IDNetwork: A deep Illness-Death Network based on multi-state event history process for versatile disease prognostication

Aziliz Cottin AZILIZ.COTTIN@3DS.COM
Nicolas Pécuchet NPT1@3DS.COM
Marine Zulian MZN3@3DS.COM
Healthcare and Life Sciences Research, Dassault Systemes, 78140 Velizy-Villacoublay, France.

Agathe Guilloux* AGATHE.GUILLOUX@UNIV-EVRY.FR
Université Paris-Saclay, CNRS, Univ Evry, Laboratoire de Mathématiques et Modélisation d'Evry, 91037, Evry-Courcouronnes, France.

Sandrine Katsahian* SANDRINE.KATSAHIAN@APHP.FR
Assistance Publique-Hôpitaux de Paris, Medical Informatics, Biostatistics and Public Health Department, Georges Pompidou, Paris, France
INSERM, UMRS 1138, Cordeliers Research Center, Paris, France
Université de Paris, Paris, France

Abstract

Multi-state models can capture the different patterns of disease evolution. In particular, the illness-death model is used to follow disease progression from a healthy state to an intermediate state and to a death-related final state. We aim to use those models in order to adapt treatment decisions according to the evolution of the disease. In state-of-the-art methods, the risks of transition are modeled via (semi-) Markov processes and transition-specific Cox proportional hazard (P.H.) models. We propose a neural network architecture called IDNetwork (Illness-Death Network) that relaxes the linear Cox P.H. assumption and integrates a large number of patients' characteristics. Our method significantly improves the predictive performance compared to state-of-the-art methods on a simulated data set, on two clinical trials for patients with colon cancer and on a real-world data set in breast cancer.

Keywords: Event history analysis, illness-death process, deep learning, neural networks

1. Introduction

Disease prognosis is of major importance for physicians when making medical decisions and requires specialized algorithms to estimate the risks of a patient. In this line of work and within event history analysis, we propose a novel algorithm for individual prognostication in a three states model: the illness-death model.

Event history analysis, also known as survival analysis, aims at predicting the time until the occurrence of a future event(s) of interest conditionally to individual covariates. Classical models rely on strong assumptions. The Cox proportional hazard (P.H.) model (Cox, 1972) relies on linear effects of the covariates and shows limitations in real-world data. To address this challenge, new machine learning algorithms have been developed.

* Co-author

For a unique event of interest, neural networks have been introduced by Faraggi and Simon (1995), and developed more recently by Katzman et al. (2018); Kvamme et al. (2019); Fotso (2018), among others, with significant improvements in predictive performance as compared to the Cox P.H. model. Lee et al. (2018) extended survival networks to handle competing events (see Kalbfleisch and Prentice (2011), chapter 8). To the best of our knowledge, no non-linear methods, especially deep neural networks, have been explicitly introduced for multi-state analysis and in particular for an illness-death process.

In the present work, we focus on the illness-death model which is a multi-state model (Webster, 2019) composed of three states: "healthy"; "relapsed" or "diseased"; "dead". Illness-death model is the most frequent structure used to follow the evolution of cancer patients as in ovarian cancer (Eulenburg et al., 2015) or in chronic myeloid leukemia (Iacobelli and Carstensen, 2013).

Most of the previous work describes the risk of transiting using (semi-) Markov processes and transition-specific Cox P.H. models (De Wreede et al., 2010). We here propose a deep learning architecture, IDNetwork (Illness-Death Network), which models the probabilities of occurrence of the transitions with no linear assumption.

We contribute to (i) derive a new form of the log-likelihood of an illness-death process, (ii) build the network architecture IDNetwork, (iii) implement in Python the pipeline of our method including performance criteria evaluation. We conduct experiments on a simulated non-linear data set and on real data sets of patients with colon cancer and with breast cancer. We show that IDNetwork achieves better performance compared to state-of-the-art methods.

2. Methodology

An illness-death process (Andersen et al., 2012) E is a continuous-time stochastic process that describes the states occupied by a patient over time with three states 0: healthy, 1: relapse, and 2: death. It is characterized by three transitions: from 0 to 1 ($0 \to 1$), from 0 to 2 ($0 \to 2$), from 1 to 2 ($1 \to 2$), where transitions from state 0 are competing, transitions $0 \to 1$ and $1 \to 2$ are successive. See an illustration in Figure 1. We assume that at time 0 all the patients are in state 0.

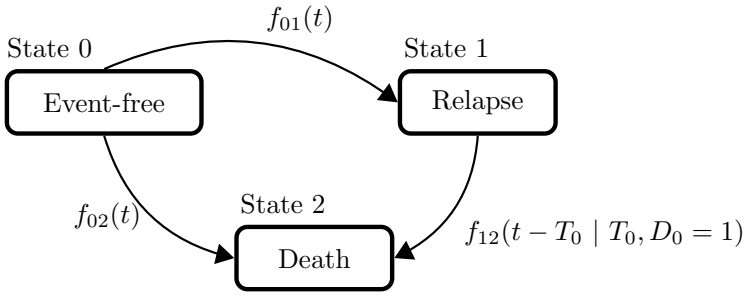


Figure 1: Illustration of an illness-death process.

2.1. Formalism

Throughout the following sections, we note $(k, l) \in \{(0, 1), (0, 2), (1, 2)\}$. The illness-death process is characterized by three random variables (r.v.) T_{kl} associated with each of the three transitions, that represent the transition times from state k to l ($k \neq l$). Subjects leaving state 0 will enter either state 1 at time T_{01} or state 2 at time T_{02}. For subjects entered in state 1 at T_{01}, they will enter in state 2 at time $T_{01} + T_{12}$. The process can be summarized by two r.v. T_0, T_2. We define the exit time from state 0

$$T_0 = \inf_{t>0}\{E(t) \neq 0\} = \min\big(T_{01}, T_{02}\big)$$

together with D_0 that indicates the entered state ($D_0 = 1$ or 2), and the entry time to state 2

$$T_2 = \inf_{t>0}\{E(t) = 2\} = T_0 + \mathbb{1}\{D_0 = 1\}T_{12}$$

that characterizes the total survival time.

In general, the process associated with each of the three transitions might depend on the time of arrival in the state (Markov process) or on the time since the entry to the state (Semi-Markov process). Here, for transition $0 \to 1$ and $0 \to 2$, we consider time non-homogeneous markovian processes. For transition $1 \to 2$, we perform a time transformation, following Andersen et al. (2012), and we consider a time homogeneous semi-markovian process (the probability of transiting from state 1 to state 2 at time t depends only on the duration $d = t - T_0$ already spent in 1). Wherever convenient, we use the duration variable d instead of t. Under these assumptions, we aim to model the transition-specific density probabilities over time. We define f_{01}, f_{02} as the infinitesimal probabilities of experiencing respectively transitions $0 \to 1$, $0 \to 2$,

$$f_{0l}(t) = \lim_{h\to 0}\frac{1}{h}\mathbb{P}\big(t \leq T_0 \leq t + h, D_0 = l\big), \text{ for } l = 1, 2.$$

We define F_{01}, F_{02} their cumulative counterparts such that

$$F_{0l}(t) = \mathbb{P}\big(T_0 \leq t, D_0 = l\big) = \int_0^t f_{0l}(t)dt, \text{ for } l = 1, 2,$$

expresses the probability that a transition $0 \to l$ occurs on or before time t. We also define $f_0(t) = f_{01}(t) + f_{02}(t)$ (resp. $F_0(t) = F_{01}(t) + F_{02}(t)$) as the infinitesimal probability of leaving state 0 at time t (resp. on or before time t). For transition $1 \to 2$, the functions of interest are defined conditionally to T_0, $D_0 = 1$. To simplify the notations, we drop this conditioning in the definitions. We define f_{12} as the infinitesimal probability of experiencing transition $1 \to 2$, such that for $f_{12}(d) := f_{12}(d \,|T_0,\ D_0 = 1)$,

$$f_{12}(d) = \lim_{h\to 0}\frac{1}{h}\mathbb{P}\big(d \leq T_2 - T_0 \leq d + h\big),$$

We define F_{12} as its cumulative counterpart such that, for $F_{12}(d) := F_{12}(d|T_0, D_0 = 1)$,

$$F_{12}(d) = \mathbb{P}\big(T_2 - T_0 \leq d\big) = \int_0^d f_{12}(d)dd,$$

expresses the probability that a transition $1 \to 2$ occurs on or before time d conditionally to $T_0, D_0 = 1$. F_{01}, F_{02} and F_{12} are commonly referred to cumulative incidence functions (CIFs) in the literature (Meira-Machado and Sestelo, 2019) and are our main functions of interest to predict.

