MSnet: A deep neural network based on piecewise-constant proposals within Multi-State event history analysis

Aziliz Cottin AZILIZ.COTTIN@3DS.COM
Marine Zulian
Healthcare and Life Sciences Research, Dassault Systemes, Velizy-Villacoublay 78140, France

Sandrine Katsahian
Inria, Université Paris Cité, Inserm, HeKA, F-75015 Paris, France
Medical Informatics, Biostatistics and Public Health Department, Georges Pompidou, Assistance Publique-Hôpitaux de Paris, Paris, France

Agathe Guilloux
Inria, Université Paris Cité, Inserm, HeKA, F-75015 Paris, France

Abstract

Multi-state models are essential to represent realistic disease trajectories in oncology, yet most existing survival and deep-learning approaches either rely on restrictive Markov assumptions or fail to provide subject-specific transition risks. We propose **MSnet**, a deep learning framework for progressive semi-Markov multi-state processes with right-censoring. MSnet models transition-specific cumulative risks as functions of sojourn time using a multi-task architecture that flexibly integrates high-dimensional clinical and omics data. Experiments on simulated data and two real-world breast cancer cohorts show that MSnet improves predictive performance while yielding clinically interpretable transition dynamics, extending deep learning–based survival analysis to more realistic, patient-centered disease processes.

Keywords: Deep survival models, Semi-Markov multi-state processes, Personalized risk prediction, Precision oncology, Transcriptomic data

Data and Code Availability All the clinical datasets used in our study are publicly available (METABRIC (Pereira et al., 2016), Breast Invasive Carcinoma TCGA PanCancer (Weinstein et al., 2013), Rotterdam breast cancer (Royston and Altman, 2013)). The code is not publicly available.

Institutional Review Board (IRB) This research was conducted using publicly available, pseudonymized and deidentified health data. As such, no identifiable personal information was involved, and IRB approval was not required.

1. Introduction

Assessing patient prognosis is a key component of clinical decision-making in oncology, enabling the identification of individuals at high risk of relapse or death. Survival analysis provides a robust statistical foundation for modelling time-to-event data (Klein and Moeschberger, 2006). However, modern oncology increasingly requires modelling sequential and competing events, calling for a more flexible framework than single-event survival models.

Multi-state processes address this need by representing disease evolution as transitions between successive states (Hougaard, 1999; Putter et al., 2006). They naturally encompass widely used settings in oncology such as competing risks (Kim, 2024; Schuster et al., 2020) and illness–death models (Putter et al., 2006), and they enable more granular descriptions of disease trajectories that align with precision medicine objectives (e.g. intermediate progression states Sopik et al. (2023)).

Classical multi-state inference is typically based on transition-specific Cox proportional hazards (P.H.) models (Cox, 1972; De Wreede et al., 2010). While effective in low-dimensional settings, these models rely on restrictive assumptions, including log-linear covariate effects, pro-

portional hazards-P.H., and prior specification of the interactions, and they do not scale well to high-dimensional multimodal data.

A further limitation is the common reliance on the *Markov* assumption, where transition risks depend solely on the current state. In oncology, however, the risk of progression or mortality is strongly influenced by the time already spent in a state (e.g., relapse-free duration, treatment exposure). Relaxing this constraint through *semi-Markov* modelling is therefore crucial to capture clinically meaningful trajectory dynamics (Davidov, 1999; Andersen et al., 2012; Meira-Machado and Sestelo, 2019).

These challenges motivate the use of deep learning (DL) and machine learning (ML) methods for multi-state survival modelling. DL models flexibly learn non-linear relationships and can integrate high-dimensional multimodal representations from genomics, imaging, and electronic health records while enabling *subject-specific* prediction (Kourou et al., 2015; Qu et al., 2023; Swanson et al., 2023; Zhang et al., 2023b,a; Wiegrebe et al., 2024). While a few studies have addressed more general multi-state processes, they often rely on the restrictive Markov assumption (Groha et al., 2020; Rahman and Purushotham, 2023). This motivates the development of DL-based methods for semi-Markov multi-state survival modelling.

2. Related Work

Traditional ML survival models include the LASSO-penalized Cox model (Tibshirani, 1997), random survival forests (Ishwaran et al., 2008), boosting-based approaches (Bühlmann and Hothorn, 2007) and support vector machines (Khan and Zubek, 2008). However, these models struggle with non-linear relationships and high-dimensional multimodal inputs.

Competing risks. DL-based competing risks models were introduced by Biganzoli et al. (2006). DeepHit (Lee et al., 2018) and subsequent extensions adopt event-specific subnetworks to estimate cumulative incidence functions, and are comprehensively reviewed in Monterrubio-Gómez et al. (2024).

Illness–death models. IDNetwork (Cottin et al., 2022) extends DeepHit to the three-state illness–death setting under a semi-Markov assumption, using transition-specific subnetworks and piecewise-constant risk function modelling.

General multi-state models. SurvNODE (Groha et al., 2020) pioneered DL-based modelling for general multi-state processes using neural ODEs, but it relies on the *Markov* assumption (preventing explicit modelling of sojourn-time effects) and implies a high computational time (Gholami et al., 2019). MSPseudo (Rahman and Purushotham, 2023) enables non-Markov estimation using pseudo-values but does not directly model transition risks, limiting subject-level prediction and high-dimensional learning.

Despite these advances, no existing DL framework jointly supports semi-Markov transition dynamics, subject-specific risk estimation, and multimodal data integration under right-censoring, and while remaining computationally feasible. This represents a key gap in the current literature.

Contributions We introduce **MSnet**, a deep learning algorithm for *progressive semi-Markov* multi-state survival with right-censoring. Our key contributions are:

1. **Semi-Markov modelling.** MSnet models transition risks as functions of both covariates and sojourn time, removing the Markov restriction of SurvNODE (Groha et al., 2020) and allowing realistic oncology trajectories.

2. **Individualised risk estimates.** Unlike MSPseudo (Rahman and Purushotham, 2023), MSnet directly outputs personalised risks and event probabilities, enhancing interpretability and clinical relevance.

3. **Scalable multi-task architecture.** Transition-specific subnetworks capture complex non-linear effects without imposing Cox-type constraints, and readily incorporate high-dimensional, multimodal data.

4. **Computational efficiency.** On simulations, MSnet outperforms the predictive performance of competing methods while remaining computationally efficient.

5. **Validation in precision oncology.** In a four-state breast-cancer application,MSnet

not only improves predictive performance but also capture clinically meaningful signals.

3. Methods

3.1. Background on multi-state processes

3.1.1. Classical formalism

We consider a progressive multi-state model with $K+1$ states. State 0 is the initial state and states $\{1, \cdots, K\}$ represent the clinical events that a patient can experience. Examples of multi-state processes are displayed in Appendix A - Figure 1.

Formally, a multi-state process $\{E(t),\ 0 \leq t \leq +\infty\}$ is a continuous-time stochastic process that takes its values in $\{0, \cdots, K\}$ (Andersen et al., 2012): $E(t) = k$ indicates that the patient is in state k at time t ($t > 0$). We assume that all subjects are in state 0 at time $t = 0$ (i.e., $\mathbb{P}(E(0) = 0) = 1$).

We consider in this paper only the case of progressive multi-state processes, whose sequence of events is increasing (i.e., with irreversible transitions) (De Wreede et al., 2010). We note $k \to l$ the transition from state k to state l, for $(k, l) \in \{0, \cdots, K\}$, and that is possible only if $k < l$. Since we consider progressive multi-state processes, the transition $k \to l$ exists only if $k < l$. We note $S = \{(k, l)\}_{(k,l) \in \{0,\cdots,K\},\ k<l}$ the space of possible transitions of the process E.

3.1.2. Transition times

The evolution of a multi-state process is implicitly characterized by the latent random variables T_{kl} associated with each transition $k \to l$ (with $k < l, (k, l) \in S$ for a progressive process) and representing the transition time from state k to state l. The law of these transition times is identifiable only under certain conditions for competing transitions (Tsiatis, 1975; Beyersmann et al., 2009).

We consider the times T_k, $0 \leq k \leq K-1$ (i.e. K is an absorbing state), to be truly observable and whose distribution law is identifiable. We define T_k as the exit time from state k and denote by S_k the space of possible arrival states from state k, i.e., $S_k = \{l\}_{l \in \{0,\cdots,K\},\ k<l}$. For $k = 0, \ldots, K-1$, we define the exit time from a state k as:

$$T_k = \inf_{t>0} \{E(t) > k\} = \min\left(T_{kl}\right)_{l \in S_k}.$$

Together we define $D_k \in S_k$ which indicates the state of arrival from of state k. If a state k^* ($k^* \in 1, \cdots, K-1$) is an absorbing state, then $T_{k^*} = +\infty$ and D_{k^*} is not defined, e.g. $T_K = +\infty$. We also get $T_k = +\infty$ in cases where a state k is not visited by an individual.

We introduce C a positive and continuous random variable of (right-)censoring that prevents the observation of T_k (i.e., so that the process E is observed only on the set $\{t \leq C\}_{t>0}$). We define the exit time of the actually observed state k as

$$\tilde{T}_k = \min\left(T_k, C\right).$$

Together with these event times, we observe a vector of covariates X and we assume $C \perp\!\!\!\perp T_k | X$. $\tilde{T}_k$ is jointly observed with the binary label

$$\delta_{kl} = \mathbb{1}\{D_k = l, T_k \leq C\}$$

that indicates the status of the transition ($\delta_{kl} = 1$ indicates the observation of a transition from state k to l, $\delta_{kl} = 0$ indicates a censoring).

Together with these event times, we observe a vector of covariates X and we assume $C \perp\!\!\!\perp T_k | X$.

3.1.3. Transition intensities

A multi-state process is conventionally associated with a set of transition intensities, or transition-specific hazard functions, noted α_{kl}, $k < l, (k, l) \in S$. For $t > 0$,

$$\alpha_{kl}(t|X) = \lim_{dt \to 0} \frac{\mathbb{P}\big(E(t+dt) = l \mid E(t) = k; X\big)}{dt}$$

represents the instantaneous risk of transitioning to state l at time $t + dt$, $dt \to 0$, conditionally on being in state k at time $t-$ (Hougaard, 1999; Andersen and Perme, 2008). We also define their cumulative counterparts as:

$$A_{kl}(t|X) = \int_0^t \alpha_{kl}(s|X) ds. \tag{1}$$

3.1.4. Log-likelihood

The log-likelihood of a multi-state process, on a time window $[0, \tau]$, and conditionally on the set

of covariates X, is traditionally defined as (Jacod and Memin, 1976; Hougaard, 1999; Andersen et al., 2012):

$$\log \mathcal{L} = \sum_{(k,l)\in S} \left[P_1^{kl} + P_2^{kl}\right],$$

with

$$P_1^{kl} = \delta_{kl} \log\left(\alpha_{kl}(\tilde{T}_k|X)\right),$$
$$P_2^{kl} = -\int_0^\tau \alpha_{kl}\left(t|X\right) Y_k\left(t\right) \mathbb{1}\left\{C \leq t\right\} dt, \quad (2)$$

and where $Y = (Y_0, \ldots, Y_K)$ is defined, for $t > 0$ and $k \in \{0, \cdots, K\}$, as $Y_k(t) = \mathbb{1}\left\{E\left(t-\right) = k\right\}$.

3.2. MSnet

3.2.1. Our theoretical formalism

New notations In the previous section we defined the general framework of Markovian multi-state processes. For the semi-Markovian case, we introduce a new notation Z_k, $0 \leq k \leq K$, defined as the entry time into a state k with $Z_0 = 0$ and, for $1 \leq k \leq K$,

$$Z_k = \min_{t>0}\left\{E(t) = k\right\},$$

so that $T_k - Z_k$ is the sojourn time in state k. We get $Z_k = +\infty$ in cases where a state k is not visited.

Because of right-censoring that prevents the observation of Z_k, we define the entry time of the actually observed state k as

$$\tilde{Z}_k = \min\left(Z_k, C\right)$$

and we assume $C \perp\!\!\!\perp Z_k|X$. $\tilde{Z}_k$ is jointly observed with the binary label δ_{Z_k}, with $\delta_{Z_0} = 1$ and, for $1 \leq k \leq K$,

$$\delta_{Z_k} = \mathbb{1}\left\{\sum_{q<k} \delta_{qk} \geq 1\right\}$$

($\delta_{Z_k} = 1$ indicates an observed entry in state k).

Quantities of interest Following the approach proposed in Cottin et al. (2022), the semi-Markovian assumption involves changes in the definitions of the densities and in the likelihood by considering event the duration spent in the state instead of event times (Hougaard, 1999; Eulenburg et al., 2015; Meira-Machado and Sestelo, 2019; Saint Pierre, 2021). For a transition $k \to l$ $(k < l, (k,l) \in S)$, the density function is then written conditionally on Z_k, and such that:

$$f_{kl}(t) = \lim_{dt\to 0} \frac{\mathbb{P}\big(t \leq T_k - Z_k \leq t + dt, D_k = l | Z_k\big)}{dt}.$$

We define the cumulative incidence function (CIF) for a transition $k \to l$ as follows

$$F_{kl}(t) = \mathbb{P}\big(T_k - Z_k \leq t, D_k = l | Z_k\big),$$

and the cumulative distribution function of T_k as follows

$$F_k(t) = \sum_{l\in S_k} F_{kl}(t). \quad (3)$$

We use piecewise constant (PC) proposals for the density functions, dividing the time axis into J disjoint intervals up to a finite time horizon τ. These intervals are defined as $v_1 = [a_0, a_1)$, $\cdots$, $v_J = [a_{J-1}, a_J)$, where $a_0 = 0$ and $a_J = \tau$. Thus, for any $t \in [0, \tau)$,

$$f_{kl}(t) = \sum_{j=1}^{J} f_{kl,j}^{\text{PC}} \times \mathbb{1}\{t \in v_j\}.$$

For $t \in [0, \tau[$, the CIFs are rewritten as piecewise linear functions as, for $j(t)$ the interval j so that $t \in v_j$, $|v_j| = a_j - a_{j-1}$ the length of interval j and $|v_{j(t)}| = t - a_{j(t)-1}$,

$$F_{kl}^{\text{PC}}\big(t\big) = \sum_{j=1}^{j(t)-1} |v_j|\ f_{kl,j}^{\text{PC}} + |v_{j(t)}|\ f_{kl,j(t)}^{\text{PC}}. \quad (4)$$

Log-likelihood The PC assumption simplifies log-likelihood optimisation. As compared to continuous-time (that need the Cox partial likelihood (Cox, 1972) and computationally intensive stochastic-gradient schemes (Kvamme et al., 2019; Achab et al., 2015)) or discrete-time models (that suffer from approximation error because they ignore the exact exposure time within each interval), the PC approach retains the subject's actual time-at-risk inside intervals (eliminating the discretisation bias) while preserving the computational efficiency of a discrete formulation.

We rewrite the log-likelihood of Equation (2) with these PC proposals and in terms of the time-homogeneous semi-Markvov property so that, for

$\mathbb{1}_k = \mathbb{1}\{k \text{ is not an absorbing state}\}$,

$$\log \mathcal{L}_J^{\text{MSnet}} = \sum_{k=0}^{K-1} \mathbb{1}_k \times \delta_{Z_k} \left(Q_1^k + Q_2^k\right) \quad (5)$$

with, for $\mathbb{1}_k^{v_j} = \mathbb{1}\left\{\left(\tilde{T}_k - \tilde{Z}_k\right) \in v_j\right\}$,

$$Q_1^k = \sum_{j=1}^{J} \mathbb{1}_k^{v_j} \times \sum_{l \in S_k} \delta_{kl} \log\left(f_{kl,j}^{\text{PC}}(X)\right),$$
$$Q_2^k = \left(1 - \sum_{l \in S_k} \delta_{kl}\right) \times \log\left(1 - F_k^{\text{PC}}\left(\tilde{T}_k - \tilde{Z}_k | X\right)\right).$$

See Appendix C for a formal proof.

