Received: 8 February 2021 | Revised: 28 October 2021 | Accepted: 17 December 2021

DOI: 10.1002/sim.9310

RESEARCH ARTICLE

Statistics in Medicine WILEY

IDNetwork: A deep illness-death network based on multi-state event history process for disease prognostication

Aziliz Cottin[1] | **Nicolas Pecuchet**[1] | **Marine Zulian**[1] | **Agathe Guilloux**[2,3] | **Sandrine Katsahian**[4,5,6,7]

[1]Healthcare and Life Sciences Research, Dassault Systemes, Velizy-Villacoublay, France

[2]CNRS, Université Paris-Saclay, Paris, France

[3]Laboratoire de Mathématiques et Modélisation d'Evry, Université d'Evry, Evry-Courcouronnes, France

[4]AP-HP, Hôpital Européen Georges Pompidou, Unité de Recherche Clinique, APHP Centre, Paris, France

[5]Inserm, Centre d'Investigation Clinique 1418 (CIC1418) Epidémiologie Clinique, Paris, France

[6]Inserm, Centre de recherche des Cordeliers, Sorbonne Université, Université de Paris, Paris, France

[7]HeKA, INRIA PARIS, Paris, France

Correspondence
Aziliz Cottin, Healthcare and Life Sciences Research, Dassault Systemes, 78140 Velizy-Villacoublay, France.
Email: aziliz.cottin@3ds.com

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 of the disease 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 between the states are modeled via (semi-) Markov processes and transition-specific Cox proportional hazard (P.H.) models. The Cox P.H. model assumes that each variable makes a linear contribution to the model, but the relationship between covariates and risks can be more complex in clinical situations. To address this challenge, we propose a neural network architecture called illness-death network (IDNetwork) that relaxes the linear Cox P.H. assumption within an illness-death process. IDNetwork employs a multi-task architecture and uses a set of fully connected subnetworks in order to learn the probabilities of transition. Through simulations, we explore different configurations of the architecture and demonstrate the added value of our model. IDNetwork 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
deep learning, illness-death process, neural networks, stratified medicine, survival analysis

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-state 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 and is used in multiple areas including healthcare, economy, finance, and engineering. In particular, survival analysis is widely used in healthcare to model patient survival outcome in order to understand disease progression. In clinical practice, clinicians may be more interested in the complete evolution of a disease and not only in a unique or composite event. The multi-state[1] approach has been developed as a generalization of survival analysis when multiple events can occur successively over time. In the present work, we focus on the illness-death model which is a multi-state model 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 through an intermediate non-fatal relapse state and a death state as in ovarian cancer[2] or in chronic myeloid leukemia.[3] Other applications of the illness-death model include Alzheimer's disease[4] and cardiovascular disease.[5]

There are two main literature streams of event history analysis. The first is based on traditional statistic theories including three approaches. (i) The nonparametric approaches include the Kaplan Meier[6] and the Nelson-Aalen[7] estimators and do not enable individual prediction. (ii) The parametric models impose a precise form of the hazard rate and the influence of covariates on it but allow individual prediction. (iii) Finally, semi-parametric models can be viewed as a compromise between nonparametric and parametric approaches. They include the widely used Cox[8] proportional hazard (P.H.) model and have been extended to multi-state analysis.[9] However, these models style assume strong assumptions on the relation between the covariates and the event times distribution. These assumptions have shown limitations in many real-world applications and have been found to be violated especially in clinical areas.

To address these challenges, a second variety of literature proposes new machine learning algorithms. In particular, neural networks have been developed to extend the Cox P.H. model in a statistical assumption-free framework. One of the advantages of neural networks is that they can fit highly nonlinear patterns in the data by using multiple layers and nonlinear activation functions. In addition, the use of fully connected networks allows to take into consideration possible interactions between covariates with no prior assumption. Hence, traditional artificial neural networks have been successfully introduced for survival analysis by Faraggi and Simon.[10] More recently, deep neural networks have been extended by Luck et al,[11] Katzman et al,[12] Fotso,[13] Kvamme et al,[14] among others.[15-17] By employing best state-of-the-art deep learning methods and larger clinical data sets, they show significant improvements in predicting patients' survival as compared to the Cox P.H. model. Regardless, their approaches are still limited to the case of a unique clinical event. Lee et al[18] extended deep neural networks to handle competing events.[19] To the best of our knowledge, no nonlinear methods, especially deep neural networks, have been explicitly extended for multi-state analysis and in particular for an illness-death process. Thus, addressing the linear limitation of the Cox P.H. model within the illness-death modeling framework is one of the focuses of this work.

While the Faraggi and Simon[10] approach uses neural networks to parameterize the Cox's linear regression function, most of the recent methods directly predict a discrete-time distribution of the event times as an output of the neural network. As an approximation for continuous-time survival data, they all perform a division of the continuous time scale into discrete-time intervals. Alternatively to discrete-time approach, piecewise survival models[20] perform a discretization of the time scale but each subject's duration of exposure during the interval is taken into account. Hence, the approximation error that arises when a discrete-time method is used can be reduced with piecewise approximations. See the work of Kvamme and Borgan[21] for more details. Thus, developing a well approximated continuous-time method instead of a discrete-time method is the second objective of this article.

There are few recommendations for dividing intervals and selecting the interval cutpoints in piecewise survival models. This has a significant impact on the model performance and can cause either over-fitting (for a larger number of intervals) or under-fitting. However, Kvamme and Borgan[21] conduct a simulation study by varying the interval cutpoints determination method in a discrete time approach. Inspired by their work, we discuss some methods for determining the optimal interval cutpoints. In addition, we propose a regularization method in order to minimize the risk of over-fitting related to the number of intervals.

In the present work, we propose a deep learning architecture, illness-death network (IDNetwork), for illness-death model that encompasses a multi-task neural network including one subnetwork shared for all the transitions and three transition-specific subnetworks. After a presentation of the classical illness-death model, we (i) derive a new form of the log-likelihood of a piecewise constant illness-death process, (ii) build the network architecture IDNetwork, (iii) implement in Python the pipeline of our method including performances criteria evaluation. We finally conduct experiments on a simulated nonlinear data set and on real data sets of patients with colon cancer and with breast cancer.

2 | THE ILLNESS-DEATH MODEL

In this section, we introduce the traditional illness-death model, see the work of Andersen et al[22] for a complete presentation.

2.1 | Notation

Multi-state models are a generalization of survival models when multiple events of interest can occur over time. A multi-state process $(E(t), 0 < t \leq +\infty)$ is a continuous-time stochastic process that describes the states occupied by a patient over time. In this article, we consider an illness-death multi-state process with three states: state 0 is the initial state "Event-free," state 1 is an intermediate state "Relapse," state 2 is an absorbing state "Death." The illness-death process is characterized by three irreversible 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. It assumes that all subjects are in state 0 at time $t = 0$ (ie, $\mathbb{P}(E(0) = 0) = 1$).

The evolution of an illness-death process can be characterized by three random variables (r.v.) T_{ql}, for $(q, l) \in \{(0,1),(0,2),(1,2)\}$, associated with each of the three transitions. They represent the transition times from state q to state l $(q \neq l)$. A subject leaving state 0 will enter either state 1 at time T_{01} or state 2 at time T_{02}. A subject having state 1 at T_{01} will enter in state 2 at time $T_{01} + T_{12}$. This can be summarized as follows. T_0, the exit time from state 0 is

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

and is recorded together with $D_0 \in \{1, 2\}$ which indicates the entered state. T_2 the entry time to state 2

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

and characterizes the total survival time.

In clinical settings, the true transition times are commonly partially observed because of right-censoring. To model this phenomenon, we introduce C a non-negative censoring r.v. 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 event times, we observe a vector of covariates X of dimension P and we assume that $C \perp\!\!\!\perp (T_0, T_2) | X$. We also observe the binary labels $\delta_{0l} = \mathbb{1}\{D_0 = l, T_0 \leq C\}$ $(l = 1, 2)$, $\delta_{12} = \delta_{01}\mathbb{1}\{T_2 \leq C\}$ that indicate the status of the transitions, where $\delta_{ql} = 1$ indicates an entry in state l from q and $\delta_{ql} = 0$ indicates a censored transition.

2.2 | Transition intensities and the Cox P.H. model

Illness-death processes are traditionally described with counting processes.[23] In an illness-death model, the observation of process E is equivalent to the observation of the three-variate process $t \geq 0 \mapsto N(t) = (N_{01}(t), N_{02}(t), N_{12}(t))$. where

$$N_{ql}(t) = \text{card}\{0 \leq s \leq t : E(s-) = q, E(s) = l\}, \quad \text{for } (q, l) \in \{(0,1),(0,2),(1,2)\},$$

the transition from q to l happened before t. To take into account the presence of censoring, we define in addition two previsible processes Y_0 and Y_1 as

$$Y_q(t) = \mathbb{1}\{E(t-) = q\}, \quad \text{for } q = 0, 1, \ t > 0,$$

they indicate if the patient is in state q before time t.

The three-variate counting process N is conventionally associated with a set of transition intensities, or transition-specific hazard functions. We define the three processes by making specific assumptions on cancer evolution over time.[24] For the transitions $0 \to 1$ and $0 \to 2$, we consider time nonhomogeneous Markovian processes. For transition $1 \to 2$, we perform a time transformation, following Anderson et al,[22] 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 $t - T_0$ already spent in 1). Wherever convenient, we use the duration variable $d = t - T_0$ instead of the time variable t for the transition $1 \to 2$. Hence, we define

$$\alpha^*_{0l}(t|X) = \lim_{h \to 0} \frac{1}{h} \mathbb{P}(E(t+h) = l \mid E(t-) = 1, X), \quad \text{for } l = 1, 2,$$
$$\alpha^*_{12}(d|X) = \lim_{h \to 0} \frac{1}{h} \mathbb{P}(E(d+h) = l \mid E(d-) = 1, X),$$

that express the instantaneous risk to transit from a state to another at a specific time for the covariates X.

Transition specific Cox P.H. models have been proposed to model multi-state processes in order to evaluate covariates effects on each transition.[9] Covariates effects are introduced in the transition intensities by means of log linear risk function as follows:

$$\alpha^*_{0l}(t|X) = \alpha^{*,0}_{0l}(t)\exp\left(X\beta^*_{0l}\right), \quad \text{for } l = 1, 2, \tag{1}$$

$$\alpha^*_{12}(d|X) = \alpha^{*,0}_{12}(d)\exp\left(X\beta^*_{12}\right), \tag{2}$$

where, for $(q, l) \in \{(0, 1), (0, 2), (1, 2)\}$, $\alpha^{*,0}_{ql}$ is the unknown baseline transition intensity related to the transition $q \to l$ (ie, the underlying hazard when all the covariates are equal to zero) and β^*_{ql} is the vector of regression coefficients to be estimated.

In general, the process associated with each of the three transitions may depend on the time t of arrival in the state (Markov process) or on the time d since the entry to the state (semi-Markov process). The choice of the time scale[3] is an important step in disease modeling because it will induct how the disease will evolve over time. We discuss our choice in Section 3.1.

2.3 | The log-likelihood

The log-likelihood associated with the observation of the three-dimensional censored counting process $t \mapsto N(t \wedge C) = (N_{01}(t \wedge C), N_{02}(t \wedge C), N_{12}(t \wedge C))$ on $[0, \tau]$, where τ is the horizon time, is given by the log product of the three transition-specific likelihoods:

$$\log \mathcal{L} = \log\left(\mathcal{L}^{0\to1} \times \mathcal{L}^{0\to2} \times \mathcal{L}^{1\to2}\right),$$

where $\mathcal{L}^{0\to1}$, $\mathcal{L}^{0\to2}$, $\mathcal{L}^{1\to2}$ are the likelihoods associated with each transition, see Andersen et al,[22] and the covariates X. They are given, for the proposals α_{0l} and α_{12}, for $l = 1, 2$, by the following equations:

$$\log \mathcal{L}^{0\to l} = \int_0^\tau \log\left(\alpha_{0l}(t|X)\right) Y_0(t)\mathbb{1}\{C \geq t\}dN_{0l}(t) - \int_0^\tau \alpha_{0l}(t|X)Y_0(t)\mathbb{1}\{C \geq t\}dt, \quad \text{for } l = 1, 2, \tag{3}$$

$$\log \mathcal{L}^{1\to2} = \int_0^\tau \log\left(\alpha_{12}(d|X)\right) Y_1(t)\mathbb{1}\{C \geq d\}dN_{12}(t) - \int_0^\tau \alpha_{12}(d|X)Y_1(t)\mathbb{1}\{C \geq d\}dd. \tag{4}$$

The log-likelihood of Equation (3) is traditionally used to estimate the coefficients β^*_{ql} while the unknown functions $\alpha^{*,0}_{ql}$ can be estimated with Nelson-Aalen[9] or spline estimators.[25] Other models have been proposed for the transition intensities,[26] the simplest assuming constant transition intensities $\alpha^*_{ql}(t) = \alpha^*_{ql}$. Finally, models with time varying coefficients, with proposals defined in Equations (1) and (2), have been studied, see Martinussen and Scheike[27] for a complete review. Among these proposals, Murphy and Sen[28] considered piecewise constant estimators that allow for time-dependent covariate effects.