We formulate our illness-death process through these probabilities. However illness-death processes are conventionally defined through transition intensities (Andersen and Borgan, 1984). Both formalisms are related.

2.2. A piecewise-constant approach

We now assume that the transition-specific density probabilities are piecewise constant, following Kvamme and Borgan (2019) and Friedman et al. (1982). This is an approximation and it is well-established that the approximation error can be bounded when the true functions are smooth (Triebel, 1983). Most of the existing survival networks assume discrete time modelization in order to simplify the optimization of the likelihood (see Section 3.2). However, the approximation error that arise when a discrete-time method is used can be reduced with piecewise constant approximations, with no supplementary computational costs.

First, we define τ as the maximum horizon time window. We divide the time axis into K disjoint time intervals: $v_1 = [a_0, a_1), \cdots, v_K = [a_{K-1}, a_K)$, with $a_0 = 0$ and $a_K = \tau$. For any time $s \in [0, \tau[$, we denote by $v_{k(s)}$ the time interval to which s belongs. Assuming that the density probabilities are constant within each interval, we can express f_{0l} $(l = 1, 2)$, f_{12} as step functions such that $f_{0l}(t) = f_{0l}(v_{k(t)})$ and $f_{12}(d) = f_{12}(v_{k(d)})$. We refer the reader to Appendix A for the exact expressions of F_{0l} $(l = 1, 2)$, F_{12} under the piecewise constant assumption.

In real clinical data the r.v. T_0, T_2 can take values after τ (a patient can leave state 0 or state 1 after τ). Hence, under the piecewise constant assumption, the following equations are satisfied

$$\sum_{k=1}^{K} f_0(v_k)|v_k| + 1 - F_0(\tau) = 1, \ \sum_{k=1}^{K} f_{12}(v_k)|v_k| + 1 - F_{12}(\tau) = 1, \tag{1}$$

where $|v_k|$ is the length of interval v_k. See Lee et al. (2018); Kvamme and Borgan (2019) for similar remarks.

2.3. Illness-death data

In clinical settings, patient characteristics are observed as P-dimensional covariates. In addition, right-censoring has to be taken into account. Let C be a non-negative censoring r.v. independent of (T_0, T_2) that precludes its observation. Let $\tilde{T}_0 = \min(T_0, C)$ and $\tilde{T}_2 = \min(T_2, C)$ be the observed event times. Together with these times, we observe binary labels indicating the status of the transitions: $\delta_{0l} = \mathbb{1}\{D_0 = l, T_0 \leq C\}$ $(l = 1, 2)$, $\delta_{12} = \delta_{01}\mathbb{1}\{T_2 \leq C\}$, where $\delta_{kl} = 1$ indicates an entry in state l from k and $\delta_{kl} = 0$ indicates a censored transition.

We observe n independent and identically distributed r.v. in $\mathbb{R}^P \times \mathbb{R}_+ \times \{0, 1\} \times \{0, 1\} \times \mathbb{R}_+ \times \{0, 1\}$:

$$\mathcal{D}_i = \left\{ X_i, (\tilde{T}_0^i, \delta_{01}^i, \delta_{02}^i), (\tilde{T}_2^i, \delta_{12}^i) \right\}_{1 \leq i \leq n},$$

where $X_i = \left(X_{i1}, \cdots, X_{iP}\right)^T$ is a vector of P covariates observed at baseline. From these observations and for each subject i, we aim to estimate the true transition specific density probabilities conditionally to the clinical features X_i in order to predict the individual CIFs. We note $f_{0l}(.|X_i)$ $(l = 1, 2)$, $f_{12}(.|X_i)$ and their cumulative counterpart $F_{0l}(.|X_i)$ $(l = 1, 2)$, $F_{12}(.|X_i)$.

2.4. Definition of the log-likelihood

Under the assumption of a piecewise constant model, we show that the conventional illness-death log-likelihood (Andersen et al., 2012) ℓ can be rewritten in terms of the density probabilities. We define ℓ by dividing the contributions in two distinct parts:

$$\ell = \frac{1}{n} \sum_{i=1}^{n} \left[\ell_0^i + \ell_1^i\right]. \tag{2}$$

ℓ_0^i is the log contribution of patient i from state 0. From state 0, patient i with an event at time $\tilde{T}_0^i$ can contribute in three ways. He can (i) experience a transition $0 \to 1$; (ii) experience a transition $0 \to 2$; (iii) be censored at $\tilde{T}_0^i$. Thus ℓ_0^i is given by

$$\ell_0^i = \sum_{l=1,2} \left\{\delta_{0l}^i \log\left(f_{0l}\left(v_{k(\tilde{T}_0^i)} \,|X_i\right)\right)\right\} + \left(1 - (\delta_{01}^i + \delta_{02}^i)\right) \log\left(1 - F_0(\tilde{T}_0^i \,|X_i)\right).$$

On the other hand, ℓ_1^i is the log contribution of patient i from the time he has entered state 1 (only for i such that $\delta_{01}^i = 1$). Following the previous reasoning, ℓ_1^i is given by

$$\ell_1^i = \delta_{01}^i \delta_{12}^i \log\left(f_{12}\left(v_{k(\tilde{T}_2^i - \tilde{T}_0^i)} \,|X_i\right)\right) + \delta_{01}^i (1 - \delta_{12}^i) \log\left(1 - F_{12}(\tilde{T}_2^i - \tilde{T}_0^i \,|X_i)\right).$$

3. Description of IDNetwork

IDNetwork is a deep learning architecture tuned to estimate the step probability functions f_{01}, f_{02}, f_{12} over the interval $[0, \tau]$ by capturing possible non-linear relations between covariates and transition probabilities.

3.1. Network architecture

Inspired by the work of Lee et al. (2018), we develop an architecture (see an illustration in Figure 2) with three task-specific sub-networks that are related to the three transitions of an illness-death process. Multi-task learning is done with hard parameter sharing (Ruder, 2017) in order to extract common and specific patterns from the patient's characteristics (ie. the baseline covariates). It is composed of a first subnetwork shared between the three transitions and of three transition-specific subnetworks. Two different softmax output layers are used to transform the transition-specific subnetworks outputs into time-dependent probabilities.

Input layer The input layer is composed of the matrix $\mathbf{X}$ of P baseline covariates for the n individuals.

