patients treated with the CAR-T product liso-cel. This decision was not motivated by biological heterogeneity, but by limited sample size and immature follow-up: as a more recently introduced product in our cohort, liso-cel was available for only a small number of patients, most of whom had short follow-up. To avoid unstable estimation and overly uncertain performance assessment driven by sparse data and heavy censoring, liso-cel recipients were therefore removed from the analysis.

Given the multicentric nature of this study, which captures the inherent heterogeneity in case mix and varying clinical practices across different institutions, I reserved patients from the hospital of "Roma Cattolica" as an independent external validation set. This design is essential to rigorously assess the model's cross-site transportability and to ensure that predictive performance remains robust despite potential covariate shifts or institutional differences in diagnostic and treatment procedures. True validation of a prognostic model necessitates its application to an independent third dataset to confirm that the results are generalizable and not merely a result of overfitting to center-specific patterns [34]. With the remaining data, I allocated 20% of patients to the test set. The remaining 80% was used either as a single training set, or further split into 70% of training and 10 % validation if required by the model. To preserve a comparable event/censoring composition across splits, I used stratified sampling based on the censoring indicator of PFS. Survival probability curves were generated for each set using the Kaplan–Meier estimator implemented in the lifelines package [29]. Differences between survival curves across the cohorts were assessed using the log-rank test, under the null hypothesis of equal survival functions. This comparison was used to verify that the data splits did not lead to marked differences in survival patterns across the partitions. Then all preprocessing steps were fitted only on the training set and then applied unchanged to the remaining datasets to prevent data leakage [35].

Since some of the models considered in the subsequent analyses cannot be applied directly in the presence of missing values, handling missingness represents one of the first methodological issues to be addressed during preprocessing [36].
Imputation methods can work up to 50% of missing data, but as it increases lower will be its preformances for this reason I choose as threshold 30% as acceptable [37] . With this rationale, features with >30% of missing data in the training set were excluded from the study.
In our pipeline, for most models I used scikit-learn's IterativeImputer, which implements a multivariate procedure and can be seen as conceptually derived from the MICE framework, although it produces a single imputed dataset rather than multiple imputations. [35]. The Multiple Imputation by Chained Equations (MICE) is a multivariate and multiple imputation approach, and is reported as one of the most effective imputation approaches that can handle together categorical and numerical variables [38, 39]. A full MICE-based procedure is explored as an additional experiment, which is described in a dedicated section later in this work.

Categorical and numerical features were preprocessed using separate procedures. For categorical variables, rare levels were merged into broader categories to reduce sparsity and improve model stability. These categorical features were then encoded according to their typology: ordinal variables were transformed using ordinal encoding, binary variables were mapped to 0/1 indicators, and nominal variables were encoded via one-hot encoding. For numerical variables, several transformations were considered to address right-skewness, including quantile transformation and power transformations (Box–Cox and Yeo–Johnson). However, quantile transformation may require substantial data to yield stable estimates, while power transformations can increase the risk of overfitting [35, 40]. I therefore adopted a simple logarithmic transformation, which is less aggressive in enforcing normality but is straightforward and robust. Finally, all numerical features were standardized using a StandardScaler fitted on the training set and applied unchanged to validation, test, and external validation data.
Finally, highly correlated features (with a Pearson correlation higher of 0.8 in the training set) were removed, with the final decision guided by clinical input rather than an automatic rule.

Multiple Imputation approach Apart from the imputation procedure, all preprocessing steps remained consistent with the main analysis, including data splits and feature selection
A full MICE) procedure was implemented using the miceforest package [41]. Specifically, $M = 20$ imputed versions of the training dataset were generated. The fitted MICE model was then applied to the test and external validation sets to obtain M corresponding imputations for each split, ensuring that no information from these datasets influenced the imputation process.
Following imputation, feature preprocessing was applied in a globally consistent manner across imputations.
For categorical variables, the merging of rare levels was identical to that adopted in the single-imputation pipeline. The encoding structure (i.e., the set of resulting categories) was therefore fixed and shared across all imputed datasets to ensure structural comparability.
After imputation, feature preprocessing was applied in a globally consistent manner across imputations. Rare categorical levels were merged exactly as in the single-imputation pipeline, ensuring a shared encoding structure across all datasets. For numerical variables, right-skewed features were log-transformed as previously described. Standardization was then performed using mean and standard deviation estimated on the merged imputed training datasets (i.e., stacking the M imputations), and the resulting scaling parameters were applied unchanged to the corresponding test and external validation sets. This strategy ensures structural consistency across imputations while isolating the impact of imputation variability from preprocessing-induced differences.

2.7. Static Prediction

In this phase models will be fitted using only data measured at leukapheresis with the objective to determine how a possible patient will respond to CAR-T therapy. Not all survival models are able to predict for a patient a Survival or Hazard function, but from all I can obtain a risk score. The models rank patients by risk: higher values indicate a higher predicted risk of experiencing the event earlier (here, progression or death when using PFS as the target). Since the score is inherently relative and lies on an arbitrary scale with no defined units, model performance and comparison are often assessed using rank-based concordance measures[35].

The most standard metric is the Frank Harrell's corncordance index (C-index). It takes the output of the model (often a risk score) and it uses it to compare all patients pair; labels them as concordant when the predicted ordering agrees with the observed ordering of event times (and discordant otherwise). The C-index is the proportion of concordant couples with respect to all the pair available:

$$\text{C-index} \;=\; \frac{\sum_i \sum_j \mathbb{I}(T_j < T_i)\, \mathbb{I}(\eta_j > \eta_i)\, \delta_j}{\sum_i \sum_j \mathbb{I}(T_j < T_i)\, \delta_j}, \tag{3}$$

where η_i is the predicted risk score for patient i, T_i is the observed follow-up time, δ_i is the event indicator ($\delta_i = 1$ if the event is observed and $\delta_i = 0$ if censored), and $\mathbb{I}(\cdot)$ denotes the indicator function. For interpretation, a perfect model achieves a C-index of 1.0, while 0.5 corresponds to random ordering. For context, a value with perfect perfomances score a 1.0 of C-index, meanwhile a 0.5 is equal to random prediction. With right-censored data, not all pairs are comparable because censoring can prevent us from knowing which patient truly would have had the event first. Harrell's C-index handles this by discarding pairs that are incomparable due to censoring and computing concordance only on the remaining comparable pairs. This lead to the limitation of this metrics: Harrell C-index can became dependent on the study specific censoring distribution. Also, this metric is defined as a global or "time indipendent" metric, because it returns a single value for the entire follow-up the period, which

means it does not take in account for how the risk (and model discrimination) may change over time [36, 42, 43].

To evaluate how the model discrimination changes over time a time-dependant metric such as the time-dependent AUC, tvAUC(t)). At a given time t, the outcome is treated as a time-specific binary status by defining cumulative cases as subjects who experienced the event by time t ($(T \leq t)$) and dynamic controls as subjects who remained event-free up to time t ($(T > t)$). Using the predicted risk scores, one can then construct the ROC curve at time t by varying the decision threshold and computing the corresponding time-dependent sensitivity and specificity; $tvAUC(t)$ is the area under this ROC curve and quantifies how well the model separates subjects who fail by t from those who fail after t. In the presence of right censoring, these quantities are typically estimated using inverse probability of censoring weights (IPCW), which adjust the contribution of observed subjects to account for the fact that some event statuses at time t are unobserved due to censoring [36, 42, 43]. The tvAUC at time t can be written as:

$$\widehat{\mathrm{AUC}}(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I(y_j > t)\, I(y_i \leq t)\, \omega_i\, I\Big(\hat{f}(x_j) \leq \hat{f}(x_i)\Big)}{\left(\sum_{i=1}^{n} I(y_i > t)\right)\left(\sum_{i=1}^{n} I(y_i \leq t)\, \omega_i\right)} \tag{4}$$

where $I(\cdot)$ denotes the indicator function, $\hat{f}(x_i)$ is the predicted risk score for subject i, y_i is the observed survival time, and ω_i are inverse probability of censoring weights (IPCW). The censoring distribution $\hat{G}(\cdot)$ was estimated exclusively on the training set using the Kaplan–Meier estimator and subsequently applied unchanged to the test and external validation sets.[36]

If I are interested not only in how well a model discriminates but also in how well it is calibrated, I consider the Brier Score $BS(t)$. This is a time-dependent extension of the mean squared error and measures the inaccuracy of predicted survival probabilities at a given time point t. In its basic form, it is computed as the mean squared difference between the predicted probability and the observed outcome:

$$\widehat{\mathrm{BS}}(t) = \frac{1}{n} \sum_{i=1}^{n} \left[I(y_i \leq t \wedge \delta_i = 1) \frac{(0 - \hat{\pi}(t \mid x_i))^2}{\hat{G}(y_i)} + I(y_i > t) \frac{(1 - \hat{\pi}(t \mid x_i))^2}{\hat{G}(t)} \right] \tag{5}$$

where $\hat{\pi}(t \mid x_i)$ is the predicted probability of remaining event-free up to time t for subject i, y_i is the observed time, δ_i is the event indicator, and $\hat{G}(\cdot)$ denotes the estimated survival function of the censoring time. A lower Brier score indicates better predictive performance. To account for right censoring in survival data, also $BS(t)$ is typically estimated using IPCW. The time-dependent BS is only applicable for models that are able to estimate a survival function. For instance, it cannot be used with Survival Support Vector Machines[36, 43].

2.7.1 ML Approach

The ML Approach is inspired by [44] and relies on established machine learning survival models implemented in the `scikit-survival` package [36], namely Gradient Boosting Survival (GBS), Random Survival Forest (RSF), Extra Survival Trees (EST), and Survival Support Vector Machines (SSVM).

Despite being a statistical model, the CPH can also be used for prediction, alongside ML models. In practice, once the regression coefficients β are estimated, each patient can be assigned an individual risk score through the linear predictor $(h_\beta(x)) = x_i\beta$ (or equivalently, the partial hazard $\exp(h_\beta(x))$. This score ranks individuals according to their relative risk across time [45]. The CPH model was fitted using the lifelines Python package [29].

The training pipeline consisted of two main steps: first, Univariate Feature Selection was performed to identify relevant predictors; then, GridSearch with 5-fold cross-validation was used to optimize model hyperparameters (Table 11). Cross-validation splits were stratified based on the censoring label to ensure balanced representation [44].

2.7.2 Multiple Imputation with CPH

Within the multiple imputation framework, a CPH model was fitted separately on each of the M imputed training datasets.
Rather than pooling the M sets of predicted risks, I aimed to obtain a single model that can be directly applied to new subjects. This strategy is feasible for regression-based models whose risk function depends linearly on estimated coefficients. In particular, the CPH model specifies the log-partial hazard as a linear function of covariates, making it possible to combine models at the parameter level. Specifically, the estimated regression

coefficients (log-partial hazard coefficients) were averaged across the M fitted models, resulting in a single pooled set of parameters that defines one final risk scoring function [46, 47].

For model evaluation, the test and external validation set were imputed M times using the MICE procedure fitted on the training data. The pooled model was then applied to each of the M imputed test/external validation datasets to generate predictions and compute the performance measure. The resulting M performance estimates were pooled across imputations using Rubin's rules.
The Rubin's rule state that the pooled point measure $\hat{\theta}$ is the aritmetic mean. meanwhile the total variance accounts for both sampling uncertainty and the uncertainty due to missing data:

$$\widehat{\mathrm{Var}}(\hat{\theta}) \;=\; \bar{W} \;+\; \left(1+\frac{1}{M}\right) B$$

where $\bar{W}$ is the within-imputation $\bar{W} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathrm{Var}}(\hat{\theta}_m)$, meanwhile B is the between-imputation variance $B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \hat{\theta}\right)^2$. (The Combined standard error is $\widehat{\mathrm{SE}}(\hat{\theta}) = \sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}$) [46, 47].

2.7.3 DeepSurv

As an alternative to ML models, a Deep Learning (DL) approach was explored. DeepSurv is a feed-forward neural network that models the effect of a patient's covariates on the hazard by learning a nonlinear risk function parameterized by the network weights θ. [31] The model was implemented using the TorchSurv package. [48]
The network architecture consists of two fully connected hidden layers with 16 and 8 neurons, respectively, each followed by a ReLU activation and dropout with rate 0.1. The network then collapses to a single output neuron. The network output $h_\theta(x)$ is a scalar with linear activation and represents the estimated log-risk (linear predictor) in the CPH model (Eq. 1).
I train the network by setting the objective function to be the negative log partial likelihood of Eq. 2:

$$\ell(\theta) := -\sum_{i:E_i=1}\left(h_\theta(x_i) - \log \sum_{j\in\mathcal{R}(T_i)} e^{h_\theta(x_j)}\right) \tag{6}$$

The network parameters are optimized by minimizing Eq. (6) using Adam with learning rate 10^{-3} and weight decay 10^{-4} [31, 48]. A learning-rate scheduler is applied, reducing the learning rate when the validation C-index reaches a plateau (i.e., when no improvement is observed for a fixed number of epochs). Finally, early stopping is performed based on the validation C-index with a patience of 50 epochs; the model parameters corresponding to the best validation C-index are retained and restored at the end of training.

2.8. Dynamic Prediction

In the study setting considered here, using only data available at leukapheresis produces a purely static prediction and therefore ignores both the changes that may occur during follow-up and the additional information collected over time that could be exploited for model training. Dynamic prediction overcomes this limitation by estimating and continuously updating a patient's prognosis as new clinical information becomes available during follow-up (e.g., longitudinal biomarker measurements). Unlike static leukapheresis-only models, this approach leverages the patient's longitudinal history to reflect the evolving nature of the individual's health status [49].
Formally, dynamic prediction is conditional on the information available up to a landmark time l, given that no event has occurred up to l. Let $\mathcal{H}_i(l)$ denote the subject-specific information observed up to l The event-free probability, i.e. the survival probability at future time t conditioned on remaining event-free up to l, is defined as [50–52]:

$$S_i(t \mid l, \mathcal{H}_i(l)) \;=\; \Pr(T_i^* > t \,|\, T_i^* > l,\; \mathcal{H}_i(l))\,, \qquad t > l. \tag{7}$$

Some models instead estimate the hazard function $h(t)$, which is the instantaneous risk that the event will occur. The hazard function is related to the survival function, so that :

$$S(t) = \exp\left(-\int_0^t h(u)\,du\right) \tag{8}$$

Because dynamic prediction aims to update risk over follow-up time, purely global ranking measures such as Harrell's C-index become less informative: they collapse performance into a single number and do not indicate

when (or at which horizons) the model discriminates well, nor how performance evolves as predictions are updated. For this reason, evaluation in a dynamic setting typically relies on time-dependent metrics, such as the tvAUC(t) and the BS(t), which assess discrimination and prediction error at specific time points or horizons. In the presence of right censoring, these quantities are commonly estimated using IPCW-based estimators to ensure that performance at time t is not biased by incomplete outcome information.