3.2.2. Architecture

The architecture of MSnet, presented in Appendix B - Figure 2, is composed of a set of fully-connected neural networks: a first subnetwork shared by all transitions belonging to S, then, as many specific subnetworks as there are transitions, and finally K state-specific output layers.

The input layer of MSnet is composed of the covariate matrix X. We note $\mathbf{z}$ the output of the shared-subnetwork, taking as input the matrix X, and defined as

$$\mathbf{z} = g^{\text{input}}(X)$$

with g^{input} a non-linear activation function. We define

$$\mathbf{y}_{kl} = g^{kl}(\mathbf{z})$$

the output of the subnetwork associated with the transition $k \to l$, taking as input $\mathbf{z}$, with g^{kl} a non-linear activation function. We define ϕ_{kl}, a transformation of the output of the specific subnetwork associated with the transition $k \to l$, taking as input $\mathbf{y}_{kl}$, and defined as

$$\phi_{kl} = g^{\text{linear}}(\mathbf{y}_{kl})$$

with g^{linear} a linear activation function defined such that

$$\begin{array}{rccc} g^{\text{linear}} & : & \mathbb{R}^{n \times L^{kl}} & \to & \mathbb{R}^{n \times (J+1)} \\ & & \mathbf{y}_{kl} & \mapsto & g^{\text{linear}}(\mathbf{y}_{kl}) = \phi_{kl} \end{array}$$

with L^{kl} the number of neurons of the last hidden layer of the subnetwork associated with transition $k \to l$.

The output layer k ($k = 0, \cdots, K-1$) takes as input the set $\{\phi_{kl}\}_{l \in S_k}$ and outputs $\mathbf{f}_k$, the piecewise-constant density functions associated with the transitions from state k,

$$\mathbf{f}_k = \left(\{\mathbf{f}_{kl}\}_{l \in S_k}\right)^T = \sigma_k\left(\left(\{\phi_{kl}\}_{l \in S_k}\right)^T\right). \quad (6)$$

where $\mathbf{f}_{kl} = \left\{f_{kl,j}^{\text{PC}}(X)\right\}_{j=1,\cdots J+1} \in \mathbb{R}^{n \times (J+1)}$. σ_k is a weighted *softmax* function applied on the set $\{\phi_{kl}\}_{l \in S_k}$ such that:

$$f_{kl,j}^{\text{PC}}(X) = \frac{\exp\left[\phi_{kl}^j(X)\right]}{\sum_{j=1}^{J+1}\left(\sum_{l \in S_k} \exp\left[\phi_{kl}^j(X)\right]\right) \times |v_j|},$$

with ϕ_{kl}^j the column j of the matrix ϕ_{kl} defined previously.

In the equations defined above, we define an additional time interval $v_{J+1} = [\tau, +\infty)$. In fact, in clinical settings, a patient could leave a state after the interval v_J (the horizon window τ), with the consequence that:

$$\sum_{l \in S_k} \int_0^{\tau} f_{kl}(t|X) = \sum_{l \in S_k} \sum_{j=1}^{J} f_{kl,j}^{\text{PC}}(X) \times |v_j| < 1.$$

We can however write

$$\sum_{l \in S_k} \sum_{j=1}^{J} f_{kl,j}^{\text{PC}}(X) \times |v_j| + 1 - F_k^{\text{PC}}(\tau|X) = 1.$$

The trick is then to consider an additional time interval v_{J+1} verifying $1 - F_k^{\text{PC}}(\tau|X) = \sum_{l \in S_k} f_{kl,J+1}^{\text{PC}}(X)$. See Kvamme and Borgan (2021); Cottin et al. (2022) for similar remarks.

3.2.3. Interval cutpoints

We considered two strategies for placing the interval cut-points in the output layer. The simplest option is to use J equally spaced cutpoints, yielding uniform intervals of identical length. Alternatively, we can let the data dictate the cutpoints by exploiting the empirical distribution of transition times from each state. In this case we estimate the J quantiles of the marginal sojourn-time distribution for a given state using the non-parametric Aalen–Johansen estimator (Aalen and Johansen, 1978), which results in state-specific intervals.

3.2.4. Loss function

The loss function of MSnet is defined as

$$\ell_{\text{MSnet}} = -\frac{1}{n}\log \mathcal{L}_{J+1}^{\text{MSnet}} + P_\lambda$$

where $\log \mathcal{L}_{J+1}^{\text{MSnet}}$ is the generalized log-likelihood defined in Equation (5) evaluated on $J+1$ time intervals.

P_λ is a penalty term inspired by the fused-LASSO penalisation (Tibshirani et al., 2005) and aims at stabilising the output-layer parameters when the number of intervals J is misspecified (too large J leading to over-fitting). For a given transition $k \to l$, the output layer contains s^{kl} neurons; let $w_{s,j}^{kl}$ and b_j^{kl} denote respectively the weight and bias associated with node s and interval v_j $(j = 1, \dots, J)$. We penalise the first-order differences between consecutive intervals:

$$\Delta_{w_{s,j}^{kl}} = w_{s,j+1}^{kl} - w_{s,j}^{kl}, \Delta_{b_l^{kl}} = b_{j+1}^{kl} - b_j^{ql}.$$

The fused-Lasso penalty therefore reads

$$P_\lambda = \sum_{(k,l)\in S} \left(\lambda_w^{kl} \sum_{s=1}^{s^{kl}} \sum_{j=1}^{J} \left| \Delta_{w_{s,j}^{kl}} \right| + \lambda_b^{kl} \sum_{j=1}^{J} \left| \Delta_{b_j^{kl}} \right| \right)$$

with $\lambda_w^{kl} > 0$ and $\lambda_b^{kl} > 0$ transition-specific regularisation constants. When $\lambda_w^{kl} \to \infty$ (resp. $\lambda_b^{kl} \to \infty$) all weight (resp. bias) differences are forced to zero, yielding piecewise-constant parameters across the time intervals. This encourages smooth, parsimonious densities while still allowing the model to adapt to abrupt changes when supported by the data.

3.2.5. Predictions

From the output of MSnet (i.e. the step functions $f_{kl,j}^{\text{PC}}(.)$ for $(k,l) \in S$ and $j = 1, \cdots, J+1$), the CIF for a transition $k \to l$, noted $\hat{F}_{kl}^{\text{PC}}(t|X_i)$, $t \in \left]0, \tau\right]$, is computed following Equation (4). The cumulative distribution function related to state k, noted $\hat{F}_k^{\text{PC}}(t|X_i)$, is computed following Equation (3). For computing the cumulative transition intensities defined in Equation (1), see the equivalence formula given in Appendix D.

4. Experiments

4.1. Data

We conducted experiments on two real-world datasets and on a set of simulated datasets. .

4.1.1. Real-world datasets

Rotterdam (Royston and Altman, 2013) This dataset includes 2,982 patients with primary breast cancer. We modeled disease progression using an illness–death process (initial state, relapse, death). The dataset comprises clinical, histopathological, and molecular covariates.

METABRIC (Pereira et al., 2016) We selected 1,903 breast cancer patients and modeled disease progression using the four-state (initial state, local relapse, distant relapse, death) process illustrated in Figure 4. The dataset contains clinical, histopathological, molecular, and gene expression features. Following a univariate gene-screening approach inspired by Cottin et al. (2022), we selected 1,419 cancer-related genes. Two models were then constructed: M_0 : incorporating clinical, histopathological, and molecular features; M_{1419}: extending M_0 with the 1,419 selected genes.

Further details on these datasets are provided in Appendix F.

4.1.2. Simulated datasets

We generated the five following semi-Markov simulation datasets to evaluate the proposed methodology: (S1) a *non-linear survival* model (see illustration in Figure 1.a) ; (S2) a *non-linear four-state* model (see illustration in Figure 1.d) ; (S3) a *high-dimensional non-linear four-state* model ; (S4) a *linear four-state* model ; (S5) a *non-linear non-P.H illness–death* model (see illustration in Figure 1.c). For each dataset, we simulated $n = 5000$ observations and we consider each reversible transition of the process.

See details on the simulations in Appendix E.

4.2. Evaluation setting

Metrics To evaluate the predictive performance and benchmark MSnet, we considered time-dependent discrimination (AUC, Uno et al. (2007)) and calibration (Brier score, BS, Gerds and Schumacher (2006)) for each transition at time $t \in [0, \tau]$. We computed these metrics using inverse probability of censoring weighted (IPCW) estimators (Blanche et al., 2013) (see Appendix G for the exact definitions). Time-dependent measures were summarized using the integrated AUC (iAUC, higher the better) and integrated Brier

score (iBS, lower the better) over a fixed time horizon.

Evaluation was performed via cross-validation, with Monte Carlo median-based estimators obtained by randomly splitting each dataset $M = 20$ times (see Appendix H and Figure 6).

Competing methods We compared MSnet against the following multi-state models: msCox (De Wreede et al., 2010) (a multi-state Cox P.H. model implemented in the R library *mstate*) and SurvNODE Groha et al. (2020) (a neural network approach assuming a Markov property). For both methods, CIFs were computed from the hazard functions using the equivalence formula in Appendix D - Equation (8).

For simulation (S1) (*non-linear survival*), we additionally evaluated DeepHit and Random Survival Forests (RSF).

Hyperparameters for each method were optimized via random search (Bergstra and Bengio, 2012) (see Appendix I for details).

Clinical interpretability We performed a clinical interpretation of MSnet on the M_0 model trained on the METABRIC cohort. To quantify feature contributions in this deep-learning multi-state setting, we applied the interpretability algorithm MS-CPFI introduced by Cottin et al. (2024). The resulting importance scores were then examined in the context of the breast-cancer literature.

Clinical decision support To illustrate how the predicted transition cumulative incidence functions of MSnet could support downstream clinical decisions, we derived risk groups based on the predicted probability of the event at a clinically relevant time horizon. Patients were stratified into three categories (low-, intermediate-, and high-risk) according to predefined thresholds of the predicted risk. Kaplan–Meier curves were then estimated for each risk group to assess the separation in observed outcomes. This analysis was performed specifically on the "relapse" transition (i.e., $0 \to 1$) within the Rotterdam cohort.

4.3. Results

In this section, we give results on iAUC. Results on iBS are given in Appendix J (Tables 9 and 10). Performance of MSnet against the other approaches were statistically compared using paired bilateral Wilcoxon signed rank tests: $\cdot$ indicates a p-value less than 0.1, † less than 0.05, ‡ less than 0.01, * less than 0.001. Bold values denotes the best algorithm. In each table, we display the number of (No.) observed events for each transition.

4.3.1. Real-world datasets

Predictive performance For the Rotterdam cohort, results of iAUC are given in Table 1(*a*). MSnet significantly outperforms msCox and SurvNODE for transitions $0 \to 1$ and $1 \to 2$. For transition $0 \to 2$, msCox outperforms MSnet. In this particular case, this is explained by the fact that very few events are observed for this transition.

Table 1(*b*) reports the iAUC values for the METABRIC cohort. In the low-dimensional model (M_0) MSnet attains the highest iAUC for the early transitions $0 \to 1$, $0 \to 2$ and $1 \to 3$outperforming both msCox and SurvNODE. SurvNODE is best for transition $0 \to 3$ and msCox for $2 \to 3$.

When the high-dimensional gene set is added (M_{1419}), msCox fails to converge, so we replace it with transition-specific Lasso-Cox models (tsLassoCox). In this setting MSnet markedly improves for early transitions: $0 \to 1$ (iAUC=0.767) and $0 \to 2$ (iAUC=0.722), significantly surpassing all baselines and demonstrating its capacity to embed information from thousands of genes for predicting local and distant relapses. SurvNODE again dominates $0 \to 3$, but its overall performance declines compared with M_0, indicating poor scalability to high-dimensional data. For $0 \to 3$ and $2 \to 3$, tsLassoCox attains the highest iAUC (fitted separately per transition because current implementations do not support multi-state analysis; see the resulting limitations in Schuster et al. (2020)). The weaker performance of MSnet on the death-related transitions likely stems from using baseline prognostic factors that change after relapse; incorporating updated clinical and treatment information should further improve prediction for these states.

Clinical interpretation Graphical interpretation of model M_0 for each transition of the four-state process in the METABRIC cohort is given in supplemental figures 8 to 12. To assess

Table 1: Real-world datasets: Benchmark of iAUCs (median ± sd).

(a) Rotterdam dataset.

Alg.	**Transitions** (No. observed events)		
	0 → 1 (1518)	**0 → 2** (195)	**1 → 2** (1077)
msCox	$0.695 \pm 0.020^{\dagger}$	$\mathbf{0.865 \pm 0.024}^{*}$	$0.630 \pm 0.048^{\dagger}$
SurvNODE	$0.573 \pm 0.021^{*}$	$0.474 \pm 0.088^{*}$	$0.631 \pm 0.100^{\cdot}$
MSnet	$\mathbf{0.710 \pm 0.013}$	0.788 ± 0.041	$\mathbf{0.646 \pm 0.044}$

(b) METABRIC dataset - Models M_0 and M_{1419}.

Mod.	**Alg.**	**Transitions** (No. observed events)				
		0 → 1 (272)	**0 → 2** (403)	**0 → 3** (507)	**1 → 3** (212)	**2 → 3** (383)
M_0	msCox	$0.686 \pm 0.051^{\dagger}$	0.680 ± 0.063	$0.700 \pm 0.071^{\dagger}$	0.590 ± 0.064	$\mathbf{0.605 \pm 0.033}$
	SurvNODE	$0.567 \pm 0.037^{*}$	$0.493 \pm 0.028^{*}$	$\mathbf{0.734 \pm 0.035}^{*}$	$0.526 \pm 0.040^{*}$	$0.502 \pm 0.016^{*}$
	MSnet	$\mathbf{0.722 \pm 0.055}$	$\mathbf{0.699 \pm 0.042}$	0.664 ± 0.050	$\mathbf{0.599 \pm 0.087}$	0.582 ± 0.035
M_{1419}	tsLassoCox	$0.620 \pm 0.324^{*}$	0.712 ± 0.045	$0.672 \pm 0.103^{*}$	$\mathbf{0.650 \pm 0.075}^{\ddagger}$	$\mathbf{0.618 \pm 0.050}^{*}$
	SurvNODE	$0.584 \pm 0.026^{*}$	$0.517 \pm 0.043^{*}$	$\mathbf{0.687 \pm 0.049}^{*}$	$0.520 \pm 0.091^{*}$	$0.465 \pm 0.046^{*}$
	MSnet	$\mathbf{0.767 \pm 0.049}$	$\mathbf{0.722 \pm 0.044}$	0.557 ± 0.113	0.592 ± 0.085	0.576 ± 0.046

the clinical relevance of the identified patterns we compared them with the established breast-cancer literature. Main observations on the top 3 features are given on the following paragraph.

- **Cancer grade.** Grade is a histological measure of proliferation: grade 1 tumours grow slowly and have a low metastatic potential, whereas grade 3 tumours proliferate rapidly and are associated with a high likelihood of metastasis (Rakha et al., 2010). MSnet identifies grade 3 as a significant risk factor for the relapsing transitions and for death after relapse (01, 02, 13, 23), while grade 1 appears protective—exactly the pattern reported in the literature.
- **Tumour size (T).** The AJCC staging system states that $T1$ tumours ($\leq$2cm) are associated with a favourable prognosis, whereas $T2$ tumours (2-5cm) confer a higher risk of recurrence and death; $T3$ tumours (>5cm) are even more adverse but are rare in our data set (Giuliano et al., 2017). MSnet reproduces this hierarchy: $T1$ is protective, $T2$ is a risk factor for all transitions, and $T3$ shows no statistically significant effect, likely because of its low frequency.
- **Hormone-receptor status.** ER (estrogen receptor)-positive and/or PR (progesterone receptor)-positive tumours tend to grow more slowly and have a better short-term prognosis, whereas ER-negative/PR-negative cancers relapse early and are associated with poorer outcomes (Allison et al., 2020). MSnet flags ER^+/PR^+ as a favourable factor and ER^-/PR^- as an adverse factor across all five transitions, in perfect agreement with the clinical evidence.