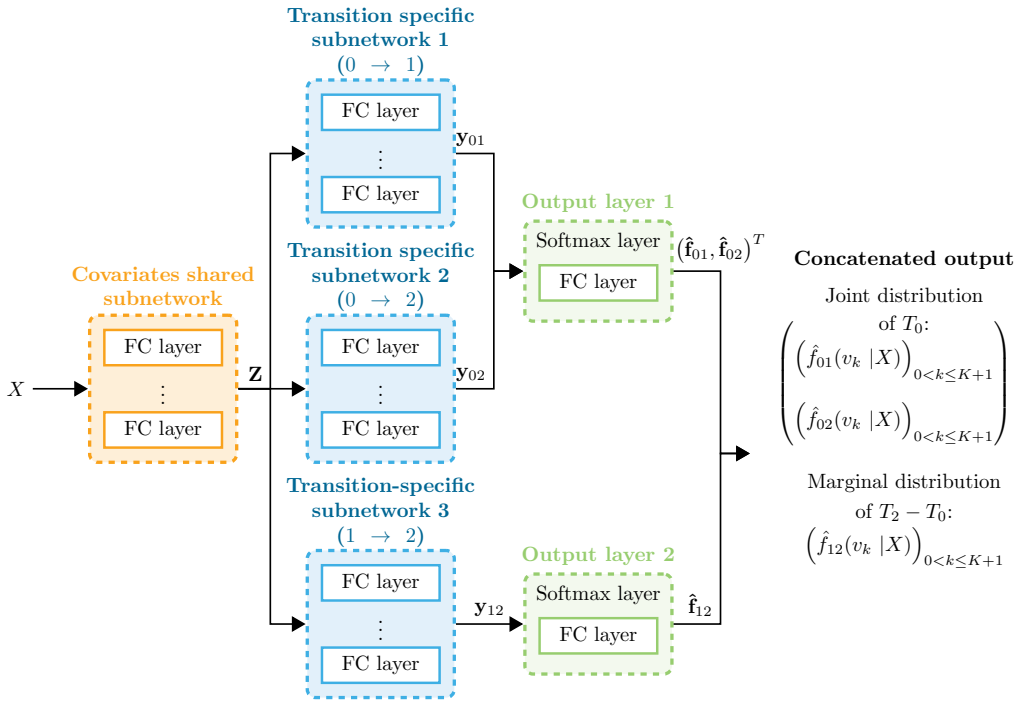


Figure 2: IDNetwork architecture. "FC layer" refers to fully-connected layer.

Covariates shared subnetwork The shared subnetwork takes as input the *input layer* and contains L fully connected hidden layers with l units. Its output is a vector $\mathbf{z} = g^{\text{input}}(X)$ in $\mathbb{R}^l$ that captures shared patterns between the three transitions (g^{input} is a non linear activation function).

Transition-specific subnetworks Each transition-specific subnetwork takes as input $\mathbf{z}$ and contains L^{kl} fully-connected hidden layers with l^{kl} units. Its output is a vector $\mathbf{y}_{kl} = g^{kl}(\mathbf{z})$, that is a transition-specific transformation of the shared features (g^{kl} is a non linear activation function). Given the unbalanced number of observations for the three transitions that may exist in real data, the range of model complexity is different for each of the three transitions. To find the best configuration, we set the structure of each subnetwork independently (see Appendix F).

Probabilistic output layers The output of the network is composed of two probabilistic layers that map the transition-specific outcomes $\mathbf{y}_{kl}$ into time-dependent probabilities. The first layer is related to the exit from state 0 (ie. the competing transitions $0 \to 1$ and $0 \to 2$). The second layer is related to the marginal transition $1 \to 2$. The two layers are independent as the first layer learns the joint distribution of T_0 and the second layer learns the marginal distribution of $T_2 - T_0 | T_0, D_0 = 1$.

Under the constraint of Equation (1), we consider a supplementary interval $v_{K+1} = [\tau, +\infty)$. Consequently, we fulfill the constraint by defining $1 - F_0(\tau) = f_0(v_{K+1})$, $1 - F_{12}(\tau) = f_{12}(v_{K+1})$. Hence, each output of the network is a fully-connected layer which

(i) use a linear activation function g^{linear} to transform $(\mathbf{y}_{01},\ \mathbf{y}_{02})^T$ (resp. $\mathbf{y}_{12}$) into a vector $\phi_0 = (\phi_{01},\ \phi_{02})^T \in \mathbb{R}^{2(K+1)}$ (resp. a vector $\phi_{12} \in \mathbb{R}^{K+1}$), and then (ii) use a weighted softmax activation function σ_0 (resp. σ_2) to transform ϕ_0 (resp. ϕ_{12}) into probabilities and provide an estimation of f_{01}, f_{02} (resp. f_{12}) in each time interval. Thus, the output layers are characterized by the vectors

$$\hat{\mathbf{f}}_0 = \left(\hat{\mathbf{f}}_{01}, \hat{\mathbf{f}}_{02}\right)^T = \sigma_0\left(g^{\text{linear}}\left((\mathbf{y}_{01}, \mathbf{y}_{02})^T\right)\right) = \sigma_0\left((\phi_{01},\ \phi_{02})^T\right),$$
$$\hat{\mathbf{f}}_{12} = \sigma_2\left(g^{\text{linear}}\left(\mathbf{y}_{12}\right)\right) = \sigma_2\left(\phi_{12}\right),$$

where $\hat{\mathbf{f}}_{kl} = \left(\hat{f}_{kl}(v_k\ |X)\right)_{0<k\leq K+1}$ and $\sigma_0,\ \sigma_2$ are two softmax functions weighted by the length of the time intervals such that

$$\hat{f}_{0l}(v_k\ |X) = \frac{\exp\left[\phi^k_{0l}(X)\right]}{\sum_{j=1}^{K+1}\left(\exp\left[\phi^j_{01}(X)\right] + \exp\left[\phi^j_{02}(X)\right]\right)|v_j|}, \text{ for } l = 1, 2,$$
$$\hat{f}_{12}(v_k\ |X) = \frac{\exp\left[\phi^k_{12}(X)\right]}{\sum_{j=1}^{K+1}\exp\left[\phi^j_{12}(X)\right]|v_j|}.$$

3.2. Loss function and mitigation of the number of time intervals effect via penalization

To learn IDNetwork parameters, we minimize a total loss function,

$$\ell_{total} = -\ell^{K+1} + P_\lambda, \tag{3}$$

that sums the negative log-likelihood and a penalization term. The first term ℓ^{K+1} is a revising of the log-likelihood ℓ defined in Equation (2) under the constraint of Equation (1). The second term P_λ is a penalization term related to ℓ^{K+1} allowing to smooth the effect of a non-optimal number of time intervals (ie. a non optimal value for K). The choice of K has a significant impact on the performance: the number of nodes grows with K, which might cause over-fitting (for large value of K) or under-fitting. The optimal selection of K can be fixed by applying a temporal smoothing technique. Following Möst (2014) and Tibshirani et al. (2005), we apply a temporal smoothness constraint by penalizing, in the weight matrices (resp. the bias vectors) of the output layers, the first order differences of the weights (resp. the bias) associated with two adjacent time intervals (see definition in Appendix B). The penalization term limits over-fitting for larger values of K.

The optimization of the log-likelihood is facilitated by the use of piecewise constant model. Indeed, a continuous model would have necessitated the use of the Cox partial log-likelihood (Cox, 1972) that significantly impact on the computational cost in the gradient descent because of the presence of two cumulative sums (see Achab et al. (2015) and Kvamme et al. (2019) for more details).

4. Prediction task and benchmark

4.1. Individual disease progression predictions

In this subsection, we define the predictions of interest according to the time scales defined below. From the output of our network (ie. the functions $\hat{f}_{0l}(.|\, X)$ $(l = 1, 2)$, $\hat{f}_{12}(.|\, X)$), we can derive the estimation of the CIFs. For a new patient j with the baseline covariates X_j, we note the estimated CIFs, derived from Equations (4) and (5) in Appendix A, as $\hat{F}_{0l}(.|\, X_j)$ $(l = 1, 2)$, $\hat{F}_{12}(.|\, X_j)$. We refer the reader to Appendix C for their exact expression.

We will use the estimated CIFs to assess the predictive performance of IDNetwork.

4.2. Predictive performance criteria

In event history analysis, commonly used performance measures are the time-dependent AUC (for discrimination) and the time-dependent Brier score (BS) (for calibration). On the basis of the transition-specific time properties defined in Section 2.1, we adapt the definitions of the time-dependent AUC (Jacqmin-Gadda et al., 2016) and the time-dependent BS (Spitoni et al., 2018). To take into account the lost of information due to censoring, we use estimators based on the inverse probability of censoring weight (IPCW) methods. The time-dependent AUC and the time-dependent BS can be extended to the interval $]0, \tau]$ by computing respectively the integrated AUC (iAUC) and the integrated BS (iBS). We refer the reader to Appendix D for the exact definitions of the criteria.