2.8.1 Joint Model

The first approach to dynamic prediction relied on a statistical framework, namely Joint Modelling (JM). The main idea is to model the longitudinal evolution of time-varying biomarkers through a mixed-effects model and subsequently link their latent trajectories to a CPH model, while adjusting for the remaining static covariates For each time-varying biomarker $k \in \{1, \dots, K\}$, I use the model of a linear mixed-effects model (LMM) to describe its longitudinal evolution. Let y_{kij} be the j-th measurement of marker k for subject i at time t_{ij}. The LMM is written as

$$y_{kij} \;=\; \mathbf{x}_{kij}^{\top}\boldsymbol{\beta}_k \;+\; \mathbf{z}_{kij}^{\top}\mathbf{b}_{ki} \;+\; \varepsilon_{kij}, \qquad \mathbf{b}_{ki} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_k), \qquad \varepsilon_{kij} \sim \mathcal{N}(0, \sigma_k^2), \tag{9}$$

Here, $\boldsymbol{\beta}_k$ is the vector of *fixed-effects* coefficients describing the average (population-level) trajectory of biomarker k over time, whereas $\mathbf{b}_{ki}$ is the vector of *subject-specific random effects* for individual i that captures departures from the average trajectory . The design vectors $\mathbf{x}_{kij}$ and $\mathbf{z}_{kij}$ collect the corresponding covariates used in the fixed and random parts of the model, respectively. This specification induces the latent (error-free) biomarker trajectory for marker k as

$$m_{ki}(t) \;=\; \mathbf{x}_{ki}(t)^{\top}\boldsymbol{\beta}_k \;+\; \mathbf{z}_{ki}(t)^{\top}\mathbf{b}_{ki}. \tag{10}$$

I then fitted a *multivariate joint model* by linking the event hazard to the *current value* of each latent biomarker trajectory, while also accounting for covariates in the survival submodel. Denoting by $h_i(t)$ the hazard for subject i, the survival component is specified as

$$h_i(t \mid \mathcal{H}_i(t)) \;=\; h_0(t)\,\exp\!\left\{\boldsymbol{\gamma}^{\top}\mathbf{w}_i \;+\; \sum_{k=1}^{K}\alpha_k\, m_{ki}(t)\right\}, \tag{11}$$

where $h_0(t)$ is the baseline hazard, $\mathbf{w}_i$ are weights for the static covariates with coefficients $\boldsymbol{\gamma}$, and α_k quantifies the association between the current latent value of biomarker k and instantaneous event risk. Under this specification, longitudinal information enters the risk model through the contemporaneous latent levels $m_{ki}(t)$, so that changes in any biomarker trajectory translate directly into updated event risk estimates as follow-up accrues [50, 53, 54].

Overall, joint models specify a joint distribution for the longitudinal Y_1 and time-to-event processes Y_2 by introducing random effects b that are shared across submodels. This is the key assumption, full conditional independence: random effect explain all interdependecies $Y_1 \perp\!\!\!\perp Y_2 \mid b$.

Joint models can be computationally demanding, particularly in multivariate settings with multiple biomarkers and high-dimensional random-effects structures. In this work, I fitted the multivariate joint model using the JMbayes2 package, which adopts a Bayesian framework and performs inference via MCMC sampling, and relied on the package default prior specifications for the model parameters [52].
Dynamic predictions are then obtained as conditional survival probabilities over a future horizon $u > t$, e.g. $\Pr(T_i > u \mid T_i > t, \mathcal{H}_i(t))$, where $\mathcal{H}_i(t)$ denotes the information accrued up to time t.

2.8.2 DynamicDeepHit

Dynamic-DeepHit (DDH) is a deep-learning approach for dynamic survival analysis that learns time-to-event predictions from longitudinal patient histories. It is designed to exploit repeated measurements collected over time, including histories observed at irregular time intervals, and it naturally extends to settings with multiple competing events and partially observed longitudinal inputs. In our application, however, I use pre-imputed longitudinal data and focus on a single event of interest, effectively applying DDH in a single-risk setting while retaining its ability to leverage the full observed history when updating predictions.

Our goal is to produce conditional event-free survival probability (Eq. 7) at a future time t , given the longitudinal history observed up to the current update time l.

In settings with competing risks, it is convenient to work with the cause-specific cumulative incidence function (CIF), from which the survival probability in (Eq. 7) can be directly recovered. Specifically, the CIF expresses the probability that cause k occurs on or before t, conditional on the same information set (which also implies event-free status up to l):

$$F(t \mid \mathcal{H}_i(l)) = P\big(T \leq t, k = 1 \mid \mathcal{H}_i(l), T > l\big) = P\big(T \leq t \mid \mathcal{H}_i(l), T > l\big) = \sum_{t^* \leq t} P\big(T = t^*, \mid \mathcal{H}_i(l), T > l\big). \quad (12)$$

The survival probability then follows as the complement of the overall cumulative incidence:

$$S_i(t \mid l, \mathcal{H}_i(l)) \;=\; 1 - \sum_k F_k(t \mid \mathcal{H}_i(l)) = 1 - F_k(t \mid \mathcal{H}_i(l). \quad (13)$$

Finally, since the *true* $F_k(t \mid \mathcal{H}_i(l)$ (and hence $S_i(t \mid l, \mathcal{H}_i(l)))$ is not observed, in practice I rely on the corresponding model-based estimates $\widehat{F}_k(t \mid \mathcal{H}_i(l)$ to obtain $\widehat{S}(t \mid l, \mathcal{H}_i(l))$ and perform dynamic risk prediction.

DDH is composed of a longitudinal network and a survival head that together produce dynamic event-time predictions. The longitudinal network processes the observed history up to the current update time l, denoted by $\mathcal{H}_i(l)$, and encodes it through a recurrent architecture (LSTM) into a latent representation which summarizes the patient's temporal trajectory and current state. This latent vector is then passed to a fully connected survival network (or survival head), which maps the encoded history into scores associated with future time intervals. The output of the model corresponds to the estimated joint probability of having event k at discrete time t, given the longitudinal history $\mathcal{H}_i(l)$:

$$o^*_{k,t} \;=\; \widehat{P}(T = t,\; k = 1 \mid \mathcal{H}_i(l)).$$

.
Therefore I can define the estimated CIF at prediction time t (conditional on being event-free up to the last measurement time l) is:

$$\widehat{F}_k(t \mid \mathcal{H}_i(l)) \;=\; \frac{\displaystyle\sum_{l < \tau \leq t} o^*_{k,\tau}}{1 - \displaystyle\sum_{n \leq l} o^*_{k,n}}\,.$$

Note that the estimated CIF is built upon the condition that this subject has survived up to the last measurement time.

The model is trained by minimizing a loss function designed to handle longitudinal measurements and right-censoring. The objective is composed of three terms. The primary component is a likelihood-based survival loss that encourages the predicted joint probability mass $o^*_{k,\tau}$ to match the observed event time and type, while properly accounting for right-censored observations. A second component is a ranking loss that improves discrimination across subjects, encouraging correct ordering of individuals according to their relative risk over time. Finally an additional term that incorporates the prediction error on trajectories of time varying covariates to capture the hidden representations of the longitudinal history and to regularize the network.

2.8.3 Landmarking

Landmarking is a fundamental strategy in dynamic survival analysis that enables risk predictions to be updated as new patient measurements become available over time. The approach consists of constructing survival models at specific, pre-defined time points, known as landmark times l. At each landmark time, only individuals who are still at risk (i.e., those who have not yet experienced the event and have not been censored) are included in the analysis, and their longitudinal history is truncated at time l, denoted as $Y_i(l)$, so that any information collected after l is ignored. This mimics a real-world clinical setting, where a physician evaluates a patient's risk at a specific visit using only the information available up to that moment. In its classical formulation, landmarking therefore requires the creation of a separate dataset for each predefined landmark time and the fitting of a distinct survival model at each landmark.While this approach is limited to predictions at fixed time points rather than arbitrary times, it aligns well with clinical practice, where follow-up assessments are typically scheduled at predefined visit times (landmarks). It also offers substantial advantages in terms of conceptual simplicity, computational efficiency, and ease of implementation, as it relies on standard survival modeling techniques without requiring joint specification of the longitudinal and survival processes [51].

Super landmarking is an extension of the traditional landmarking strategy in which, instead of fitting a separate survival model at each landmark time, one first constructs a single "super" landmarked training set by stacking multiple landmarked versions of the original dataset. Concretely, for each preselected landmark time l, the data are (i) restricted to subjects still at risk at l and (ii) truncated so that only measurements observed up to l are retained; the resulting landmarked subsets are then concatenated into one larger dataset in which the same subject can appear multiple times (once per landmark time, as long as they remain at risk), with progressively longer histories available at later landmarks. This design allows training a single model while preserving the "at-risk" conditioning that reduces the train–test mismatch typical of using full trajectories without landmarking. In the context of two-step approaches for dynamic survival analysis, the idea is to separate (a) a longitudinal stage that summarizes the truncated history up to l into an encoding (e.g., last-observation summaries, or sequence encodings from an RNN), from (b) a survival stage that maps to conditional survival/risk beyond l (e.g., CPH, RSF, or a neural survival head) [51].

Last-visit approach with RSF The first try was the Last Visit approach with RSF: at a given landmark time l, I represent each subject i using only the most recent measurements available up to l, $Z_i = Y_i(\max\{t_{ij} \mid t_{ij} \leq l\})$. These landmark-specific encodings are then used as input features to train a Random Survival Forest (RSF) on the subjects still at risk at time l, in order to estimate the conditional survival distribution beyond the landmark. The main limitation of this strategy is that it reduces the longitudinal history to a single snapshot: it discards earlier measurements and therefore cannot explicitly capture temporal patterns such as trends, variability, or changes over time that may be prognostically informative [51].

RNN with survival FNN Menahwhile in the approach using a RNN with a FNN for survival. In the used implementation, the RNN is a traditional multi-layer Elman RNN (instead of a LSTM), which gives as output a latent representation of the longitudinal history of time variant feature $h_{i,j}$ of patient i with data available up to time point j. In the adopted implementation, this latent representation is finally collapsed to a one-dimensional embedding. For the longitudinal predicition a feed-foward neural network is used to predict the input variable of the next time time point $(j+1)$: $\hat{Y}(t_{i,j+1}) = FNN(h_{i,j}, X_{i,leukapheresis}(t_{i0}))$ For the survival prediction I will use the hidden state at the last time point J for each patient i: h_{iJ_i}. Combined with the static variable I obtain the following encoding for subject i: $Z_i = [X_{i,leukapheresis}(t_{i0}), h_{iJ_i}]$. The output of this network is a vector of probabilities $\hat{P}(T = t|Z_i)$ for each subject. This indicates the probability of subject i having the event of interest during the time interval [t,t+1). From these probabilities, I can compute the failure function at time τ as follows:

$$\hat{F}(\tau) = \frac{\sum_{l<t\leq\tau} \hat{P}(T=t)}{1-\sum_{t=0}^{l} \hat{P}(T=t)} = \sum_{l<t\leq\tau} \frac{\hat{P}(T=t)}{\hat{P}(T>l)} = \sum_{l<t\leq\tau} \hat{P}(T=t \mid T>l). \tag{14}$$

and the survival function is then given by $\hat{S}(t) = 1 - \hat{F}(t)$. To train the RNN model I use a combined loss function $L = L_{long} + L_{surv}$. The longitudinal loss L_{long} is the mean squared error between the predicted input variables $\hat{Y}_i(t_{ij})$ and the data $Y_i(t_{ij+1})$. The survival loss L_{surv} is the negative log-likelihood function with the prediction $\hat{S}(t)$ of the FNN:

$$\mathcal{L}_{\text{surv}} = -\sum_{i=1}^{N} \Big[\delta_i \log\big(\hat{P}(T=T_i \mid Z_i)\big) + (1-\delta_i) \log\big(\hat{S}(T_i \mid Z_i,\, T>l)\big)\Big]. \tag{15}$$

where T_i indicates the time interval in which an event is observed for subject i [51].

RNN with CPH In the RNN–CPH setting, the recurrent encoder is first trained according to the longitudinal paradigm described above, but with the survival loss term explicitly omitted ($L_{Surv} = 0$). By focusing the optimization exclusively on the longitudinal loss (L_{long}), the RNN acts as a pure feature extractor of temporal dynamics, summarizing subject trajectories into an encoding Z_i. Once the subject-specific embedding Z_i has been learned, it is employed as the covariate vector in a CPH model,

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

In this configuration, the CPH regression coefficients β are estimated in a second stage by minimizing the CPH negative partial log-likelihood on the fixed embeddings; hence, no gradient is backpropagated through the RNN during this step. Accordingly, the RNN acts as a feature extractor that summarizes the longitudinal history into

latent representations, while the CPH model provides a semi-parametric specification of the hazard conditional on such representations [51].

3. Results

3.1. Dataset

The dataset comprises clinical data from 1,465 patients, collected across 24 participating centers over the period from 2019 to 2025. To better frame the clinical profile of the cohort, it is useful to observe, for example, that most patients were affected by DLBCL (59.7%), with smaller proportions of MCL (13.1%), HGBCL (12.1%), PMBCL (7.2%) and FL (7.1%). Prior treatment exposure was substantial: most patients had received at least two previous lines of therapy (2 lines: 51.6%; 3 lines: 18.2%; 4–5 lines: 10.1%), whereas 17.7% had undergone only one prior line. Indicators of disease burden and treatment history further showed bulky disease in 27.1% of cases and a previous ASCT in 26.3%. The overall handling of the dataset, from the initial cohort to the final analytical partitions, is summarized in the consort flow diagram in Figure 6.
Before any preprocessing, potential center effects were assessed by comparing leukapheresis covariate distributions across centers using pre-specified omnibus tests. To ensure stable center-level comparisons, centers with fewer than 20 patients (*Bari*, *Ancona*, *Milano IEO*, and *Pisa*) were pooled into a single *Other* category. Overall, most variables showed broadly comparable distributions, with no indication of systematic between-center differences. Notable deviations were observed for leukapheresis *ferritin* levels, particularly in the *Bologna* and *PA Maddalena* centers and for leukapheresis *PCR* levels in *PA Maddalena*, as shown in Figures 4 and 5, respectively.

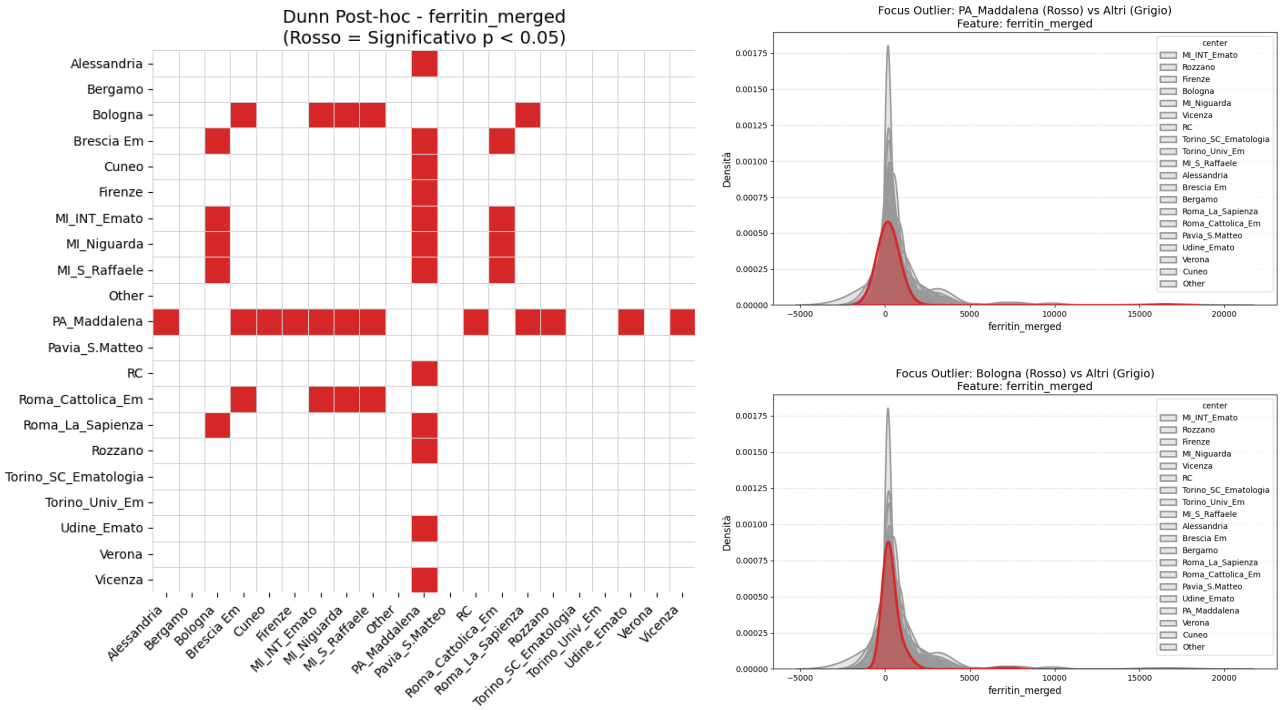

Figure 4: Between-center comparison of ferritin at leukapheresis. The left panel reports significant pairwise differences from the Dunn post-hoc test (red, $p < 0.05$), while the right panel shows the corresponding distributional shift, highlighting Bologna and PA Maddalena versus the other centers.