Full details on all the features are given in Appendix K. In every case, the MSnet-derived importance profiles are in strong agreement with well-established breast-cancer risk factors, demonstrating that the deep-learning architecture can faithfully capture clinically meaningful signals while providing transition-specific risk predictions.

Clinical decision support Kaplan Meier curves of relapse-free survival (RFS) per risk profile are given in Appendix L - Figure 7. Risk profiles clearly separate a high-risk population whose relapse-free curve lies significantly below the curves of the two other groups. Such stratification could support a simple clinical decision policy in which patients classified as high risk would be considered for intensified monitoring or more aggressive management, while low-risk patients would remain under standard care.

4.3.2. Simulated datasets

Predictive performance The simulation study consistently shows that MSnet attains the highest iAUC across a variety of settings (see Table 2 for a summary and Table 3 for details). In the scenario (S1) (*non-linear survival*), MSnet markedly outperforms every baseline, while

Table 2: Simulations: iAUCs averaged over the transitions. The weighted average (based on the number of events observed on each transition) is given in parentheses.

Alg.	**Simulations**				
	(S1)	(S2)	(S3)	(S4)	(S5)
msCox	0.502 (0.502)	0.508 (0.510)	0.522 (0.514)	**0.675** (**0.689**)	0.724 (0.713)
SurvNODE	0.537 (0.537)	0.487 (0.488)	NC	0.564 (0.570)	NC
RSF	0.745 (0.745)	NA	NA	NA	NA
DeepHit	0.717 (0.717)	NA	NA	NA	NA
MSnet	**0.766** (**0.766**)	**0.654** (**0.653**)	**0.708** (**0.699**)	0.668 (0.666)	**0.731**(**0.728**)

NC: Non convergent ; NA: Non applicable.

Table 3: Simulations: Benchmark of iAUCs (median ± sd).

(a) Simulation (S1) - *non-linear survival*.

Alg.	**Transition** (No. observed events)
	0 → 1 (3500)
msCox	$0.502 \pm 0.020*$
SurvNODE	$0.537 \pm 0.011^{*}$
RSF	$0.745 \pm 0.012^{\dagger}$
DeepHit	$0.717 \pm 0.008^{*}$
MSnet	$\mathbf{0.766 \pm 0.020}$

(b) Simulation (S2) - *non-linear four-state*.

Alg.	**Transitions** (No. observed events)					
	0 → 1 (1482)	**0 → 2** (1575)	**0 → 3** (1751)	**1 → 2** (677)	**1 → 3** (722)	**2 → 3** (1075)
msCox	$0.492 \pm 0.020^{*}$	$0.556 \pm 0.020^{*}$	$0.485 \pm 0.010^{*}$	$0.472 \pm 0.030^{*}$	$0.513 \pm 0.030^{*}$	$0.528 \pm 0.030^{*}$
SurvNODE	$0.395 \pm 0.011^{*}$	$0.547 \pm 0.021^{*}$	$0.503 \pm 0.019^{*}$	$0.444 \pm 0.033^{*}$	$0.517 \pm 0.022^{*}$	$0.514 \pm 0.018^{*}$
MSnet	$\mathbf{0.723 \pm 0.031}$	$\mathbf{0.660 \pm 0.017}$	$\mathbf{0.561 \pm 0.032}$	$\mathbf{0.648 \pm 0.053}$	$\mathbf{0.580 \pm 0.046}$	$\mathbf{0.751 \pm 0.026}$

(c) Simulation (S3) - *high-dimensional non-linear four-state*.

Alg.	**Transitions** (No. observed events)					
	0 → 1 (1541)	**0 → 2** (1545)	**0 → 3** (1732)	**1 → 2** (753)	**1 → 3** (710)	**2 → 3** (1127)
msCox	$0.506 \pm 0.070^{*}$	0.525 ± 0.011	$0.474 \pm 0.020^{*}$	0.601 ± 0.130	$0.512 \pm 0.030^{*}$	$0.514 \pm 0.040^{*}$
SurvNODE	NC	NC	NC	NC	NC	NC
MSnet	$\mathbf{0.802 \pm 0.026}$	$\mathbf{0.530 \pm 0.027}$	$\mathbf{0.645 \pm 0.025}$	$\mathbf{0.636 \pm 0.026}$	$\mathbf{0.749 \pm 0.106}$	$\mathbf{0.883 \pm 0.083}$

NC: Non convergent

(d) Simulation (S5) - *non-linear non-P.H. four-state*.

Alg.	**Transitions** (No. observed events)		
	0 → 1 (1895)	**0 → 2** (2297)	**1 → 2** (1269)
msCox	$0.752 \pm 0.010^{\dagger}$	$0.649 \pm 0.010^{\ddagger}$	$\mathbf{0.771 \pm 0.010}^{*}$
SurvNODE	NC	NC	NC
MSnet	$\mathbf{0.765 \pm 0.010}$	$\mathbf{0.696 \pm 0.034}$	0.732 ± 0.082

NC: Non convergent

msCox—unsuitable for non-linear effects—lags far behind and RSF/DeepHit trail only slightly. SurvNODE, in contrast, exhibits only modest predictive ability. To validate the robustness of MSnet under varying censoring rates, we conducted additional experiments on scenario (S1) by introducing higher variation in the censoring rate. The results confirm that MSnet maintains stable and competitive performance under higher censoring rate (see Table 11). In the simulation (S2) (*non-linear four-state*), MSnet again dominates all competitors for every transition. The advantage persists in the high-dimensional extension (S3), where the number of covariates rises from 12 to 50; MSnet remains stable, whereas SurvNODE suffers convergence instability (discussed below). In the setting (S5) (*non-linear non-P.H. illness-death*), MSnet and msCox achieve comparable iAUCs, while SurvNODE encounters severe convergence

problems due to the extreme computational burden (also described later).

Overall, these results demonstrate that MSnet is robust to non-linearity and moderate dimensionality, and only loses its edge when the underlying risk structure is genuinely linear (see additional results on simulation (S4) in supplemental Table 12).

Computational performance Running times and the hardware resources required for each simulation scenario are reported in Table 4.

In simulation (S1) (*non-linear survival*), all methods finish in a comparable time except SurvNODE, which needs $>$ 30 min, whereas MSnet and the other baselines run in $<$ 2 min.

In simulation (S2) (*non-linear four-state*), MSnet uses more RAM/CPU than msCox but still completes in a few minutes. The longer runtime of msCox is mainly due to the computation of cumulative hazard and cumulative-incidence functions in the R package *mstate*. SurvNODE also exhibits a high runtime here.

In simulation (S5) (*non-linear non-P.H. illness-death*), SurvNODE aborts after the first epoch because GPU memory quickly exceeds 48 GB. This is a generic limitation of Neural-ODE–based survival models (e.g., Gholami et al. (2019), Kim et al. (2021)) that deliberately relax the P.H. assumption by learning fully time-dependent hazard dynamics. That flexibility, however, introduces numerical stiffness, which leads to a sharp increase in computational cost and frequent solver failures (see also Rindt et al. (2022)), making benchmarking impractical in highly non-linear, high-dimensional and non-P.H. settings.

Overall, MSnet offers a favourable trade-off between predictive accuracy and computational efficiency: it runs in minutes on a standard CPU while achieving state-of-the-art performance. Full experimental details are provided in Appendix J.3.

5. Discussion

We introduced MSnet, a deep-learning architecture that models any progressive multi-state process under the semi-Markov assumption. Building on the framework of Cottin et al. (2022), MSnet employs transition-specific subnetworks and state-specific fully-connected output layers to estimate the density of transition times from individual covariates.

To the best of our knowledge, MSnet is the first DL-based method that simultaneously (i) accommodates semi-Markov transition dynamics, (ii) yields individualized predictions, and (iii) integrates multimodal data (clinical and transcriptomic) for general progressive multi-state processes. Experiments on two breast-cancer cohorts demonstrate that MSnet achieves competitive iAUC scores for a four-state semi-Markov model, while simulation studies show consistent superiority over existing multi-state survival models (msCox, SurvNODE) especially in non-linear settings. Moreover, MSnet is substantially faster than neural-ODE approaches such as SurvNODE, whose theoretical flexibility is offset by severe practical limitations (instability, prohibitive runtimes) that impede reliable benchmarking in complex healthcare scenarios.

In summary, MSnet provides a flexible and efficient interpretable approach for semi-Markov multi-state analysis and can serve as a foundation for extending deep learning to a broad range of health-care applications.

Limitations Interpretability, a traditional barrier to clinical adoption of deep models (Farah et al., 2023), is addressed by recent explainability techniques (Cottin et al., 2024). We refer the reader to the Appendix K for a demonstration on the METABRIC cohort. While this demonstration illustrates how well our model's predictions align with established clinical knowledge, it was not designed to establish causal relationships. Causal interpretation—beyond mere interpretation—remains an open challenge for future extensions.

Like all deep learning methods, MSnet requires sizable training data to surpass classical statistical approaches (Rajula et al., 2020). This limitation is evident for transitions with few at-risk patients and for late transitions where only baseline covariates are available. Incorporating updated or time-varying features is expected to further improve performance. We see two viable pathways to address this limitation: (i) Dynamic data updating: The most immediate solution involves updating the input data following a state transition. In this scenario, once a pa-

Table 4: Running times and required computational resources for the simulation scenarios for one run (hyper-parameters fixed), for simulations (S1) *non-linear survival*, (S2) *non-linear four-state*, (S5) *non-linear non-P.H. illness-death*.

Sim.	**Model**	**Time** (h:s:m)	**CPU** (%)	**GPU usage** (GB)	**RAM** (GB)
(S1)	msCox	00:01:08	99	/	0.36
	SurvNODE	00:34:22	100	0.40	1.75
	RSF	00:01:38	4400	/	10.00
	DeepHit	00:02:18	4200	/	0.74
	MSnet	00:01:56	4300	/	1.63
(S2)	msCox	00:55:26	99	/	0.44
	SurvNODE	01:16:59	100	0.90	1.90
	MSnet	00:07:09	4300	/	1.75
(S5)	msCox	00:36:31	99	/	0.57
	SurvNODE	NC	NC	NC	NC
	MSnet	00:11:05	4400	/	1.74

NC: Non convergent

tient exits a state, the updated covariates are re-entered into the network. (ii) Integration of longitudinal data: A more comprehensive solution would involve integrating full longitudinal data streams. This would necessitate adapting the covariates-shared network to handle temporal dependencies explicitly, e.g., recurrent architectures such as LSTM (Lee et al., 2019), transformer-based mechanisms (Zhang et al., 2025), or neural controlled differential equations (Bleistein et al., 2024). Additionally, the output layer would require adaptation to support dynamic predictions. We plan to explore these extensions in our future work.

Another key limitation of this work is scalability to large multi-state settings. While our architecture is flexible, its performance—as with most existing multi-state models—is fundamentally constrained by the number of states and transitions and, more critically, by the sparsity of observed events per transition. In real-world clinical datasets, this sparsity is exacerbated by censoring, competing risks, and terminal events, which substantially reduce the effective sample size available for estimating transition-specific risks. As the number of states increases (e.g., 6-10+), the number of transitions grows combinatorially, making these issues particularly acute. In our simulation study, we initially experimented a challenging scenario with 8 states/56 transitions (see additional Table 13). MSnet exhibits degraded performance (iAUC≈0.49–0.54), while msCox and SurvNode fail to converge, likely due to severe over-parameterization relative to the number of observed events per transition. This highlights a general limitation of current multi-state approaches rather than a weakness specific to our architecture. Addressing scalability in such regimes will likely require a combination of increased data availability (e.g., larger cohorts or synthetic augmentation) and architectural adaptations. In this regard, two directions appear particularly promising: (i) Weight sharing across transitions with similar hazard structures (e.g., competing or biologically related transitions), which could substantially reduce the number of parameters; and (ii) flexible transition-specific architectures, where the capacity allocated to each transition is adapted to its level of support in the data, allowing sparse transitions to rely primarily on shared representations and thereby limiting over-fitting (this is already supported in our implementation by allowing transition-specific subnetworks to be removed, e.g., by setting their hidden layers to zero neurons).

References

Odd O Aalen and Søren Johansen. An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian journal of statistics*, pages 141–150, 1978.

Massil Achab, Agathe Guilloux, Stéphane Gaïffas, and Emmanuel Bacry. Sgd with variance reduction beyond empirical risk minimization. *arXiv preprint arXiv:1510.04822*, 2015.

Kimberly H Allison, M Elizabeth H Hammond, Mitchell Dowsett, Shannon E McKernin, Lisa A Carey, Patrick L Fitzgibbons, Daniel F Hayes, Sunil R Lakhani, Mariana Chavez-MacGregor, Jane Perlmutter, et al. Estrogen and progesterone receptor testing in breast cancer: Asco/cap guideline update. *Journal of Clinical Oncology*, 38(12):1346–1366, 2020.

Per Kragh Andersen and Maja Pohar Perme. Inference for outcome probabilities in multistate models. *Lifetime data analysis*, 14(4): 405, 2008.

Per Kragh Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723, 2005.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1): 281–305, 2012.

Jan Beyersmann, Aurelien Latouche, Anika Buchholz, and Martin Schumacher. Simulating competing risks data in survival analysis. *Statistics in medicine*, 28(6):956–971, 2009.

Elia M Biganzoli, Patrizia Boracchi, Federico Ambrogi, and Ettore Marubini. Artificial neural network for the joint modelling of discrete cause-specific hazards. *Artificial intelligence in medicine*, 37(2):119–130, 2006.

Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine*, 32(30):5381–5397, 2013.

Linus Bleistein, Van-Tuan Nguyen, Adeline Fermanian, and Agathe Guilloux. Dynamical survival analysis with controlled latent states. *arXiv preprint arXiv:2401.17077*, 2024.

Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. 2007.

Aziliz Cottin, Nicolas Pécuchet, Marine Zulian, Agathe Guilloux, and Sandrine Katsahian. Idnetwork: A deep illness-death network based on multi-state event history process for disease prognostication. *Statistics in Medicine*, 41(9): 1573–1598, 2022.

Aziliz Cottin, Marine Zulian, Nicolas Pécuchet, Agathe Guilloux, and Sandrine Katsahian. Mscpfi: A model-agnostic counterfactual perturbation feature importance algorithm for interpreting black-box multi-state models. *Artificial Intelligence in Medicine*, 147:102741, 2024.

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

Ori Davidov. The steady-state probabilities for regenerative semi-markov processes with application to prevention and screening. *Applied stochastic models and data analysis*, 15(1):55–63, 1999.

Liesbeth C De Wreede, Marta Fiocco, and Hein Putter. The mstate package for estimation and prediction in non-and semi-parametric multistate and competing risks models. *Computer methods and programs in biomedicine*, 99(3): 261–274, 2010.

Christine Eulenburg, Sven Mahner, Linn Woelber, and Karl Wegscheider. A systematic model specification procedure for an illness-death model without recovery. *PloS one*, 10 (4):e0123489, 2015.

Line Farah, Juliette M Murris, Isabelle Borget, Agathe Guilloux, Nicolas M Martelli, and Sandrine IM Katsahian. Assessment of performance, interpretability, and explainability in artificial intelligence–based health technologies: what healthcare stakeholders need to know. *Mayo Clinic Proceedings: Digital Health*, 1(2):120–138, 2023.