4.3. Benchmark and validation

Predictive performances of IDNetwork in predicting the CIFs are compared in terms of discrimination (with the iAUC) and calibration (with the iBS) with two state-of-the-art statistical methods: the multi-state Cox P.H model (**msCox**) from the R library `mstate` (De Wreede et al., 2010) and a spline-based version of the multi-state Cox P.H. model (**msSplineCox**) from the R library `flexsurv` (Jackson, 2016). We also compare IDNetwork with a simplified linear version of IDNetwork (**LinearIDNetwork**) (see an illustration in Appendix E). We perform two sets of experiments on (1) a simulated data set and (2) on three real clinical data sets. We score predictive performance of the methods through internal validation (Royston and Altman, 2013) by randomly splitting $M = 50$ times each data set $\mathcal{D}_m$ $(m = 1, \cdots, M)$ into a training/testing data set $\mathcal{D}_m^{\text{train}}$ (68% for training / 12% for early stopping) and a validation data set $\mathcal{D}_m^{\text{val}}$ (20%). For each split m, we perform $R = 20$ random hyper-parameters searches (Bergstra and Bengio, 2012) and we choose the set of hyper-parameters associated with the best iAUC (averaged across the three transitions) on the held-out validation sets. Experimental details on IDNetwork's hyper-parameters tuning are given in Appendix F. We compare the median ($\pm$ standard deviation (sd)) iAUC (higher the better) and iBS (lower the better) on the validation sets. We statistically compare performances of IDNetwork over the three other methods using a bilateral Wilcoxon (Wilcoxon, 1992) signed rank test. In the results, $\cdot$ indicates a p-value less than 0.1, † less than 0.05, ‡ less than 0.01, * less than 0.001.

5. Experiments on a simulated data set

5.1. Data simulation

We generate a data set with $n = 5000$ observations and $\tau = 100$. We simulate four 2-dimensional baseline covariates, each drawn from a multivariate Gaussian distribution. We then simulate continuous illness-death times through Cox transition-specific non-linear (quadratic) risk functions, see simulation details in Appendix G. In this simulation scheme, the Cox's linear assumption doesn't hold.

5.2. Predictive performance

The predictive performances are shown in Tables 1. Detailed results per evaluation time are displayed in Appendix H. In this simulated data set, the Cox's linear assumption doesn't hold anymore. Consequently, as expected, IDNetwork significantly outperforms msCox and msSplineCox with a p-value less than 0.001 in terms of iAUC and iBS (except for the transition $1 \to 2$ where msCox outperforms IDNetwork but with no statistical difference). IDNetwork significantly outperforms the linear version LinearIDNetwork as well. IDNetwork significanlty outperforms the linear version LinearIDNetwork as well.

Table 1: Predictive performance (median $\pm$ sd) on the validation sets (internal validation) for the non-linear simulated data set. We integrate the AUC and BS measures at all the 4 equidistant time points in $[0, \tau]$ (for computational cost reasons).

Criteria	**Algorithm**	**Transition**		
		$\mathbf{0 \to 1}$	$\mathbf{0 \to 2}$	$\mathbf{1 \to 2}$
iAUC	msCox	$0.533^{*} \pm 0.03$	$0.481^{*} \pm 0.03$	$0.489^{*} \pm 0.03$
	msSplineCox	$0.532^{*} \pm 0.03$	$0.479^{*} \pm 0.03$	$0.486^{*} \pm 0.03$
	LinearIDNetwork	$0.531^{*} \pm 0.03$	$0.510^{\ddagger} \pm 0.03$	$0.504^{*} \pm 0.03$
	IDNetwork	$\mathbf{0.580 \pm 0.03}$	$\mathbf{0.525 \pm 0.03}$	$\mathbf{0.566 \pm 0.03}$
iBS	msCox	$0.234^{*} \pm 0.01$	$0.243^{*} \pm 0.01$	$\mathbf{0.159 \pm 0.01}$
	msSplineCox	$0.242^{*} \pm 0.01$	$0.251^{*} \pm 0.01$	$\mathbf{0.159 \pm 0.01}$
	LinearIDNetwork	$0.162^{*} \pm 0.01$	$0.161^{*} \pm 0.01$	$0.165^{\dagger} \pm 0.01$
	IDNetwork	$\mathbf{0.146 \pm 0.01}$	$\mathbf{0.144 \pm 0.01}$	0.160 ± 0.01

6. Application on real clinical data sets

6.1. Description of the data sets

We conduct experiments on real data from two clinical trials in colon cancer and one clinical trial in breast cancer. A more detailed description of the data sets is given in Appendix J.

Clinical trials in colon cancer We use two data sets from Phase III clinical trials evaluating endpoints relapse-free survival (RFS) and overall-survival (OS) in non-metastatic colon cancer. (1) The study NCT00079274 contains 2121 observed patients followed for 60 months for RFS and for 96 months for OS. (2) The study NCT00275210 contains 1122

patients followed over 60 months for RFS and OS. We select 9 baseline clinical covariates shared between the two studies. The study NCT00079274 will be use for internal validation (training and validation) and NCT00275210 for external validation.

Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort This data set contains clinical, histo-pathological, gene copy number and gene expression features used to determine breast cancer subgroups. We include 1903 patients followed for 360 months for the endpoints RFS and OS. We select 17 baseline clinical features and we add 100 gene expression features.

6.2. Predictive performance

The integrated predictive performances are shown in Tables 2 and 3. Results per evaluation time are displayed in Appendix J. On the colon data sets, we conduct (1) internal and (2) external validation. On the validation splits, neither methods outperforms the other for the three transitions. However, IDNetwork shows better performance on the external NCT00275210 validation study for all the transitions (excluding in terms of iAUC for the transition $1 \to 2$ where msSplineCox outperforms IDNetwork, but with no statistical significance). On the METABRIC cohort, IDNetwork significantly outperforms msCox for transitions $0 \to 1$ and $1 \to 2$. IDNetwork and LinearIDNetwork show bad performance for the transition $0 \to 2$. This is, in our opinion, due to the fact that experiencing transition $0 \to 2$ means that the patient died from another cause than cancer, but the clinical and biological features provided to the model were not related (excluding the age) to non-cancer causes of death.

7. Conclusion and discussion

We present IDNetwork a novel method to model an illness-death process and to predict two-stages evolution of a disease based on baseline covariates. IDNetwork uses a multi-task architecture to learn an estimation of the probabilities of occurrence of state transitions of an illness death process without any assumption on the relation between covariates and risks of transition.

We benchmark the predictive performance of our method with the state-of-the-art methods and show, through experiments conducted on a simulated data set and on real data sets on colon and breast cancer, that IDNetwork provides significant improvements in terms of discrimination and calibration when non-linear patterns are found in the data. The architecture of IDNetwork can learn transition-specific patterns in the data, but can suffer of over-fitting when the training size is too small. Hence, We adapt the architecture of IDNetwork to handle these limitations.

Medical decision making requires to combine heterogeneous individual features. IDNetwork can be easily adapted to integrate various types of data (as images, biological, clinical data). For the future work, it may be relevant to add an interpretability functionality to IDNetwork. It could reveal what are the patient characteristics associated with each transition and increase the understanding of the evolution of the disease.

In clinical practice, IDNetwork may be useful in personalized medicine by providing prediction of the risks of relapse and death. It could help physicians to adapt the therapeutic

Table 2: Predictive performance (median ± sd) for the data sets NCT00079274, NCT00275210 on colon cancer on (1) the validation sets (internal validation), (2) the external NCT00275210 test set (external validation). We integrate the AUC and BS measures at all the 30 equidistant time points in $[60, \tau]$ (ie. at every month from 2 months).