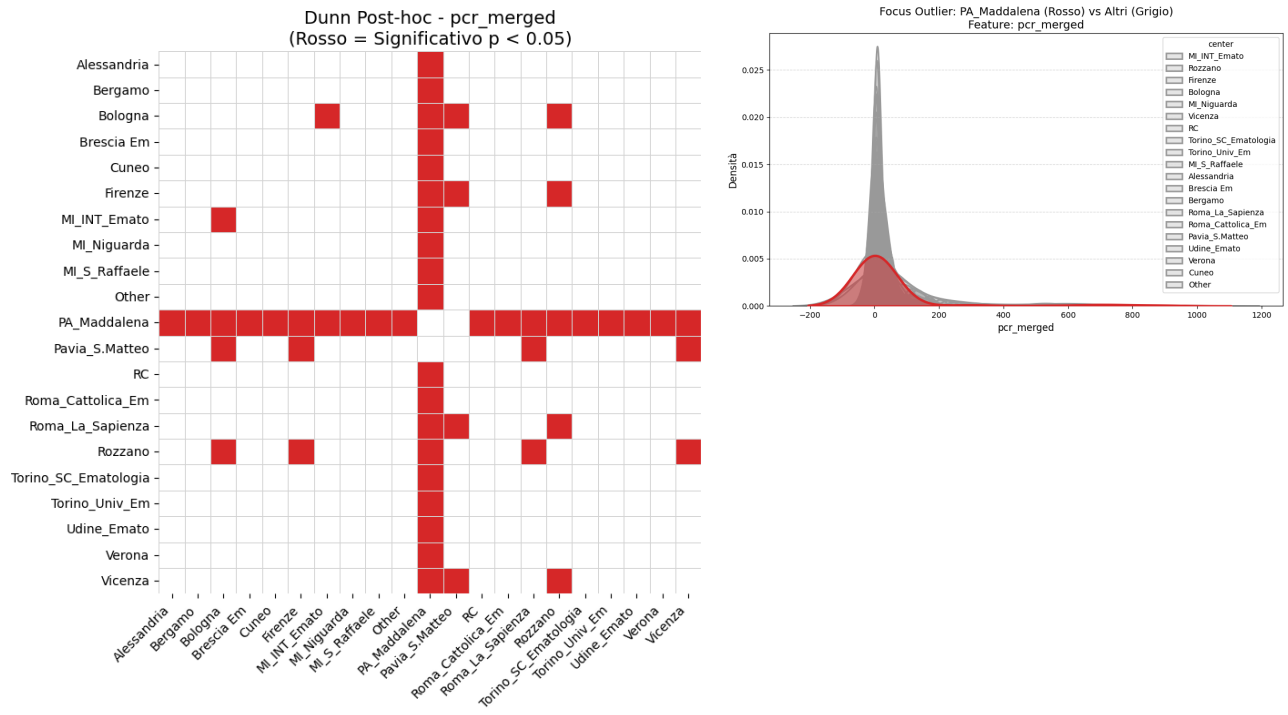

Figure 5: Between-center comparison of PCR at leukapheresis. The left panel reports significant pairwise differences from the Dunn post-hoc test (red, $p < 0.05$), and the right panel visualizes the distribution, highlighting PA Maddalena versus the other centers.

As an initial preprocessing step, I refined the analytical cohort by applying predefined exclusion criteria. Specifically, patients diagnosed with MCL and/or treated with liso-cel. Finally, records lacking the target PFS outcome were excluded, since these observations cannot contribute to supervised model training or evaluation. After applying these criteria, the final dataset comprised 1,071 patients used for model development. Features were selected based on their availability at leukapheresis; the complete list is reported in Table 10.
The next step was dataset splitting. I first reserved the "Roma Cattolica" center as an independent external validation cohort, including 67 patients (approximately 6% of the overall dataset). The remaining patients were then randomly split into a test set of 201 subjects (20%) and a training set of 803 subjects (80%). When required by the modeling pipeline, the training set was further partitioned into a training set of 702 patients (70%) and a validation set of 101 patients (10%). A comparable event/censoring composition across splits was achieved, with censoring proportions of 50.4% , 50.0%, 50.5%, and 52.2% in the training, validation, test, and external validation sets, respectively. The resulting dataset partitioning is illustrated in Figure 6. Meanwhile Figure 7 summarizes the dataset construction workflow and shows how centers were allocated across the training, validation, test, and external validation sets.
In addition, I inspected the availability of longitudinal measurements across follow-up time points. Table 1 summarizes, for each cohort, the number and percentage of patients with observed data at leukapheresis, infusion, and subsequent follow-up visits (at 30, 60, 90 and 180 days from infusion).

Figure 8 reports the estimated functions $S(t)$, together with median PFS and 95% confidence intervals (the two panels reflect the two split configurations used in this study, depending on whether a validation set was required). The results of the pairwise log-rank comparisons do not shows significant differences between the population of the sets. Finally, Table 2 provides a compact description of the post-split follow-up profile of each cohort in terms of numbers at risk, events, and censoring over time.

Categorical features were handled as described above: for *histology*, the label "FL" was used as the reference category and therefore omitted to avoid multicollinearity and redundant dimensions, while the remaining binary and ordinal features underwent binary (0/1) and ordinal encoding, respectively [35] .
To reduce sparsity and improve model stability, rare categories were grouped into broader classes. In particular, patients with ≥ 4 previous treatments were collapsed into a single "4+" category; ECOG performance status values ≥ 2 were grouped into "2+"; and the number of extranodal sites ≥ 4 was merged into "4+".
Given the substantial amount of missing data and the inability of the modeling framework to directly handle missing values, features with more than 30% missingness in the training set were excluded. This resulted in the removal of *D-dimer* (59.3% missing), *BM involvement* (43.7%), and *Albumin* (42.2%).

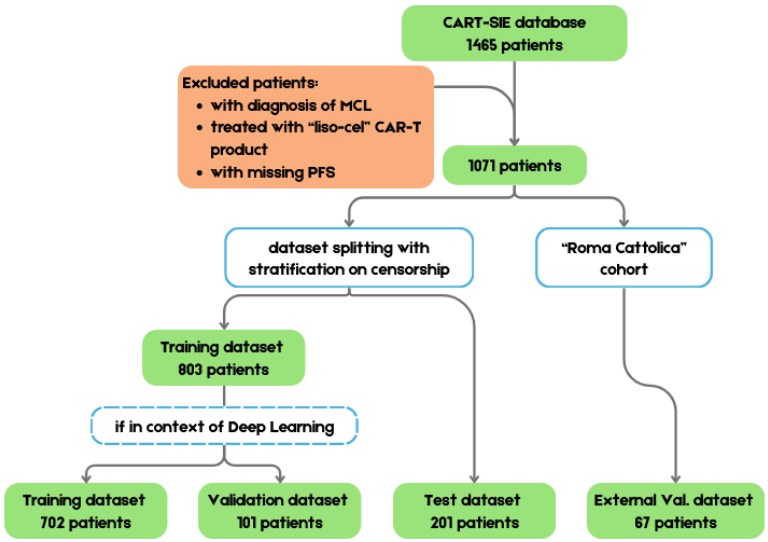

Figure 6: Study consort flow and dataset splitting workflow, including training, validation, test, and external validation partitions.

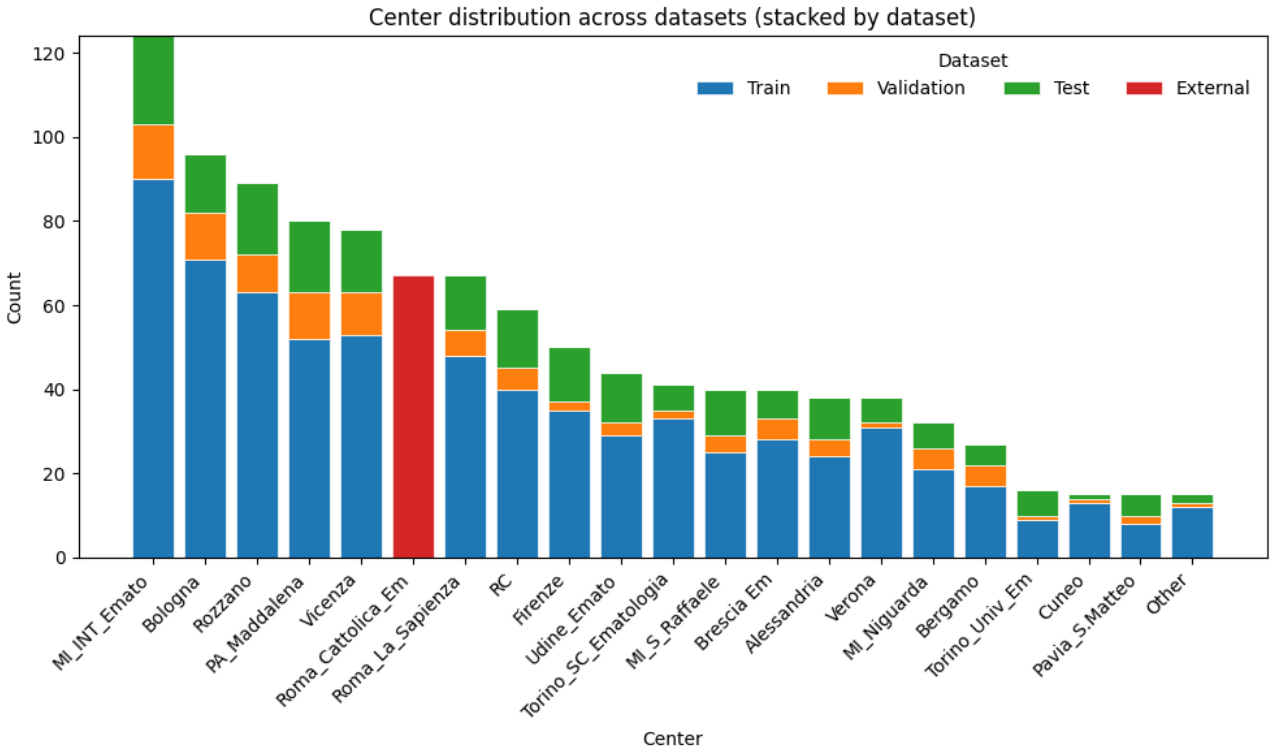

Figure 7: Distribution of participating centers across the training, validation, test, and external validation datasets.

Longitudinal data availability across follow-up visits

		leukapheresis	infusion	Day30	Day60	Day90	Day180
Train set (702)	N	695	692	602	239	433	303
	%	99%	99%	86%	34%	62%	43%
Validation set (101)	N	100	98	89	31	58	35
	%	99%	97%	88%	31%	57%	35%
Test set (201)	N	199	197	182	61	125	88
	%	99%	98%	91%	30%	62%	44%
External validation (67)	N	61	65	52	24	29	17
	%	91%	97%	78%	36%	43%	25%

Table 1: Availability of longitudinal measurements across scheduled visits in each cohort. The total number of patients in each set is indicated in the round brackets. For each time point, the table reports the number (N) and percentage (%) of patients with an observed measurement.

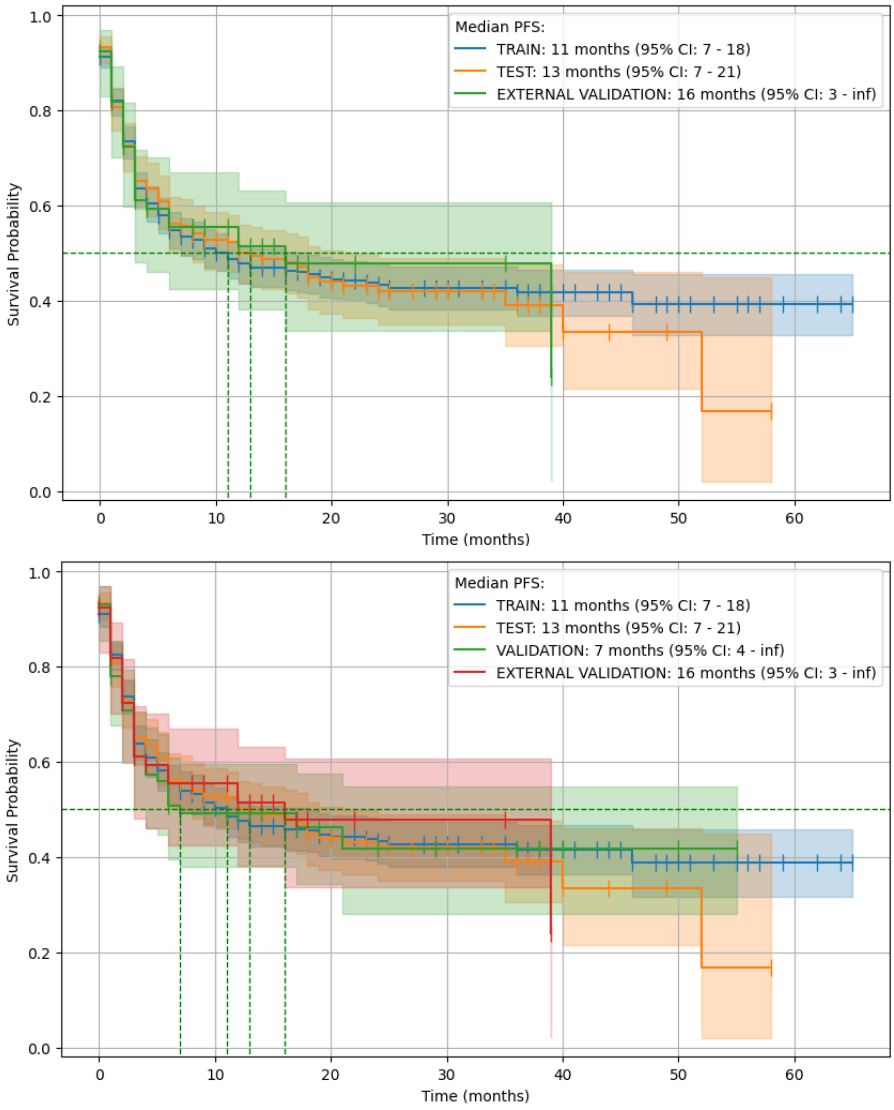

Figure 8: Kaplan–Meier curves across the data splits.

Risk table for PFS across all datasets

		leukapheresis	infusion	after 1 month	after 2 months	after 3 months	after 6 months	after 9 months	after 12 months	after 18 months	after 24 months
Train	At risk	702	702	628	542	478	341	248	217	140	87
	Censored	0	0	8	31	35	79	143	159	223	270
	Events	0	0	66	129	189	282	311	326	339	345
Test	At risk	201	201	186	155	138	105	79	68	45	31
	Censored	0	0	4	8	9	23	38	46	65	74
	Events	0	0	11	38	54	73	84	87	91	96
Validation	At risk	101	101	93	72	62	39	31	23	18	17
	Censored	0	0	2	10	10	19	25	28	33	34
	Events	0	0	6	19	29	43	45	50	50	50
Ext. Validation	At risk	67	67	59	52	45	32	29	27	9	3
	Censored	0	0	2	3	4	9	10	12	27	33
	Events	0	0	6	12	18	26	28	28	31	31

Table 2: Risk table for PFS across training, test, validation, and external validation datasets. "At risk" corresponds to the number of individuals with observed time $T \geq t$ (i.e., still event-free and under observation immediately before time t); "Censored" denotes the cumulative number of subjects with censoring time $T < t$; "Events" indicates the cumulative number of observed PFS events with $T < t$.