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

Amir Gholami, Kurt Keutzer, and George Biros. Anode: Unconditionally accurate memory-efficient gradients for neural odes. *arXiv preprint arXiv:1902.10298*, 2019.

Armando E Giuliano, James L Connolly, Stephen B Edge, Elizabeth A Mittendorf, Hope S Rugo, Lawrence J Solin, Donald L Weaver, David J Winchester, and Gabriel N Hortobagyi. Breast cancer—major changes in the american joint committee on cancer eighth edition cancer staging manual. *CA: a cancer journal for clinicians*, 67(4):290–303, 2017.

William J Gradishar, BO Anderson, J Abraham, Rebecca Aft, Doreen Agnese, KH Allison, SL Blair, HJ Burstein, C Dang, A Elias, et al. Nccn clinical practice guidelines in oncology: breast cancer. In *National Comprehensive Cancer NetworkNCCN: Plymouth Meeting, PA, USA*, 2020.

Stefan Groha, Sebastian M Schmon, and Alexander Gusev. A general framework for survival analysis and multi-state modelling. *arXiv preprint arXiv:2006.04893*, 2020.

Philip Hougaard. Multi-state models: a review. *Lifetime data analysis*, 5:239–264, 1999.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

Jean Jacod and Jean Memin. Caractéristiques locales et conditions de continuité absolue pour les semi-martingales. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 35:1–37, 1976.

Faisal M Khan and Valentina Bayer Zubek. Support vector regression for censored data (svrc): a novel tool for survival analysis. In *2008 Eighth IEEE International Conference on Data Mining*, pages 863–868. IEEE, 2008.

Haesook Teresa Kim. Competing risks data in clinical oncology. *Frontiers in Oncology*, 14: 1360266, 2024.

Suyong Kim, Weiqi Ji, Sili Deng, Yingbo Ma, and Christopher Rackauckas. Stiff neural ordinary differential equations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31 (9), 2021.

John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime data analysis*, 27: 710–736, 2021.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

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

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

Jaafar Makki. Diversity of breast carcinoma: histological subtypes and clinical relevance. *Clinical medicine insights: Pathology*, 8:CPath–S31563, 2015.

Andrew McGuire, James AL Brown, Carmel Malone, Ray McLaughlin, and Michael J Kerin. Effects of age on the detection and management of breast cancer. *Cancers*, 7(2):908–929, 2015.

Luís Meira-Machado and Marta Sestelo. Estimation in the progressive illness-death model: A nonexhaustive review. *Biometrical Journal*, 61 (2):245–263, 2019.

Karla Monterrubio-Gómez, Nathan Constantine-Cooke, and Catalina A Vallejos. A review on statistical and machine learning competing risks methods. *Biometrical Journal*, 66(2): 2300060, 2024.

Juliette Pénichoux, Thierry Moreau, and Aurélien Latouche. Simulating recurrent events that mimic actual data: a review of the literature with emphasis on event-dependence. *arXiv preprint arXiv:1503.05798*, 2015.

Bernard Pereira, Suet-Feung Chin, Oscar M Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature communications*, 7(1):11479, 2016.

Hein Putter, Jos van der Hage, Geertruida H de Bock, Rachid Elgalta, and Cornelis JH van de Velde. Estimation and prediction in a multi-state model for breast cancer. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 48(3):366–380, 2006.

Jingkun Qu, Chaofan Li, Mengjie Liu, Yusheng Wang, Zeyao Feng, Jia Li, Weiwei Wang, Fei Wu, Shuqun Zhang, and Xixi Zhao. Prognostic models using machine learning algorithms and treatment outcomes of occult breast cancer patients. *Journal of Clinical Medicine*, 12 (9):3097, 2023.

Md Mahmudur Rahman and Sanjay Purushotham. Multi-state survival analysis using pseudo value-based deep neural networks. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 757–765. SIAM, 2023.

Hema Sekhar Reddy Rajula, Giuseppe Verlato, Mirko Manchia, Nadia Antonucci, and Vassilios Fanos. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, 56(9):455, 2020.

Emad A Rakha, Jorge S Reis-Filho, Frederick Baehner, David J Dabbs, Thomas Decker, Vincenzo Eusebi, Stephen B Fox, Shu Ichihara, Jocelyne Jacquemier, Sunil R Lakhani, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast cancer research*, 12(4):207, 2010.

David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival regression with proper scoring rules and monotonic neural networks. In *International conference on artificial intelligence and statistics*, pages 1190–1205. PMLR, 2022.

Patrick Royston and Douglas G Altman. External validation of a cox prognostic model: principles and methods. *BMC medical research methodology*, 13(1):33, 2013.

Philippe Saint Pierre. Analyse de survie, modèles multi-états et processus de comptage. 2021.

Noah A Schuster, Emiel O Hoogendijk, Almar AL Kok, Jos WR Twisk, and Martijn W Heymans. Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risk analysis. *Journal of clinical epidemiology*, 122:42–48, 2020.

Victoria Sopik, David Lim, Ping Sun, and Steven A Narod. Prognosis after local recurrence in patients with early-stage breast cancer treated without chemotherapy. *Current Oncology*, 30(4):3829–3844, 2023.

Cristian Spitoni, Violette Lammens, and Hein Putter. Prediction errors for state occupation and transition probabilities in multi-state models. *Biometrical Journal*, 60(1):34–48, 2018.

Kyle Swanson, Eric Wu, Angela Zhang, Ash A Alizadeh, and James Zou. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell*, 186(8):1772–1791, 2023.

Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1):91–108, 2005.

Anastasios Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22, 1975.

Hajime Uno, Tianxi Cai, Lu Tian, and Lee-Jen Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, 2007.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3):65, 2024.

Bo Zhang, Huiping Shi, and Hongtao Wang. Machine learning and ai in cancer prognosis, prediction, and treatment selection: a critical approach. *Journal of multidisciplinary healthcare*, pages 1779–1791, 2023a.

Heng Zhang, Qianyi Xi, Fan Zhang, Qixuan Li, Zhuqing Jiao, and Xinye Ni. Application of deep learning in cancer prognosis prediction model. *Technology in Cancer Research & Treatment*, 22:15330338231199287, 2023b.

Zhiyue Zhang, Yao Zhao, and Yanxun Xu. Transformerlsr: Attentive joint model of longitudinal data, survival, and recurrent events with concurrent latent structure. *Artificial intelligence in medicine*, 160:103056, 2025.

Zhi Zhao, John Zobolas, Manuela Zucknick, and Tero Aittokallio. Tutorial on survival modeling with applications to omics data. *Bioinformatics*, 40(3):btae132, 2024.

Appendix A. Example of multi-state processes

See Figure 1 on the next page.

Appendix B. MSnet architecture

See Figure 2 on the next page.

Appendix C. Re-writing of the conventional log-likelihood

In this section, we define our log-likelihood under some of our assumptions (semi-Markov, piecewise constant proposals) and as a function of the quantities of interest defined in Section 3.2.1 for n realizations of the process E. We omit the conditional notation on X for simplicity.

We start from the definition given in Equation (2) in a general Markovian and continuous framework and we transpose this definition into the framework of semi-Markovian and time-homogeneous transition intensities (i.e., for a transition $k \to l$, $k < l, (k,l) \in S$, α_{kl} is a function of $t - Z_k$, with Z_k the entry time in state k). The log-likelihood defined in Equation (2) is then rewritten as: $\log \mathcal{L} = \sum_{i=1}^{n} \sum_{(k,l)\in S} \left[P_1^{i,kl} + P_2^{i,kl}\right]$, with

$$P_1^{i,kl} = \delta_{kl}^i \log\left(\alpha_{kl}(\tilde{T}_k^i - \tilde{Z}_k^i)\right),$$

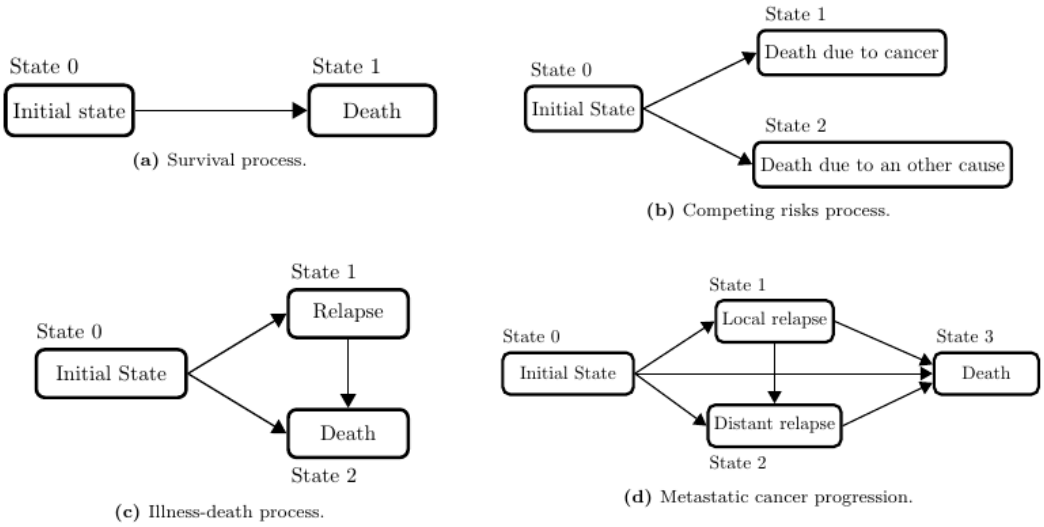
$$P_2^{i,kl} = -\int_0^\tau \alpha_{kl}\left(t - Z_k^i\right) Y_k^i\left(t\right) \mathbb{1}\{t \geq C_i\} dt.$$

Under the semi-Markovian hypothesis and considering that $\mathbb{P}\left(E(0) = 0\right) = 1$ for all i $(1 \leq i \leq n)$, the risk of transition from a state k to a state l necessarily implies the observation of a transition from a previous state to a state k (for $k > 0$) such that:

$$Y_k^i\left(t\right) \mathbb{1}\{t \geq C_i\} = \delta_{Z_k} \mathbb{1}\{\tilde{Z}_k < t \leq \tilde{T}_k\}.$$

In this case, we can write the log-likelihood terms as $\log \mathcal{L} = \sum_{i=1}^{n} \times \sum_{(k,l)\in S} \delta_{Z_k}^i \times \left[P_1^{i,kl} + P_2^{i,kl}\right]$. The term $P_2^{i,kl}$ can then be simplified as:

$$P_2^{i,kl} = -\int_{\tilde{Z}_k^i}^{\tilde{T}_k^i} \alpha_{kl}(t - \tilde{Z}_k^i) dt.$$



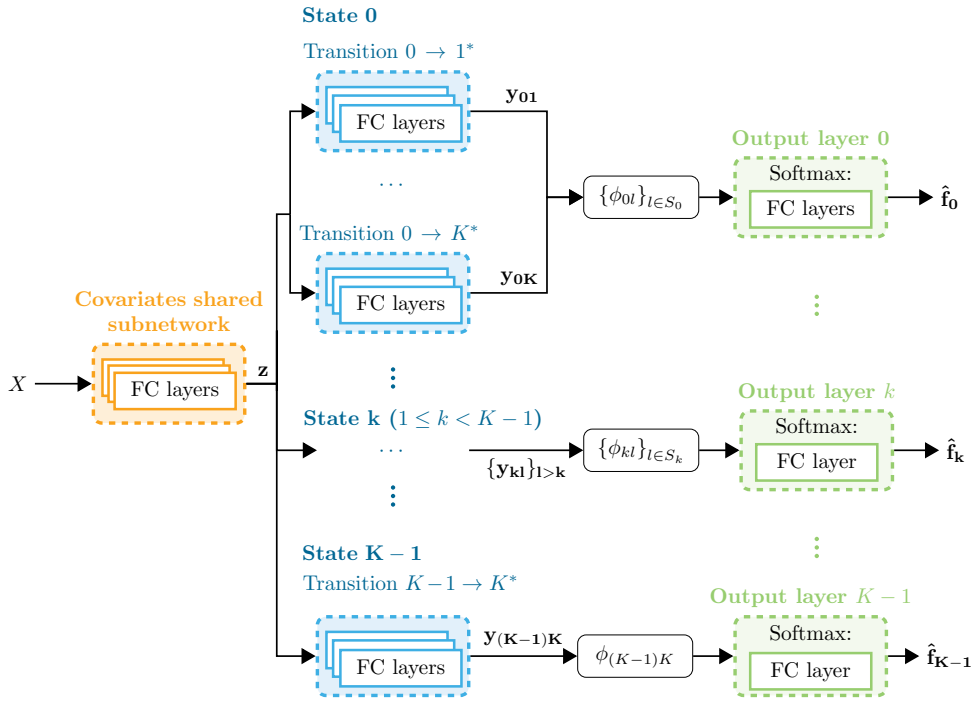
(a) Survival process.

(b) Competing risks process.

(c) Illness-death process.

(d) Metastatic cancer progression.

Figure 1: Examples of multi-state processes for an application in oncology.



*A specific subnetwork for a transition $k \to l$ $(k > l,\ (k,l) \in S)$ exists only if $l \in S_k$ and if k is not an absorbing state. Same remark for an output layer k.

Figure 2: MSnet architecture.

We now define this log-likelihood in terms of our quantities of interest defined in Section 3.2.1. To this end, we use the equivalences between transition intensities and densities (see Appendix D - Equation (8)).

Let's start with the case of continuous densities. Using the equivalences mentioned ear-

lier, we can rewrite the log-likelihood as the sum of three terms: $\log \mathcal{L} = \sum_{i=1}^{n} \times \sum_{(k,l)\in S} \delta_{Z_k}^i \times \left[P_{1.a}^{i,kl} + P_{1.b}^{i,kl} + P_2^{i,kl}\right]$, with

$$P_{1.a}^{i,kl} = \delta_{kl}^i \log\left(f_{kl}(\tilde{T}_k^i - \tilde{Z}_k^i)\right),$$
$$P_{1.b}^{i,kl} = -\delta_{kl}^i \log\left(1 - F_k(\tilde{T}_k^i - \tilde{Z}_k^i)\right). \quad (7)$$

Then, to express this likelihood as a function of the densities, we factor the transition intensities starting from the same state and rewrite the likelihood, for $\mathbb{1}_k = \mathbb{1}\{k \text{ is not an absorbing state}\}$, as follows:

$$\log \mathcal{L} = \sum_{k=0}^{K-1} \mathbb{1}_k \times \sum_{i=1}^{n} \delta_{Z_k}^i \left(Q_1^{i,k} + Q_2^{i,k}\right),$$

with

$$Q_1^{i,k} = \sum_{l\in S_k} \delta_{kl}^i \times \log\left(f_{kl}(\tilde{T}_k^i - \tilde{Z}_k^i)\right),$$
$$Q_2^{i,k} = -\left(\sum_{l\in S_k} \delta_{kl}^i\right) \times \log\left(1 - F_k(\tilde{T}_k^i - \tilde{Z}_k^i)\right) - \int_{\tilde{Z}_k^i}^{\tilde{T}_k^i} \left(\sum_{l\in S_k} \alpha_{kl}(t - \tilde{Z}_k^i)\right) dt$$

We now have the second part of the equation of $Q_2^{i,k}$ which is a function of α_{kl} to rewrite. We know that:

$$\begin{aligned}
&\exp\left[-\int_{\tilde{Z}_k^i}^{t} \left(\sum_{l\in S_k} \alpha_{kl}(s - \tilde{Z}_k^i)\right) ds\right] \\
&= 1 - \int_{\tilde{Z}_k^i}^{t} \left(\sum_{l\in S_k} f_{kl}(s - \tilde{Z}_k^i)\right) ds \\
&= 1 - \int_{\tilde{Z}_k^i}^{t} f_k(s - \tilde{Z}_k^i) ds \\
&= 1 - F_k(t - \tilde{Z}_k^i).
\end{aligned}$$