Under the assumption of a piecewise constant model, the optimization of the log-likelihood is facilitated. Indeed, they allow to perform a continuous model that would have necessitated the use of the Cox partial log-likelihood[8] and significantly impact on the computational cost when using classical stochastic gradient descent algorithm because of the presence of two cumulative sums.[14,29] Whereas a discrete-time model approximates the full log-likelihood by dividing the time axis into discrete time intervals. Piecewise constant approaches are a compromise to reduce the approximation error that arise when a discrete-time method is used by computing the cumulative functions by taking into account the subject's duration of exposure in the intervals, with no supplementary computational cost, see the Supplementary materials (Section 3) for details.

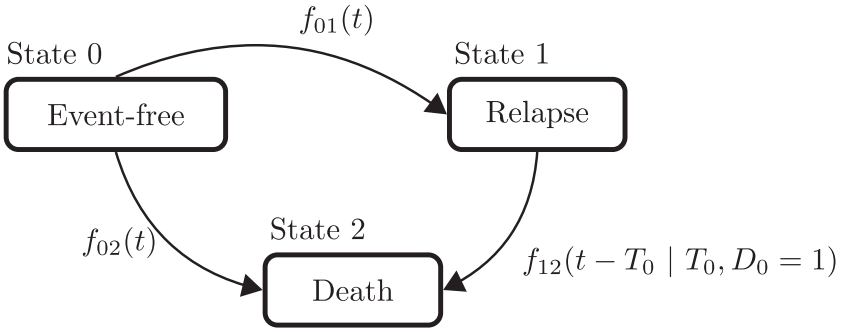


FIGURE 1 Illustration of our illness-death process

3 | OUR METHODOLOGY

In this section, we present our approach to model an illness-death process.

3.1 | Functions of interest and data

Instead of considering the transition intensities, we will write the log-likelihood in terms of transition probabilities $f^*_{ql}(.|X)$, that we define in this paragraph. See an illustration in Figure 1.

Let f^*_{01} and f^*_{02} be the infinitesimal probabilities of experiencing, respectively, transitions $0 \to 1$ and $0 \to 2$, defined as

$$f^*_{0l}(t|X) = \lim_{h\to 0}\frac{1}{h}\mathbb{P}\left(E(t+h) = l, E(t-) = 0|X\right) = \lim_{h\to 0}\frac{1}{h}\mathbb{P}\left(t \le T_0 \le t+h, D_0 = l|X\right),$$

and F^*_{01} and F^*_{02} their cumulative counterparts,

$$F^*_{0l}(t|X) = \mathbb{P}\left(E(t) \ne l, D_0 = l|X\right) = \mathbb{P}\left(T_0 \le t, D_0 = l|X\right) = \int_0^t f^*_{0l}(s|X)ds, \text{ for } l = 1,2 \text{ and } t \ge 0$$

expresses the probability that a transition $0 \to l$ occurs on or before time t. With these definitions, we define

$$f^*_0(t|X) = f^*_{01}(t|X) + f^*_{02}(t|X)$$

as the infinitesimal probability of exiting state 0 at time t and

$$F^*_0(t|X) = F^*_{01}(t|X) + F^*_{02}(t|X)$$

as the probability of having exited state 0 before time t.

For the 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,

$$f^*_{12}(d|X) := f^*_{12}\left(d\ |T_0,\ D_0 = 1,\ X\right) = \lim_{h\to 0}\frac{1}{h}\mathbb{P}\left(E(d+h) = 2, E(d-) = 1|X\right) = \lim_{h\to 0}\frac{1}{h}\mathbb{P}\left(d \le T_2 - T_0 \le d+h|X\right),$$

and F^*_{12} as its cumulative counterpart such that,

$$F^*_{12}(d|X) := F^*_{12}\left(d\ |T_0,\ D_0 = 1,\ X\right) = \mathbb{P}\left(E(d) = 2|X\right) = \mathbb{P}\left(T_2 - T_0 \le d|X\right) = \int_0^d f^*_{12}(s|X)ds, \text{ for } d \le 0.$$

F^*_{01}, F^*_{02}, and F^*_{12} are commonly referred to cumulative incidence functions (CIFs) in the literature[30] and are our main functions of interest. Indeed, for a patient with covariates X, if he/she still is state, the physician will need to estimate precisely the values $F^*_{01}(t_0|X)$ and $F^*_{02}(t_0|X)$ where t_0 is a certain horizon time, chosen according to the pathology. Even if they are formula linking the transition intensities[23] to these probabilities, see Supplementary materials (Section 1), we chose to work with transition probabilities.

Our data consist in the observation of n independent and identically distributed r.v. in $\mathbb{R}^P \times \mathbb{R}_+ \times \{0,1\} \times \{0,1\} \times \mathbb{R}_+ \times \{0,1\}$ such that

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

where $X_i = (X_{i1}, \dots, X_{iP})^T$ 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 defined in Equations (9) and (10).

3.2 | Writing of the log-likelihood in terms of the functions of interest

We show that the conventional illness-death log-likelihood defined in Equation (3) can be rewritten in terms of the density probability functions introduced in Section 3.1, see the Supplementary materials (Section 2) for a formal proof.

We define the log-likelihood ℓ_n by dividing the contributions in three distinct parts:

$$\ell_n = \log \mathcal{L}_n = \log \mathcal{L}_n^{0\to1} + \log \mathcal{L}_n^{0\to2} + \log \mathcal{L}_n^{1\to2} \tag{5}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left[\ell_i^{0\to1} + \ell_i^{0\to2} + \ell_i^{1\to2}\right], \tag{6}$$

where $\ell_i^{0\to1} + \ell_i^{0\to2}$ is the log contribution of patient i from state 0. From state 0, patient i with an event a $\tilde{T}_0^i$ can contribute in three ways. He (i) can experience a transition $0 \to 1$; (ii) can experience a transition $0 \to 2$; (iii) can be censored at $\tilde{T}_0^i$. Thus ℓ_0^i is given, for proposals f_{01}, f_{02}, F_0, and binary labels δ_{01}, δ_{02} that indicate an entry in another state or a censored transition, by

$$\ell_i^{0\to1} + \ell_i^{0\to2} = \sum_{l=1,2} \{\delta_{0l}^i \log(f_{0l}(\tilde{T}_0^i\ |X_i))\} + (1 - (\delta_{01}^i + \delta_{02}^i)) \log(1 - F_0(\tilde{T}_0^i\ |X_i)). \tag{7}$$

On the other hand, $\ell_i^{1\to2}$ 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, patient i can contribute in two ways at time $\tilde{T}_2^i - \tilde{T}_0^i$. He (i) can experience a transition $1 \to 2$; (ii) can be censored. Thus $\ell_i^{1\to2}$ is given, for proposals f_{12}, F_{12}, and the binary label δ_{12} that indicate an entry in state 2 or a censored transition, by

$$\ell_i^{1\to2} = \delta_{01}^i \delta_{12}^i \log(f_{12}(\tilde{T}_2^i - \tilde{T}_0^i\ |X_i)) + \delta_{01}^i(1 - \delta_{12}^i) \log(1 - F_{12}(\tilde{T}_2^i - \tilde{T}_0^i\ |X_i)). \tag{8}$$

The log-likelihood in Equation (5) has been derived under a time nonhomogeneous Markovian assumption for transitions $0 \to 1$ and $0 \to 2$, and a time homogeneous semi-Markovian assumption for transition $1 \to 2$. Depending on the application, time assumptions can be reformulated, it would change the log likelihood above.

3.3 | Rewriting of the log-likelihood for piecewise constant proposals

We propose in the present article to consider the class of piecewise constant transition-specific probabilities following the ideas of Murphy and Sen.[28] More specifically candidate estimators will be assumed to be constant on K time intervals. We define τ as the maximum horizon time window and we divide the time axis into K disjoint time intervals:

$v_1 = [a_0, a_1), \ldots, v_K = [a_{K-1}, a_K)$, 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. so that :

$$f_{0l}(t|X) = f_{0l}(v_{k(t)}|X), \quad \text{for } l = 1, 2,$$
$$f_{12}(t|X) = f_{0l}(v_{k(d)}|X).$$

Piecewise constant models allow to keep constant the probabilities within observed time intervals. In this model, effects of the covariates are time-dependent.[31]

We rewrite the log-likelihood of an illness-death process for piecewise constant proposals. It has to be understood as an approximation and it is well-established that the approximation error can be bounded when the true functions are smooth.[32]

Assuming a piecewise constant model allows to consider a nonhomogeneous model (see Section 3.1) while retaining the hypothesis of homogeneity within the same time interval, following the lines of Kvamme and Borgan,[21] Friedman et al.[20] Its means that the density probabilities are constant within each interval but covariates effects are parameterized to be different in each interval (see Section 2.2). Hence, we can express f_{ql} ($q = 0, 1$ and $l = 1, 2$) as step functions such that $f_{ql}(\ .\ |X) = f_{ql}(v_{k(.)}\ |X)$ that depend on the covariates and on time. Equivalently, since the density probabilities are assumed to be piecewise constant, their corresponding cumulative counterparts F_{0l} ($l = 1, 2$) and F_{12} are piecewise linear:

$$F_{0l}(t\ |X) = \sum_{k=1}^{k(t)-1} f_{0l}(v_k\ |X)|v_k| + (t - a_{k(t)-1}) f_{0l}(v_{k(t)}\ |X), \tag{9}$$

$$F_{12}(d\ |X) = \sum_{k=1}^{k(d)-1} f_{12}(v_k\ |X)|v_k| + (d - a_{k(d)-1}) f_{12}(v_{k(d)}\ |X), \tag{10}$$

where $|v_k|$ is the length of interval v_k. In Equations (9) and (10), we see that each duration of exposure during the intervals is taken into account. In comparison, only whether an event occurred or not in a given time interval is taken into account in discrete-time models, disregarding the duration of exposure in the given time interval.

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:

$$\begin{aligned} &\sum_{k=1}^{K} f_0(v_k|X)|v_k| + 1 - F_0(\tau|X) = 1, \\ &\sum_{k=1}^{K} f_{12}(v_k|X)|v_k| + 1 - F_{12}(\tau|X) = 1, \end{aligned} \tag{11}$$

where $|v_k|$ is the length of interval v_k. See Lee et al[18] and Kvamme and Borgan[21] for similar remarks.

Regarding the constraint of Equation (11), we consider a supplementary interval $v_{K+1} = [\tau, +\infty)$. Consequently, we fulfill the constraint by defining $1 - F_0(\tau|X) = f_0(v_{K+1}|X)$, $1 - F_{12}(\tau|X) = f_{12}(v_{K+1}|X)$.

Under the assumption of a piecewise constant model, we rewrite the log-likelihood in Equation (5). We rewrite the sum of the first two contributions in Equation (7) as follows:

$$\ell_i^{0\to1} + \ell_i^{0\to2} = \sum_{l=1,2} \left\{ \delta_{0l}^i \log\left(f_{0l}\left(v_{k(\tilde{T}_0^i)}\ |X_i \right) \right) \right\} + (1 - (\delta_{01}^i + \delta_{02}^i)) \log(1 - F_0(\tilde{T}_0^i\ |X_i)).$$

We rewrite the third contribution in Equation (8) as follows:

$$\ell_i^{1\to2} = \delta_{01}^i \delta_{12}^i \log(f_{12}(v_{k(\tilde{T}_2^i - \tilde{T}_0^i)}\ |X_i)) + \delta_{01}^i (1 - \delta_{12}^i) \log(1 - F_{12}(\tilde{T}_2^i - \tilde{T}_0^i\ |X_i)).$$

Estimators of f_{0l} and f_{12} will be precisely defined in Section 4.1.