With the remaining 24 features, I fitted the chosen imputation strategy on the training set and applied it to all splits. After imputation, a subset of highly right-skewed variables (which included: platelets, neutrophilis, lymphocytes, monocytes, ferritin, ldh norm., pcr, ast, alkaline phospate, bilirubin) were log-transformedn. Finally, all numerical features were standardized using a StandardScaler from the scikit-learn package, fitted on the training set and applied consistently to validation, test, and external validation data.
Finally a strong correlation between WBC and neutrophil count ($r = 0.87$) has been noted as expected since neutrophils constitute a subset of white blood cells (in Figure 11 is reported the correlation matrix of the train set). To reduce redundancy, WBC was removed and the neutrophil count was retained as the more informative variable. Overall, the final dataset used for model training comprised 23 features: sex, age at infusion, disease status, ann arbor stage, Ipi-mipi-flipi, B symptoms,ecog, bulky, extranodal merged, ASCT, number of previous treatment, hb, platelets, neutrophilis, lymphocytes, monocytes, ferritin, ldh norm., pcr, ast, alkaline phospate, bilirubin and hystology.

Multiple Imputation In the multiple imputation experiment, the same preprocessing pipeline described above was applied prior to imputation. In particular, the same set of 24 features was retained, and identical grouping of rare categorical levels and encoding procedures were used. After generating $M = 20$ imputed versions of each dataset, the same right-skewed features were log-transformed. All numerical features were then standardized using scaling parameters estimated on the merged imputed training datasets (i.e., by stacking the M training imputations); these parameters were subsequently applied unchanged to the corresponding test and external validation sets. Finally, correlation-based feature filtering on the training dataset resulted in the elimination of WBC from all datasets.

3.2. Model Evaluation

3.2.1 Static prediction

In the Standard Approach, the models CPH, GBS, RSF, EST, and SSVM were trained and evaluated on the datasets obtained through the standard single-imputation procedure. By doing univariate feature selection and Hyperparameters-search, each models ended up training on a different set of feature and parameters, reported on Table 11 . The corresponding C-index values across Train, Test, and External Validation sets are reported in Table 5. Meanwhile the DeepSurv model was trained and evaluated using Train-Validation-Test-External Validation set, but obtain with the same standard procedure as above. In this case no feature selection has been applied. The performances across the dataset are reported in Table 4. In the multiple imputation experiment, the $M = 20$ multiple imputation of the train dataset are used to train M CPH models. The resulting pooled performance estimates are summarized in Table 3.
Since the patient cohorts in the Test and External Validation sets are consistent across all experiments, I can directly compare model generalization. Among the standard approaches, the CPH model demonstrated the best overall performance (0.64 on the Test set and 0.73 on the External Validation set). Furthermore, applying Multiple Imputation slightly improved the CPH performance compared to its standard counterpart (Test: 0.666, Ext Val: 0.734). Interestingly, the deep learning approach (DeepSurv) achieved the highest C-index on the Test

set (0.67), but yielded lower performance on the External Validation set (0.69) compared to the CPH models.

C-index of Multiple Imputation Approach

Train		Test		Ext Val	
C-index	**Var**	**C-index**	**Var**	**C-index**	**Var**
0.6819	0.0002	0.6662	0.0009	0.7342	0.0029

Table 3: Pooled C-index and corresponding variances obtained under the multiple imputation evaluation strategy.

C-index of DeepSurv

	Train	Val	Test	Ext Val
DeepSurv	0.70	0.70	0.67	0.69

Table 4: C-index values for the DeepSurv model across data splits.

C-index of Standard Approach

	Train	Test	Ext Val
CPH	0.66	0.64	0.73
GBS	0.70	0.64	0.68
RSF	0.68	0.63	0.67
EST	0.66	0.62	0.71
SSVM	0.74	0.62	0.65

Table 5: C-index values for static models across training, test, and external validation sets.

The time-dependent $AUC(t)$ and $BR(t)$ over 1-2-3-6 months after infusion, were computed only when these metrics were available in the model's native package implementation, and they were used primarily to benchmark the longitudinal models. The Brier score can be evaluated only for methods that provide an estimate of the survival function; for example, it is not available for SSVM in `scikit-survival` [36]. In the meantime, for those models supporting time-dependent evaluation, I report the mean tvAUC and the Integrated BS (IBS) as global summaries of discrimination and calibration performance over the follow-up period.

	Test		External Validation	
Model	**Mean AUC**	**IBS**	**Mean AUC**	**IBS**
CPH	0.67	X	0.72	X
GBS	0.66	0.242	0.66	0.22
RSF	0.65	0.25	0.68	0.21
EST	0.67	0.25	0.74	0.22
SSVM	0.63	X	0.66	X
DeepSurv	0.69	X	0.69	X
Multiple Imputation Approach	X	X	X	X

Table 6: Mean tvAUC and IBS for static models on the test and external validation datasets.

3.2.2 Dynamic Prediction

Transitioning to the dynamic prediction framework, the preprocessing pipeline remained consistent with the standard approach. Consequently, the final feature set and the applied transformations were identical, with the notable addition of time-updated longitudinal measurements, collected at specific time landmark: leukapheresis, infusion, and follow-up visits at 30, 60, 90, and 180 days post-infusion. The models were evaluated based on their ability to leverage patient information available up to one of the landmark to predict progression-free

survival across multiple future horizons, specifically at 1, 2, 3, 6, 9, 12, 18, and 24 months after infusion. The predictive performance of these longitudinal approaches, quantified in terms of the tvAUC(t) and BS(t).

The first dynamic approach evaluated was a statistical one, specifically the JM. This model was constructed by fitting a linear mixed-effects model for time-varying features and linking it to a CPH submodel. After being fitted on the training set, the model was evaluated on both the test and external validation cohorts. Overall, the predictive performance of the JM was modest but notably stable; it achieved a tvAUC generally ranging between 0.55 and 0.65. Interestingly, held higher performances on early landmarks and/or horizion, as it can be seen in Table 12. Furthermore, while the tvAUC was successfully computed using the JMbayes2 package, the Brier score routine repeatedly encountered runtime errors despite identical inputs. As this anomaly appears to be software- or compute-related rather than a methodological flaw, I report only the tvAUC for the JM. I verified that this failure persisted across different random seeds and parallelization settings, strongly suggesting a reproducible implementation issue.

The next approach evaluated was the DDH architecture. After being trained on the training set and optimized using the validation set, the DDH model demonstrated a substantial improvement in predictive capabilities compared to the JM, as seen in Table 13. At early landmark times and short prediction horizons, the model's performance was modest across all evaluated metrics. However, as the prediction horizon extended, the discriminative performance of the model noticeably increased. More importantly, as the landmark time advanced the predictive accuracy improved even further, with the tvAUC reaching approximately 0.80. It is worth noting, however, that while the discriminative performance tvAUC improved over time, the BS also exhibited an upward trend, as reported in Table 16.

Within the super landmarking framework, the dataset preparation was extended to create a "super" landmarked training set by stacking multiple landmarked versions of the original data. Conversely, the validation, test, and external validation sets were not modified into this stacked format. Because it was integrated into this broader pipeline, the Last-visit RSF approach retained the pre-established train-validation split, even though the validation set was not strictly utilized during the forest's training procedure.

In stark contrast to the DDH model, the Last-visit RSF yielded its best discriminative performance at early landmark times and shorter prediction horizons (Table 15). Specifically, it achieved an tvAUC of approximately 0.70 when predictions were anchored at leukapheresis. However, as either the landmark time advanced or the prediction horizon extended, the model's discriminative ability progressively deteriorated. Interestingly, despite this marked decline in discrimination, the Last-visit RSF maintained a consistently low BS (Table 19).Furthermore, while the model exhibited clear signs of overfitting during training, its generalizability proved relatively robust, as the overall performance on the external test sets did not suffer a drastic decline.
Finally, the effectiveness of utilizing a RNN to extract and encode longitudinal history proved to be highly dependent on the choice of the underlying survival head. When the recurrent encoder was coupled with a FNN, the model yielded generally poor discriminative performance across the validation, test, and external validation sets, characterized by low tvAUC (Table 14) and higher BS (Table 19) compared to the Last-visit RSF. Tt can be observed in the RNN-FNN architecture suffered from severe overfitting; it achieved an tvAUC of approximately 0.90 during training but failed entirely to generalize to unseen data.

Conversely, shifting from FNN to a CPH head significantly altered the results. This hybrid configuration successfully generalized to the test and external validation cohorts, often yielding performance metrics superior to those of the training set. The model maintained a robust predictive performance, with tvAUC frequently gravitating around 0.75, particularly during the earlier landmarks (Table 7). Notably, this discriminative capability remained strongest at the leukapheresis and infusion stages; however, as anticipated, performance exhibited fluctuations and a relative decline during later landmark times, such as the 90-day mark. Furthermore, while its Brier score was slightly superior to the baseline established by the Last-visit RSF, the RNN-CPH remained adequately calibrated, maintaining an absolute prediction error below 0.20 (Table 17).

3.2.3 Static versus Dynamic Prediction

In Figures 9 and 10, I report the final comparison of discriminative performance across models using the tvAUC, evaluated on the test set and on the external validation cohort, respectively, over the follow-up horizon. Importantly, all models are trained to predict outcomes using only information available at leukapheresis; therefore, differences in performance reflect how effectively each approach extracts prognostic signal.

tvAUC of RNN-CPH

Landmark	Split	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
Leukapheresis	**Train**	0.74	0.75	0.73	0.71	0.70	0.70	0.70	0.70
	Validation	0.75	0.66	0.68	0.69	0.68	0.67	0.67	0.66
	Test	0.70	0.78	0.79	0.76	0.77	0.77	0.75	0.78
	Ext. Validation	0.93	0.76	0.84	0.76	0.75	0.75	0.80	0.76
Infusion	**Train**	0.78	0.78	0.75	0.72	0.69	0.70	0.70	0.71
	Validation	0.75	0.67	0.70	0.70	0.69	0.68	0.67	0.67
	Test	0.71	0.78	0.79	0.76	0.77	0.77	0.75	0.78
	Ext. Validation	0.93	0.77	0.83	0.76	0.75	0.75	0.79	0.76
30 day	**Train**	X	0.78	0.75	0.71	0.70	0.70	0.69	0.70
	Validation	X	0.63	0.69	0.69	0.68	0.67	0.66	0.66
	Test	X	0.79	0.80	0.76	0.77	0.77	0.75	0.78
	Ext. Validation	X	0.60	0.77	0.70	0.69	0.70	0.75	0.71
60 day	**Train**	X	X	0.72	0.69	0.70	0.68	0.68	0.69
	Validation	X	X	0.73	0.71	0.70	0.64	0.63	0.63
	Test	X	X	0.79	0.71	0.74	0.74	0.72	0.75
	Ext. Validation	X	X	0.92	0.73	0.71	0.72	0.78	0.75
90 day	**Train**	X	X	X	0.68	0.66	0.67	0.67	0.68
	Validation	X	X	X	0.69	0.67	0.64	0.63	0.63
	Test	X	X	X	0.65	0.70	0.71	0.68	0.73
	Ext. Validation	X	X	X	0.58	0.59	0.60	0.70	0.66
180 day	**Train**	X	X	X	X	0.63	0.65	0.66	0.67
	Validation	X	X	X	X	0.50	0.54	0.54	0.54
	Test	X	X	X	X	0.75	0.76	0.69	0.75
	Ext. Validation	X	X	X	X	0.61	0.61	0.79	0.73

Table 7: tvAUC values for the RNN–CPH model at different landmark times and prediction horizons, stratified by data split (train, validation, test, and external validation).

tvAUC on Test Set across longitudinal models

Model	Landmark	1 month after inf.	2 months after inf.	3 months after inf.	6 months after inf.	9 months after inf.	12 months after inf.	18 months after inf.	24 months after inf.
JM	**Leukapheresis**	0.52	0.68	0.71	0.69	0.67	0.65	0.60	0.61
	Infusion	0.52	0.69	0.70	0.65	0.62	0.61	0.57	0.56
	Day 30	X	0.67	0.67	0.57	0.53	0.53	0.52	0.55
	Day 60	X	X	0.67	0.56	0.52	0.49	0.46	0.45
	Day 90	X	X	X	0.57	0.54	0.55	0.51	0.49
	Day 180	X	X	X	X	0.45	0.48	0.44	0.44
DDH	**Leukapheresis**	0.71	0.69	0.69	0.70	0.73	0.71	0.70	0.76
	Infusion	0.71	0.68	0.69	0.70	0.73	0.67	0.68	0.74
	Day 30	X	0.76	0.77	0.76	0.76	0.72	0.75	0.80
	Day 60	X	X	0.81	0.80	0.76	0.77	0.72	0.77
	Day 90	X	X	X	0.80	0.80	0.77	0.76	0.80
	Day 180	X	X	X	X	0.81	0.77	0.76	0.22
Last-visit RSF	**Leukapheresis**	0.74	0.73	0.71	0.67	0.66	0.66	0.64	0.66
	Infusion	0.70	0.71	0.72	0.68	0.69	0.68	0.68	0.69
	Day 30	X	0.70	0.71	0.64	0.64	0.63	0.61	0.63
	Day 60	X	X	0.58	0.60	0.56	0.57	0.57	0.60
	Day 90	X	X	X	0.57	0.55	0.57	0.55	0.58
	Day 180	X	X	X	X	0.54	0.53	0.48	0.55
RNN–FNN	**Leukapheresis**	0.55	0.68	0.69	0.61	0.52	0.50	0.50	0.51
	Infusion	0.53	0.68	0.69	0.52	0.50	0.50	0.49	0.50
	Day 30	X	0.65	0.68	0.52	0.50	0.51	0.49	0.50
	Day 60	X	X	0.53	0.46	0.46	0.47	0.45	0.47
	Day 90	X	X	X	0.49	0.47	0.48	0.45	0.46
	Day 180	X	X	X	X	0.60	0.70	0.64	0.70
RNN–CPH	**Leukapheresis**	**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	**0.79**	**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**

Table 8: tvAUC evaluated at multiple prediction horizons for each longitudinal model and landmark time.

tvAUC on External Validation Set across longitudinal models

Model	Landmark	1 month after inf.	2 months after inf.	3 months after inf.	6 months after inf.	9 months after inf.	12 months after inf.	18 months after inf.	24 months after inf.
JM	**Leukapheresis**	0.88	0.72	0.73	0.66	0.61	0.60	0.54	0.53
	Infusion	0.87	0.73	0.71	0.62	0.56	0.55	0.54	0.47
	Day 30	X	0.66	0.80	0.64	0.58	0.58	0.60	0.59
	Day 60	X	X	0.88	0.61	0.63	0.61	0.65	0.63
	Day 90	X	X	X	0.58	0.63	0.64	0.69	0.67
	Day 180	X	X	X	X	0.85	0.90	0.79	0.79
DDH	**Leukapheresis**	0.75	0.63	0.65	0.67	0.68	0.75	0.80	0.77
	Infusion	0.75	0.68	0.68	0.69	0.70	0.76	0.76	0.85
	Day 30	X	0.74	0.73	0.71	0.72	0.74	0.79	0.85
	Day 60	X	X	0.78	0.79	0.79	0.70	0.81	0.89
	Day 90	X	X	X	0.78	0.79	0.82	0.81	0.89
	Day 180	X	X	X	X	0.80	0.83	0.81	0.09
Last-visit RSF	**Leukapheresis**	0.80	0.74	0.78	0.71	0.69	0.71	0.78	0.80
	Infusion	0.83	0.65	0.72	0.60	0.62	0.61	0.68	0.61
	Day 30	X	0.68	0.76	0.67	0.64	0.65	0.77	0.77
	Day 60	X	X	0.81	0.70	0.68	0.71	0.73	0.81
	Day 90	X	X	X	0.54	0.56	0.57	0.73	0.74
	Day 180	X	X	X	X	0.49	0.50	0.79	0.82
RNN–FNN	**Leukapheresis**	0.62	0.55	0.42	0.48	0.43	0.44	0.51	0.60
	Infusion	0.63	0.53	0.40	0.47	0.43	0.42	0.47	0.56
	Day 30	X	0.50	0.35	0.49	0.44	0.45	0.56	0.68
	Day 60	X	X	0.36	0.54	0.50	0.50	0.66	0.81
	Day 90	X	X	X	0.61	0.56	0.60	0.71	0.84
	Day 180	X	X	X	X	0.66	0.67	0.79	0.79
RNN–CPH	**Leukapheresis**	**0.93**	**0.76**	**0.84**	**0.76**	**0.75**	**0.75**	**0.80**	**0.76**
	Infusion	**0.93**	**0.77**	**0.83**	**0.76**	**0.75**	**0.75**	**0.79**	**0.76**
	Day 30	X	**0.60**	**0.77**	**0.70**	**0.69**	**0.70**	**0.75**	**0.71**
	Day 60	X	X	**0.92**	**0.73**	**0.71**	**0.72**	**0.78**	**0.75**
	Day 90	X	X	X	**0.58**	**0.59**	**0.60**	**0.70**	**0.66**
	Day 180	X	X	X	X	**0.61**	**0.61**	**0.79**	**0.73**

Table 9: tvAUC evaluated at multiple prediction horizons for each longitudinal model and landmark time.