Following this point and factoring, we can write the term $Q_2^{i,k}$ as follows:

$$Q_2^{i,k} = \left(1 - \sum_{l\in S_k} \delta_{kl}^i\right) \times \log\left(1 - F_k(\tilde{T}_k^i - \tilde{Z}_k^i)\right).$$

Let us now consider the writing of the terms $Q_1^{i,k}$ and $Q_2^{i,k}$ in the case of piecewise constant densities and the corresponding CIFs (Equation (4)):

$$Q_1^{i,k} = \sum_{j=1}^{J} \sum_{l\in S_k} \delta_{kl}^i \log\left(f_{kl,j}^{\text{PC}}\right) \mathbb{1}\left\{(\tilde{T}_k^i - \tilde{Z}_k^i) \in v_j\right\},$$
$$Q_2^{i,k} = \left(1 - \sum_{l\in S_k} \delta_{kl}^i\right) \times \log\left(1 - F_k^{\text{PC}}(\tilde{T}_k^i - \tilde{Z}_k^i)\right).$$

Appendix D. Cumulative hazard function based on piecewise constant density probabilities

To compute the cumulative transition intensities defined in Equation (1), someone interested must used the following equivalence formula:

$$\alpha_{kl}(.) = \frac{f_{kl}(.)}{1 - F_k(.)}. \quad (8)$$

Then, for the new instance i, the cumulative transition intensity must be computed as a piecewise-linear function. For $d > 0$, $|v_j| = a_j - a_{j-1}$ the length of interval j and $|v_{j(d)}| = d - a_{j(d)-1}$, we get:

$$\hat{A}_{kl}^{\text{PC}}(d|X_i) = \sum_{j=1}^{j(d)-1} |v_j| R_{kl,j}^i + |v_{j(d)}| R_{kl,j(d)}^i$$

with

$$R_{kl,j}^i = \frac{\hat{f}_{kl,j}^{\text{PC}}(X_i)}{1 - \hat{F}_k^{\text{PC}}(v_j|X_i)},$$
$$R_{kl,j(d)}^i = \frac{\hat{f}_{kl,j(d)}^{\text{PC}}(X_i)}{1 - \hat{F}_k^{\text{PC}}(v_{j(d)}|X_i)}.$$

Appendix E. Simulated datasets

We generated the five following semi-Markov simulation datasets to evaluate the proposed methodology: (S2) a *non-linear survival* model; (S2) a *non-linear four-state* model; (S3) a *high-dimensional non-linear four-state* model ; (S4) a *linear four-state* model ; (S5) a *non-linear and non-P.H illness–death* model.

E.1. Covariates

For each simulation scenario, $n = 5000$ observations were generated. For K fixed states, we

consider the last state K as absorbent. For each state k ($k = 0, \cdots, K-1$), we consider all the progressive transitions (i.e., $(K-1)-k$ transitions). For each transition $k \to l$ ($l > k$, $l \in S_k$), we consider that two covariates have an effect on that transition (e.g., for $K = 4$, we simulate $P = \left(\sum_{k=0}^{K-1}((K-1)-k)\right) \times 2 = 12$ covariates ; for $K = 2$: $P = 2$; for $K = 3$: $P = 6$). Each covariate is drawn from a gaussian distribution $\mathcal{N}(0,1)$.

An exception is made in simulation (S3) where $P = 50$ covariates are generated with a multivariate normal distribution with a mean of zero and a variance-covariance matrix Σ, with Σ being the variance-covariance matrix of the gene expression data of the first 50 genes in the METABRIC cohort. For each transition, we still consider that only two covariates have an effect on that transition (i.e., 12 active variables out of the 50 generated).

E.2. Event times

Continuous semi-markovian event times (i.e., $\tilde{T}_k$) were generated using a Weibull survival model and following the simulation procedure given by Bender et al. (2005). Properties of the competing and successive transitions were also guaranteed by following the procedures outlined by (Beyersmann et al., 2009) and (Pénichoux et al., 2015), respectively. See pseudo code in Algorithm 1. See an illustration of the generated times in Figure 3.

Appendix F. Real-world datasets

F.1. METABRIC cohort

The METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset is a large breast cancer cohort combining clinical information with high-dimensional molecular data. It includes approximately 2,000 patients with primary breast cancer, with detailed clinical and pathological variables, together with gene expression profiles measured using microarray technology. Time-to-event outcomes are available for both overall survival and relapse-free survival (local or distant relapse) so that a four-state process can be constructed (see Figure 4).

Algorithm 1: Procedure of simulation for semi-markovian event times

Consider a multi-state process with K states. For each state k with the set of transitions (k,l) for $l \in S_k$:

- Choose a distribution for the baseline hazard $\alpha_{0,k}(t)$, e.g., weibull(λ_k, γ_k): $\alpha_{0,k}(t) = \lambda_k \gamma_k t^{\gamma_k - 1}$.
- Choose a (non-linear) risk function $g_{kl}(X, \beta_{kl})$ with β_{kl} a vector of scalar.
- Compute the hazard function of state k:

$$\alpha_k(t|X) = \alpha_{0,k}(t)\left(\sum_{l \in S_k} \exp\left(g_{kl}(X, \beta_{kl})\right)\right).$$

- Compute the survival time T_k using the inverse method (Bender et al., 2005). There are two cases.
  (a) If $k = 0$: $T_k = \left[-\frac{\log(U)}{\left(\sum_{l \in S_k} \exp(g_{kl}(X,\beta_{kl}))\right) \times \lambda_k}\right]^{1/\gamma_k}$;
  (b) If $k > 0$: $T_k = \left[Z_k^{\gamma_k} - \frac{\log(U)}{\left(\sum_{l \in S_k} \exp(g_{kl}(X,\beta_{kl}))\right) \times \lambda_k}\right]^{1/\gamma_k}$;
  with $U \sim \mathcal{U}[0,1]$. This guaranties the semi-markovian property (Pénichoux et al., 2015).
- Then, for each transition kl ($l \in S_k$):
  - Compute $\alpha_{kl}(T_k|X) = \alpha_{0,k}(T_k) \times \exp\left(g_{kl}(X, \beta_{kl})\right)$.
  - Simulate the probability that the exit from state k at time T_k is linked to cause l: $D_k = \arg\max_l(\mathcal{M}_{kl})$, with $\mathcal{M}_{kl}$ a multinomial distribution:

$$\mathcal{M}_{kl} = \mathcal{M}\left(\frac{\alpha_{kl}(T_k|X)}{\sum_{m \in S_k} \alpha_{km}(T_k|X)}\right).$$

    This guaranties properties of the competing transitions (Beyersmann et al., 2009).
- Simulate censoring C_k through an exponential or a uniform distribution.

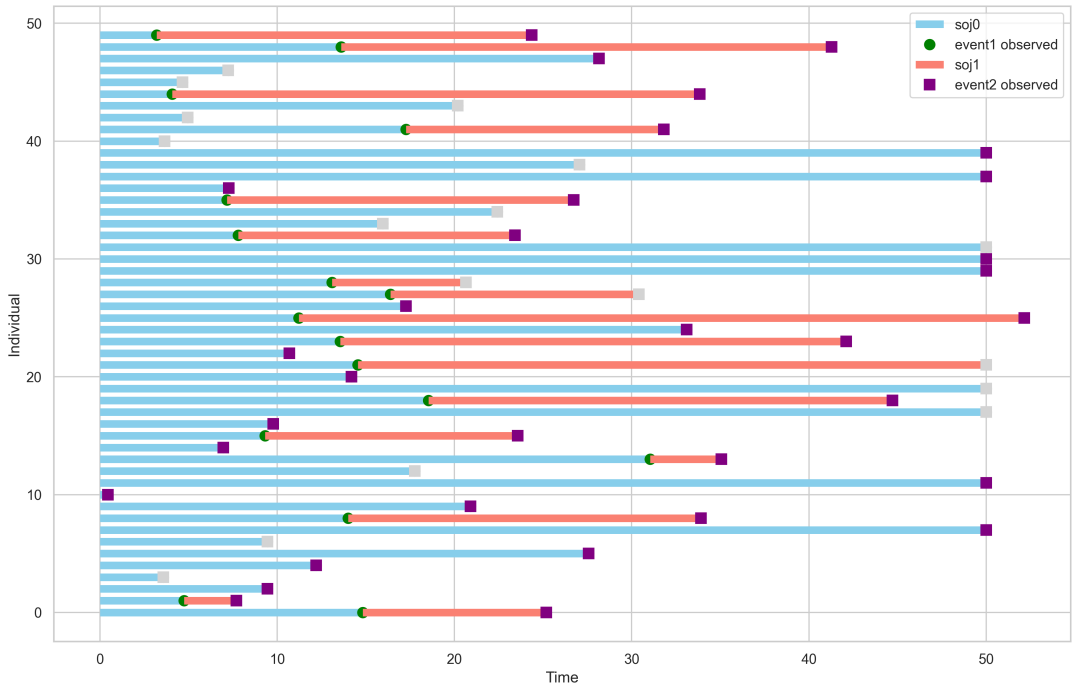


Figure 3: Simulation (S5) *non-linear non-P.H. illness-death*: Sojourn times (soj) per state for the first 50 observations. The color grey indicates a censored event

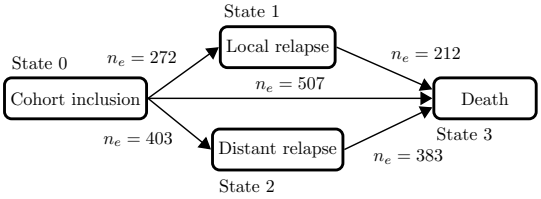


Figure 4: METABRIC: Four-state process with the number of observed events (n_e).

To train MSnet, we fix a uniform length for the time intervals to one month such that the time interval for month j, $v_j = [j-1, j)$, includes all the events that occurred on the daily time interval $[(j-1)\times 30.5, j\times 30.5)$. We set $J = 120$, i.e., 120 time intervals of length 30.5 days so that the horizon window τ is 120 months (10 years). We set the event times after τ in a last interval v_{121}.

F.1.1. Clinical features

Clinical features include: age at diagnosis, inferred menopausal status, tumor size, positive lymph node, cancer grade, tumor histology, cellularity, Her2 copy number, Her2 expression, oestrogen receptor expression, progesterone receptor expression, breast surgery, chemotherapy, hormonal therapy, radiotherapy. For all the features, missing values were imputed by the median value for numerical features and by the mode for categorical features. We apply dummy encoding on categorical features and standardize numerical features with a min-max scaler.

F.1.2. Gene selection

The METABRIC dataset includes transcriptomic data with approximatively 24,000 raw gene expression features. Prior to integrating these features into our prognostic algorithm, gene selection is necessary. To this end, we implemented a gene selection approach inspired by the methodology in Cottin et al. (2022). This approach is based on gene-univariate screening - that is a standard practice in high-dimensional omics to reduce dimensionality and stabilize model fitting (Zhao et al., 2024) - extracting a ranked list

of cancer-related genes based on their adjusted p-values from gene-specific Cox P.H. models:

- For each gene: (1) Fit a Cox P.H. model including the clinical, histo-pathological and molecular features and each gene expression feature. (2) Compute the p-value from a Wald test on the estimated coefficient related to the specific gene.
- Adjust and rank the p-values with a Benjamini-Hochberg multi-test.

This procedure was applied using an independent dataset to control for selection bias: the Breast Invasive Carcinoma TCGA PanCancer dataset (Weinstein et al., 2013). The analysis was performed for both endpoints, OS (Overall survival) and RFS (relapse-free survival), leading to the identification of two corresponding lists of genes that were then merged. Several versions of the METABRIC dataset were initially constructed by considering different values of the selection threshold (noted α). Higher-than-usual values of α were explored due to the large number of initial genes, which induces strong constraints on adjusted p-values. This strategy aimed to avoid excluding genes that might still have a potential prognostic effect on OS or RFS. Based on this exploration, the final analyses were conducted using the model corresponding to $\alpha = 0.3$, yielding a total of 1 419 selected genes (285 for RFS and 1 223 for OS).

F.2. Rotterdam cohort

The Rotterdam dataset is a real-world clinical cohort of primary breast cancer patients collected through the Rotterdam tumor bank (Royston and Altman, 2013) and included in the *survival* package in R. It comprises 2,982 patients with information on demographic and tumor characteristics such as age at surgery, menopausal status, tumor size and grade, number of positive lymph nodes, hormone receptor levels, and treatment indicators, along with time-to-event outcomes for RFS and OS (see Figure 5).

To train MSnet, we fix a uniform length for the time intervals to one month such that the time interval for month j, $v_j = [j-1, j)$, includes all the events that occurred on the daily time interval $[(j-1)\times 30.5, j\times 30.5)$. We set $J = 120$, i.e., 120 time intervals of length 30.5 days so that the horizon window τ is 120 months (10 years). We set the event times after τ in a last interval v_{121}.

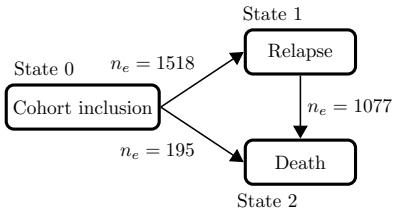


Figure 5: Rotterdam: Illness-death process with the number of observed events (n_e).

To train MSnet, we fix a uniform length for the time intervals to one month such that the time interval for month j, $v_j = [j-1, j)$, includes all the events that occurred on the daily time interval $[(j-1)\times 30.5, j\times 30.5)$. We set $J = 100$, i.e., 100 time intervals of length 30.5 days so that the horizon window τ is 5000 days. We set the event times after τ in a last interval v_{101}.

Appendix G. Performance metrics

G.1. IPCW estimator of the transition specific and time-dependent AUC

For two patients i and j, the transition-specific time-dependent AUC measures (Uno et al., 2007) the probability that a patient i who experienced the transition $k \to l$ before time t has greater probability of occurrence of the transition than a patient j who has survived to the transition. Commonly, the AUC is defined as the integration of the ROC curve opposing specificity (Sp) and 1-sensitivity (1-Se). From the IPCW estimators proposed by Blanche et al. (2013), an IPCW estimator of the AUC, for a transition $k \to l$ and a time $t > 0$ generalizes for a time homogeneous semi-Markov multi-state algorithm from the following estimators of the sensitivity (Se) and specificity (Sp):

$$\widehat{\mathrm{Se}}_t^{kl}(c) = \frac{\sum_{i:\delta^i_{Z_k}=1} \delta^i_{kl} \times Q^{t,\bar{c}}_{kl,i} \times \mathbb{1}^t_{k,i} \times \hat{w}_k(\tilde{T}^i_k - \tilde{Z}^i_k)}{\sum_{i:\delta^i_{Z_k}=1} \delta^i_{kl} \times \mathbb{1}^t_{k,i} \times \hat{w}_k(\tilde{T}^i_k - \tilde{Z}^i_k)},$$

$$\widehat{\mathrm{Sp}}_t^{kl}(c) = \frac{\sum_{i:\delta^i_{Z_k}=1} Q^{t,c}_{kl,i} \times \mathbb{1}^{\bar{t}}_{k,i}}{\sum_{i:\delta^i_{Z_k}=1} \mathbb{1}^{\bar{t}}_{k,i}},$$

with

$$\mathbb{1}^{t}_{k,i} = \mathbb{1}\{\tilde{T}^i_k - \tilde{Z}^i_k \leq t\},$$
$$\mathbb{1}^{\bar{t}}_{k,i} = \mathbb{1}\{\tilde{T}^i_k - \tilde{Z}^i_k > t\},$$
$$Q^{t,\bar{c}}_{kl,i} = \mathbb{1}\{\hat{F}^{\text{PC}}_{kl}(t|X_i) > c\},$$
$$Q^{t,c}_{kl,i} = \mathbb{1}\{\hat{F}^{\text{PC}}_{kl}(t|X_i) \leq c\},$$

and for $\hat{w}_k(.) = 1/\hat{S}^k_c(.)$, $\hat{S}^k_c(.)$ a non-parametric estimator of the survival function C from state k computed on the set of at risk observations (e.g., the Aalen and Johansen estimator (Aalen and Johansen, 1978)).