Evaluation	**Criteria**	**Algorithm**	**Transition**		
			0 → 1	**0 → 2**	**1 → 2**
(1) Internal	iAUC	msCox	$\mathbf{0.686 \pm 0.03}$	$0.633^{\ddagger} \pm 0.09$	0.670 ± 0.04
		msSplineCox	$\mathbf{0.686 \pm 0.03}$	$0.629^{\ddagger} \pm 0.10$	$\mathbf{0.679 \pm 0.04}$
		LinearIDNetwork	0.671 ± 0.03	$0.637^{\ddagger} \pm 0.10$	$0.663^{\cdot} \pm 0.04$
		IDNetwork	0.676 ± 0.03	$\mathbf{0.673 \pm 0.09}$	0.677 ± 0.04
	iBS	msCox	$\mathbf{0.152 \pm 0.01}$	$0.029^{\ddagger} \pm 0.01$	0.196 ± 0.02
		msSplineCox	$\mathbf{0.152 \pm 0.01}$	$0.029^{\ddagger} \pm 0.01$	$\mathbf{0.192^{\cdot} \pm 0.03}$
		LinearIDNetwork	0.153 ± 0.01	$\mathbf{0.026 \pm 0.01}$	0.207 ± 0.03
		IDNetwork	$\mathbf{0.152 \pm 0.01}$	$\mathbf{0.026 \pm 0.01}$	0.203 ± 0.03
(2) External	iAUC	msCox	$0.669^{*} \pm 0.00$	$0.607^{*} \pm 0.02$	0.562 ± 0.02
		msSplineCox	$0.669^{\ddagger} \pm 0.00$	$0.605^{*} \pm 0.03$	$\mathbf{0.565 \pm 0.01}$
		LinearIDNetwork	0.672 ± 0.01	0.694 ± 0.03	$0.543^{*} \pm 0.02$
		IDNetwork	$\mathbf{0.674 \pm 0.01}$	$\mathbf{0.698 \pm 0.05}$	0.558 ± 0.03
	iBS	msCox	$0.157^{*} \pm 0.00$	$0.013^{*} \pm 0.00$	$0.182^{*} \pm 0.01$
		msSplineCox	$0.159^{*} \pm 0.00$	$0.013^{*} \pm 0.00$	$0.172^{*} \pm 0.01$
		LinearIDNetwork	$\mathbf{0.154 \pm 0.00}$	$\mathbf{0.011 \pm 0.00}$	$\mathbf{0.146 \pm 0.01}$
		IDNetwork	$\mathbf{0.154 \pm 0.01}$	$\mathbf{0.011 \pm 0.00}$	$\mathbf{0.146 \pm 0.01}$

Table 3: Predictive performance (median ± sd) on the validation sets (internal validation) for the METABRIC data set. We integrate the AUC and BS measures at all the 30 equidistant time points in $[90, \tau]$ (ie. at every month from 3 months).

Criteria	**Algorithm**	**Transition**		
		0 → 1	**0 → 2**	**1 → 2**
iAUC	msCox	0.702 ± 0.03	$\mathbf{0.734^{*} \pm 0.05}$	$0.689^{*} \pm 0.04$
	msSplineCox	0.706 ± 0.03	$0.718^{\dagger} \pm 0.06$	$0.693^{\ddagger} \pm 0.04$
	LinearIDNetwork	$0.697^{\ddagger} \pm 0.03$	$0.625^{*} \pm 0.04$	0.718 ± 0.03
	IDNetwork	$\mathbf{0.711 \pm 0.03}$	0.672 ± 0.04	$\mathbf{0.728 \pm 0.04}$
iBS	msCox	$0.150^{*} \pm 0.01$	$0.067^{*} \pm 0.01$	$0.180^{*} \pm 0.01$
	msSplineCox	$0.152^{\ddagger} \pm 0.02$	$0.068^{*} \pm 0.01$	$0.186^{*} \pm 0.01$
	LinearIDNetwork	$0.150^{*} \pm 0.01$	0.060 ± 0.01	0.168 ± 0.02
	IDNetwork	$\mathbf{0.142 \pm 0.01}$	$\mathbf{0.057 \pm 0.01}$	$\mathbf{0.165 \pm 0.02}$

guidelines for a specific patient. IDNetwork is a flexible method developed for an illness-death process and can readily be applied in many cancers to predict two-stages evolution. It can be generalized to embrace more complex disease evolution patterns by adapting the states and transitions (Jackson, 2007).

Acknowledgement

This publication is based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere. Neither Project Data Sphere nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this publication.

References

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

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

Per Kragh Andersen and Ørnulf Borgan. Counting process models for life history data: A review. *Preprint series. Statistical Research Report http://urn. nb. no/URN: NBN: no-23420*, 1984.

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.

Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.

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

Liesbeth C De Wreede, Marta Fiocco, and Hein Putter. The mstate package for estimation and prediction in non-and semi-parametric multi-state 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.

David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*, 2018.

Michael Friedman et al. Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10(1):101–113, 1982.

Simona Iacobelli and Bendix Carstensen. Multiple time scales in multi-state models. *Statistics in medicine*, 32(30):5315–5327, 2013.

Christopher Jackson. Multi-state modelling with r: the msm package. *Cambridge, UK*, pages 1–53, 2007.

Christopher H Jackson. flexsurv: a platform for parametric survival modeling in r. *Journal of statistical software*, 70, 2016.

Hélène Jacqmin-Gadda, Paul Blanche, Emilie Chary, Célia Touraine, and Jean-François Dartigues. Receiver operating characteristic curve estimation for time to event with semicompeting risks and interval censoring. *Statistical methods in medical research*, 25 (6):2750–2766, 2016.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

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

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724*, 2019.

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 R Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

Stephanie Möst. *Regularization in discrete survival models*. PhD thesis, lmu, 2014.

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

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

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.

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.

Hans Triebel. Theory of function spaces, vol. 78 of. *Monographs in mathematics*, 1983.

Wessel N Van Wieringen, David Kun, Regina Hampel, and Anne-Laure Boulesteix. Survival prediction using gene expression data: a review and comparison. *Computational statistics & data analysis*, 53(5):1590–1603, 2009.

Anthony J Webster. Multi-stage models for the failure of complex systems, cascading disasters, and the onset of disease. *PloS one*, 14(5):e0216422, 2019.

Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

Safoora Yousefi, Fatemeh Amrollahi, Mohamed Amgad, Chengliang Dong, Joshua E Lewis, Congzheng Song, David A Gutman, Sameer H Halani, Jose Enrique Velazquez Vega, Daniel J Brat, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1):1–11, 2017.

Appendix A. Exact definitions of the CIFs under a piecewise constant model

Under the assumption of a piecewise constant model, the cumulative incidence functions F_{0l} $(l = 1, 2)$ and F_{12} are piecewise linear and have the following expressions

$$F_{0l}(t) = \sum_{k=1}^{k(t)-1} f_{0l}(v_k)|v_k| + (t - a_{k(t)-1}) f_{0l}(v_{k(t)}), \tag{4}$$

$$F_{12}(d) = \sum_{k=1}^{k(d)-1} f_{12}(v_k)|v_k| + (d - a_{k(d)-1}) f_{12}(v_{k(d)}). \tag{5}$$

where $|v_k|$ is the length of interval v_k. See Kvamme and Borgan (2019) (Section 4.3) for a similar remark in the at-most one event case.