On the test set (Figure 9), two main patterns emerge. First, the best performance is obtained by the RNN-CPH model, which remains the top or among the top methods across most landmark times, despite a weaker behavior at the earliest horizon (1 month after infusion). In contrast, not all dynamic approaches systematically improve discrimination: both the RNN-FNN and the JM display only moderate performance and, at several horizons, are comparable to (or worse than) strong static model. Among the remaining methods, Last-Visit RSF shows a consistently good performance, especially at early horizons, while DDH tends to improve at later horizons, reaching its strongest discrimination in the long-term follow-up. The second pattern that seem to emerge is the distinct advantage of models incorporating deep learning techniques, in fact within the static framework, where DeepSurv consistently outperforms traditional machine learning approachs.

The performance dynamics shift noticeably on the external validation set (Figure 10). While the RNN-CPH remains highly competitive, it is noteworthy that the majority of the dynamic models (particularly DDH and Last-visit RSF) demonstrate an impressive ability to maintain, or even increase, their discriminative power at extended future horizons (e.g., 18 to 24 months). In stark contrast, most static models show strong initial performance that steadily degrades as the prediction horizon extends. A notable outlier among the static approaches is the EST and CPH models, which exhibit an unusual trajectory by recovering and maintaining strong predictive performance at the latest evaluation horizons.

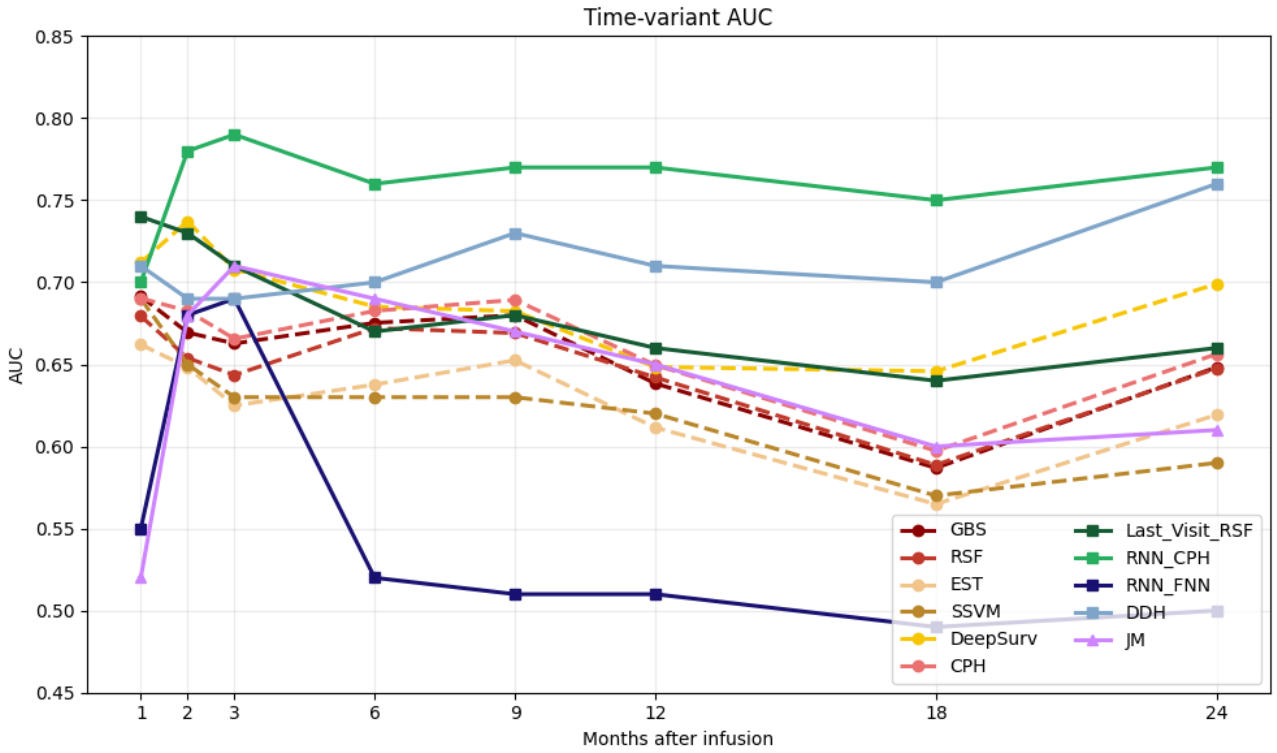

Figure 9: tvAUC on test set over months after infusion. Comparison between longitudinal models (solid lines) and static benchmarks (dashed lines) using exclusively data at leukapheresis.

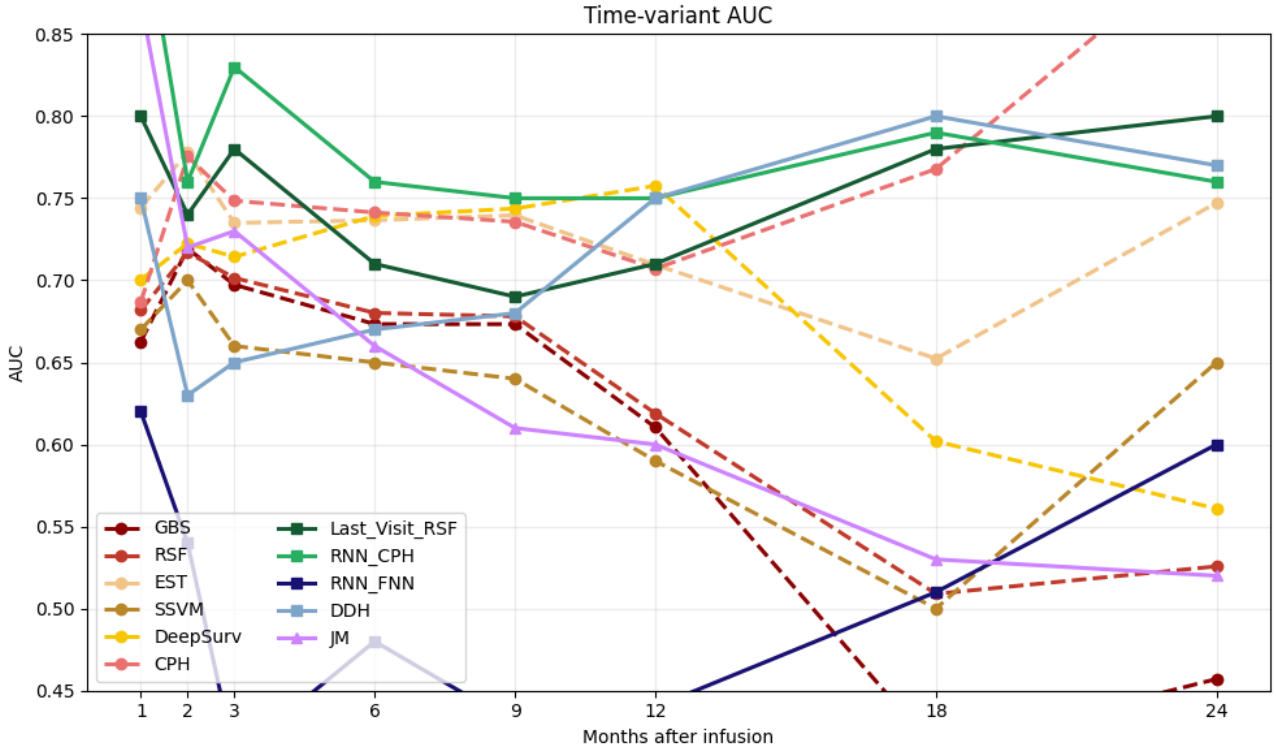

Figure 10: tvAUC on external validation set over months after infusion. Comparison between longitudinal models (solid lines) and static benchmarks (dashed lines) using exclusively data at leukapheresis.

4. Discussion

This study aimed to develop and evaluate longitudinal machine learning models for predicting the efficacy of CAR-T cell therapy in patients diagnosed with B-Cell Lymphoma. The dataset was derived from CART-SIE, a prospective, multicenter study incorporating real-world data from various Italian hematologic centers. The final analytical cohort comprised 1,071 patients with relapsed or refractory B-cell lymphomas (specifically including subtypes such as DLBCL, HGBCL, PMBCL, and FL) assessed for eligibility according to current regulatory criteria.

Initially, a statistical analysis was conducted to assess potential differences between centers and evaluate the homogeneity of the patient population selected for CAR-T therapy. While the descriptive and statistical analyses revealed no substantial systematic differences across the majority of variables, it is noteworthy that certain centers, specifically Bologna and PA Maddalena, exhibited differences in a limited number of features . The distribution of centers across datasets appeared relatively balanced (Figure 7), and no center-specific patterns were evident in data at leukapheresis or outcome incidence that could clearly explain variations in predictive performance.

An additional crucial observation pertains to the model's sensitivity to the random initialization state, commonly controlled by the random seed. While a certain degree of variance in performance and convergence across different seed permutations was observed across all evaluated models, this volatility was exacerbated in the architectures trained within the super-landmarking framework. Specifically, modifying the random seed in these super-landmarked models led to fluctuations in training convergence and a heightened susceptibility to severe overfitting.

A critical methodological consideration in our dataset curation involved the temporal alignment of clinical events, specifically the interval between leukapheresis and infusion. In real-world clinical practice, the time elapsed between these two milestones varies substantially across patients. Specifically, while for approximately 90% of the patients the infusion occurs within the second month after leukapheresis, this interval can be notably extended (sometimes even exceeding one year) due to factors such as the patient's response to intermediate bridging treatments or the inherent variability of the disease course. Consequently, indexing PFS follow-up strictly from leukapheresis does not yield directly comparable measurement times across subjects. For instance,

at a fixed three-month mark post-leukapheresis, some patients might still be awaiting infusion, whereas others could already be one or two months into their post-infusion recovery. Because the infusion itself represents a pivotal clinical intervention that fundamentally alters the subsequent disease trajectory, aligning patients along a common temporal reference was imperative to ensure the comparability of longitudinal information.
To achieve harmonized time windows and facilitate dynamic modeling, I standardized the timeline by treating the infusion as a fixed reference anchor, assuming it occurred exactly one month after leukapheresis for all subjects. While this assumption introduces a degree of artificial temporal smoothing, it was a necessary methodological compromise. It aligns the risk sets across the cohort, enabling the models architectures to learn consistent clinical and temporal patterns without being confounded by logistical variations or administrative delays in CAR-T cell manufacturing.

It is also important to acknowledge how the predictive use of the CPH model compares with "traditional" machine learning approaches. The CPH model provides strong interpretability and a principled way to handle censoring through partial likelihood, but it relies on rigid structural assumptions (most notably proportional hazards and strict functional-form choices for covariate effects) that, if violated, may harm calibration or discrimination. However, despite these theoretical limitations, our empirical results demonstrated that the standard CPH model performed comparable to machine learning ones, such as RSF and SSVM [42]. Considering this equivalent performance, alongside the fact that it is a simpler and more interpretable framework, the CPH model represents the most logical choice among static models for this clinical application. Therefore, I adopted the CPH model as the reference static baseline to benchmark our dynamic architectures across all evaluated time horizons using tvAUC. Advanced deep learning methods can theoretically overcome CPH's limitations by natively capturing non-proportional hazards and complex non-linearities [55]. Building upon this premise, our internal evaluation showed that the DeepSurv architecture achieved slightly superior discriminative performance compared to the standard CPH model on both the training and internal test sets. Interestingly, however, this dynamic was reversed during external evaluation. On the independent external validation cohort, the traditional CPH approach exhibited better generalizability, achieving slightly better results than the deep learning alternative regardless of whether standard single imputation or Multiple Imputation was used. Generally, deep learning models are most effective when applied to highly complex datasets with a very large number of features. Since our analysis relied on a relatively focused set of 23 clinical variables, it is likely that simpler models were sufficient to capture the underlying prognostic signal, explaining why more complex architectures did not yield a significant advantage in this specific context
Furthermore, regarding the handling of missing data, an evaluation of the imputation strategies revealed important practical considerations. While the implementation of MICE yielded a marginal improvement in the predictive metrics of the static models, this gain was ultimately negligible. Considering the substantial computational overhead and the intricate complexity it introduced into the data preprocessing and cross-validation pipelines, the minor performance bump did not justify its application. Ultimately, standard single imputation techniques provided a far more favorable and pragmatic trade-off between predictive accuracy and overall pipeline efficiency.

Transitioning to the dynamic prediction framework, it is essential to evaluate the trade-offs between different longitudinal modeling strategies. I initially explored JM, which simultaneously model the longitudinal and survival processes through shared random effects. While joint models can theoretically achieve high predictive accuracy when the association between biomarkers and survival is correctly specified, our empirical results highlighted significant practical limitations. In joint modeling, strict assumptions are required regarding the joint distribution of the longitudinal and survival processes [51]. Furthermore, they become computationally demanding and prone to software implementation failures, as observed in our pipeline with the JMbayes2 package [52], particularly in the presence of multiple longitudinal variables, due to the high dimensionality of the random-effects structure.

Conversely, purely deep-learning-based dynamic architectures, such as the DDH model, circumvent these statistical constraints and do not require explicit distributional assumptions. However, our evaluation revealed a distinct temporal imbalance in its predictive profile. The DDH model exhibited suboptimal discriminative performance at early landmark times and short prediction horizons; yet, as the landmark time advanced, its performance increased substantially. While achieving high accuracy at later horizons, this unbalanced behavior (coupled with a rising BS) suggests a potential over-reliance on late-stage longitudinal history at the expense of early robust predictions [51].