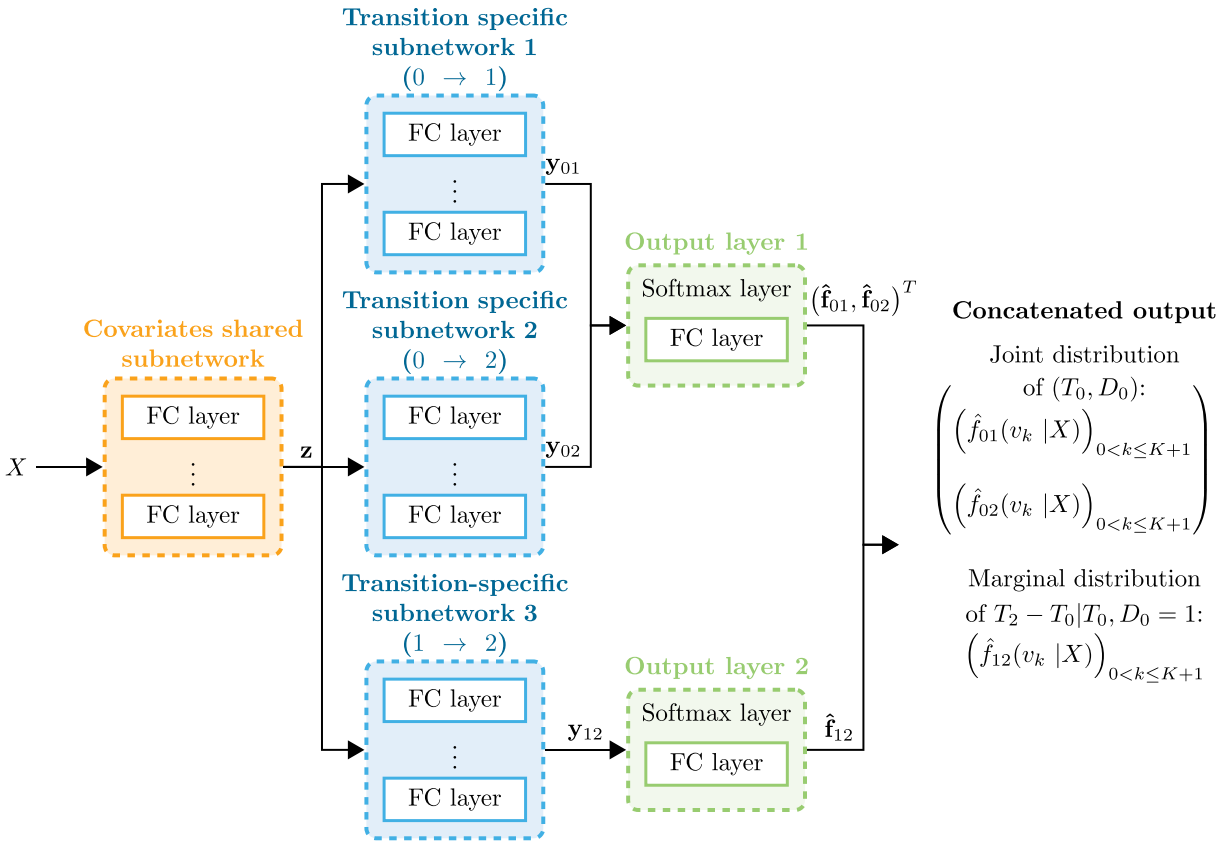


FIGURE 2 IDNetwork architecture. "FC layer" refers to fully connected layer

4 | DESCRIPTION OF IDNETWORK

In this section, we describe how we use a new deep learning approach to parameterize the step probability functions f_{01},f_{02}, f_{12} over the interval $[0, \tau]$. Our deep learning architecture, called IDNetwork, model the relationship between covariates and these transition probabilities with no linear assumption. Unlike classical methods,[9,33] IDNetwork is divided into transition-specific tasks and uses nonlinear activation functions to capture nonlinearity between covariates and transition probabilities. We propose a loss function that encompasses the negative log-likelihood of Equation (5) and that is tuned to automatically choose a good time division K (ie, the number of time intervals) in order to minimize the risk of over-fitting. In addition, we propose two methods to select the interval cutpoints.

4.1 | Network architecture

Inspired by the work of Lee et al[18] and Fotso,[13] we develop an architecture (see an illustration in Figure 2) with three task-specific subnetworks that are related to the three transitions of an illness-death process. Multi-task learning is done with hard parameter sharing[34] 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. One softmax layer is related to the exit from state 0 (ie, the transitions $0 \to 1$ and $0 \to 2$), the other to the exit from state 1 (ie, the transition $1 \to 2$).

4.1.1 | Input layer

The input layer is composed of the matrix X of P baseline covariates for the n individuals.

4.1.2 | Covariates shared subnetwork

The shared subnetwork takes as input the *input layer* and contains L fully connected (FC) 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 nonlinear activation function).

4.1.3 | Transition-specific subnetworks

Each transition-specific subnetwork takes $\mathbf{z}$ as input and contains L^{ql} fully connected hidden layers with l^{ql} units. Its output is a vector $\mathbf{y}_{ql} = g^{ql}(\mathbf{z})$, that is a transition-specific transformation of the shared features (g^{ql} is a nonlinear 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. If the architecture is too complex for a transition, the model will poorly be generalist on new data (over-fitting). If the architecture is not complex enough for a transition, the model will not capture all the information in the data (under-fitting). To find the best configuration that will result in the best model performance for each of the three transitions, we will set the structure of each subnetwork independently. Hence, a transition with a high number of observations (oftenly the transition $0 \to 1$) can support more hidden layers than a transition with less observations (oftenly the transition $0 \to 2$).

4.1.4 | Probabilistic output layers

The output of the network is composed of two probabilistic layers that map the transition-specific outcomes $\mathbf{y}_{ql}$ into time-dependent probabilities. Each output of the network is a fully connected layer. The first output layer is related to transitions $0 \to 1$ and $0 \to 2$, the second output layer is related to transition $1 \to 2$. Each output layer is built in two steps:

1. Each uses primarily a fully connected layer with a linear activation function, noted g^{linear}, to transform the transition-specific outputs $\mathbf{y}_{ql}$ into vectors of lengths $K+1$ (ie, the number of time intervals). The first output layer (resp. the second output layer) transform the transition-specific outputs $(\mathbf{y}_{01},\ \mathbf{y}_{02})^T$ (resp. the transition-specific $\mathbf{y}_{12}$) into vectors $(\phi_{01},\ \phi_{02})^T$ (resp. ϕ_{12}), each of length $K+1$.
2. Subsequently, each output layer uses a fully connected layer with a weighted softmax activation function. The first output layer (resp. the second output layer) uses a weighted softmax activation function σ_0 (resp. σ_2) to transform $(\phi_{01},\ \phi_{02})^T$ (resp. ϕ_{12}) into probabilities and provide an estimation of f^*_{01}, f^*_{02} (resp. f^*_{12}). The use of softmax activation functions ensure the fulfillment of the model constraint defined in Equation (11). Each softmax function is weighted by the length of the time intervals to provide an estimation of the density probabilities under the assumption of a piecewise constant model.
3. Finally the output layers are characterized by the vectors:

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

where

$$\hat{\mathbf{f}}_{ql} = (\hat{f}_{ql}(v_k\ |X))_{0<k\le K+1},\ \text{ for } (q,l) \in \{(0,1),(0,2),(1,2)\},$$

and

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

4.2 | Loss function and mitigation of the number of time intervals effect via penalization

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

$$\ell_{\text{total}} = -\ell^{K+1} + P_\lambda, \tag{12}$$

that sums the negative log-likelihood and a penalization term.

The first term $-\ell^{K+1}$ is a revising of the negative log-likelihood $-\ell$ defined in Equation (5) considering a supplementary interval $(K+1)$ in accordance with the constraint described in Equation (11).

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. Lee et al[35(Section4)] suggest to choose a large K but prevents over-fitting by using L1 regularization over weights in the output layer. Kvamme and Borgan[21(Section4.1)] suggest to fix a small value for K in order to reduce the size of the output layers as much as possible. However, the automatic selection of K can be fixed by applying a temporal smoothing technique. Following Möst[36] and Tibshirani et al,[37] 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. Let's consider $W = (W^1, W^{12})^T,\ B = (B^1, B^{12})^T$ the weight and bias parameters associated with the two output layers, with $W^1 = (W^{01}, W^{02})^T \in \mathbb{R}^{(l^{01}+l^{02})\times 2(K+1)},\ W^{12} \in \mathbb{R}^{l^{12}\times(K+1)}$ and $B^1 = (B^{01}, B^{02})^T \in \mathbb{R}^{2(K+1)},\ B^{12} \in \mathbb{R}^{K+1}$. For $k = 1, \dots, K$, we compute

$$\Delta_{w^{ql}_{j,k}} = w^{ql}_{j,k+1} - w^{ql}_{j,k},\ \Delta_{b^{ql}_k} = b^{ql}_{k+1} - b^{ql}_k,$$

the weight and bias differences associated with the transition $q \to l$, neuron j and adjacent time intervals $v_k,\ v_{k+1}$. Then the penalty term of our loss function in Equation (12) has the form

$$P_\lambda(B, W) = \sum_{ql}\left(\lambda^{ql}_w \sum_{j=1}^{l^{ql}}\sum_{k=1}^{K}\left|\Delta_{w^{ql}_{j,k}}\right| + \lambda^{ql}_b \sum_{k=1}^{K}\left|\Delta_{b^{ql}_k}\right|\right),$$

where λ^{ql}_w and λ^{ql}_b are transition-specific positive constants determining the amount of smoothing to be applied for each transition. For $\lambda^{ql}_w \to +\infty$ (respectively $\lambda^{ql}_b \to +\infty$), 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.

4.3 | Selection of the interval cutpoints a_k's

Under the piecewise constant assumption (see Section 3.3), the definition of the density probabilities requires time to be on the form $0 = a_0 < a_1 < \cdots < a_K = \tau$. Hence, we need to perform a division of the time scale and the selection of the interval cutpoints a_k's $(k = 1, \dots, K)$. On the one hand, we would like to select sufficiently wide interval cutpoints to retain enough information in each interval (ie, keeping a sufficient number of observed transitions within each interval). On the other hand, we would like to select sufficiently narrow cutpoints to ensure that significant temporal changes in the density probabilities can be identified.

To select the cutpoints, the most obvious way would be to choose K equidistant cutpoints in $[0, \tau]$ (ie, uniform intervals each of length K/τ). An alternative is to select the cutpoints based on the distribution of the transition times. This approach inducts different cutpoints $a^{*,1}_k$, $a^{*,2}_k$ for the two output layers of IDNetwork. In that case, for the first output layer related to the transitions $0 \to l\,(l = 1, 2)$, we can select the cutpoints by estimating the K quantiles $q^{*,1}_k$ $(k = 1, \dots, K)$ of the marginal distribution of the sojourn duration in state 0 (ie, $1 - F^*_0(.)$). In the same way, for the second output layer related to the transition $1 \to 2$, we can select the cutpoints based on the K quantiles $q^{*,2}_k$ $(k = 1, \dots, K)$ of the distribution of the sojourn duration in state 1 among patients at risk (ie, $1 - F^*_{12}(.)$). The quantiles $q^{*,1}_k,\ q^{*,2}_k$ (for $k = 1, \dots, K$) verify:

$$1 = 1 - F_0^*(0) = q_0^{*,1} > q_1^{*,1} > \cdots > q_K^{*,1} = 1 - F_0^*(\tau),$$
$$1 = 1 - F_{12}^*(0) = q_0^{*,2} > q_1^{*,2} > \cdots > q_K^{*,2} = 1 - F_{12}^*(\tau),$$

and can be found by estimating the duration distributions F_0^* and F_{12}^* via the nonparametric Aalen-Johansen[38] estimator. This approach ensures that each time interval has the same decrease in the duration distribution estimates, such that

$$q_{k+1}^{*,1} - q_k^{*,1} = \left(1 - F_0^*(\tau)\right)/K,$$
$$q_{k+1}^{*,2} - q_k^{*,2} = \left(1 - F_{12}^*(\tau)\right)/K.$$

Finally, by denoting $\hat{F}_0^{\text{AJ}}$, $\hat{F}_{12}^{\text{AJ}}$ the Aalen-Johansen estimators of F_0^*, F_{12}^*, the interval cutpoints are found by solving

$$1 - \hat{F}_0^{\text{AJ}}(a_k^1) = q_k^1,$$
$$1 - \hat{F}_{12}^{\text{AJ}}(a_k^2) = q_k^2.$$

See the work of Kvamme and Borgan[21] for a similar approach in the case of a unique event.