Appendix B. Definition of the penalizaton term P_λ

Let's consider

$$W = (W^1, W^{12})^T, \ B = (B^1, B^{12})^T$$

the weight and bias parameters associated with both output layers, with

$$W^1 \in \mathbb{R}^{(l^{01}+l^{02})\times 2(K+1)}, \ W^{12} \in \mathbb{R}^{l^{12}\times(K+1)}$$

the weights matrices,

$$B^1 \in \mathbb{R}^{2(K+1)}, \ B^{12} \in \mathbb{R}^{K+1}$$

the bias vectors. For $k = 1 \ldots, K$ we compute

$$\Delta_{w_{j,k}^{kl}} = w_{j,k+1}^{kl} - w_{j,k}^{kl},\ \Delta_{b_k^{kl}} = b_{k+1}^{kl} - b_k^{kl},$$

the weight and bias differences associated with the transition $k \to l$, the neuron j and the adjacent time intervals $v_k,\ v_{k+1}$. Then the penalty term of our loss function in (3) has the form

$$P_\lambda(B, W) = \sum_{kl} \Big(\lambda_w^{kl} \sum_{j=1}^{l^{kl}} \sum_{k=1}^{K} \Big| \Delta_{w_{j,k}^{kl}} \Big| + \lambda_b^{kl} \sum_{k=1}^{K} \Big| \Delta_{b_k^{kl}} \Big| \Big),$$

where λ_w^{kl} and λ_b^{kl} are transition-specific positive constants determining the amount of smoothing to be applied for each transition. For $\lambda_w^{kl} \to +\infty$ (respectively $\lambda_b^{kl} \to +\infty$), all differences will be set to zero resulting in constant weights (respectively constant bias). This penalization term allows to minimize the risk of over-fitting, for the three transitions independently, for larger values of K.

Appendix C. Exact expressions of our predictions $\hat{F}_{0l}(.|\ X_j)$ $(l = 1, 2)$ and $\hat{F}_{12}(.|\ X_j)$

For a new patient j, with the set of covariates X_j, the exact expression of the estimated CIFs is given by the following equations:

$$\begin{aligned}
\hat{F}_{0l}\big(t|X_j\big) &= \sum_{k=1}^{k(t)-1} \hat{f}_{0l}\big(v_k|X_j\big)|v_k| + (t - a_{k(t)-1})\hat{f}_{0l}\big(v_{k(t)}|X_j\big), \text{ for } l = 1, 2, \\
\hat{F}_{12}\big(d|X_j\big) &= \sum_{k=1}^{k(d)-1} \hat{f}_{12}\big(v_k|X_j\big)|v_k| + (d - a_{k(d)-1})\hat{f}_{12}\big(v_{k(d)}|X_j\big),
\end{aligned}$$

this is just a rewriting of Equations (4) and (5) with estimation.

Appendix D. The predictive performance criteria

The transition-specific time-dependent AUC measures, for two patients i and j, is the probability that a patient i who experienced the transition kl before time t has greater probability of occurrence of the transition than a patient j who has survived to the transition. It is defined as the integration of the ROC curve opposing specificity (Sp) and sensitivity (Se):

$$\begin{aligned}
\text{AUC}^{0l}(t) &= \mathbb{P}\Big(F_{0l}\left(t\ |X_i\right) > F_{0l}\left(t\ |X_j\right)\ |\ T_0^i \le t,\ T_0^j > t,\ D_0^i = l \Big) \\
&= \int_0^t \text{Se}_t^{0l}\left(\left(1 - \text{Sp}_t^{0l}\right)^{-1}(p) \right) \mathrm{d}p, \text{for } l = 1, 2, \\
\text{AUC}^{12}(d) &= \mathbb{P}\Big(F_{12}\left(d\ |\ X_i\right) > F_{12}\left(d\ |\ X_j\right) |\ T_2^i - T_0^i \le d,\ T_2^j - T_0^j > d, D_0^i = 1,\ D_0^j = 1 \Big) \\
&= \int_0^d \text{Se}_d^{12}\left(\left(1 - \text{Sp}_d^{12}\right)^{-1}(p) \right) \mathrm{d}p.
\end{aligned}$$

The transition-specific time-dependent Brier score measures the difference between the predicted probability of occurrence of the transition at time t and the status of the transition:

$$\mathrm{BS}^{0l}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{1}\{T_0^i > t\} - F_{0l}\left(t \mid X_i\right)\right]^2 \text{ for } l = 1, 2,$$
$$\mathrm{BS}^{12}(d) = \frac{1}{n_{12}}\sum_{i:D_0^i=1}\left[\mathbb{1}\{T_2^i - T_0^i > d\} - F_{12}\left(d \mid X_i\right)\right]^2,$$

where n_{12} is the number of patients at risk for transition $1 \to 2$.

The time-dependent AUC and the time-dependent BS can be extended to the interval $]0, \tau]$ by computing respectively the integrated AUC (iAUC) and the integrated Brier score (iBS) as follows:

$$\mathrm{iAUC}^{kl} = \frac{1}{\tau}\int_0^\tau \mathrm{AUC}^{kl}(t)dt, \ \mathrm{iBS}^{kl} = \frac{1}{\tau}\int_0^\tau \mathrm{BS}^{kl}(t)dt.$$

Appendix E. LinearIDNetwork : a linear version of IDNetwork

The architecture of LinearIDNetwork, a simplified linear version of IDNetwork with no covariates shared subnetwork and no transition-specific subnetworks, is illustrated in Figure 3.

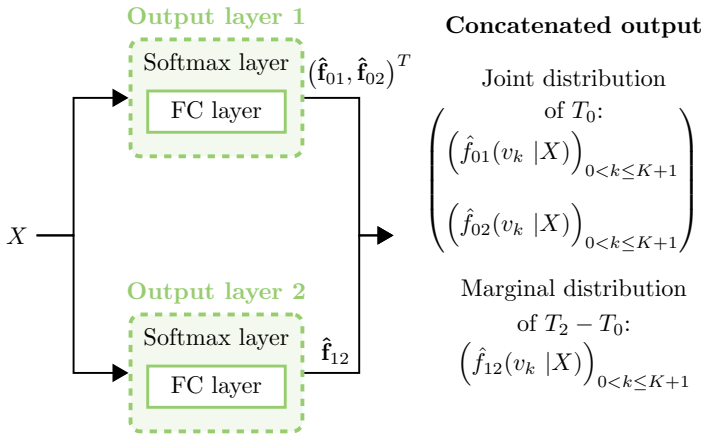


Figure 3: LinearIDNetwork architecture. "FC layer" refers to fully-connected layer.

Appendix F. Experimental details of IDNetwork

In this section, we give additional details on the implementation and the optimization settings of IDNetwork. IDNetwork is implemented in Python within a Tensorflow environment[1]. It uses standard deep learning techniques as L_2 regularized layers to avoid over-fitting, L_1 regularized output layers, Xavier Gaussian initialization schemes, Adam optimizer, mini-batch learning, learning rate weight decay, early stopping. IDNetwork is

1. https://www.tensorflow.org/

optimized in two steps including the hyper-parameters optimization and the parameters optimization.

Hyper-parameters of IDNetwork include number of nodes and hidden layers in each subnetworks, regularization penalty parameters, output-specific penalization parameters, activation functions for each subnetwork, etc. A complete list of hyper-parameters are given in Table 4. Hyper-parameters are tuned using Random Search (Bergstra and Bengio, 2012). For each hyper-parameter, we fix a discrete space of search by manual search. For each model, the sets of search spaces are adjusted according to the data set in input.

Definition*	**Discrete search space**
Initialization	Xavier initialization
Optimizer	Adam Optimizer
Learning rate	10^{-4}
Mini-batch size	{8, 16, 32, 64, 128}
Nodes per shared hidden layer (l)	{15, 30, 50}
No. shared hidden layers (L)	{1, 2, 3}
Nodes per transition-specific hidden layer (l^{kl})	{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	$\{1^{-7}, 1^{-6}, 1^{-5}, 1^{-4}, 1^{-3}, 1^{-2}\}$
Decay for gradient descent	$\{1^{-3}, 1^{-2}, 0.1, 0.4, 0.6, 0.8, 1.0\}$
Parameter difference for early stopping	$\{1^{-6}, 1^{-5}, 1^{-4}\}$
Output-specific penalization (λ_w^{kl}, λ_b^{kl})	$\{1^{-5}, 1^{-4}, 1^{-3}, 1^{-2}, 0.1\}$

*references to notations in the paper: transition $(k,l) \in \{(01),(02),(12)\}$.

Table 4: Hyper-parameters of IDNetwork.