To reconcile the computational tractability issues of JM with the temporal imbalances of architectures like DDH, the landmarking framework provides a highly effective and flexible alternative. If a model is trained on complete longitudinal trajectories, it can end up relying mostly on the most recent measurements, which are often collected very close to the event (or censoring time). In real dynamic prediction, however, these late

measurements are not yet available when the clinician asks for a risk estimate at an earlier visit, so the model may behave unpredictably when only limited history is provided. Moreover, patients who fail early still affect model estimation during training, but they are absent from the evaluation set at later landmarks because only event-free subjects are considered. As a result, the model is fitted on the full cohort but evaluated on a subset of survivors, creating a train–evaluation population mismatch.
When predictive models are trained on complete longitudinal trajectories, they implicitly condition on future survival, as the availability of subsequent measurements inherently implies that the patient has not yet experienced the event. This creates a fundamental discrepancy with real-world clinical practice, where risk predictions must be made at a specific visit using only the history accrued up to that moment, without any knowledge of future outcomes.
As highlighted in the literature, the landmarking framework effectively resolves this by artificially truncating the longitudinal history at predefined landmark times and restricting the analysis exclusively to the "risk set", those patients who are still event-free at that specific time. In our study, the super-landmarking approach was utilized to further optimize this solution. By stacking multiple landmark-specific datasets, it enables the training of a single, robust global model that rigorously preserves this essential "at-risk" conditioning. Consequently, this strategy ensure that the model is trained under the exact conditions of temporal uncertainty it will encounter during clinical deployment [51]. It is worth to mention to be aware that by stacking multiple truncated trajectories of the same individuals (i.e. the dataset repeatedly includes the same subjects) it may introduce a high degree of intra-subject correlation. Ultimately, this approach allowed our advanced models (such as the RNN-CPH) to fully exploit the longitudinal history while maintaining robust, consistent performance across all evaluated timeframes.

The most significant finding of this study emerges from the global comparison between static and longitudinal architectures. When evaluating discriminative performance with tvAUC using only the information available at the leukapheresis (Figure 9 and Figure 10), the longitudinal models in general seem to offer an increment in term of prediction. In particular, the RNN-CPH model (within the super-landmarking framework) seems to show superior and consistent performance compared to all other models. Meanwhile static models (and even some dynamic ones like the JM or RNN-FNN) often displayed moderate or declining accuracy as the prediction horizon extended. In the RNN-CPH the recurrent encoder likely learned to model the biological evolution of patient features over time. Consequently, when presented with a new patient whose data is limited to the leukapheresis stage, the model can effectively leverage these learned patterns to anticipate how clinical variables might evolve, leading to a significantly more accurate and reliable prediction.
From a clinical perspective, this longitudinal model offers a two fold advantage for personalized medicine:

- Optimal Patient Selection: By providing a more precise risk estimate before the start of the therapy, it allows for the early identification of subjects at high risk of disease progression, ensuring that the treatment is directed toward those who will truly benefit from it.
- Dynamic Risk Adaptation: Beyond initial screening, the model enables continuous, longitudinal monitoring once the therapy has begun. By periodically updating risk estimates as new clinical measurements accrue, clinicians can anticipate unfavorable evolutions and consider timely adjustments in supportive care or additional therapeutic interventions.

Ultimately, the integration of sequential measurements represents a substantial step forward in continuous risk stratification, maximizing the chances of successful CAR-T therapy and providing a truly personalized management strategy for the patient.

A technical observation is that the improvements in predictive performance were noted when using RNN architectures, the CPH model, or their combination. While this does not necessarily imply a definitive causal link, a consistent trend was observed where these components (whether used individually or integrated into a hybrid system) are associated with enhanced modeling capabilities compared to the other methodologies tested.
As a final observation, the higher performance observed at early horizons in the external validation (Figure 10) suggests that the "Roma Cattolica" cohort possesses a stronger prognostic signal. This makes it intrinsically easier to stratify responders and non-responders. Despite this, one month after infusion, the overall predictive trend and the hierarchical ranking of the models still show that dynamic architectures consistently maintain long-term accuracy compared to static models, showing off their ability to generalize.

In light of the findings discussed, several avenues for future research emerge, primarily driven by the inherent limitations of the current study that must be addressed. Although an external validation cohort was included, it shared a similar underlying distribution with the training set; therefore, the generalizability of the proposed models to populations and clinical workflows with different characteristics is not guaranteed, and broader validation on fully independent cohorts is required. Furthermore, future developments must account for the stacking mechanism in the super-landmarking framework, which introduces intra-subject correlation and partial redundancy. The precise impact of this correlation should be thoroughly investigated, and, if necessary,

appropriate modeling strategies should be adopted to mitigate it. Additionally, while the temporal alignment used to harmonize follow-up windows was necessary for comparability, it may not fully reflect individual clinical timelines, suggesting that alternative temporal modeling strategies should be explored. Finally, real-world longitudinal observational data are inherently affected by missingness and irregular sampling, and the recurrent architectures employed remain only partially interpretable, potentially limiting their immediate clinical adoption. Future work should therefore focus on broader external validation, improving model interpretability, and demonstrating clinical utility in prospective settings.

5. Conclusion

The primary objective of this thesis was the development and evaluation of ML and DL models capable of integrating longitudinal data to predict PFS in patients diagnosed with lymphomas undergoing CAR-T therapy enrolled in the multicentric CART-SIE study. By comparing static and dynamic architectures, the work aimed to quantify the added value of information collected during follow-up relative to assessments performed at the leukaferesis stage alone.
The analysis identified the hybrid RNN-CPH architecture, implemented within a super-landmarking framework, as the most effective predictive tool among those examined. This model demonstrated a remarkable ability to maintain stable performance across the entire follow-up horizon, successfully anticipating clinical evolution even when data were limited to the leukaferesis stage.

From the perspective of future clinical applications, the proposed framework has the potential to enable patient management oriented toward personalized medicine. The ability to dynamically update the risk profile allows for informed adaptive management strategies, anticipating disease progression and supporting timely decisions regarding additional treatments. Ultimately, this study confirms that the integration of sequential clinical measurements through hybrid recurrent architectures represents a fundamental step forward in continuous risk stratification, maximizing the potential of CAR-T therapy in a real-world setting.

References

[1] E. Silkenstedt, G. Salles, E. Campo, and M. Dreyling. B-cell non-hodgkin lymphomas. *Lancet (London, England)*, 403(10438):1791–1807, 2024. doi: 10.1016/S0140-6736(23)02705-8. URL `https://doi.org/10.1016/S0140-6736(23)02705-8`.

[2] P. Dabas and A. Danda. Revolutionizing cancer treatment: a comprehensive review of car-t cell therapy. *Medical oncology (Northwood, London, England)*, 40(9):275, 2023. doi: 10.1007/s12032-023-02146-y. URL `https://doi.org/10.1007/s12032-023-02146-y`.

[3] M. Ayala Ceja, M. Khericha, C. M. Harris, C. Puig-Saus, and Y. Y. Chen. Car-t cell manufacturing: Major process parameters and next-generation strategies. *The Journal of experimental medicine*, 221(2): e20230903, 2024. doi: 10.1084/jem.20230903. URL `https://doi.org/10.1084/jem.20230903`.

[4] Joseph McGuirk, Edmund K Waller, Muna Qayed, Sunil Abhyankar, Solveig Ericson, Peter Holman, Christopher Keir, and G Douglas Myers. Building blocks for institutional preparation of ctl019 delivery. *Cytotherapy*, 19(9):1015–1024, 2017.

[5] 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.

[6] Hamed Salem S Albalwei. Maximization of resources in health care facilities: simple review article. *Haya Saudi J Life Sci*, 7(12):344–351, 2022.

[7] Emma Rabinovich, Kith Pradhan, R Alejandro Sica, Lizamarie Bachier-Rodriguez, Ioannis Mantzaris, Noah Kornblum, Aditi Shastri, Kira Gritsman, Mendel Goldfinger, Amit Verma, et al. Elevated ldh greater than 400 u/l portends poorer overall survival in diffuse large b-cell lymphoma patients treated with cd19 car-t cell therapy in a real world multi-ethnic cohort. *Experimental Hematology & Oncology*, 10(1): 55, 2021.

[8] Sandeep S Raj, Teng Fei, Shalev Fried, Andrew Ip, Joshua A Fein, Lori A Leslie, Ana Alarcon Tomas, Doris Leithner, Jonathan U Peled, Magdalena Corona, et al. An inflammatory biomarker signature of response to car-t cell therapy in non-hodgkin lymphoma. *Nature medicine*, 31(4):1183–1194, 2025.

[9] L Vercellino, R Di Blasi, S Kanoun, B Tessoulin, C Rossi, M D'Aveni-Piney, L Obéric, C Bodet-Milin, P Bories, P Olivier, et al. Predictive factors of early progression after car t-cell therapy in relapsed/refractory diffuse large b-cell lymphoma. blood adv. 2020; 4 (22): 5607–15. *Blood Advances*, 2023.

[10] Sunday Ayoola Oke. A literature review on artificial intelligence. *International journal of information and management sciences*, 19(4):535–570, 2008.

[11] Pavel Hamet and Johanne Tremblay. Artificial intelligence in medicine. *metabolism*, 69:S36–S40, 2017.

[12] Alberto Boretti. Improving chimeric antigen receptor t-cell therapies by using artificial intelligence and internet of things technologies: A narrative review. *European Journal of Pharmacology*, 974:176618, 2024.

[13] Alireza Naghizadeh, Wei-chung Tsao, Jong Hyun Cho, Hongye Xu, Mohab Mohamed, Dali Li, Wei Xiong, Dimitri Metaxas, Carlos A Ramos, and Dongfang Liu. In vitro machine learning-based car t immunological synapse quality measurements correlate with patient clinical outcomes. *PLoS computational biology*, 18(3): e1009883, 2022.

[14] Kyle G Daniels, Shangying Wang, Milos S Simic, Hersh K Bhargava, Sara Capponi, Yurie Tonai, Wei Yu, Simone Bianco, and Wendell A Lim. Decoding car t cell phenotype using combinatorial signaling motif libraries and machine learning. *Science*, 378(6625):1194–1200, 2022.

[15] Alex Bogatu, Magdalena Wysocka, Oskar Wysocki, Holly Butterworth, Manon Pillai, Jennifer Allison, Donal Landers, Elaine Kilgour, Fiona Thistlethwaite, and Andre Freitas. Meta-analysis informed machine learning: supporting cytokine storm detection during car-t cell therapy. *Journal of Biomedical Informatics*, 142:104367, 2023.

[16] Xiaofeng Yu, Qingqing Wang, Tangnuran Halimulati, Jiling Lv, Kai Zhou, Guilai Chen, Li Yin, Yulin Liu, Jingwang Bi, Zhuo Xiang, et al. Machine learning-based predictive model for high-grade cytokine release syndrome in chimeric antigen receptor t-cell therapy. *Frontiers in Immunology*, 16:1692892, 2025.

[17] Anna Cascarano, Jordi Mur-Petit, Jeronimo Hernandez-Gonzalez, Marina Camacho, Nina de Toro Eadie, Polyxeni Gkontra, Marc Chadeau-Hyam, Jordi Vitria, and Karim Lekadir. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. *Artificial Intelligence Review*, 56 (Suppl 2):1711–1771, 2023.

[18] P. Corradini. Italian observational study on car-t therapy for lymphoma (cart-sie). ClinicalTrials.gov. URL `https://clinicaltrials.gov/study/NCT06339255`. Identifier: NCT06339255. Updated April 1, 2024. Accessed February 17, 2026.

[19] David Lebwohl, Andrea Kay, William Berg, Jean Francois Baladi, and Ji Zheng. Progression-free survival: gaining on overall survival as a gold standard and acceleratingdrug development. *The Cancer Journal*, 15 (5):386–394, 2009.

[20] Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

[21] Alexis Dinno. Nonparametric pairwise multiple comparisons in independent groups using dunn's test. *The Stata Journal*, 15(1):292–300, 2015.

[22] A. Hazra and N. Gogtay. Biostatistics series module 3: Comparing groups: Numerical variables. *Indian Journal of Dermatology*, 61:251–260, 2016. doi: 10.4103/0019-5154.182416.

[23] Joel Juarros-Basterretxea, Gema Aonso-Diego, Álvaro Postigo, Pelayo Montes-Álvarez, Álvaro Menéndez-Aller, and Eduardo García-Cueto. Post-hoc tests in one-way anova: The case for normal distribution. *Methodology*, 20(2):84–99, 2024. doi: 10.5964/meth.11721.

[24] Tae Kyun Kim. Understanding one-way anova using conceptual figures. *Korean journal of anesthesiology*, 70(1):22, 2017.

[25] Viv Bewick, Liz Cheek, and Jonathan Ball. Statistics review 10: Further nonparametric methods. *Critical Care*, 8(3):196–199, 2004. doi: 10.1186/cc2857. URL `http://ccforum.com/content/8/3/196`.

[26] A. Hazra and N. Gogtay. Biostatistics series module 4: Comparing groups - categorical variables. *Indian Journal of Dermatology*, 61(4):385–392, 2016. doi: 10.4103/0019-5154.185700.

[27] Mary L. McHugh. The chi-square test of independence. *Biochemia Medica*, 23(2):143–149, 2013. doi: 10.11613/BM.2013.018. URL `http://dx.doi.org/10.11613/BM.2013.018`.

[28] Mike J Bradburn, Taane G Clark, Sharon B Love, and Douglas Graham Altman. Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3): 431–436, 2003.

[29] Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40): 1317, 2019. doi: 10.21105/joss.01317. URL `https://doi.org/10.21105/joss.01317`.

[30] Samar Abd ElHafeez, Graziella D'Arrigo, Daniela Leonardis, Maria Fusaro, Giovanni Tripepi, and Stefanos Roumeliotis. Methods to analyze time-to-event data: The cox regression analysis. *Oxidative Medicine and Cellular Longevity*, page Article ID 1302811, 2021. doi: 10.1155/2021/1302811. URL `https://doi.org/10.1155/2021/1302811`.

[31] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

[32] E. J. Oh, Ravi B. Parikh, C. Chivers, and Jinbo Chen. Two-stage approaches to accounting for patient heterogeneity in machine learning risk prediction models in oncology. *JCO clinical cancer informatics*, 5: 1015–1023, 2021. doi: 10.1200/cci.21.00077.

[33] Eashwar Somasundaram, P. Johnson, A. El-Jawahri, Andrea Pusic, Jason B. Liu, T. M. Mulvey, Joseph A. Greer, and Thomas J. Roberts. Predictive modeling by tumor type: Comparing performance and feature importance of machine learning–based predictive models in patients with hematologic and solid tumors. *Journal of Clinical Oncology*, 2025. doi: 10.1200/jco.2025.43.16_suppl.e13638.

[34] F. Cabitza, A. Campagner, F. Soares, L. García de Guadiana-Romualdo, F. Challa, A. Sulejmani, M. Seghezzi, and A. Carobene. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Computer methods and programs in biomedicine*, 208: 106288, 2021. doi: 10.1016/j.cmpb.2021.106288. URL https://doi.org/10.1016/j.cmpb.2021.106288.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[36] Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL http://jmlr.org/papers/v21/20-729.html.

[37] M. W. Heymans and J. W. R. Twisk. Handling missing data in clinical research. *Journal of Clinical Epidemiology*, 151:185–188, 2022. doi: 10.1016/j.jclinepi.2022.08.016.

[38] S. M. Z. Memon, R. Wamala, and I. H. Kabano. A comparison of imputation methods for categorical data. *Informatics in Medicine Unlocked*, 42:101382, 2023. doi: 10.1016/j.imu.2023.101382.

[39] Peter C. Austin, Ian R. White, Douglas S. Lee, and Stef van Buuren. Missing data in clinical research: A tutorial on multiple imputation. *Canadian Journal of Cardiology*, 37:1322–1331, 2021. doi: 10.1016/j.cjca.2020.11.010.

[40] Frank E Harrell et al. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, volume 608. Springer, 2001.