Finally, the estimated AUC for transition $k \to l$ at time t is computed as follows:

$$\widehat{\text{AUC}}^{kl}(t) = \int_0^t \widehat{\text{Se}}^{kl}_t \left(\left(1 - \widehat{\text{Sp}}^{kl}_t\right)^{-1}(p) \right) \text{d}p.$$

G.2. IPCW estimator of the transition specific and time-dependent BS

Similarly, we can generalize the IPCW estimator of the BS (Gerds and Schumacher, 2006) for a transition $k \to l$ and a time $t > 0$ based on the work of Spitoni et al. (2018) as follows:

$$\widehat{\text{BS}}_{kl}(t) = \frac{1}{n_k} \sum_{i:\delta^i_{Z_k}=1} \left(\mathbb{1}^{\bar{t}}_{k,i} - \hat{F}^{\text{PC}}_{kl}(t|X_i) \right)^2 \times \hat{w}^*_k(\tilde{T}^i_k - \tilde{Z}^i_k, t),$$

with $\mathbb{1}^{\bar{t}}_{k,i} = \mathbb{1}\{\tilde{T}^i_k - \tilde{Z}^i_k > t\}$, n_k the sum of the observations i so that $\delta^i_{Z_k} = 1$ and, for $S^k_c(.)$ defined previously:

$$\hat{w}^*_k(\tilde{T}^i_k - \tilde{Z}^i_k, t) = \frac{\mathbb{1}^{\bar{t}}_{k,i} \times \delta^i_{kl}}{\hat{S}^k_c(\tilde{T}^i_k - \tilde{Z}^i_k)} + \frac{\mathbb{1}^{\bar{t}}_{k,i}}{\hat{S}^k_c(t)}.$$

G.3. Integrated performance metrics

An estimator of the integrated AUC (iAUC) is computed as follows:

$$\widehat{\text{iAUC}}_{kl} = \frac{1}{\tau} \int_0^\tau \widehat{\text{AUC}}^{kl}(t)dt.$$

Similarly, and estimator of the integrated BS (iBS) is computed as follows:

$$\widehat{\text{iBS}}_{kl} = \frac{1}{\tau} \int_0^\tau \widehat{\text{BS}}^{kl}(t)dt.$$

Appendix H. Cross validation

The cross validation (CV) procedure used to estimate performances metrics for MSnet is described in Figure 6. For tsRSF, the same CV procedure was used except for the hyper-parameters that were optimized through grid search. For tsLassoCox, the same CV procedure was used except for the hyper-parameter of penalisation that was optimized with the function *cv.glmnet()* from the R package *glmnet* (see Appendix I).

Appendix I. Hyper-parameters

I.1. MSnet

MSnet is implemented in Python with *Tensorflow*. It uses standard deep learning techniques as L_1 and L_2 regularized layers as well as dropout to avoid over-fitting, mini-batch learning, learning rate weight decay and early stopping. A complete list of hyper-parameters of MSnet are given in Table 5. Hyper-parameters are tuned using random search. For each model, the sets of search spaces are adjusted according to the number of covariates. At each iteration of the CV procedure, the best set of hyper-parameters was used as the one maximizing the averaged iAUCs on the transitions.

I.2. tsLassoCox

Transition-specific Lasso Cox P.H. models were trained with the R package *glmnet*. The unique hyper-parameter in tsLassoCox is the regularisation parameter that was optimized with the function *cv.glmnet()* from the R package *glmnet*.

I.3. tsRSF

Transition-specific RSFs were trained with the Python library *scikit_survival*. A complete list of hyper-parameters are given in Table 6. Hyper-parameters were tuned using grid search. At each iteration of the CV procedure, the best set of hyper-parameters was used as the one maximizing the averaged iAUCs on the transitions.

I.4. SurvNODE

The algorithm SurvNODE proposed by Groha et al. (2020) was trained based on the Python code available at https://github.

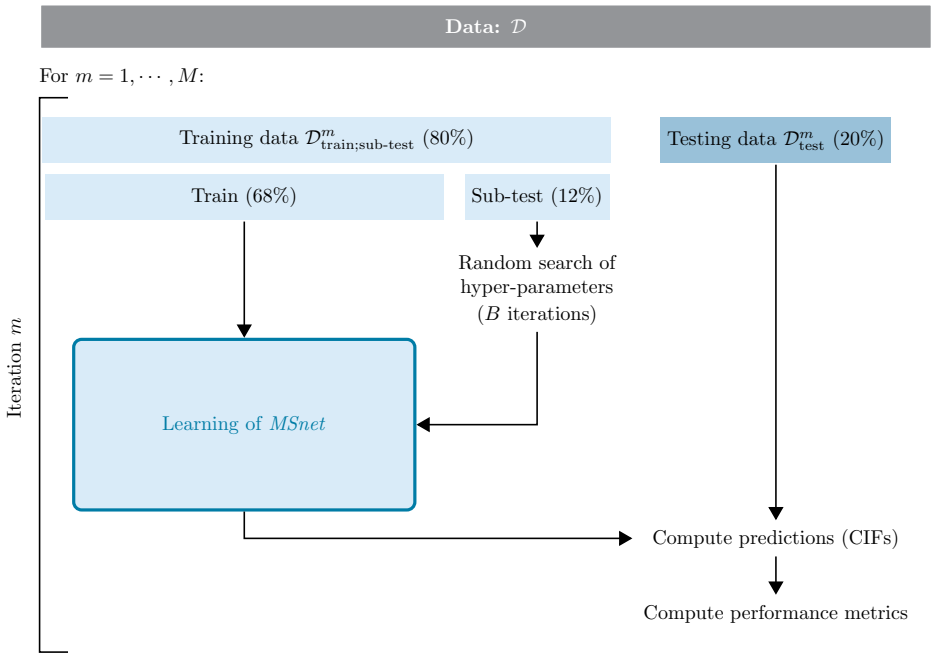


Figure 6: Validation process.

com/stefangroha/SurvNODE. List of hyper-parameters are given in Table 7. Values were given based on the values used in their code.

I.5. DeepHit

The algorithm DeepHit proposed by Lee et al. (2018) was trained based on the Python code available at https://github.com/chl8856/DeepHit. List of hyper-parameters are given in Table 8 (based on the values used in their code). Hyper-parameters were tuned using random search.

Appendix J. Additional results

This appendix reports additional results based in terms on predictive performance, complementing the main results presented in the paper. In these results, performance of MSnet against the other approaches were statistically compared using paired bilateral Wilcoxon signed rank tests: · indicates a p-value less than 0.1, † less than 0.05, ‡ less than 0.01, * less than 0.001. Bold values denotes the best algorithm.

J.1. Real-world datasets

We display additional results on the iBS on the two real-world datasets.

For the Rotterdam cohort, results of iBS are given in Table 9(a). For transitions $0 \to 1$ and $0 \to 2$, MSnet outperforms msCox and SurvNODE. For transition $1 \to 2$, msCox outperforms SurvNODE and MSnet, but with no statistical difference.

For the METABRIC cohort, results of iBS are given in Table 9(b) for models M_0 and M_{1419}. For transition $0 \to 1$, MSnet outperforms the other algorithms. For the other transitions, the best predictive performance is either for msCox/tsLassoCox or for SurvNODE.

J.2. Simulated datasets

We display additional results on the iBS on the simulated datasets in Table 10.

Table 5: Hyper-parameters of MSnet.

Hyper-parameter name	Discrete search space
Initialization	Uniform initialization
Optimizer	Adam Optimizer
Learning rate	$\{10^{-4}, 10^{-5}\}$
Mini-batch size	{8, 16, 32, 64, 128}
Keep probability for dropout	{0.5, 0.6, 0.7, 0.8, 0.9}
Nodes per shared hidden layer (l)	{3, 5, 15, 30, 50, 100}
No. shared hidden layers (L)	{1, 2, 3}
Nodes per transition-specific hidden layer (l^{kl})	{3, 5, 15, 30, 50}
No. transition-specific hidden layers (L^{kl})	{0, 1, 2, 3}
Non linear activation function (g^{input}, g^{kl})	{ReLU, ELU, Leaky-ReLU}
L_1/L_2 regularization parameter	$\{1^{-5}, 1^{-4}, 1^{-3}, 1^{-2}, 0.1\}$
Decay for gradient descent	$\{1^{-3}, 1^{-2}, 0.1, 0.4, 0.6, 0.8, 1.0\}$
Parameter difference for early stopping	$\{1^{-6}, 1^{-5}\}$
Output-specific penalization (λ_w^{kl}, λ_b^{kl})	$\{1^{-5}, 1^{-4}, 1^{-3}, 1^{-2}, 0.1\}$

Table 6: Hyper-parameters of tsRSF.

Hyper-parameter name	Discrete search space
min_samples_split	10
max_features	"sqrt"
n_estimators	{50, 100, 500, 1000}
min_samples_leaf	{20, 60, 100}

Table 7: Hyper-parameters of SurvNODE.

Hyper-parameter name	Value
Learning rate	$\{10^{-4}, 10^{-5}\}$
Batch size	128
layers_encoder	{50, 100, 200, 400}
dropout_encoder	{0.0, 0.1, 0.2}
layers_odefunc	{30, 400, 1000}
num_latent	{30, 100}

Table 8: Hyper-parameters of DeepHit.

Hyper-parameter name	Discrete search space
Learning rate	1^{-4}
Initialization	Uniform initialization
Batch size	{32, 64, 128}
Number of layers	{1, 2, 3, 5}
Number of nodes	{50, 100, 200, 300}
Activation function	{ReLU, ELU, tanh}
Keep probability for dropout	0.6
alpha	{0.1, 0.5, 1.0, 3.0, 5.0}
beta	{0.1, 0.5, 1.0, 3.0, 5.0}
gamma	{0.1, 0.5, 1.0, 3.0, 5.0}

To validate the robustness of our algorithm under varying censoring rates, we conducted an additional simulation study (Scenario (S1) - *survival*). While the initial submission used a 30% censoring rate, we increased this rate to 60% to better mimic complex clinical scenarios. The results confirm that MSnet maintains stable and competitive performance even at this higher censoring level, as summarized in Table 11.

We display additional results on the iAUC and iBS on the simulation scenario (S4) (*linear four-state*) in Table 12. In these results, the advantage of MSnet disappears, as expected: the log-linear assumption of msCox holds, and its performance matches that of MSnet.

Table 9: Real-world datasets: Benchmark of iBSs (median ± sd).

(a) Rotterdam dataset.

Alg.	**Transitions** (No. observed events)		
	0 → 1 (1518)	**0 → 2** (195)	**1 → 2** (1077)
msCox	0.193 ± 0.006	0.042 ± 0.006	**0.146 ± 0.016**
SurvNODE	0.334 ± 0.009	0.064 ± 0.005	0.259 ± 0.012
MSnet	**0.179 ± 0.004**	**0.036 ± 0.004**	0.150 ± 0.017

(b) METABRIC dataset - Models M_0 and M_{1419}.

Mod.	**Alg.**	**Transitions** (No. observed events)				
		0 → 1 (272)	**0 → 2** (403)	**0 → 3** (507)	**1 → 3** (212)	**2 → 3** (383)
M_0	msCox	0.038 ± 0.006$^{\cdot}$	**0.045 ± 0.006***	0.029 ± 0.004	0.214 ± 0.027	**0.216 ± 0.014**
	SurvNODE	0.039 ± 0.006$^{\cdot}$	0.051 ± 0.006*	**0.017 ± 0.004**	**0.136 ± 0.010***	0.185 ± 0.016*
	MSnet	**0.035 ± 0.006**	0.069 ± 0.008	0.028 ± 0.007	0.213 ± 0.040	0.224 ± 0.046
M_{1419}	tsLassoCox	0.038 ± 0.005*	0.072 ± 0.007	0.027 ± 0.004$^{\cdot}$	0.185 ± 0.032*	**0.153 ± 0.020***
	SurvNODE	0.039 ± 0.026*	**0.051 ± 0.006***	**0.017 ± 0.004***	**0.136 ± 0.036***	0.243 ± 0.018‡
	MSnet	**0.032 ± 0.004**	0.065 ± 0.008	0.029 ± 0.005	0.217 ± 0.035	0.223 ± 0.025

Table 10: Simulations: Benchmark of iBSs (median ± sd).

(a) Simulation (S2) - *non-linear survival*.

Alg.	**Transition** (No. observed events)
	0 → 1 (3500)
msCox	0.137 ± 0.005*
SurvNODE	0.123 ± 0.004
RSF	**0.117 ± 0.004**‡
DeepHit	0.118 ± 0.004‡
MSnet	0.126 ± 0.005

(b) Simulation (S2) - *non-linear four-state*.

Alg.	**Transitions** (No. observed events)					
	0 → 1 (14821)	**0 → 2** (1575)	**0 → 3** (1751)	**1 → 2** (677)	**1 → 3** (722)	**2 → 3** (1075)
msCox	0.229 ± 0.010*	**0.099 ± 0.010***	0.221 ± 0.010*	0.363 ± 0.020*	**0.105 ± 0.010***	**0.095 ± 0.003**†
SurvNODE	0.344 ± 0.017*	0.516 ± 0.024*	0.133 ± 0.009‡	0.560 ± 0.022*	0.108 ± 0.005*	0.109 ± 0.005†
MSnet	**0.131 ± 0.014**	0.130 ± 0.008	**0.117 ± 0.010**	**0.152 ± 0.020**	0.132 ± 0.016	0.102 ± 0.003

(c) Simulation (S3) - *high-dimensional non-linear four-state*.

Alg.	**Transitions** (No. observed events)					
	0 → 1 (1541)	**0 → 2** (1545)	**0 → 3** (1732)	**1 → 2** (753)	**1 → 3** (710)	**2 → 3** (1127)
msCox	0.383 ± 0.050*	**0.138 ± 0.030**	0.277 ± 0.010*	0.393 ± 0.010*	**0.080 ± 0.010***	**0.078 ± 0.003***
SurvNODE	NC	NC	NC	NC	NC	NC
MSnet	**0.124 ± 0.010**	0.145 ± 0.009	**0.120 ± 0.010**	**0.162 ± 0.010**	0.140 ± 0.013	0.107 ± 0.009

NC: Non convergent

(d) Simulation (S5) - *non-linear non-P.H. four-state*

Alg.	**Transitions** (No. observed events)		
	0 → 1 (1895)	**0 → 2** (2297)	**1 → 2** (1269)
msCox	0.168 ± 0.010*	0.171 ± 0.004*	**0.131 ± 0.004***
SurvNODE	NC	NC	NC
MSnet	**0.129 ± 0.005**	**0.137 ± 0.005**	0.146 ± 0.012

NC: Non convergent

Table 13 displays supplemental results on a simulation with 8 states and 56 transitions. In this particular case, performance of MSnet degrades (iAUC≈0.49–0.54), while msCox and SurvNode fail to converge (likely due to severe overparameterization with limited events

Table 11: Variation in the censoring rate (r) - Simulation (S1) - *non-linear survival*: Benchmark of iAUCs and iBSs (median ± sd).

(a) iAUCs (median ± sd).

Alg.	**Transition** (No. observed events)	
	0 → 1	
	(3500)	(2000)
	r=30%	r=60%
msCox	$0.502 \pm 0.020*$	$0.504 \pm 0.020^{*}$
SurvNODE	$0.537 \pm 0.011^{*}$	$0.524 \pm 0.018^{*}$
RSF	$0.745 \pm 0.012^{\dagger}$	$0.695 \pm 0.019^{*}$
DeepHit	$0.717 \pm 0.008^{*}$	$0.679 \pm 0.024^{*}$
MSnet	$\mathbf{0.766 \pm 0.020}$	$\mathbf{0.753 \pm 0.020}$

(b) iBSs (median ± sd).