Appendix G. Additional details on data simulation

For each observation i $(1 \leq i \leq n)$ we simulate four 2-dimensional baseline variables, each drawn from a multivariate Gaussian distribution with mean 0 and a matrix of variance covariance Σ_p:

$$X_i = \left(X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, X_i^{(4)}\right)^T,$$

where $X_i^{(p)} \in \mathbb{R}^2 \sim \mathcal{N}(0, \Sigma_p)$ $(1 \leq p \leq 4)$ and the entries of the matrix $\Sigma_p^{1/2}$ are simulated from i.i.d. uniform variables on $[0,1]$.

We aim to generate the processes T_0 (together with D_0) and T_2, such that the illness-death times T_{kl}, for $(k,l) \in \{(01),(02),(12)\}$, are simulated through Cox transition-specific hazard functions, noted $\alpha_{kl}(.)$, as done traditionally (Bender et al., 2005):

$$T_{kl}^i \sim \alpha_{kl}(t|X_i) = \alpha_{kl}^0(t) \exp\left(g_{kl}\left(X_i, \beta_{kl}\right)\right)$$

where $g_{kl}(.)$ is a transition-specific risk function,

$$\beta_{kl} = \left(\beta_{kl}^{(1)}, \beta_{kl}^{(2)}, \beta_{kl}^{(3)}, \beta_{kl}^{(4)}\right)^T$$

with $\beta_{kl}^{(p)} \in \mathbb{R}^2$ for $1 \leq p \leq 4$, are fixed effect coefficients, and α_{kl}^0 is a transition-specific baseline hazard function. We generate the three baseline hazard functions as follows: $\alpha_{kl}^0(.) \sim \text{Weibull}\,(\text{scale } = 0.01,\ \text{shape } = 1.2)$.

We set the transition-specific risk functions to be non linear using quadratic functions, in the spirit of Lee et al. (2018), as

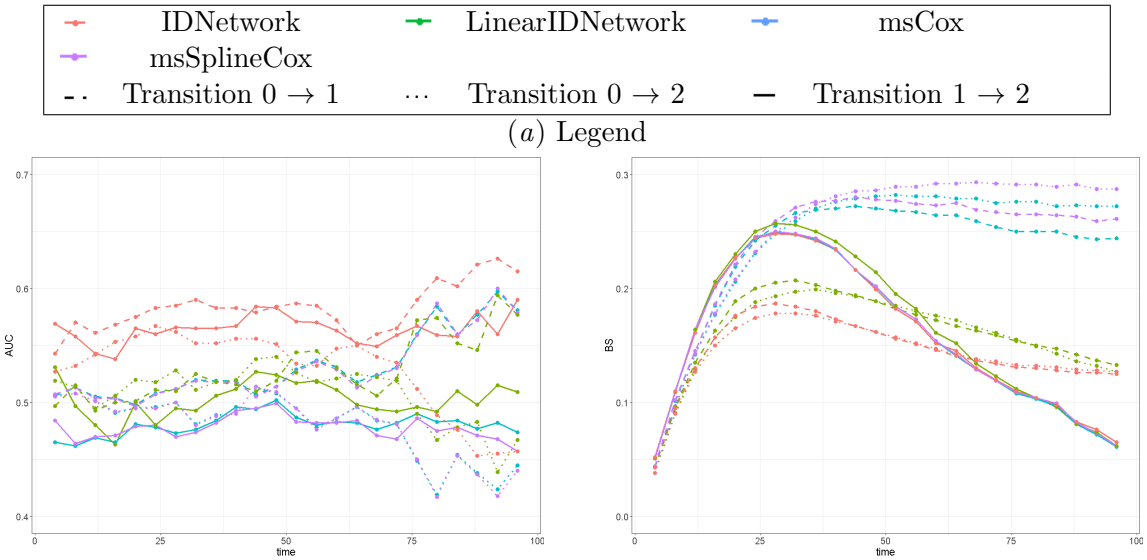
$$g_{01}\left(X_i, \beta_{01}\right) = \left(X_i^{(1)}\beta_{01}^{(1)} + X_i^{(2)}\beta_{01}^{(2)}\right)^2,$$
$$g_{02}\left(X_i, \beta_{02}\right) = \left(X_i^{(2)}\beta_{02}^{(2)} + X_i^{(3)}\beta_{02}^{(3)}\right)^2,$$
$$g_{12}\left(X_i, \beta_{12}\right) = \left(X_i^{(3)}\beta_{12}^{(3)} + X_i^{(4)}\beta_{12}^{(4)}\right)^2.$$

We fix arbitrary values for the fixed effects coefficients. Hence, in this simulation scheme, the Cox's linear assumption doesn't hold anymore.

We fix the censoring rate r to $r = 30\%$ such that 30% of patients from state 0 are censored, and 30% of patients at risk for transition $1 \to 2$ are censored from state 1.

Appendix H. Additional results on the non-linear simulated data set

Additional results on the non-linear simulated data set are displayed in Figure 4.



(a) Legend

(b) AUCs per evaluation time

(c) BSs per evaluation time

Figure 4: AUCs and BSs per evaluation time for the different models for the non linear simulated data set.

Appendix I. Description of the real data sets

Descriptive statistics of the data sets are shown in Table 5.

Data set	No. observations (%)					Total
	$0 \to 1$	$0 \to 2$	$0 \to$ cens.	$1 \to 2$	$1 \to$ cens.	
NCT00079274	623 (29%)	81 (4%)	1427 (67%)	276 (44%*)	347 (56%*)	2121
NCT00275210	279 (25%)	14 (1%)	829 (74%)	132 (47%*)	147 (53%*)	1122
METABRIC	677 (36%)	509 (27%)	717 (38%)	593 (88%*)	84 (12%*)	1903

* among patients at risk

Table 5: Descriptive statistics of illness-death data from real clinical data sets.

The data on colon cancer is composed of two data sets from two Phase III clinical trials evaluating the outcomes RFS and OS on non-metastatic colon cancer: (1) The **study NCT00079274**[2] contains 2121 observed patients followed for 60 months (5 years) for RFS and for 96 months (8 years) for OS. It presents 67% of censoring from state 0 and 56% from state 1 among patients at risk. (2) The **study NCT00275210**[3] contains 1122 patients followed over 60 months for RFS and OS. The data set presents 74% of censoring from state 0 and 53% from state 1 among patients at risk. The preprocessing of this two data sets requires a preliminary evaluation of the compatible features (same covariates and same distributions of the covariates) and to adjust the length of follow-up between both data sets. Thus, we finally restrict our attention to 9 baseline clinical covariates (8 categorical and 1 numerical) including the following features: BMI, sex, race, age, tumor histological type, number of positive lymph nodes, cancer stage, ECOG performance status, presence of bowel obstruction/perforation. In the study NCT00079274, outcome RFS has been right-censored at 5 years and outcome OS at 8 years. Whereas in the study NCT00275210, both outcomes have been right-censored at 5 years. We adjust the length of follow-up of both studies choosing a value for τ compatible with both.

For the **METABRIC**[4] data set, we include 1903 patients followed for 360 months (30 years) for relapse-free survival (RFS) and overal-survival (OS), with 38% of censoring from state 0 and 13% from state 1 among patients at risk (see Table 5). Based on the literature, we select 17 baseline clinical and molecular covariates (12 categorical, 5 numerical) including the following features: age, inferred on the menopausal status, Nottingham Prognostic Index (NPI), immunohistochemical oestrogen-receptor (ER) status, cumber of positive lymph nodes, cancer grade, tumor size, tumor histological type, cellularity, Her2 copy number by SNP6, Her2 expression, ER Expression, progesterone (PR) expression, type of breast surgery, cancer molecular subtype (pam50 subgroup, integrative cluster), chemotherapy regime, hormone regime, radiotherapy regime. We also use gene expression data. Several approaches have been reported to integrate gene expression data into survival models, based either on dimension reduction, on genes or metagenes selection (Van Wieringen et al., 2009) or, more recently, on the use of a large number of gene expression values (> 1000) with the development of deep learning methods (Yousefi et al., 2017; Ching et al., 2018). We preprocess gene expression data based on a selection approach in order to extract

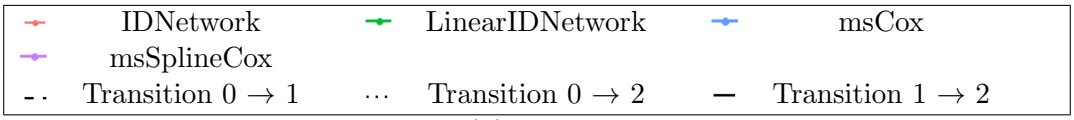
2. The clinical trial NCT00079274 is available under request at https://data.projectdatasphere.org/projectdatasphere/html/content/161
3. The clinical trial NCT00275210 is available under request at https://data.projectdatasphere.org/projectdatasphere/html/content/128
4. The METABRIC data set is available at https://www.nature.com/articles/s41586-019-1007-8#Sec22

cancer-related information from a massive amount of features. To extract this information, we select the 100 genes with the largest standard deviation.