[41] Samuel Von Wilson, Bogdan Cebere, James Myatt, and Samuel Wilson. Anothersamwilson/miceforest: Release for zenodo doi. *Zenodo*.

[42] 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.

[43] Seo Young Park, Ji Eun Park, Hyungjin Kim, and Seong Ho Park. Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). *Korean Journal of Radiology*, 22(10):1697–1707, 2021. doi: 10.3348/kjr.2021.0223.

[44] Sara Ferri. Shap-driven explainability in machine learning models applied to urothelial cancer real world data. Master's thesis, Politecnico di Milano, 2024. Tesi di Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica, Academic year 2023–2024.

[45] Gerd Assmann, Paul Cullen, and Helmut Schulte. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular münster (procam) study. *Circulation*, 105:310–315, 2002.

[46] Edouard F Bonneville, Matthieu Resche-Rigon, Johannes Schetelig, Hein Putter, and Liesbeth C de Wreede. Multiple imputation for cause-specific cox models: assessing methods for estimation and prediction. *Statistical Methods in Medical Research*, 31(10):1860–1880, 2022.

[47] Angela M Wood, Patrick Royston, and Ian R White. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biometrical Journal*, 57(4):614–632, 2015.

[48] Mélodie Monod, Peter Krusche, Qian Cao, Berkman Sahiner, Nicholas Petrick, David Ohlssen, and Thibaud Coroller. Torchsurv: A lightweight package for deep survival analysis. *Journal of Open Source Software*, 9 (104):7341, 2024. doi: 10.21105/joss.07341. URL https://doi.org/10.21105/joss.07341.

[49] Hans Van Houwelingen and Hein Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 2011.

[50] 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.

[51] 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.

[52] Dimitris Rizopoulos, Pedro Miranda-Afonso, and Grigorios Papageorgiou. *JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data*, 2026. URL `https://github.com/drizopoulos/jmbayes2`. R package version 0.6-0.

[53] Dimitris Rizopoulos. The R package JMbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, 72(7):1–45, 2016. doi: 10.18637/jss.v072.i07.

[54] Katya Mauff, Ewout Steyerberg, Isabella Kardys, Eric Boersma, and Dimitris Rizopoulos. Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach. *Statistics and Computing*, 30(4):999–1014, 2020.

[55] I. Rossi, F. Sartori, C. Rollo, G. Birolo, P. Fariselli, and T. Sanavia. Beyond cox models: Assessing the performance of machine-learning methods in non-proportional hazards and non-linear survival analysis. *Computers in biology and medicine*, 198 Pt B:111176, 2025. doi: 10.1016/j.compbiomed.2025.111176. URL `https://doi.org/10.1016/j.compbiomed.2025.111176`.

A. Appendix A

Table 10: List of features collected at the leukapheresis time point, including their description, data type, and percentage of missing values. Features highlighted in cyan indicate variables that are longitudinally tracked and collected at subsequent time points: infusion, and 30, 60, 90, and 180 days after infusion.

Feature Name	Description	Data type	Missing[%]
Sex	Sex of the patient	Categorical, Binary: Male, Female	0.0
Age at infusion	Age of the patient	Numerical	10.7
Disease Status	Disease status, relapsed or refractory	Categorical, Binary: relapsed/refractory	4.1
Ann Arbor stage	Anatomical extent of lymphoma	Categorical, Ordinal: I - IV	4.6
IPI-MIPI-FLIPI	Histology-specific prognostic index, IPI for DLBCL or PMBCL , MIPI for MCL, FLIPI for FL	Categorical, Ordinal: low, intermediate, high	10.1
B symptoms	Indicator of systemic lymphoma-related symptoms	Categorical, Binary: True,False	4,98
ECOG	Determines ability of patient to tolerate therapies in serious illness	Categorical, Ordinal: 0-4	3,8
Bulky	Indicator of presence of a large lymphoma mass	Categorical, Binary:Yes,No	4,7
Extranodal Sites	Number of lymphoma localizations outside lymph nodes, reflecting disease spread.	Caategorical, Ordinal :0-inf	9.8
Bone Marrow involvement	Indicator of lymphoma infiltration of the bone marrow	Categorical, Binary: Yes,No	41,0
ASCT	Indicator if the patient previously underwent Autologous Stem Cell Transplant	Categorical, Binary: Yes, No	2,8
num. of previous treatment	Count of prior therapy lines received before CAR-T	Categorical, Ordinal: 0-inf	2,4
Histology	Patient's diagnosed subtype of lymphoma	Categorical, Nominal: DLBCL, PMBCL, FL or MCL	0,8
HB	Count of hemoglobin	Numerical	4,4
Platelets	Count of platelets	Numerical	4,4
WBC	Count of White Blood Cells	Numerical	4.4
Neutrophils	Count of neutrophils	Numerical	4.8
Lymphocytes	Count of lymphocytes	Numerical	5.6
Monocytes	Count of monocytes	Numerical	14,5
Ferritin	Count of ferritin	Numerical	24,2
LDH norm.	Lactate DeHydrogenase express as a ratio to the laboratory upper limit of normal	Numerical	19,3
PCR	Count of C-Reactive Protein	Numerical	23,6
AST	Count of aspartate aminotransferase	Numerical	14,6

Continued on next page...

Continuation of Table 10

Feature Name	Description	Data type	Missing[%]
Alkaline phosphatase	Count of alkaline phosphatase	Numerical	23,6
D dimer	Count of D dimer	Numerical	60,8
Bilirubin	Count of bilirubin	Numerical	11,3
Albumin	Count of albumin	Numerical	41,5

End of Table 10

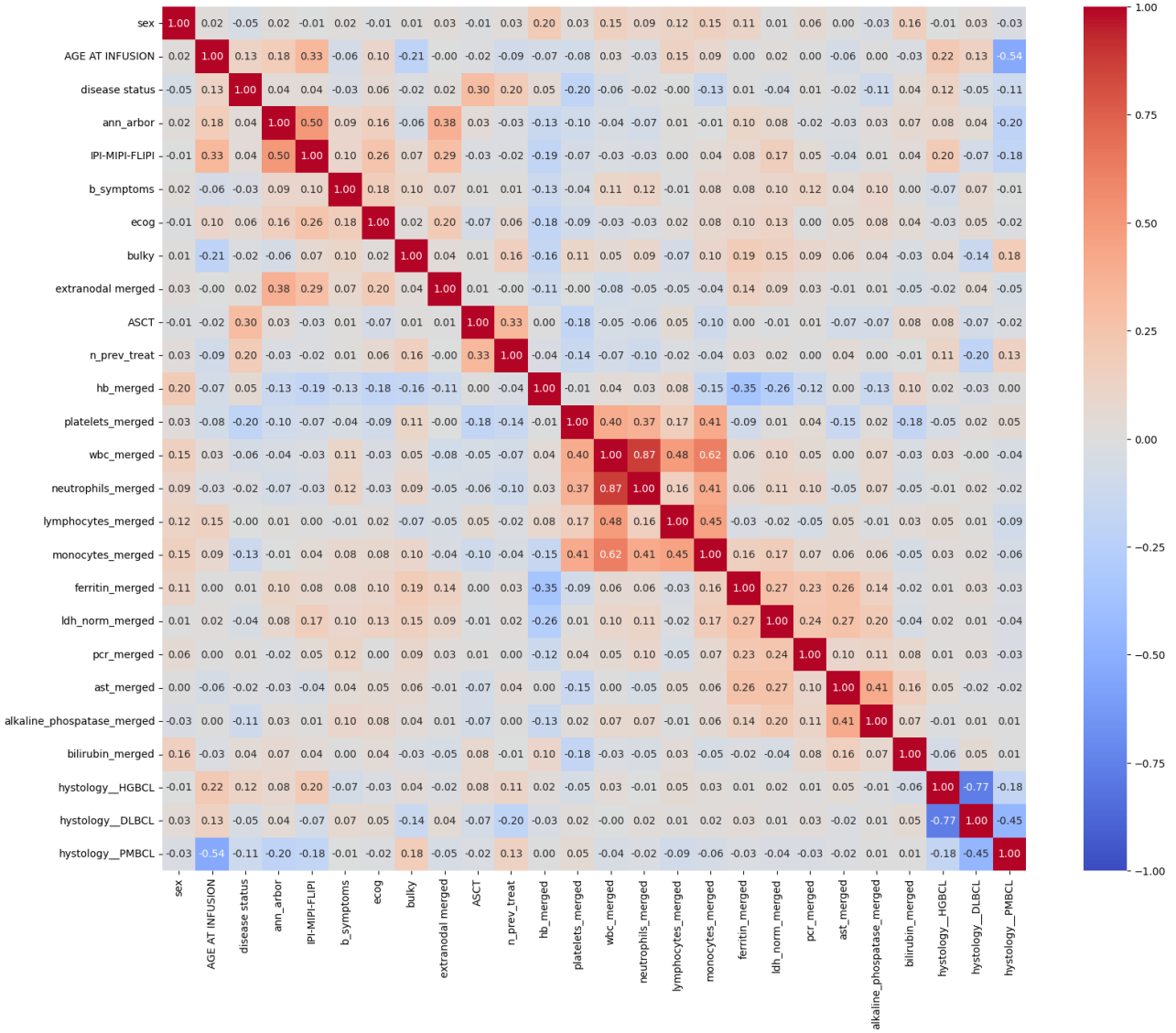

Figure 11: Correlation matrix of the train set.

Model	Hyperparameters	Selected Features
CPH	baseline estimation method: breslow, penalizer: 0.3, l1 ratio:0.0	hb, ldh norm, ferritin, pcr, extranodal sites, IPI-MIPI-FLIPI, alkaline phospatase, ann arbor, lymphocytes, monocytes
GBS	learning rate: 0.02, max depth:2, min samples leaf: 30, num estimators:300	ldh norm, hb ,ferritin, pcr, extranodal sites, IPI-MIPI-FLIPI, ann arbor
RSF	max depth: 2, min samples split:30, min samples leaf:30 , n estimators:100	hb, ldh norm, extranodal sites, ann arbor
EST	max depth: 3, min samples leaf: 30, min samples split: 30, n estimators: 200	ldh norm, hb, ferritin, ann arbor, extranodal sites, monocytes
SSVM	alpha: 0.001, gamma: 0.05, 'kernel':rbf	hb, ldh norm, ferritin, platelets, bilirubin, monocytes, extranodal sites, pcr, IPI-MIPI-FLIPI

Table 11: Selected features and optimal hyperparameters under the Standard Approach

tvAUC of JM

Landmark	Split	1 month after inf.	2 months after inf.	3 months after inf.	6 months after inf.	9 months after inf.	12 months after inf.	18 months after inf.	24 months after inf.
Leukapheresis	**Train**	0.69	0.67	0.68	0.63	0.63	0.63	0.57	0.52
	Test	0.52	0.68	0.71	0.69	0.67	0.65	0.60	0.61
	Ext. Validation	0.88	0.72	0.73	0.66	0.61	0.60	0.54	0.53
Infusion	**Train**	0.70	0.67	0.67	0.62	0.62	0.61	0.56	0.47
	Test	0.52	0.69	0.70	0.66	0.65	0.62	0.57	0.56
	Ext. Validation	0.87	0.73	0.71	0.62	0.56	0.55	0.54	0.47
30 day	**Train**	X	0.65	0.68	0.63	0.59	0.57	0.58	0.57
	Test	X	0.67	0.67	0.57	0.53	0.53	0.52	0.55
	Ext. Validation	X	0.66	0.80	0.64	0.58	0.58	0.60	0.59
60 day	**Train**	X	X	0.68	0.63	0.60	0.59	0.58	0.58
	Test	X	X	0.67	0.56	0.52	0.49	0.46	0.45
	Ext. Validation	X	X	0.88	0.61	0.63	0.61	0.65	0.63
90 day	**Train**	X	X	X	0.60	0.58	0.57	0.57	0.56
	Test	X	X	X	0.57	0.54	0.55	0.51	0.49
	Ext. Validation	X	X	X	0.58	0.63	0.64	0.69	0.67
180 day	**Train**	X	X	X	X	0.50	0.51	0.55	0.55
	Test	X	X	X	X	0.45	0.48	0.44	0.44
	Ext. Validation	X	X	X	X	0.85	0.90	0.79	0.79

Table 12: tvAUC values for the JM at different landmark times and prediction horizons, stratified by data split.

tvAUC of DDH

Landmark	Split	1 month after inf.	2 months after inf.	3 months after inf.	6 months after inf.	9 months after inf.	12 months after inf.	18 months after inf.	24 months after inf.
Leukapheresis	**Train**	0.68	0.65	0.65	0.66	0.68	0.68	0.68	0.71
	Validation	0.68	0.66	0.68	0.70	0.72	0.68	0.66	0.78
	Test	0.71	0.69	0.69	0.70	0.73	0.71	0.70	0.76
	Ext. Validation	0.75	0.63	0.65	0.67	0.68	0.75	0.80	0.77
Infusion	**Train**	0.67	0.67	0.67	0.69	0.70	0.70	0.70	0.73
	Validation	0.66	0.70	0.73	0.72	0.74	0.71	0.69	0.82
	Test	0.71	0.68	0.69	0.70	0.73	0.67	0.68	0.74
	Ext. Validation	0.75	0.68	0.68	0.69	0.70	0.76	0.76	0.85
30 day	**Train**	X	0.73	0.72	0.73	0.73	0.73	0.74	0.75
	Validation	X	0.78	0.78	0.76	0.79	0.75	0.73	0.86
	Test	X	0.76	0.77	0.76	0.76	0.72	0.75	0.80
	Ext. Validation	X	0.74	0.73	0.71	0.72	0.74	0.79	0.85
60 day	**Train**	X	X	0.74	0.74	0.74	0.73	0.74	0.74
	Validation	X	X	0.82	0.76	0.78	0.75	0.73	0.83
	Test	X	X	0.81	0.80	0.76	0.77	0.72	0.77
	Ext. Validation	X	X	0.78	0.79	0.79	0.70	0.81	0.89
90 day	**Train**	X	X	X	0.74	0.74	0.74	0.74	0.76
	Validation	X	X	X	0.76	0.77	0.73	0.71	0.81
	Test	X	X	X	0.80	0.80	0.77	0.76	0.80
	Ext. Validation	X	X	X	0.78	0.79	0.82	0.81	0.89
180 day	**Train**	X	X	X	X	0.74	0.74	0.74	0.23
	Validation	X	X	X	X	0.74	0.71	0.68	0.22
	Test	X	X	X	X	0.81	0.77	0.76	0.22
	Ext. Validation	X	X	X	X	0.80	0.83	0.81	0.09

Table 13: tvAUC values for the DDH model at different landmark times and prediction horizons, stratified by data split (train, validation, test, and external validation).