Alg.	**Transition** (No. observed events)	
	0 → 1	
	(3500)	(2000)
	r=30%	r=60%
msCox	$0.137 \pm 0.005^{*}$	$0.165 \pm 0.010^{*}$
SurvNODE	0.123 ± 0.004	$\mathbf{0.146 \pm 0.005}$
RSF	$\mathbf{0.117 \pm 0.004}^{\ddagger}$	$0.497 \pm 0.056^{*}$
DeepHit	$0.118 \pm 0.004^{\ddagger}$	0.150 ± 0.007
MSnet	0.126 ± 0.005	0.148 ± 0.011

Table 12: Simulation (S4) - *linear four-state*: Benchmark of iAUCs and iBSs (median ± sd).

(a) iAUCs (median ± sd).

Alg.	**Transitions** (No. observed events)					
	0 → 1 (1164)	**0 → 2** (1356)	**0 → 3** (1744)	**1 → 2** (377)	**1 → 3** (558)	**2 → 3** (1198)
msCox	0.671 ± 0.020	$\mathbf{0.647 \pm 0.030}$	$\mathbf{0.659 \pm 0.020}^{*}$	$0.588 \pm 0.020^{\ddagger}$	$\mathbf{0.686 \pm 0.020}^{*}$	$\mathbf{0.831 \pm 0.010}^{*}$
SurvNODE	$0.561 \pm 0.021^{*}$	$0.573 \pm 0.014^{*}$	$0.570 \pm 0.124^{*}$	$0.501 \pm 0.025^{*}$	$0.587 \pm 0.026^{*}$	$0.591 \pm 0.013^{*}$
MSnet	$\mathbf{0.674 \pm 0.018}$	0.625 ± 0.023	0.615 ± 0.029	$\mathbf{0.655 \pm 0.047}$	0.644 ± 0.028	0.795 ± 0.018

(b) iBSs (median ± sd).

Alg.	**Transitions** (No. observed events)					
	0 → 1 (1164)	**0 → 2** (1356)	**0 → 3** (1744)	**1 → 2** (377)	**1 → 3** (558)	**2 → 3** (1198)
msCox	$0.175 \pm 0.010^{*}$	$\mathbf{0.090 \pm 0.004}^{*}$	$0.136 \pm 0.004^{*}$	$0.201 \pm 0.020^{*}$	$\mathbf{0.111 \pm 0.010}^{*}$	$\mathbf{0.094 \pm 0.004}^{*}$
SurvNODE	$0.273 \pm 0.011^{*}$	$0.368 \pm 0.011^{*}$	$0.200 \pm 0.010^{*}$	$0.343 \pm 0.023^{*}$	$0.164 \pm 0.007^{*}$	$0.208 \pm 0.008^{*}$
MSnet	$\mathbf{0.129 \pm 0.007}$	0.119 ± 0.005	$\mathbf{0.106 \pm 0.004}$	$\mathbf{0.157 \pm 0.020}$	0.134 ± 0.010	0.148 ± 0.011

per transition). This illustrates a shared limitation of current multi-state models—not a weakness specific to our architecture.

J.3. Time and computational resources

In the experiments given in Table 4, we run each algorithm once on a split of the data with a fixed set of hyper-parameters. We have adjusted the hyper-parameters of the neural network-based algorithms (i.e., MSnet, DeepHit, SurvNODE) so that they contain the same number of layers and neurons, and that the computation time is not influenced by the size of the models.

In simulation (S5) *non-linear non-P.H. illness-death*, the computation time of MSnet is longer than in simulation (S2) *non-linear four-state* (11 min. vs. 7 min.) because we parameterized the size of the output layer of MSnet in simulation (S5) as being larger (i.e., more time intervals: $J = 400$ vs. $J = 100$) thus allowing the model to easily adapt to the non-P.H. assumption).

Appendix K. METABRIC cohort: Clinical interpretation

Graphical interpretation of model M_0 for each transition of the four-state process in the METABRIC cohort is given in figures 8, 9, 10, 11 and 12, below.

Table 13: Additional simulation with 8 states and 56 transitions: Benchmark of iAUCs and iBSs (median ± sd). In this table, we gave the predictive performance for each transition: competing transitions are averaged per state (e.g., 0 → 1-7 means the transitions from state 0 to states 1 to 7).

(*a*) iAUCs (median ± sd).

Alg.	**Transitions** (No. observed events)						
	0 → 1-7	1 → 2-7	2→3-7	3→4-7	4→5-7	5→6-7	6→7
msCox	CE	CE	CE	CE	CE	CE	CE
SurvNODE	CE	CE	CE	CE	CE	CE	CE
MSnet	0.509 ± 0.015	0.490 ± 0.046	0.536 ± 0.035	0.515 ± 0.031	0.512 ± 0.015	0.504 ± 0.018	0.535 ± 0.055

CE: Convergence error

(*b*) iBSs (median ± sd).

Alg.	**Transitions** (No. observed events)						
	0 → 1-7	1 → 2-7	2→3-7	3→4-7	4→5-7	5→6-7	6→7
msCox	CE	CE	CE	CE	CE	CE	CE
SurvNODE	CE	CE	CE	CE	CE	CE	CE
MSnet	0.088 ± 0.005	0.102 ± 0.003	0.111 ± 0.002	0.128 ± 0.005	0.141 ± 0.005	0.147 ± 0.004	0.130 ± 0.004

CE: Convergence error

Detailed clinical interpretation is given in the following paragraphs.

Cancer grade (Rakha et al., 2010) Tumor grade is a prognostic classification based on the proliferation rate of cancer cells. A grade 1 tumor grows slowly and is associated with a low likelihood of metastasis, while a grade 3 tumor grows rapidly and is associated with a high likelihood of metastasis. Grade 2 tumors grow faster than grade 1 tumors but slower than grade 3 tumors and have an intermediate probability of metastasis. Therefore, patients with grade 1 and grade 2 tumors generally have a better prognosis than those with grade 3 tumors. Our analysis confirms these findings, as we observe that grade 3 is a risk factor for transitions 01, 02, 13 and 23. It is not significantly associated with transition 03 (death with no relapse), which is quite logical because this transition is rather associated with deaths not due to cancer.

TNM classification (Giuliano et al., 2017) The TNM (Tumor size, Nodes involvement, presence of Metastasis) classification from the American Joint Committee on Cancer (AJCC) is used to establish tumor stage of patients with breast cancer. The tumor size (T) is a classification from T0 (no evidence of primary tumor) to T3 (tumor size greater than 5 cm), and T4 (tumor size growing into the chest wall and/or skin). The number of invaded lymph nodes (N) is a classification from N0 (cancer has not spread to nearby lymph nodes), to N3 (cancer has spread to more than 10 auxiliary lymph nodes). There features are well-known risk factors of breast cancer progression. Interpretability of MSnet reveal that for all the transitions T1 is a protective factor while T2 is a risk factor. T3 is not significant, that is probably due to the fact that the frequency of observations of T3 is low against frequencies of T1 and T2. For all transitions, N0 is a protective factor, N1/N2/N3 are risk factors (except for transition13 where N2/N3 are not significant). These findings align with the existing literature, indicating that a T0 (or N0) cancer is a protective factor, while a higher T (respectively a higher N) increases the risks of relapse and death.

Hormone receptor status (Allison et al., 2020) Breast cancers can be classified in two groups: hormone receptor-positive versus hormone receptor-negative cancers. In hormone receptor-positive breast cancers, female sex hormones (estrogen - ER - and/or progesterone - PR) stimulate tumor growth. These cancers usually grow slower than hormone receptor negative cancers and have a better short-term prognosis, but may have a higher risk of late recurrence. In hormone receptor-negative breast cancers, female sex hormones do not affect cancer cells growth. They have a greater risk of relapse in the first years after the end of treatment. In our results, we found that hormone receptor-positive (ER+ and/or PR+) cancers are favourable prognostic factors, while hormone receptor negative (ER- and/or PR-) cancers are unfavorable prognostic

factors for all five transitions. This means that hormone receptor-positive cancers have a lower risk of relapse or death than hormone receptor negative cancers; these conclusions are consistent with the literature.

Her2 status (Allison et al., 2020) Breast cancers can also be classified as Her2-positive (Her2+) or Her2-negative (Her2-) cancers. Her2+ breast cancers have a higher level of the protein Her2 (Human epidermal growth factor receptor 2). This protein increases the growth of cancer cells, which makes the cancer more aggressive than Her2- cancers. However, Her2 targeted treatments are very effective, that makes Her2+ cancers having a very good prognosis for relapse and death. In ours results, we can see that, for the transitions related to death (03, 13, 23), Her2- is a risk factor, Her2+ is a protective factor; for transitions related to distant and local relapses (01 and 02), there is no significant effect of the protein Her2.

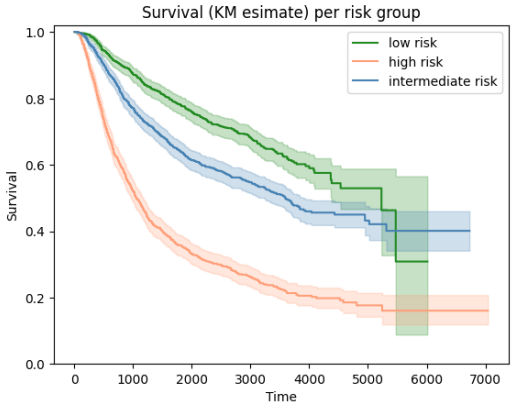
Age at diagnosis (McGuire et al., 2015) Age is a known risk factor of breast cancers: incidence increases with age. However, patients diagnosed at a young age (less than 40 years old) have a poorer prognosis than patients aged from 40 to 60 years old. They present a higher probability to have a triple-negative cancer (i.e., ER-, PR- and Her2-) because they are about 2 to 3 times more likely to have the BRCA1 mutation that is linked to the development of aggressive breast cancer as the triple-negative breast cancer. These cancers have a poorer remission prognosis than other sub-types of breast cancer. Consequently, this age group has a higher recurrence rate than the others age groups. Patients over 70 years old have the lowest survival; their survival is affected by their age and their risk of developing medical comorbidities. However, cancers of elderly patients are mainly hormone receptor-positive cancers; therefore they present a low risk of cancer-related death. In our the results of interpretability of MSnet, for the transitions related to distant and local relapses (01 and 02), younger ages ($\leq$ 50 years old) are risk factors, while older ages (>50) are protective factors. These results are consistent with the literature; younger patients have a higher probability of relapse; older patients have a good prognosis of cancer recurrence as they can be treated effectively with hormonotherapy. For the transitions related to death after relapse (13 and 23), older ages (greater than 50 years old) are risk factors, while younger ages (less than 50) are protective factors. We can also see that patients aged of 90 years old a higher risk for these transitions than patients aged of 70 years old. These results are consistent as the risk of death for patients over 70 years old is also strongly correlated to comorbidities, and this risk increases when age increases.

Therapy and surgery (Gradishar et al., 2020) Breast cancer treatment is based on surgery (with either breast conserving surgery or either mastectomy) and therapies (radiotherapy and/or chemotherapy and/or hormonotherapy). Surgery is most often performed first in order to remove tissues affected by cancer cells, and then it is followed by therapies. Indications for radiotherapy depend on the cancer stage. Indications for chemotherapy depend on the cancer stage and on the risk factors for cancer recurrence, it aims to increase relapse prognosis. In our results on MSnet, for the five transitions, having a chemotherapy is a very high risk factor; this is because indication for this type of treatment concerns patients who are most at risk of relapse, and therefore who have a poor remission prognosis after surgery. Alongside, radiotherapy is indicated for less aggressive cancers with better prognosis; that can explain why radiotherapy is a protective factor, and why chemotherapy is a risk factor. The majority of patients in the METABRIC cohort have hormonal cancer, i.e. hormone receptor positive cancers. Hormonotherapy is indicated in that case and is very effective for improving patients' prognosis. Therefore, as we can see in our results, be treated with hormonotherapy is a protective factor for relapse and death. Finally, we can see that breast surgery is a significant prognostic factors for transitions 13 and 23: a breast conserving surgery (respectively a mastectomy) is a protective factor (respectively a risk factor); this result is coherent because a mastectomy is realized for larger tumors, i.e. with a high T. The reader should nevertheless note that therapies and surgery can't really be interpreted as prognostic factors, as their effect is correlated to the targeted patients that have initially a poorer or a better prognosis.

Histological classification (Makki, 2015) Breast cancers can be classified with their histological type. We preprocess the histological classification feature in the METABRIC dataset by categorizing the histological sub-types in 5 modalities based on the literature and on a M.D expertise: IDC, IDC rare, ILC, mixed IDC-ILC, others. We distinguish two sub-types of IDC (Invasive Ductual Carcinoma) cancers that represent about 80% of breast cancers. Modality IDC represents the majority, i.e. 75%, of IDC cancers; modality IDC rare characterizes a special type of breast carcinoma representing about 2% of IDC cancers. ILC (Invasive Lobular Carcinoma) cancers are the second main type of invasive breast cancers, representing about 5 to 15% of invasive breast cancers. Mixed IDC-ILC cancers can be categorized as IDC and ILC cancers. Modality others represents breast cancers with no specific sub-type, i.e. cancers that haven't sufficient histological characteristics to enter one of the previous sub-type. According to the literature, IDC cancers have a high probability of developing metastasis, and therefore have a poor prognosis. IDC rare cancers have a good prognosis. ILC cancers tend to develop metastasis as well; they are harder to detect than IDC cancers; they are more likely to affect both breasts and older women; they have a poor prognosis. Mixed IDC-ILC cancers have an uncertainty prognosis and an uncertainty treatment response, but they generally present a better prognostic than ILC cancers. In our results, we can see that the modality IDC rare is a high protective factor for transitions related to recurrence. Modality IDC is a risk factor for transitions related to death. IDC-ILC is a risk factor for transitions related to death, while it is a protective factor for transition related to cancer recurrence.

Appendix L. Rotterdam cohort: Clinical decision support

Kaplan Meier curves of relapse-free survival (RFS) per risk profile are given in Figure 7 below. Risk profiles clearly separate a high-risk population whose relapse-free curve lies significantly below the curves of the two other groups. Such stratification could support a simple clinical decision policy in which patients classified as high risk would be considered for intensified monitoring or more aggressive management, while low-risk patients would remain under standard care.



low risk								
At risk	965	819	620	340	93	10	1	0
Censored	0	25	122	349	564	641	648	649
Events	0	121	223	276	308	314	316	316
high risk								
At risk	956	479	271	160	62	15	5	2
Censored	0	27	66	127	197	239	248	251
Events	0	450	619	669	697	702	703	703
intermediate risk								
At risk	1061	793	564	362	133	42	2	0
Censored	0	27	101	248	435	522	560	562
Events	0	241	396	451	493	497	499	499

Figure 7: Kaplan Meier curves of relapse-free survival per risk profile on the Rotterdam cohort patients.