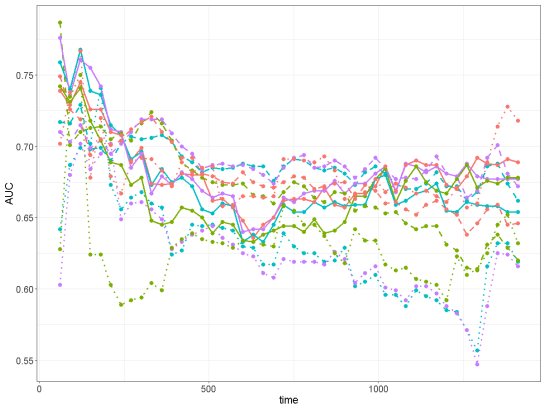
For the three data sets, missing values were imputed by the median value for numerical features and by the mode for categorical features. We apply one-hot encoding on categorical features and standardize numerical features with the Z-score. We fix a length for the time intervals to 1 month such that the time interval for month j, $v_j = [j-1, j)$ includes all the events that occurred in the interval $[(j-1) \times 30.5, j \times 30.5)$. We fix respectively $K = 48$ and $K = 120$ for the METABRIC data set and we set the event times after respectively 48 (months) and 120 (months) in supplementary last intervals v_{49} and v_{121}.

Appendix J. Additional results on the real data sets

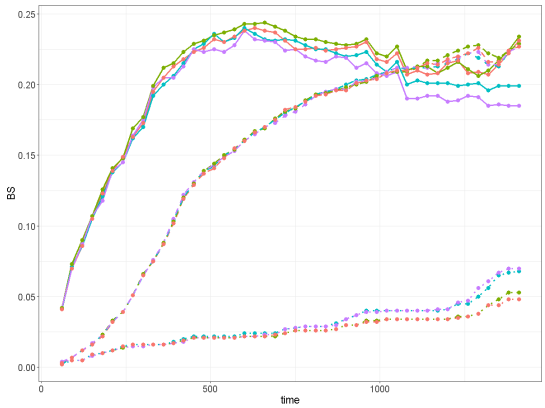
Additional results per evaluation time on the colon cancer data sets are displayed in Figures 5 and 6. Additional results on the METABRIC cohort are presented in Figure 7.



(a) Legend

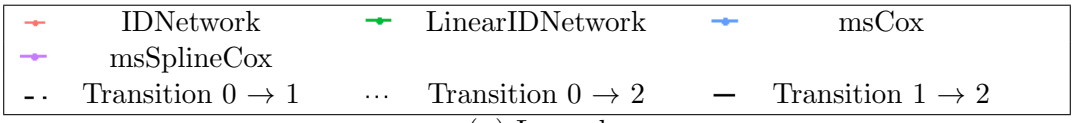


(b) AUCs per evaluation time

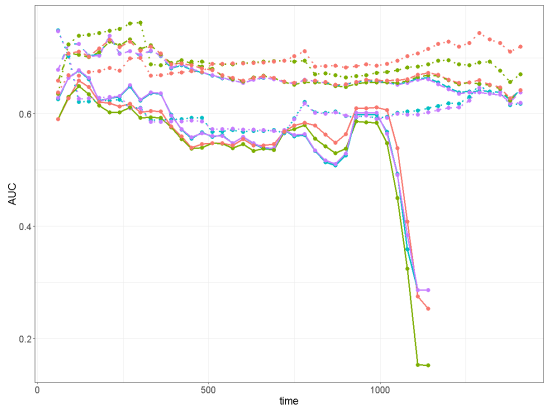


(c) BSs per evaluation time

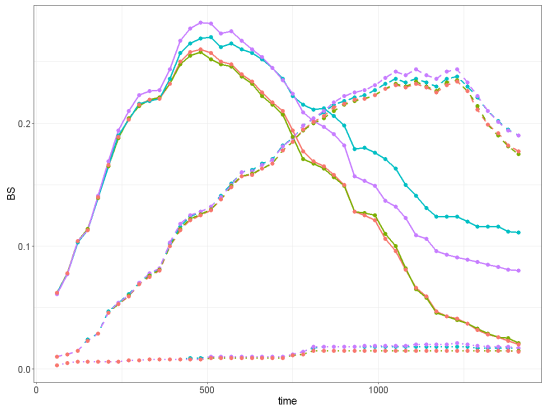
Figure 5: AUCs and BSs per evaluation time for the different algorithms for the colon data set (internal validation).



(*a*) Legend

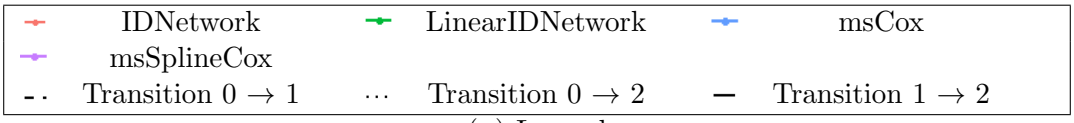


(*b*) AUCs per evaluation time

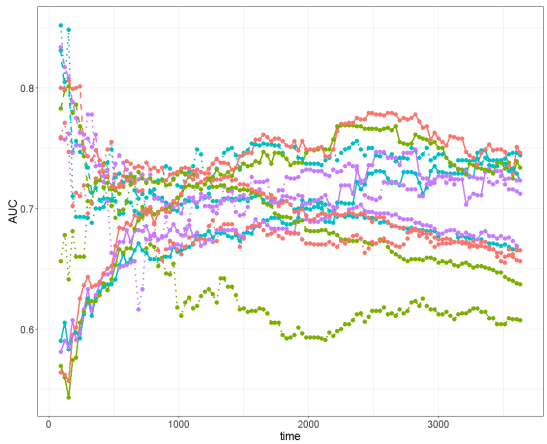


(*c*) BSs per evaluation time

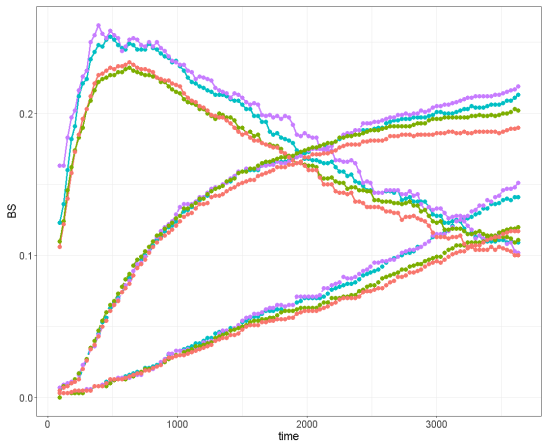
Figure 6: AUCs and BSs per evaluation time for the different algorithms for the colon data set (external validation).



(*a*) Legend



(*b*) AUCs per evaluation time



(*c*) BSs per evaluation time

Figure 7: AUCs and BSs per evaluation time for the different algorithms for the METABRIC cohort (internal validation).