tvAUC of RNN-FNN

Landmark	Split	1 month after inf.	2 months after inf.	3 months after inf.	6 months after inf.	9 months after inf.	12 months after inf.	18 months after inf.	24 months after inf.
Leukapheresis	**Train**	0.96	0.96	0.96	0.84	0.85	0.86	0.86	0.86
	Validation	0.46	0.48	0.54	0.61	0.63	0.68	0.69	0.70
	Test	0.55	0.68	0.69	0.61	0.52	0.50	0.50	0.51
	Ext. Validation	0.62	0.55	0.42	0.48	0.43	0.44	0.51	0.60
Infusion	**Train**	0.97	0.96	0.84	0.86	0.86	0.87	0.87	0.86
	Validation	0.48	0.44	0.54	0.63	0.59	0.64	0.65	0.66
	Test	0.53	0.68	0.69	0.52	0.50	0.50	0.49	0.50
	Ext. Validation	0.63	0.53	0.40	0.47	0.43	0.42	0.47	0.56
30 day	**Train**	X	0.98	0.86	0.87	0.88	0.88	0.88	0.88
	Validation	X	0.46	0.57	0.62	0.62	0.67	0.69	0.69
	Test	X	0.65	0.68	0.52	0.50	0.51	0.49	0.50
	Ext. Validation	X	0.50	0.35	0.49	0.44	0.45	0.56	0.68
60 day	**Train**	X	X	0.88	0.88	0.89	0.89	0.89	0.89
	Validation	X	X	0.60	0.63	0.66	0.71	0.71	0.71
	Test	X	X	0.53	0.46	0.46	0.47	0.45	0.47
	Ext. Validation	X	X	0.36	0.54	0.50	0.50	0.66	0.81
90 day	**Train**	X	X	X	0.99	0.99	0.99	0.99	0.99
	Validation	X	X	X	0.67	0.67	0.70	0.71	0.71
	Test	X	X	X	0.49	0.47	0.48	0.45	0.46
	Ext. Validation	X	X	X	0.61	0.56	0.60	0.71	0.84
180 day	**Train**	X	X	X	X	0.99	0.99	0.99	0.99
	Validation	X	X	X	X	0.12	0.52	0.71	0.79
	Test	X	X	X	X	0.60	0.70	0.64	0.70
	Ext. Validation	X	X	X	X	0.66	0.67	0.79	0.79

Table 14: tvAUC values for the RNN-FNN model at different landmark times and prediction horizons, stratified by data split (train, validation, test, and external validation).

tvAUC of Last-visit RSF

Landmark	Split	1 month after inf.	2 months after inf.	3 months after inf.	6 months after inf.	9 months after inf.	12 months after inf.	18 months after inf.	24 months after inf.
Leukapheresis	**Train**	0.96	0.95	0.95	0.97	0.98	0.98	0.98	0.99
	Validation	0.92	0.85	0.88	0.89	0.89	0.88	0.89	0.89
	Test	0.74	0.73	0.71	0.67	0.66	0.66	0.64	0.66
	Ext. Validation	0.80	0.74	0.78	0.71	0.69	0.71	0.78	0.80
Infusion	**Train**	0.93	0.92	0.92	0.94	0.95	0.98	0.97	0.98
	Validation	0.64	0.65	0.66	0.73	0.76	0.80	0.81	0.89
	Test	0.70	0.71	0.72	0.68	0.69	0.68	0.68	0.69
	Ext. Validation	0.83	0.65	0.72	0.60	0.62	0.61	0.68	0.61
30 day	**Train**	X	0.98	0.97	0.98	0.99	0.99	0.99	0.99
	Validation	X	0.72	0.76	0.78	0.81	0.82	0.82	0.83
	Test	X	0.70	0.71	0.64	0.64	0.63	0.61	0.63
	Ext. Validation	X	0.68	0.76	0.67	0.64	0.65	0.77	0.77
60 day	**Train**	X	X	0.99	0.99	0.99	0.99	0.99	0.99
	Validation	X	X	0.72	0.81	0.83	0.83	0.83	0.83
	Test	X	X	0.58	0.60	0.56	0.57	0.57	0.60
	Ext. Validation	X	X	0.81	0.70	0.68	0.71	0.73	0.81
90 day	**Train**	X	X	X	0.99	0.99	0.99	0.99	0.99
	Validation	X	X	X	0.83	0.82	0.78	0.80	0.82
	Test	X	X	X	0.57	0.55	0.57	0.55	0.58
	Ext. Validation	X	X	X	0.54	0.56	0.57	0.73	0.74
180 day	**Train**	X	X	X	X	0.99	0.99	0.99	0.99
	Validation	X	X	X	X	0.30	0.59	0.62	0.66
	Test	X	X	X	X	0.54	0.53	0.48	0.55
	Ext. Validation	X	X	X	X	0.49	0.50	0.79	0.82

Table 15: tvAUC values for the last-visit RSF model at different landmark times and prediction horizons, stratified by data split (train, validation, test, and external validation).

BS score of DDH

Landmark	Split	1 month after inf.	2 months after inf.	3 months after inf.	6 months after inf.	9 months after inf.	12 months after inf.	18 months after inf.	24 months after inf.
Leukapheresis	**Train**	0.11	0.16	0.18	0.22	0.23	0.24	0.24	0.24
	Validation	0.23	0.26	0.27	0.28	0.28	0.27	0.26	0.28
	Test	0.22	0.25	0.24	0.26	0.26	0.25	0.27	0.25
	Ext. Validation	0.23	0.26	0.27	0.28	0.28	0.27	0.26	0.28
Infusion	**Train**	0.11	0.17	0.19	0.22	0.23	0.24	0.24	0.24
	Validation	0.25	0.27	0.28	0.29	0.29	0.27	0.26	0.28
	Test	0.28	0.27	0.29	0.28	0.27	0.29	0.27	-
	Ext. Validation	0.25	0.27	0.28	0.29	0.29	0.27	0.26	0.28
30 day	**Train**	X	0.17	0.20	0.23	0.24	0.25	0.25	0.25
	Validation	X	0.29	0.30	0.31	0.32	0.30	0.28	0.28
	Test	X	0.29	0.31	0.30	0.29	0.31	0.29	-
	Ext. Validation	X	0.29	0.30	0.31	0.32	0.30	0.28	0.28
60 day	**Train**	X	X	0.21	0.24	0.25	0.26	0.26	0.26
	Validation	X	X	0.33	0.34	0.34	0.32	0.30	0.29
	Test	X	X	0.29	0.31	0.30	0.29	0.31	0.29
	Ext. Validation	X	X	0.33	0.34	0.34	0.32	0.30	0.29
90 day	**Train**	X	X	X	0.27	0.27	0.28	0.28	0.27
	Validation	X	X	X	0.37	0.38	0.35	0.32	0.31
	Test	X	X	X	0.34	0.33	0.32	0.33	0.31
	Ext. Validation	X	X	X	0.37	0.38	0.35	0.32	0.31
180 day	**Train**	X	X	X	X	0.31	0.32	0.31	0.30
	Validation	X	X	X	X	0.45	0.41	0.37	0.34
	Test	X	X	X	X	0.40	0.37	0.38	0.35
	Ext. Validation	X	X	X	X	0.45	0.41	0.37	0.34

Table 16: BS values for the DDH model at different landmark times and prediction horizons, stratified by data split (train, validation, test, and external validation).

BS score of RNN-CPH

Landmark	Split	1 month after inf.	2 months after inf.	3 months after inf.	6 months after inf.	9 months after inf.	12 months after inf.	18 months after inf.	24 months after inf.
Leukapheresis	**Train**	0.03	0.07	0.11	0.20	0.23	0.24	0.24	0.24
	Validation	0.05	0.15	0.19	0.20	0.20	0.19	0.21	0.28
	Test	0.05	0.15	0.18	0.20	0.20	0.21	0.23	0.23
	Ext. Validation	0.07	0.13	0.15	0.18	0.22	0.23	0.12	0.09
Infusion	**Train**	0.03	0.07	0.11	0.18	0.20	0.22	0.26	0.21
	Validation	0.06	0.16	0.21	0.23	0.22	0.22	0.22	0.25
	Test	0.05	0.16	0.20	0.22	0.22	0.22	0.22	0.23
	Ext. Validation	0.08	0.15	0.18	0.20	0.21	0.22	0.14	0.13
30 day	**Train**	X	0.04	0.09	0.17	0.20	0.20	0.22	0.21
	Validation	X	0.13	0.20	0.23	0.22	0.23	0.23	0.26
	Test	X	0.13	0.18	0.21	0.22	0.22	0.24	0.23
	Ext. Validation	X	0.09	0.16	0.21	0.22	0.22	0.16	0.14
60 day	**Train**	X	X	0.06	0.14	0.19	0.20	0.22	0.29
	Validation	X	X	0.12	0.19	0.21	0.23	0.24	0.28
	Test	X	X	0.09	0.15	0.19	0.20	0.22	0.23
	Ext. Validation	X	X	0.09	0.18	0.20	0.20	0.16	0.14
90 day	**Train**	X	X	X	0.13	0.17	0.19	0.21	0.21
	Validation	X	X	X	0.17	0.19	0.22	0.23	0.28
	Test	X	X	X	0.12	0.17	0.18	0.21	0.23
	Ext. Validation	X	X	X	0.15	0.18	0.19	0.17	0.15
180 day	**Train**	X	X	X	X	0.08	0.12	0.16	0.18
	Validation	X	X	X	X	0.06	0.17	0.19	0.19
	Test	X	X	X	X	0.10	0.12	0.17	0.21
	Ext. Validation	X	X	X	X	0.07	0.08	0.13	0.11

Table 17: BS values for the RNN-CPH at different landmark times and prediction horizons, stratified by data split.

BS score of RNN-FNN

Landmark	Split	1 month after inf.	2 months after inf.	3 months after inf.	6 months after inf.	9 months after inf.	12 months after inf.	18 months after inf.	24 months after inf.
Leukapheresis	**Train**	0.01	0.02	0.03	0.04	0.05	0.05	0.08	0.12
	Validation	0.08	0.21	0.28	0.27	0.25	0.18	0.18	0.23
	Test	0.07	0.18	0.25	0.31	0.32	0.33	0.35	0.38
	Ext. Validation	0.09	0.18	0.33	0.37	0.42	0.44	0.30	0.20
Infusion	**Train**	0.01	0.02	0.03	0.04	0.05	0.06	0.08	0.11
	Validation	0.08	0.21	0.29	0.27	0.26	0.20	0.19	0.24
	Test	0.07	0.18	0.25	0.31	0.32	0.34	0.37	0.38
	Ext. Validation	0.10	0.18	0.33	0.37	0.42	0.46	0.26	0.20
30 day	**Train**	X	0.01	0.02	0.04	0.05	0.05	0.08	0.11
	Validation	X	0.16	0.26	0.26	0.25	0.20	0.18	0.2
	Test	X	0.16	0.24	0.31	0.33	0.33	0.37	0.38
	Ext. Validation	X	0.12	0.30	0.38	0.42	0.44	0.27	0.19
60 day	**Train**	X	X	0.02	0.03	0.04	0.05	0.07	0.10
	Validation	X	X	0.18	0.30	0.29	0.22	0.22	0.26
	Test	X	X	0.18	0.30	0.32	0.34	0.38	0.42
	Ext. Validation	X	X	0.21	0.30	0.40	0.41	0.31	0.16
90 day	**Train**	X	X	X	0.02	0.03	0.03	0.05	0.07
	Validation	X	X	X	0.26	0.27	0.21	0.19	0.22
	Test	X	X	X	0.21	0.27	0.30	0.37	0.40
	Ext. Validation	X	X	X	0.30	0.35	0.35	0.32	0.19
180 day	**Train**	X	X	X	X	0.00	0.00	0.01	0.02
	Validation	X	X	X	X	0.09	0.17	0.13	0.13
	Test	X	X	X	X	0.13	0.16	0.22	0.25
	Ext. Validation	X	X	X	X	0.08	0.10	0.18	0.16

Table 18: BS values for the RNN-FNN at different landmark times and prediction horizons, stratified by data split.

BS score of Last Visit RSF

Landmark	Split	1 month after inf.	2 months after inf.	3 months after inf.	6 months after inf.	9 months after inf.	12 months after inf.	18 months after inf.	24 months after inf.
Leukapheresis	**Train**	0.03	0.06	0.09	0.12	0.13	0.13	0.12	0.11
	Validation	0.05	0.15	0.19	0.19	0.18	0.17	0.17	0.20
	Test	0.05	0.14	0.18	0.20	0.22	0.22	0.24	0.25
	Ext. Validation	0.08	0.14	0.17	0.21	0.23	0.23	0.17	0.13
Infusion	**Train**	0.03	0.06	0.09	0.14	0.15	0.15	0.14	0.14
	Validation	0.06	0.15	0.19	0.19	0.19	0.18	0.18	0.21
	Test	0.05	0.15	0.18	0.21	0.22	0.23	0.24	0.26
	Ext. Validation	0.07	0.13	0.17	0.21	0.24	0.25	0.15	0.13
30 day	**Train**	X	0.04	0.07	0.11	0.11	0.11	0.10	0.10
	Validation	X	0.13	0.20	0.24	0.23	0.22	0.20	0.22
	Test	X	0.13	0.19	0.22	0.23	0.24	0.26	0.27
	Ext. Validation	X	0.10	0.16	0.22	0.24	0.24	0.16	0.15
60 day	**Train**	X	X	0.05	0.10	0.11	0.11	0.10	0.09
	Validation	X	X	0.12	0.19	0.21	0.21	0.20	0.22
	Test	X	X	0.09	0.15	0.22	0.22	0.25	0.27
	Ext. Validation	X	X	0.10	0.19	0.21	0.21	0.17	0.14
90 day	**Train**	X	X	X	0.08	0.09	0.09	0.09	0.08
	Validation	X	X	X	0.17	0.18	0.21	0.20	0.22
	Test	X	X	X	0.12	0.17	0.20	0.23	0.27
	Ext. Validation	X	X	X	0.15	0.19	0.20	0.17	0.15
180 day	**Train**	X	X	X	X	0.04	0.06	0.06	0.06
	Validation	X	X	X	X	0.06	0.18	0.17	0.18
	Test	X	X	X	X	0.11	0.13	0.19	0.24
	Ext. Validation	X	X	X	X	0.07	0.08	0.13	0.12

Table 19: BS values for the Last Visit RSF at different landmark times and prediction horizons, stratified by data split.

Abstract in lingua italiana

La terapia con cellule T con recettore chimerico dell'antigene (CAR-T) ha rivoluzionato il trattamento del linfoma a cellule B recidivato o refrattario, ma non tutti i pazienti raggiungono una remissione duratura. L'identificazione precoce dei soggetti ad alto rischio di progressione della malattia è cruciale. Questo studio valuta architetture statiche e dinamiche di Machine Learning (ML) e Deep Learning (DL) per predire la Progression-Free Survival (PFS) in una coorte multicentrica di 1.465 pazienti arruolati nello studio CART-SIE. Abbiamo utilizzato variabili cliniche e di laboratorio raccolte alla leucaferesi, all'infusione e nei successivi follow-up standardizzati (tra 30 e 180 giorni dopo l'infusione). Per la predizione statica, i modelli di Cox Proportional Hazards (CPH) e i modelli di ML sono stati confrontati con reti neurali profonde (DeepSurv). Sebbene DeepSurv abbia ottenuto una performance discriminativa interna superiore, il modello CPH standard ha mantenuto un C-index più elevato nella coorte indipendente di validazione esterna (0,73 contro 0,69), dimostrando una maggiore generalizzabilità. Per sfruttare i dati longitudinali, sono stati sviluppati framework predittivi dinamici, confrontando Joint Models, Dynamic-DeepHit e strategie di super-landmarking basate su Recurrent Neural Networks (RNN). I risultati dimostrano che codificare l'intera traiettoria del paziente attraverso un'architettura ibrida RNN-CPH offre le migliori capacità predittive. I confronti tra tutti i modelli sono stati effettuati utilizzando la AUC tempo-variante (tvAUC) alla leucaferesi sia sul set di test sia sul set di validazione esterna. In queste valutazioni, l'architettura RNN-CPH ha ottenuto risultati superiori, probabilmente perché il suo encoder ricorrente ha appreso con successo dall'evoluzione temporale delle traiettorie dei pazienti. In conclusione, mentre il CPH standard si conferma un modello solido per le predizioni statiche, l'integrazione di misurazioni cliniche sequenziali tramite architetture ricorrenti ibride all'interno di un framework di super-landmarking offre un vantaggio sostanziale per una stratificazione del rischio continua e dinamica nella terapia CAR-T.

Parole chiave: Terapia CAR-T, Analisi di sopravvivenza dinamica, Machine Learning, Super-landmarking, Reti Neurali Ricorrenti, Linfoma a grandi cellule B