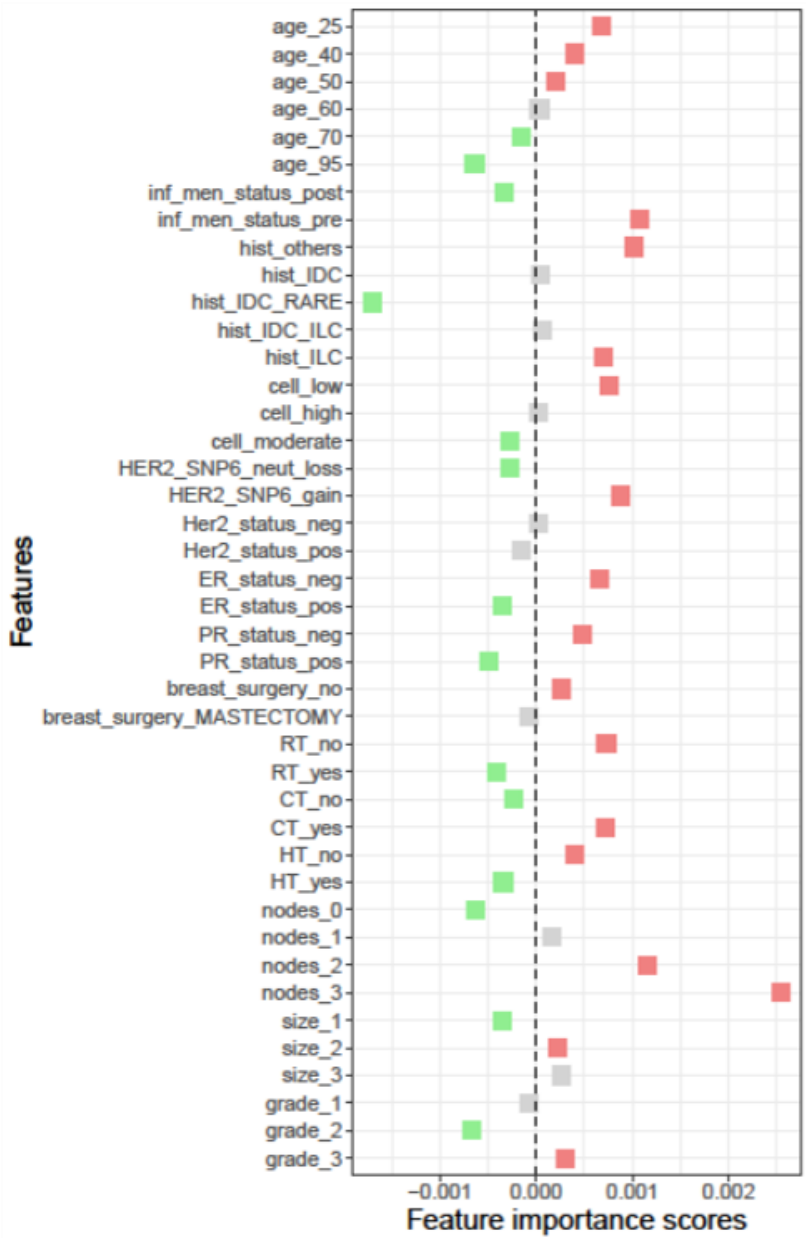


Figure 8: Interpretation for transition $0 \rightarrow 1$ with the algorithm MS-CPFI.

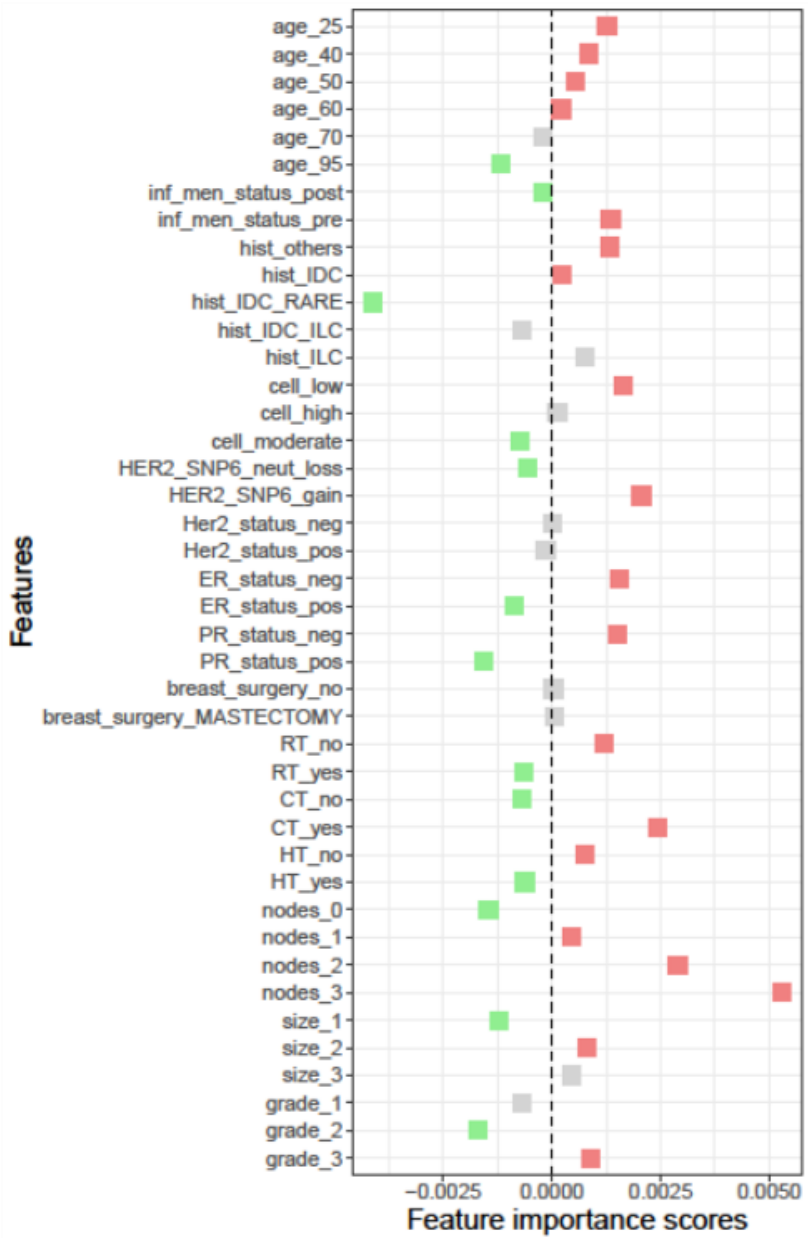


Figure 9: Interpretation for transition $0 \to 2$ with the algorithm MS-CPFI.

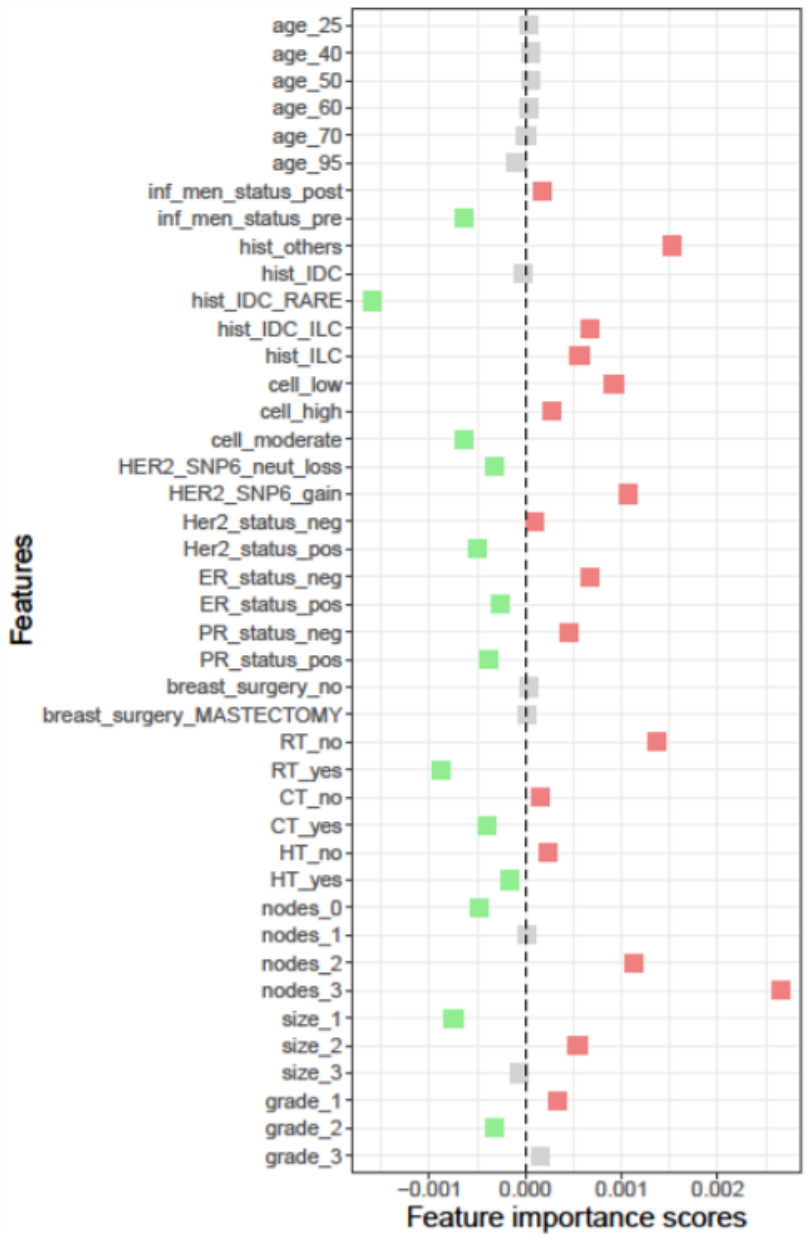


Figure 10: Interpretation for transition $0 \rightarrow 3$ with the algorithm MS-CPFI.

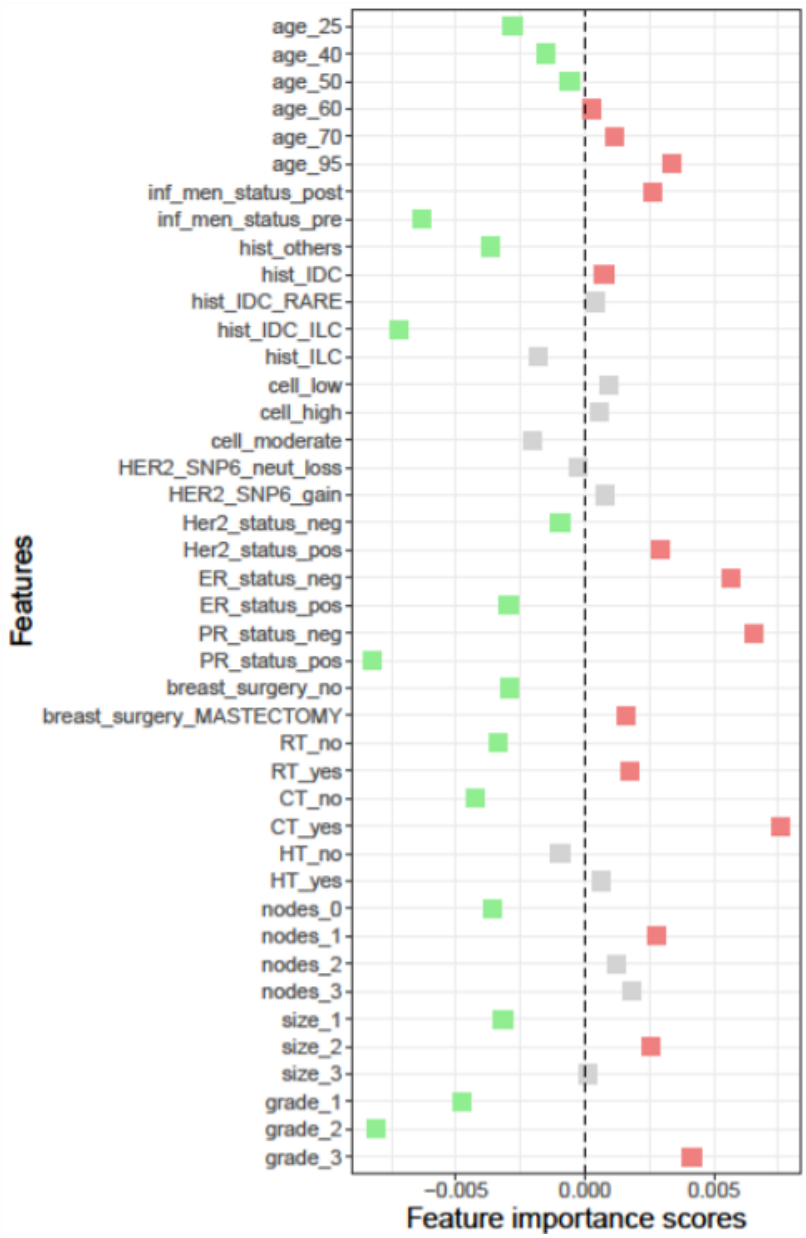


Figure 11: Interpretation for transition $1 \to 3$ with the algorithm MS-CPFI.

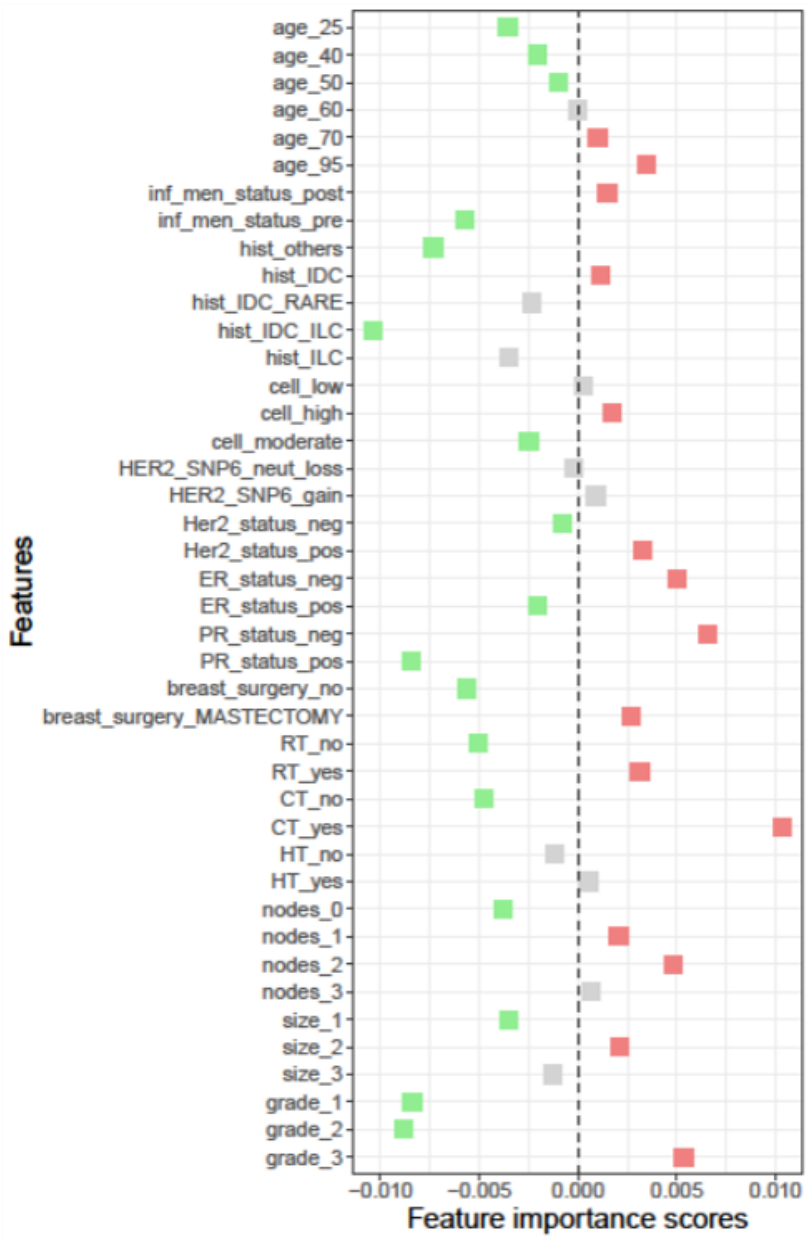


Figure 12: Interpretation for transition 2 → 3 with the algorithm MS-CPFI.