5 | PREDICTION TASK AND BENCHMARK

5.1 | Prediction of the individual CIFs

In this subsection, we define the predictions of interest according to the time scales defined below. From the output of our network (ie, the step functions $\hat{f}_{0l}(\,.\mid X)\ (l = 1, 2), \hat{f}_{12}(\,.\mid X)$), 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 (9) and (10), as $\hat{F}_{0l}(\,.\mid X_j)\ (l = 1, 2), \hat{F}_{12}(\,.\mid X_j)$, such that:

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

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

5.2 | Predictive evaluation 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 3.1, we adapt the definitions of the time-dependent AUC[39] and the time-dependent Brier score.[40]

To evaluate predictive performances related to transitions $0 \to 1$ and $0 \to 2$, all the patients are considered. For the transition $1 \to 2$, predictions of interest are formulated conditionally to be in state 1 at time T_0 and from the duration $t - T_0$. Hence only the patients at risk for experiencing the transition are considered (ie, only the patients who have already experienced a transition $0 \to 1$).

For two patients i and j, the transition-specific time-dependent AUC measures the probability that a patient i who experienced the transition ql before time t has greater probability of occurrence of the transition than a patient j who has survived to the transition. For, on the one hand the transitions $0 \to 1, 0 \to 2$, and on the other hand the transition $1 \to 2$, we define

$$\text{AUC}^{0l}(t) = \mathbb{P}(F_{0l}^*(t\mid X_i) > F_{0l}^*\left(t\mid X_j\right) \mid T_0^i \le t,\ T_0^j > t,\ D_0^i = l), \text{for } \ l = 1, 2,$$
$$\text{AUC}^{12}(d) = \mathbb{P}(F_{12}^*(d \mid X_i) > F_{12}^*\left(d \mid X_j\right) \mid T_2^i - T_0^i \le d,\ T_2^j - T_0^j > d, D_0^i = 1,\ D_0^j = 1).$$

Commonly, the AUC is defined as the integration of the ROC curve opposing specificity (Sp) and sensitivity (Se).

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

$$\mathrm{BS}^{0l}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{1}\left\{T_0^i > t\right\} - F_{0l}^*(t\,|X_i)\right]^2 \text{ for } l = 1, 2,$$

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

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

Several estimators have been developed taking into account the loss of information due to censoring by using the inverse probability of censoring weight (IPCW) methods. To evaluate predictive performances related to transitions $0 \to 1$ and $0 \to 2$, we use classical IPCW estimators already developed in the literature.[39,40] For the transition $1 \to 2$, we rewrite the classical estimators under the semi-Markovian assumption considering only the patients at risk for experiencing the transition. Exact expressions of the estimators that we have implemented are given in the Supplementary materials (Section 4).

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:

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

5.3 | Softwares and benchmark

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), that is defined in Section 2.2, from the R library `mstate`[9] and a spline-based version of the Cox multi-state model (msSplineCox) from the R library `flexsurv`.[25] We also compare IDNetwork with a simplified linear version of IDNetwork (LinearIDNetwork), see details in the Supplementary materials (Section 5). We implement IDNetwork and LinearIDnetwork in Python within a Tensorflow environment.*

5.4 | Validation

We perform two sets of experiments on (1) simulated data sets and on (2) three real clinical data sets. For the two sets of experiments, we score predictive performances of the methods through internal validation.[41] We employ Monte Carlo Simulations (MCS) to validate experiments on simulations by generating M data sets. We employ Monte Carlo Cross Validation (MCCV) to validate model performance and estimate model variance on the real data sets by randomly splitting M times each data set. For the two sets of experiments, we set $M = 20$ due to a high computational time. We validate performance and estimate variance of IDNetwork as follows:

1. For each iteration m ($m = 1, \ldots, M$) (either the data set m for simulations, or the data set from split m for the real data sets):
   - We split the data set $\mathcal{D}_m$ into $\mathcal{D}_m^{\text{train}}/\mathcal{D}_m^{\text{test}}/\mathcal{D}_m^{\text{validation}}$ (70% for training, 10% for early stopping and hyper-parameters tuning, 20% for validation).
   - IDNetwork hyper-parameters are tuned by performing $B = 60$ random searches on $\mathcal{D}_m^{\text{train}}$. Each random set of hyper-parameters is evaluated on the set $\mathcal{D}_m^{\text{test}}$.
   - We choose the set of hyper-parameters maximizing the iAUCs (averaged across the three transitions) on the set $\mathcal{D}_m^{\text{test}}$.
   - With the optimal set of hyper-parameters, we estimate model parameters on $\mathcal{D}_m^{\text{train}}$.

- We compute model performance on the external validation set $\mathcal{D}_m^{\text{val}}$.

2. We estimate model performance by computing the median (± standard deviation (SD)) iAUC (higher the better) and iBS (lower the better) on the validation sets $\mathcal{D}_m^{\text{val}}$ $(m = 1, \dots, M)$.
3. We estimate predictive performance of the other methods in the same way (excluding the hyper-parameters tuning).
4. We statistically compare performances of IDNetwork over the other methods using a bilateral Wilcoxon[42] signed rank test. In the results, · indicates a P-value less than 0.1, † less than 0.05, ‡ less than 0.01, * less than 0.001.
5. In the results, bold values indicate the best model performance.

Experimental details on IDNetwork's hyper-parameters tuning are given in the Supplementary materials (Section 6). The pseudocode for the validation process of IDNetwork is given in the Supplementary materials (Section 7). On real clinical data sets, an external validation[41] of the predictive performance of the models can be conducted by computing for each iteration m $(m = 1, \dots, M)$ the criteria on an independent data set as well.

6 | EXPERIMENTS ON SIMULATED DATA SETS

Through simulations, we aim to first illustrate the effects of IDNetwork parameterization on the predictive performance and to secondly compare IDNetwork performance with other methods cited above.

6.1 | Data simulation

For the first set of experiments, we conduct Monte Carlo simulations by generating $M = 20$ data sets with the same parameters. We generate continuous-time illness-death data and we fix the horizon-time window at $\tau = 100$. We generate three sets of data sets: $\mathcal{D}^{2000}_{\text{nonlin.}}$, $\mathcal{D}^{5000}_{\text{nonlin.}}$ $\mathcal{D}^{20\,000}_{\text{nonlin.}}$ varying, respectively, the training sample size in $n_{\text{tr}} = 2000,\ 5000,\ 20\,000$ (80% the data sets) and the validation sample size in $n_{\text{val}} = 500,\ 1250,\ 5000$ (20% the data sets). For each observation i $(1 \leq i \leq n)$ we generate four 2-dimensional baseline variables,[18] each drawn from a multivariate Gaussian distribution with mean 0 and a matrix of variance covariance Σ_p:

$$X_i = \left(X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, X_i^{(4)}\right)^T \text{ with } X_i^{(p)} \in \mathbb{R}^2 \sim \mathcal{N}\left(0, \Sigma_p\right),\ 1 \leq p \leq 4,$$

where 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, D_0) and T_2, such that the illness-death times T_{ql}, for $(q, l) \in \{(01), (02), (12)\}$, are simulated through Cox transition-specific hazard functions:[43]

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

where $g_{ql}(.)$ is a transition-specific risk function, $\beta_{ql}^* = \left(\beta_{ql}^{*,(1)}, \beta_{ql}^{*,(2)}, \beta_{ql}^{*,(3)}, \beta_{ql}^{*,(4)}\right)^T$ with $\beta_{ql}^{*,(p)} \in \mathbb{R}^2$ for $1 \leq p \leq 4$ are fixed effect coefficients, and $\alpha_{ql}^{*,0}(.)$ is the baseline hazard function. We generate the three baseline hazard functions as follows: $\alpha_{ql}^{*,0}(.) \sim \text{Weibull}\,(\text{scale} = 0.01,\ \text{shape} = 1.2)$.

We set the transition-specific risk functions to be nonlinear using quadratic functions, in the spirit of Lee et al,[18] as

$$g_{01}\left(X_i, \beta_{01}^*\right) = \left(X_i^{(1)}\beta_{01}^{*,(1)} + X_i^{(2)}\beta_{01}^{*,(2)}\right)^2,\ g_{02}\left(X_i, \beta_{02}^*\right) = \left(X_i^{(2)}\beta_{02}^{*,(2)} + X_i^{(3)}\beta_{02}^{*,(3)}\right)^2,\ g_{12}\left(X_i, \beta_{12}^*\right) = \left(X_i^{(3)}\beta_{12}^{*,(3)} + X_i^{(4)}\beta_{12}^{*,(4)}\right)^2. \tag{13}$$

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

From the simulated T_{ql}, the simulation of the processes (T_0, D_0) and T_2 has to respect constraints of the model (ie, T_{01} and T_{02} are competing, T_{01} and T_{12} are recurrent) in order to generate an identifiable Cox model . We fix $r = 30\%$ 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, see Table 1. We refer the reader to the Supplementary materials (Section 8) for more details on the simulation.

TABLE 1 Descriptive statistics on the number of (No.) observations in the simulated data sets

	No. observations (%)					
Data set	**0 → 1**	**0 → 2**	**0 → cens.**	**1 → 2**	**1 → cens.**	**Total**
$\mathcal{D}^{2000}_{\text{nonlin.}}$	720 (36%)	680 (34%)	600 (30%)	504 (70%[a])	216 (30%[a])	2000
$\mathcal{D}^{5000}_{\text{nonlin.}}$	1817 (36%)	1683 (34%)	1500 (30%)	1272 (70%[a])	545 (30%[a])	5000
$\mathcal{D}^{20\ 000}_{\text{nonlin.}}$	7234 (36%)	6766 (34%)	6000 (30%)	5064 (70%[a])	2170 (30%[a])	20 000

[a]Among patients at risk.

6.2 | Simulation study

6.2.1 | Understanding the effect of K and n

To get a better understanding of the methodologies discussed in Sections 4.2 and 4.3, we perform a simulation study where we vary the size n of the data sets, the number K of time intervals used for the piecewise approximations and the selection methods of the interval cutpoints. Gensheimer and Narasimhan[17] performed a similar study in a discrete-time approach by varying the value K with the conclusion that there is no difference in the predictive performance. Kvamme and Borgan[21] performed a similar study as well in both discrete-time and piecewise approaches with the conclusion that there is no difference in predictive performances between the methods to select the interval cutpoints. However, they concluded that smaller values for K are better for smaller values of n.

For evaluation, we use the internal validation process described in Section 5.3. We use the iAUC and iBS measures, in addition to the transition-specific integrated mean absolute error (iMAE$_{ql}$) between the estimated CIFs $\hat{F}_{ql}(.|X_i)$ and the true CIFs $F^*_{ql}(.|X_i)$:

$$\text{iMAE}_{ql} = \frac{1}{\tau}\int_0^\tau \frac{1}{n_{\text{val}}}\sum_{i=1}^{n_{\text{val}}}\left|\hat{F}_{ql}(t|X_i) - F^*_{ql}(t|X_i)\right|, \quad \text{for } (q,l) \in \{(0,1),(0,2),(1,2)\},$$

where n_{val} is the number of subjects in the validation sets.

For the discretization of the time scale in the three data sets, we vary the number K of time intervals in $K = 25$, 50, 100, 200. We applied two methods to select the interval cutpoints with either equidistant cutpoints (uniform) or cutpoints obtained with the Aalen-Johansen quantiles (quantilesAJ).

In Figure 3, we plot the transition-specific validated criteria (iAUCs, iBSs, iMAEs) of the cutpoints selection methods versus the values of K and n. We integrate the AUC and BS measures at all the 4 equidistant time points in $[0, \tau]$ (ie, at times $t = 4, 8, 12, \ldots, 96$, for computational cost reasons). We integrate the MAE measure at all the 100 discrete time points in $[0, \tau]$ (ie, at times $t = 1, 2, 3 \ldots, 100$). We can see that the selection methods of the interval cutpoints give similar performances in terms of iAUC and iMAE. However, in term of iBS, the uniform selection method give slightly poorer performance for a smaller value of K (for $K = 25$) than the quantilesAJ selection method. In terms of iAUC, iBS and iMAE, it is evident to see that a larger value of n increases the performances for all the values of K. However, for transition $0 \to 2$, we can see that the performance differences in terms of iAUC are very unstable when varying simultaneously the value of n, the value of K and the methods for selecting the interval cutpoints.

6.2.2 | Understanding the role of P_λ

To get a better understanding of the role of the penalization term P_λ in the loss function in Equation (12), we perform a second study on set of simulated data sets $\mathcal{D}^{5000}_{\text{nonlin.}}$ where we vary the effect of P_λ, the number K of time intervals ($K = 25$, 50, 100, 200) and the selection methods of the interval cutpoints (uniform, quantilesAJ).

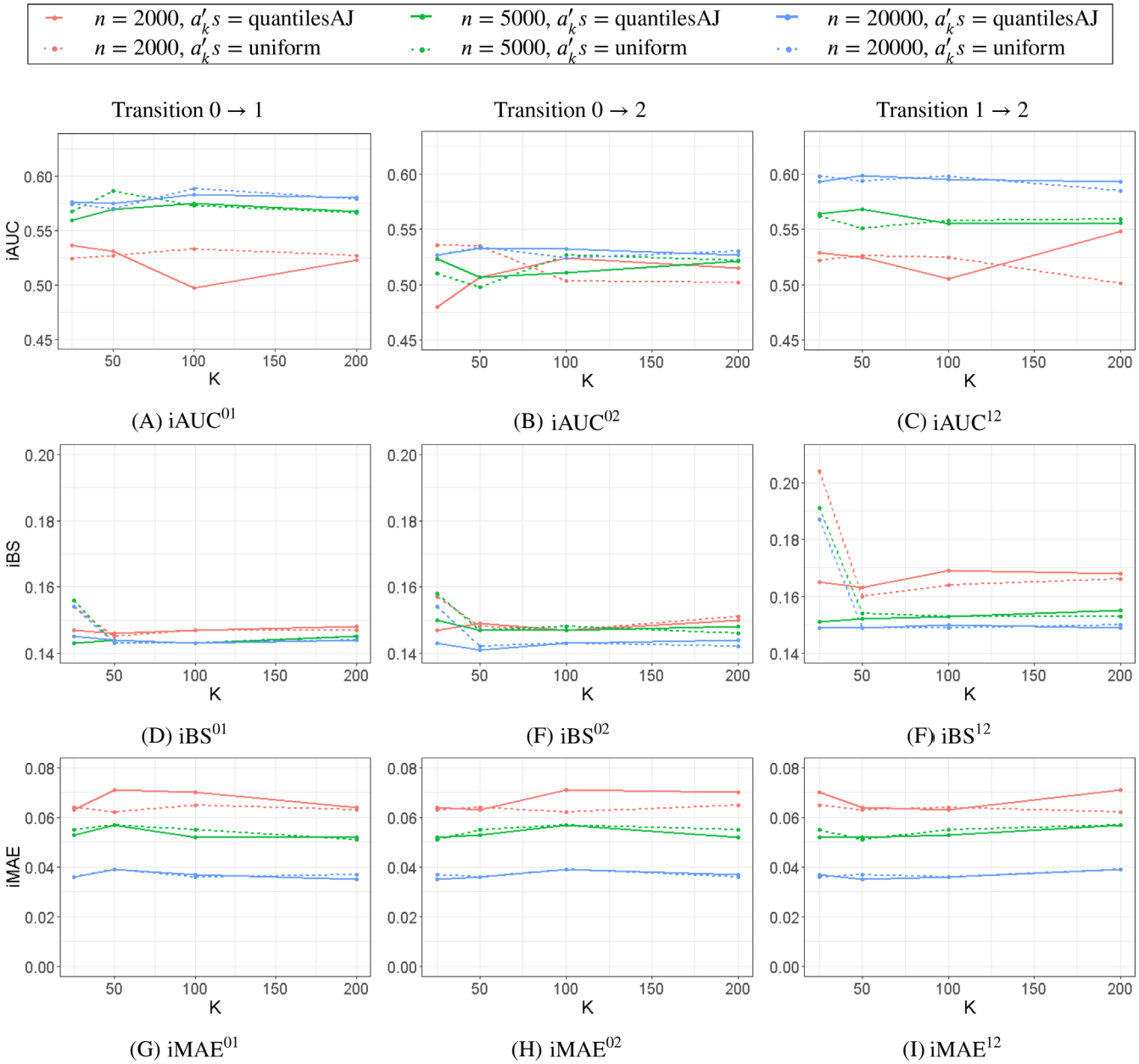


FIGURE 3 Median iAUCs, iBSs, and iMAEs per transition, for each specified values of K and n in the simulation study. The full lines represent the method quantilesAJ, while the dotted lines represent the method uniform

In Figure 4, we plot the transition-specific validated criteria (iAUCs, iBSs, iMAEs) versus the values of K. The full lines represent the method quantilesAJ, while the dotted lines represent the method uniform. The green color represents the performance of IDNetwork with $P_\lambda \neq 0$, while the brown color represents the performance of IDNetwork with $P_\lambda = 0$. In terms of iAUCs, the penalized model (ie, with $P_\lambda \neq 0$) outperforms the unpenalized model (ie, with $P_\lambda = 0$) for all the values of K for the transitions $0 \to 1$ and $1 \to 2$. In terms of iBSs, the penalized and the unpenalized models display similar performances for all the values of K for transitions $0 \to 1$ and $0 \to 2$; the penalized model gives slightly better performance for transition $1 \to 2$. In terms of iMAEs, the penalized model outperforms the unpenalized model for all the values of K for the three transitions. Subsequently, we can see an interaction between the value of K and the effect of the penalization regarding the iAUCs and the iMAEs. For the smaller value of K (ie, $K = 25$), the penalized and the unpenalized models are equivalent. For the larger value of K (ie, $K = 200$), the penalized model outperforms the unpenalized model. Thus, it is evident that when the value of K increases, the performance of the unpenalized model decreases (in particular for the method quantilesAJ) while the performance of the penalized model

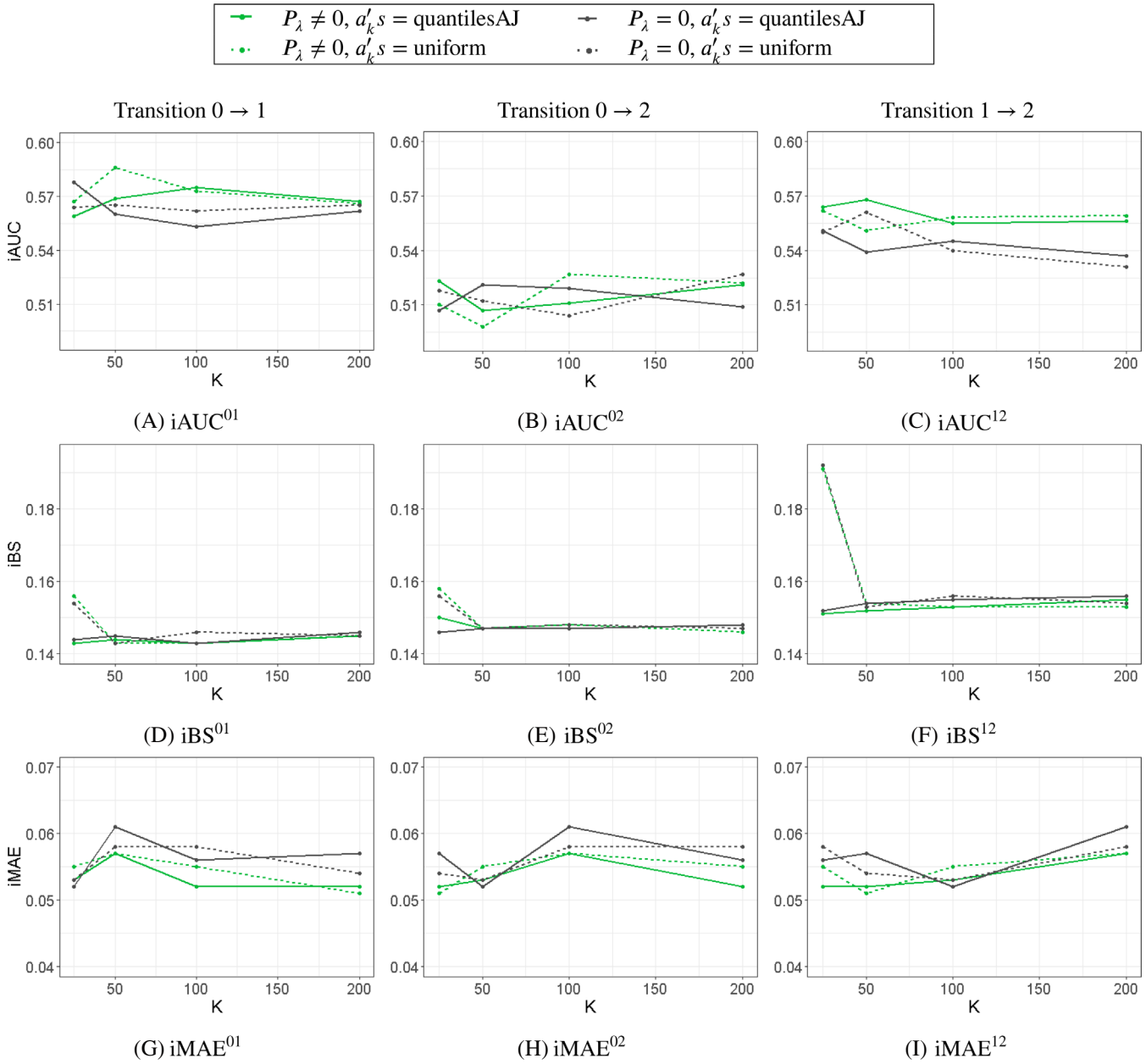
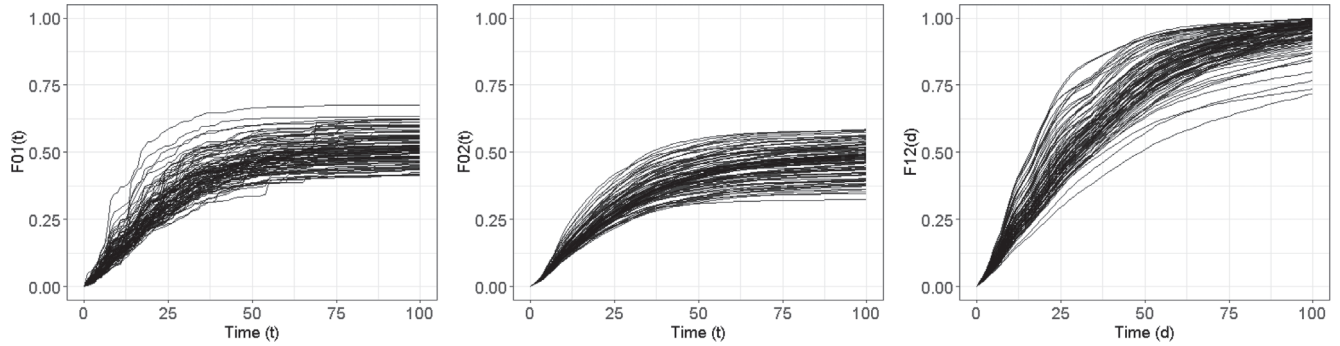


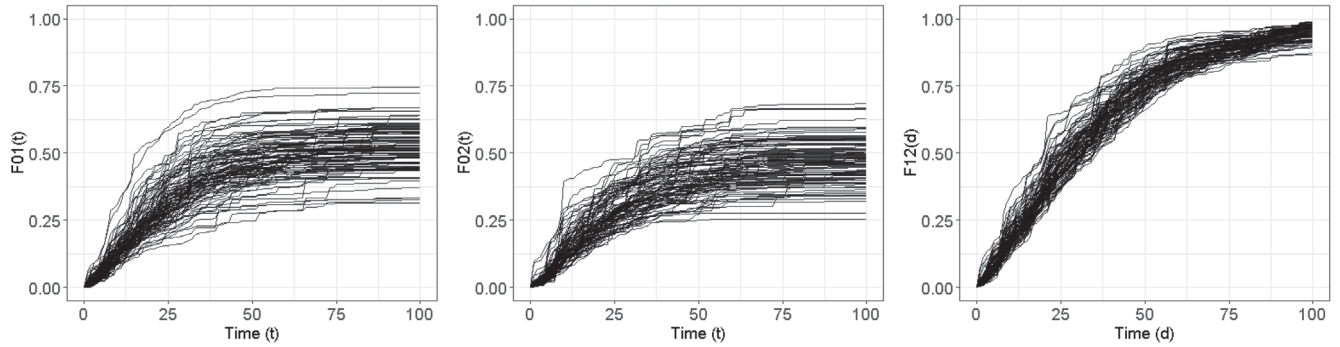
FIGURE 4 Illustration of the effect of the penalization term P_λ in the loss function in Equation (12) on the set of simulated data sets $D^{5000}_{\text{nonlin.}}$: median iAUCs, iBSs, and iMAEs per transition, for each specified values of K. The full lines represent the method quantilesAJ, while the dotted lines represent the method uniform. The red color represents the performance of IDNetwork with $P_\lambda \neq 0$, while the brown color represents the performance of IDNetwork with $P_\lambda = 0$

is stable. Hence, the penalization term P_λ allows to mitigate the effect of a too large value of K on the IDNetwork performance.

In addition, the role of the penalization term P_λ is to provide smooth estimates of the density probability functions. Indeed, it may be reasonable to assume that large temporal variations in the density functions over successive time intervals should be smoothed. For a too large value of K, it will result in the emergence of high jumps when computing the CIFs. This problem can be fixed by applying a temporal smoothing method as done by P_λ. To illustrate our point, we use the simulated data set $D^{5000}_{\text{nonlin.}}$ and generate 100 additional observations. We estimate the CIFs of the additional observations with IDNetwork in the case of $P_\lambda = 0$ (unpenalized model) and in the case of $P_\lambda \neq 0$ (penalized model). Results per transition are shown in Figure 5. For the transitions $0 \to 1$ and $0 \to 2$, the penalization term smoothes the estimates when high jumps are estimated between adjacent time intervals. For the transition $1 \to 2$, estimates of IDNetwork with



(A) $\hat{F}_{01}(.|X)$ estimated with IDNetwork. (B) $\hat{F}_{02}(.|X)$ estimated with IDNetwork. (C) $\hat{F}_{12}(.|X)$ estimated with IDNetwork.



(D) $\hat{F}_{01}(.|X)$ estimated with IDNetwork($P_\lambda = 0$). (E) $\hat{F}_{02}(.|X)$ estimated with IDNetwork($P_\lambda = 0$). (F) $\hat{F}_{12}(.|X)$ estimated with IDNetwork($P_\lambda = 0$).

FIGURE 5 Estimated $\hat{F}_{ql}(.|X)$ in $[0, \tau]$ on the set of simulated data sets $\mathcal{D}^{5000}_{\text{nonlin.}}$ for 100 additional simulated patients $((q,l) \in \{(0,1),(0,2),(1,2)\})$

$P_\lambda = 0$ result in a loss of the variability that is corrected by setting $P_\lambda \neq 0$. We refer the reader to the true functions in the Supplementary materials (Section 8.2); it is evident that the penalization term smooth very well the jumps that are nonexistent in the true curves.

6.3 | Benchmark

We use the set of nonlinear simulated data sets $\mathcal{D}^{5000}_{\text{nonlin.}}$ to compare the predictive performance of IDNetwork with the state-of-the-art methods (see Section 5.3). To benchmark IDNetwork, we set $K = 100$ and divide the time scale into intervals of uniform length.

The integrated predictive performances are shown in Table 2. Detailed results per evaluation times are displayed in the Supplementary materials (Section 9.1). In these simulations, the Cox's linear assumption does not hold anymore. Consequently, as expected, IDNetwork significantly outperforms msCox and msSplineCox with a P-value less than 0.001 for the three transitions in terms of iAUC and iBS. (except for the transition $1 \to 2$ where msCox and msSplineCox outperform IDNetwork but with no statistical difference). IDNetwork significantly outperforms the linear version LinearIDNetwork with a P-value less than 0.001 as well. Moreover, we note that LinearIDNetwork outperforms msCox and msSplineCox. This illustrates the effect of a deep learning approach as compared with a statistical approach. We also evaluate the predictive performances of IDNetwork on a linear simulated data set. Results are shown in the Supplementary materials (Section 9.2).

7 | APPLICATION ON REAL CLINICAL DATA SETS

We conduct experiments on real illness-death data from two clinical trials in colon cancer and one clinical trial in breast cancer.

TABLE 2 Predictive performance (median ± SD) on the validation sets (internal validation with Monte Carlo simulations) for the set of nonlinear simulated data sets $D^{5000}_{\text{nonlin.}}$, with $K = 100$ and a uniform subdivision of the time scale

		Transition			
Criteria	**Algorithm**	**0 → 1**	**0 → 2**	**1 → 2**	**Average**
iAUC	msCox	0.508* ± 0.04	0.489· ± 0.05	0.499* ± 0.04	0.499* ± 0.02
	msSplineCox	0.509* ± 0.05	0.488· ± 0.05	0.501* ± 0.04	0.498* ± 0.01
	LinearIDNetwork	0.517* ± 0.05	0.499 ± 0.06	0.509* ± 0.03	0.502* ± 0.02
	IDNetwork	**0.573 ± 0.03**	**0.527 ± 0.03**	**0.558 ± 0.04**	**0.545 ± 0.01**
iBS	msCox	0.229* ± 0.01	0.241* ± 0.01	**0.152 ± 0.01**	0.206* ± 0.01
	msSplineCox	0.238* ± 0.01	0.249* ± 0.01	**0.152 ± 0.01**	0.211* ± 0.01
	LinearIDNetwork	0.161* ± 0.01	0.160* ± 0.01	0.158 ± 0.01	0.160* ± 0.01
	IDNetwork	**0.143 ± 0.01**	**0.148 ± 0.01**	0.153 ± 0.01	**0.148 ± 0.01**

TABLE 3 Descriptive statistics on the number of (No.) observations in the colon cancer data sets

	No. observations (%)					
Data set	**0 → 1**	**0 → 2**	**0 → cens.**	**1 → 2**	**1 → cens.**	**Total**
NCT00079274	623 (29%)	81 (4%)	1427 (67%)	276 (44%[a])	347 (56%[a])	2121
NCT00275210	279 (25%)	14 (1%)	829 (74%)	132 (47%[a])	147 (53%[a])	1122

[a]Among patients at risk.

TABLE 4 Descriptive statistics on the transition times (in months) in the colon cancer data sets

	Transition								
	0 → 1			**0 → 2**			**1 → 2**		
Data set	**Q1**	**Median**	**Q3**	**Q1**	**Median**	**Q3**	**Q1**	**Median**	**Q3**
NCT00079274	11	15	23	6	14	32	17	25	33
NCT00275210	10	14	22	4	7	23	16	22	30

7.1 | Data on colon cancer

7.1.1 | Description of the data sets

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 (5 years) for RFS and for 96 months (8 years) for OS.[44] It presents 67% of censoring from state 0 and 56% of censoring from state 1 among patients at risk. (2) The study NCT00275210[‡] contains 1122 patients followed over 60 months for RFS and OS.[45] The data set presents 74% of censoring from state 0 and 53% from state 1 among patients at risk. Descriptive statistics of the colon data sets are shown in Tables 3 and 4.

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. Then for the experimentation,

TABLE 5 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), with a uniform subdivision of the time scale in $K = 48$ (months)

			Transition			
Evaluation	**Criteria**	**Algorithm**	**0 → 1**	**0 → 2**	**1 → 2**	**Average**
(1) Internal	iAUC	msCox	**0.677 ± 0.03**	0.658 ± 0.08	0.682 ± 0.04	**0.663 ± 0.03**
		msSplineCox	0.672 ± 0.03	0.600 ± 0.09	**0.692**† **± 0.04**	0.659 ± 0.03
		LinearIDNetwork	0.666 ± 0.03	0.617 ± 0.09	0.665$^{\cdot}$ ± 0.05	0.637 ± 0.04
		IDNetwork	0.669 ± 0.03	**0.660 ± 0.09**	0.662 ± 0.06	0.651 ± 0.03
	iBS	msCox	**0.153 ± 0.01**	0.032‡ ± 0.00	0.201$^{\cdot}$ ± 0.02	0.130 ± 0.01
		msSplineCox	0.156 ± 0.01	0.032$^{\cdot}$ ± 0.01	**0.192**† **± 0.03**	**0.126 ± 0.01**
		LinearIDNetwork	0.154 ± 0.01	0.028 ± 0.00	0.207 ± 0.03	0.130 ± 0.01
		IDNetwork	**0.153 ± 0.01**	**0.027 ± 0.00**	0.212 ± 0.03	0.129 ± 0.01
(2) External	iAUC	msCox	0.669† ± 0.00	0.601* ± 0.02	0.559 ± 0.02	0.610* ± 0.01
		msSplineCox	0.668† ± 0.00	0.598* ± 0.03	**0.566 ± 0.01**	0.613* ± 0.00
		LinearIDNetwork	0.670 ± 0.01	0.694 ± 0.03	0.533† ± 0.02	0.642† ± 0.01
		IDNetwork	**0.673 ± 0.01**	**0.713 ± 0.05**	0.562 ± 0.03	**0.651 ± 0.02**
	iBS	msCox	0.157† ± 0.00	0.013* ± 0.00	0.184* ± 0.01	0.118* ± 0.00
		msSplineCox	0.159† ± 0.00	0.013* ± 0.00	0.176* ± 0.01	0.116* ± 0.00
		LinearIDNetwork	**0.154 ± 0.00**	**0.011 ± 0.00**	**0.146 ± 0.01**	**0.103 ± 0.00**
		IDNetwork	**0.154 ± 0.01**	**0.011 ± 0.00**	0.150 ± 0.01	0.105 ± 0.00

Note: We integrate the AUC and BS measures at all the 30 equidistant time points in $[60, \tau]$ (ie, at every month from 2 months).

the study NCT00079274 will be used for internal validation (training and validation) and the study NCT00275210 for external validation.[41]

For the two 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. For each data set, even though IDNetwork is a continuous-time method, we still need to subdivide the time axis into time intervals. We fix a uniform 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 on the daily time interval $[(j-1) \times 30.5, j \times 30.5)$. We set $K = 48$ (ie, $\tau = 48$ months $= 4$ years) and subdivide the time axis into monthly intervals between 0 and 48 months and set event times after 48 months in a last interval v_{49}.

7.1.2 | Benchmark

We conduct (1) internal and (2) external validation. The integrated predictive performances are shown in Table 5. Results per evaluation times are detailed in the Supplementary materials (Section 10). On the validation splits, neither method outperforms the other for the three transitions independently. In addition, all the models are equivalent (ie, no significant differences) on average. However, IDNetwork shows significant better performance on the external NCT00275210 validation study for all the transitions (excluding in terms of iAUC for transition 1 → 2 where msSplineCox outperforms IDNetwork but with no statistical significance, in terms if iBS for transition 1 → 2 where LinearIDNetwork display better performance but with no statistical significance). On average, IDNetwork displays significant better iAUC and iBS than msCox and msSpline Cox with a P-value less than 0.001, a significant better iAUC than LinearIDNetwork with a P-value less than 0.05.

TABLE 6 Descriptive statistics on the number of (No.) observations in the METABRIC data set

	No. observations (%)					
Data set	**0 → 1**	**0 → 2**	**0 → cens.**	**1 → 2**	**1 → cens.**	**Total**
METABRIC	677 (36%)	509 (27%)	717 (38%)	593 (88%[a])	84 (12%[a])	1903

[a]Among patients at risk.

TABLE 7 Descriptive statistics on the transition times (in months) in the METABRIC data set

	Transition								
	0 → 1			**0 → 2**			**1 → 2**		
Data set	**Q1**	**Median**	**Q3**	**Q1**	**Median**	**Q3**	**Q1**	**Median**	**Q3**
METABRIC	20	39	81	65	113	121	36	61	110

7.2 | Data on breast cancer

7.2.1 | Description of the data set

We use the data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort.[46] For the METABRIC[§] data set, we include 1903 patients followed for 360 months (30 years) for RFS and OS, with 38% of censoring from state 0 and 13% of censoring from state 1 among patients at risk. Descriptive statistics of the data set are shown in Tables 6 and 7.

This data set contains clinical, histo-pathological, gene copy number and gene expression features used to determine breast cancer subgroups. Based on the literature, we select 17 baseline clinical, histo-pathological and gene copy number features (12 categorical, 5 numerical). It includes the following features: age, inference on the menopausal status, Nottingham Prognostic Index (NPI), immunohistochemical oestrogen-receptor (ER) status, number 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 regimen, hormonal regimen, radiotherapy regimen.

The data set contains around 24 000 gene expression features. Before applying methods on this data set, we conduct a preprocessing step by using statistical methods, in order to select a subset of relevant genes to be included in IDNetwork. Several approaches have been reported to integrate gene expression data into survival models. These approaches are based either on dimension reduction, on genes or metagenes selection[47] or, more recently, on the use of a large number of gene expression values (>1000) with the development of deep learning methods.[48,49] We choose a selection approach in order to extract a ranked list of cancer-related genes based on their *P*-values from independent transition-specific Cox P.H. models. The genes are selected using an independent data set in order to control for selection bias; we use the Breast Invasive Carcinoma TCGA PanCancer[¶] data set. We describe our selection process in two steps as follows:

1. For each gene expression feature:
   - We fit independent transition-specific Cox P.H. models including the clinical covariates and each gene expression feature.
   - For each of the three transitions, we compute the *P*-value from a Wald test on the estimated coefficient related to the specific gene.
2. For each transition:
   - We adjust the *P*-values with a Benjamini-Hochberg multi-test correction for each transition independently. See the Supplementary materials (Section 11) for details on the *P*-values per transition.
   - We rank the *P*-values for each transition.

TABLE 8 Number of (No.) genes selected per transition on the Breast Invasive Carcinoma TCGA PanCancer data set for each threshold

	No. genes selected per transition			
Threshold (α)	**0 → 1**	**0 → 2**	**1 → 2**	**Total[a]**
0.001	8	0	1	9
0.005	17	1	1	19
0.01	26	2	1	29
0.05	76	6	3	84
0.1	110	36	11	156

[a]Total number of unique genes selected.

- We use different thresholds, noted α, of the P-values for gene selection. We present the number of gene selected for each threshold in Table 8. For the transition $0 \to 1$, several genes are selected (up to 110 genes for the threshold 0.1). While for the transitions $0 \to 2$ and $1 \to 2$, very few genes are selected. In order to assess the biological meaning of the selected genes, we perform a gene set enrichment analysis[50] for each transition-specific list of genes against the Hallmark[51] collection of the Molecular Signatures Database (MSigDB). For the transitions $0 \to 2$ and $1 \to 2$, there is no significant enriched terms. For the transition $0 \to 1$, there are 9 significant enriched terms. We refer the reader to the Supplementary materials (Section 12) for detailed results on the analysis. In particular, the most significant enriched term "Epithelial Mesenchymal Transition" is related to a core biological component of the metastatic process in breast cancer[52] and therefore well related to a primary cancer progression through a state relapse. Hence, the gene selected on the transition $0 \to 1$ display a coherent functional profile associated to cancer relapse. The addition of the selected genes in an illness-death model is thus expected to improve its performance for the transition $0 \to 1$ more than the others.

For all the features, 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 uniform 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 on the daily time interval $[(j-1) \times 30.5, j \times 30.5)$. We fix $K = 120$ (ie, $\tau = 120$ months $=$ 10 years) and subdivide the time axis into monthly intervals between 0 and 120 months and set event times after 120 months in a last interval v_{121}.

7.2.2 | Benchmark

With the Metabric data set we aim to compare the scalability performance of IDNetwork with other methods when including a large number of features and varying the amount of features. The integrated predictive performances are shown in Tables 9 and 10. Detailed results per evaluation times are displayed in the Supplementary materials (Section 13). We compare different models with different lists of genes varying the threshold α of the P-values. For each of the four compared algorithms, the models show different performances in terms of iAUC. The algorithms msCox and msSplineCox display similar iAUC for transition $0 \to 1$, better iAUC for transition $0 \to 2$ and poorer iAUC for transition $1 \to 2$ when the value of α increases. The algorithms IDNetwork and LinearIDNetwork improve their iAUC for transition $0 \to 1$ and $1 \to 2$ when the value of α increases. As expected, IDNetwork and LinearIDNetwork display similar iAUC for the transitions $0 \to 2$ and $1 \to 2$ when the value of α increases, as most of the genes included are informative on transition $0 \to 1$. Hence, the two deep learning methods have a better iAUC for the transition $0 \to 1$ thanks to the addition of the selected gene features in the modelization. This result is consistent with the gene set enrichment analysis realized in Section 7.2.1 in which we have highlighted than the list of genes selected for the transition $0 \to 1$ is particularly related to the metastatic process whereas there are no enriched terms in the genes selected for the transitions $0 \to 2$ and $1 \to 2$. In terms of iBS, we can see a decreasing of the msCox and msSpline performances when the value of α increases. Whereas, LinearIDNetwork and IDNetwork performances remain stable. Hence it seems that the deep learning methods are less sensitive in terms of iBS to the increase in the number of features. Finally, we chose the model $M_{0.1}^{BH}$ as the best model to predict transitions

TABLE 9 Integrated AUCs (median ± SD) on the validation sets (internal validation) for the METABRIC data set, with a uniform subdivision of the time scale in $K = 120$ (months)

			Transition			
Criteria	**Model**	**Algorithm**	**0 → 1**	**0 → 2**	**1 → 2**	**Average**
iAUC	M_0	msCox	**0.686 ± 0.02**	**0.702**† **± 0.05**	0.725 ± 0.04	**0.711**† **± 0.02**
		msSplineCox	**0.686 ± 0.02**	**0.702**† **± 0.05**	0.729 ± 0.04	0.710† ± 0.02
		LinearIDNetwork	0.649‡ ± 0.04	0.609* ± 0.05	0.705 ± 0.04	0.657* ± 0.03
		IDNetwork	0.681 ± 0.03	0.671 ± 0.05	**0.732 ± 0.05**	0.689 ± 0.03
	$M^{BH}_{0.001}$	msCox	**0.697 ± 0.03**	0.740* ± 0.03	0.728 ± 0.02	0.720* ± 0.01
		msSplineCox	0.694 ± 0.03	**0.751*** **± 0.04**	**0.734 ± 0.01**	**0.726**‡ **± 0.01**
		LinearIDNetwork	0.657‡ ± 0.03	0.591* ± 0.06	0.718 ± 0.03	0.651* ± 0.03
		IDNetwork	0.689 ± 0.03	0.659 ± 0.05	0.717 ± 0.04	0.692 ± 0.03
	$M^{BH}_{0.005}$	msCox	**0.702**† **± 0.02**	**0.750*** **± 0.03**	0.722 ± 0.03	**0.724*** **± 0.01**
		msSplineCox	0.690 ± 0.02	0.756* ± 0.04	**0.723 ± 0.02**	0.720‡ ± 0.02
		LinearIDNetwork	0.656† ± 0.03	0.629‡ ± 0.07	0.703$^{\cdot}$ ± 0.03	0.653* ± 0.03
		IDNetwork	0.684 ± 0.03	0.677 ± 0.04	0.722 ± 0.03	0.692 ± 0.02
	$M^{BH}_{0.01}$	msCox	0.693 ± 0.02	0.753* ± 0.03	0.710 ± 0.04	0.713* ± 0.02
		msSplineCox	**0.694 ± 0.02**	**0.763*** **± 0.03**	**0.721 ± 0.04**	**0.724**‡ **± 0.02**
		LinearIDNetwork	0.665† ± 0.03	0.615$^{\cdot}$ ± 0.06	0.715 ± 0.03	0.654‡ ± 0.02
		IDNetwork	0.685 ± 0.03	0.649 ± 0.06	0.717 ± 0.03	0.690 ± 0.03
	$M^{BH}_{0.05}$	msCox	0.696 ± 0.03	**0.743*** **± 0.03**	0.692‡ ± 0.03	0.709* ± 0.02
		msSplineCox	**0.716 ± 0.03**	0.722‡ ± 0.03	**0.729 ± 0.05**	**0.739**$^{\cdot}$ **± 0.04**
		LinearIDNetwork	0.665* ± 0.03	0.627 ± 0.05	0.720 ± 0.02	0.667‡ ± 0.02
		IDNetwork	0.704 ± 0.03	0.640 ± 0.05	0.722 ± 0.02	0.686 ± 0.02
	$M^{BH}_{0.1}$	msCox	0.681* ± 0.04	**0.722*** **± 0.03**	0.671* ± 0.04	0.691 ± 0.02
		msSplineCox	0.686 ± 0.01	0.683 ± 0.02	0.688$^{\cdot}$ ± 0.02	0.686 ± 0.01
		LinearIDNetwork	0.670* ± 0.03	0.634 ± 0.05	0.719 ± 0.03	0.679‡ ± 0.02
		IDNetwork	**0.709 ± 0.03**	0.661 ± 0.05	**0.732 ± 0.03**	**0.697 ± 0.02**

Note: We integrate the AUC and BS measures at all the 30 equidistant time points in $[90, \tau]$ (ie, at every month from 3 months). The model M_0 uses only the clinical features and each model M^{BH}_{α}, for $\alpha \in [0.001, 0.005, 0.01, 0.05, 0.1]$, uses the clinical features and the genes selected with a corrected P-value less than α.

$0 \to 1$ and $1 \to 2$. With the model $M^{BH}_{0.1}$, IDNetwork outperforms significantly the other algorithms with an averaged iAUC of 0.697 and an averaged iBS of 0.124, an iAUC of 0.714 and an iBS of 0.142 for transition $0 \to 1$, an iAUC of 0.730 and an iBS of 0.167 for transition $1 \to 2$. IDNetwork show poorer performance for 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.

8 | DISCUSSION

We present IDNetwork a novel deep learning method to model an illness-death process and to predict two-stage evolution of a disease based on baseline covariates. To the best of our knowledge, it is the first deep learning architecture developed in the context of multi-state analysis. It outperforms standard methodologies in this context. In clinical practice, IDNetwork may be useful in personalized medicine by providing predictions of the risks of relapse and death. It could help physicians to adapt the therapeutic guidelines for a specific patient.

TABLE 10 Integrated BSs (median ± SD) on the validation sets (internal validation) for the METABRIC data set, with a uniform subdivision of the time scale in $K = 120$ (months)

			Transition			
Criteria	**Model**	**Algorithm**	**0 → 1**	**0 → 2**	**1 → 2**	**Average**
iBS	M_0	msCox	**0.145 ± 0.01**	0.071* ± 0.01	**0.156 ± 0.01**	0.124 ± 0.01
		msSplineCox	0.146 ± 0.01	0.073* ± 0.01	0.161 ± 0.01	0.126† ± 0.01
		LinearIDNetwork	0.149 ± 0.01	0.061 ± 0.01	0.175† ± 0.02	0.127‡ ± 0.01
		IDNetwork	**0.145 ± 0.01**	**0.059 ± 0.01**	0.164 ± 0.01	**0.122 ± 0.01**
	$M^{BH}_{0.001}$	msCox	**0.145 ± 0.01**	0.067* ± 0.01	**0.159† ± 0.01**	**0.124 ± 0.00**
		msSplineCox	0.146 ± 0.01	0.065† ± 0.01	0.165 ± 0.01	0.125 ± 0.01
		LinearIDNetwork	0.146 ± 0.01	0.060 ± 0.01	0.173 ± 0.01	0.126 ± 0.01
		IDNetwork	0.146 ± 0.01	**0.059 ± 0.01**	0.166 ± 0.02	**0.124 ± 0.01**
	$M^{BH}_{0.005}$	msCox	**0.145 ± 0.01**	0.068* ± 0.01	**0.164 ± 0.01**	0.126 ± 0.00
		msSplineCox	0.150 ± 0.01	0.071‡ ± 0.00	0.169 ± 0.01	0.131 ± 0.01
		LinearIDNetwork	0.146 ± 0.01	0.060 ± 0.01	0.178· ± 0.01	0.128 ± 0.01
		IDNetwork	0.146 ± 0.01	**0.058 ± 0.01**	0.168 ± 0.02	**0.125 ± 0.01**
	$M^{BH}_{0.01}$	msCox	0.146 ± 0.01	0.068* ± 0.01	**0.167 ± 0.01**	0.126 ± 0.00
		msSplineCox	0.148 ± 0.01	0.069* ± 0.00	0.174 ± 0.01	0.130 ± 0.01
		LinearIDNetwork	0.144 ± 0.01	**0.060 ± 0.01**	0.175 ± 0.01	0.127 ± 0.01
		IDNetwork	**0.145 ± 0.01**	**0.060 ± 0.01**	0.171 ± 0.02	**0.124 ± 0.01**
	$M^{BH}_{0.05}$	msCox	0.149† ± 0.01	0.071* ± 0.01	0.185† ± 0.02	0.134* ± 0.01
		msSplineCox	0.144 ± 0.01	0.068· ± 0.00	**0.165 ± 0.03**	0.126 ± 0.01
		LinearIDNetwork	0.144 ± 0.01	0.060 ± 0.01	0.175† ± 0.01	0.127† ± 0.01
		IDNetwork	**0.142 ± 0.01**	**0.059 ± 0.01**	0.167 ± 0.01	**0.123 ± 0.01**
	$M^{BH}_{0.1}$	msCox	0.157* ± 0.01	0.072* ± 0.01	0.207* ± 0.01	0.145* ± 0.01
		msSplineCox	0.162† ± 0.01	0.066 ± 0.01	0.201‡ ± 0.01	0.143‡ ± 0.00
		LinearIDNetwork	**0.145 ± 0.01**	**0.060 ± 0.01**	0.178† ± 0.02	0.127† ± 0.01
		IDNetwork	**0.145 ± 0.01**	**0.060 ± 0.01**	**0.165 ± 0.01**	**0.124 ± 0.01**

Note: We integrate the AUC and BS measures at all the 30 equidistant time points in $[90, \tau]$ (ie, at every months from 3 months). The model M_0 uses only the clinical features and each model M^{BH}_{α}, for $\alpha \in [0.001, 0.005, 0.01, 0.05, 0.1]$, uses the clinical features and the genes selected with a corrected P-value less than α.

Prognostication of diseases is a key momentum in the medical decision process of various diseases, for example to identify population at risk of cardiovascular complications or to identify population at risk of cancer relapse. The most used approaches are based on nomograms and scores computed based on a few clinical and biological features. For example, the CHAD2-DS2-VASC score is commonly used by physicians to identify patients who require anticoagulant treatment following the diagnosis of a cardiac atrial arrhythmia.[53] In the same way, the RSClin tool that combines clinical, pathological and genetic information has been developed in oncology to predict the risk of breast cancer relapse and to determine more precisely which patients need chemotherapy in addition to surgery.[54] But a lot of medical information contained in the patients' records is left unexploited. In parallel to the development of new biomarkers relying on highly specialized technologies, another approach is to focus on the optimization of prognostic models based on large but accessible information. In the area of digital medicine which aims to capture and combine a huge amount of data, healthcare workers will need specialized algorithms to support their practices based on standardized guidelines on the one hand, and on personalized assessment on the other hand. IDNetwork has been developed in that sense.

IDNetwork uses a multi-task architecture and transition-specific subnetworks to learn an estimation of the density probabilities of occurrence of state transitions of an illness death process. It uses piecewise approximations to provide

accurate predictions of the cumulative probabilities of state transitions. IDNetwork uses multiple fully connected layers and nonlinear activation functions to model the relationships between covariates and risks of transitions without any assumption. It is trained by minimizing a loss function designed to both capture the relationships between covariates and risks of transitions and provide smooth piecewise approximations of the density probabilities.

Through experiments on simulated data sets, we investigate different configurations of IDNetwork and illustrate the benefit of our loss function. We compare the predictive performances of our method with the state-of-the-art methods using discrimination and calibration criteria. We evaluate IDNetwork in predicting the cumulative probabilities of state transitions on a simulated data set and on real data sets on colon and breast cancer. We show that IDNetwork provides significant improvements compared with the others methods.

Furthermore, medical decision-making requires to combine heterogeneous individual features. On the real data set on breast cancer, we illustrate how IDNetwork can be easily adapted to integrate various types of data (as clinical, biological, molecular, gene expression) and displays in this case coherent and significant improvements in comparison with statistical methods.

Developing a reliable multi-state model with a small training set and few events can be challenging and can result in a poor predictive model. Time-to-event data collected in a clinical setting can suffer from a high censoring rate, especially for rare events. Applying survival models on data sets with fewer events than censored observations can impair the risks of events estimates. In addition, deep learning methods require large amount of training data. This is rarely the case with data sets from clinical trials where the number of patients is relatively low compared to the databases commonly used in deep learning. We adapt the architecture of IDNetwork to handle these limitations by using hyper-parameters to simplify the architecture when available training data may require it for each of the three transitions. Moreover, given the increasing popularity of using real-world data (RWD) collection, such as electronic health records (EHRs) or disease registries, we aim to exploit these data in the future for our problematic to make more efficient and reliable decision-making.

Explaining predictions in deep neural networks is challenging but essential in clinical applications where interpretability and reliability are evaluated to support medical decisions. For the future work, it may be relevant to add a model interpretability functionality to IDNetwork in order to help clinicians in establishing the patient prognosis. It could reveal what are the patient characteristics associated with each transition and increase the understanding of the evolution of the disease. Some methods have already been developed to explain predictions in deep neural networks.[55] But these methods are mainly used in the domain of image classification and are rarely developed in contexts like ours. It will be therefore necessary to adapt these methods for multi-state analysis in order to understand the relationship between covariates and risks of transitions. It would be interesting to illustrate the interaction effects between covariates in IDNetwork as well. In particular, it could reveal the role of gene interactions in the variation of the risks.

IDNetwork is a flexible method developed for an illness-death process and can readily be applied in many cancers and cardiovascular diseases to predict two-stage evolution. It can be generalized to embrace more complex disease evolution patterns by adapting the states and transitions.[56] For example, the evolution of a disease should be modeled with multiple intermediate states to reflect the long-term patient journey composed of sequential treatments. Therefore, IDNetwork could provide a more realistic prediction of disease evolution through different phases. At the end, it could support medical decisions not only based on a single event as done by traditional approaches, but by anticipating relevant outcomes closed to the disease course of a real patient.

In the same way, IDNetwork could be generalized to model time-varying covariates (rather than only baseline covariates) by adapting the model architecture and the loss function. Incorporating time-varying features could provide dynamic updated predictions of the cumulative incidence functions based on all the patients' characteristics measured since the beginning of their clinical follow-up.

ACKNOWLEDGEMENTS

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.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

We use two data sets from Phase III clinical trials in non-metastatic colon cancer. (1) The data from the clinical trial NCT00079274[44] is available under request at https://data.projectdatasphere.org/projectdatasphere/html/content/161.

(2) The data from the clinical trial NCT00275210[45] is available under request at https://data.projectdatasphere.org/projectdatasphere/html/content/128.

We use also the data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort.[46] The METABRIC data set is available at https://www.cbioportal.org/study/clinicalData?id=brca_metabric.

ENDNOTES

*https://www.tensorflow.org/.

†The clinical trial NCT00079274 data set is available under request at https://data.projectdatasphere.org/projectdatasphere/html/content/161.

‡The clinical trial NCT00275210 data set is available under request at https://data.projectdatasphere.org/projectdatasphere/html/content/128.

§The METABRIC data set is available at https://www.cbioportal.org/study/clinicalData?id=brca_metabric.

¶The Breast Invasive Carcinoma TCGA PanCancer data set is available at https://www.cbioportal.org/study/clinicalData?id=brca_tcga_pan_can_atlas_2018.

ORCID

Aziliz Cottin https://orcid.org/0000-0003-4394-7002

REFERENCES

1. Webster AJ. Multi-stage models for the failure of complex systems, cascading disasters, and the onset of disease. *PLoS One*. 2019;14(5):e0216422.
2. Eulenburg C, Mahner S, Woelber L, Wegscheider K. A systematic model specification procedure for an illness-death model without recovery. *PLoS One*. 2015;10(4):e0123489.
3. Iacobelli S, Carstensen B. Multiple time scales in multi-state models. *Stat Med*. 2013;32(30):5315-5327.
4. Commenges D, Joly P, Letenneur L, Dartigues J-F. Incidence and mortality of Alzheimer's disease or dementia using an illness-death model. *Stat Med*. 2004;23(2):199-210.
5. Ramezankhani A, Blaha MJ, Mirbolouk M, Azizi F, Hadaegh F. Multi-state analysis of hypertension and mortality: application of semi-Markov model in a longitudinal cohort study. *BMC Cardiovasc Disord*. 2020;20(1):1-13.
6. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481.
7. Aalen O. Nonparametric inference for a family of counting processes. *Ann Stat*. 1978;6(4):701-726.
8. Cox DR. Regression models and life-tables. *J Royal Stat Soc Ser B (Methodol)*. 1972;34(2):187-202.
9. De Wreede LC, Fiocco M, Putter H. The mstate package for estimation and prediction in non-and semi-parametric multi-state and competing risks models. *Comput Methods Prog Biomed*. 2010;99(3):261-274.
10. Faraggi D, Simon R. A neural network model for survival data. *Stat Med*. 1995;14(1):73-82.
11. Luck M, Sylvain T, Cardinal H, Lodi A, Bengio Y. Deep learning for patient-specific kidney graft survival analysis; 2017. arXiv preprint arXiv:1705.10245.
12. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*. 2018;18(1):24.
13. Fotso S. Deep neural networks for survival analysis based on a multi-task framework; 2018. arXiv preprint arXiv:1801.05512.
14. Kvamme H, Borgan Ø, Scheel I. Time-to-event prediction with neural networks and Cox regression. *J Mach Learn Res*. 2019;20(129):1-30.
15. Giunchiglia E, Nemchenko A, Schaar M. Rnn-surv: a deep recurrent model for survival analysis. Proceedings of the International Conference on Artificial Neural Networks; 2018:23-32; Springer, Cham.
16. Ren K, Qin J, Zheng L, et al. Deep recurrent survival analysis; 2019:4798-4805.
17. Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. *PeerJ*. 2019;7:e6257.
18. Lee C, Zame WR, Yoon J, Schaar M. Deephit: a deep learning approach to survival analysis with competing risks; 2018.
19. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: John Wiley & Sons; 2011.
20. Friedman M. Piecewise exponential models for survival data with covariates. *Ann Stat*. 1982;10(1):101-113.
21. Kvamme H, Borgan Ø. Continuous and discrete-time survival prediction with neural networks; 2019. arXiv preprint arXiv:1910.06724.
22. Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Berlin, Germany: Springer Science & Business Media; 2012.
23. Andersen PK, Borgan Ø. Counting process models for life history data: a review. Preprint series statistical research report NBN: no-23420. 1984. http://urn nb no/URN.
24. Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C, Andersen PK. Multi-state models for the analysis of time-to-event data. *Stat Methods Med Res*. 2009;18(2):195-222.
25. Jackson CH. flexsurv: a platform for parametric survival modeling in R. *J Stat Softw*. 2016;70:1-25.
26. Andersen PK, Perme MP. Inference for outcome probabilities in multi-state models. *Lifetime Data Anal*. 2008;14(4):405.
27. Martinussen T, Scheike TH. *Dynamic Regression Models for Survival Data*. Berlin, Germany: Springer Science & Business Media; 2007.
28. Murphy SA, Sen PK. Time-dependent coefficients in a Cox-type regression model. *Stoch Process Appl*. 1991;39(1):153-180.
29. Achab M, Guilloux A, Gaïffas S, Bacry E. SGD with variance reduction beyond empirical risk minimization; 2015. arXiv preprint arXiv:1510.04822.
30. Meira-Machado L, Sestelo M. Estimation in the progressive illness-death model: a nonexhaustive review. *Biom J*. 2019;61(2):245-263.

31. Wey A, Salkowski N, Kremers W, Ahn YS, Snyder J. Piecewise exponential models with time-varying effects: estimating mortality after listing for solid organ transplant. *Stat*. 2020;9(1):e264.
32. Triebel H. *Theory of Function Spaces*. Monographs in Mathematics. Vol 78; Switzerland: Birkhäuser Basel; 1983.
33. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26(11):2389-2430.
34. Ruder S. An overview of multi-task learning in deep neural networks; 2017. arXiv preprint arXiv:1706.05098.
35. Lee C, Yoon J, Van Der Schaar M. Dynamic-deephit: a deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans Biomed Eng*. 2019;67(1):122-133.
36. Möst S. *Regularization in Discrete Survival Models*. PhD thesis. LMU; 2014.
37. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. *J Royal Stat Soc Ser B (Stat Methodol)*. 2005;67(1):91-108.
38. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand J Stat*. 1978;5(3):141-150.
39. Jacqmin-Gadda H, Blanche P, Chary E, Touraine C, Dartigues J-F. Receiver operating characteristic curve estimation for time to event with semicompeting risks and interval censoring. *Stat Methods Med Res*. 2016;25(6):2750-2766.
40. Spitoni C, Lammens V, Putter H. Prediction errors for state occupation and transition probabilities in multi-state models. *Biom J*. 2018;60(1):34-48.
41. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol*. 2013;13(1):33.
42. Wilcoxon F. Individual comparisons by ranking methods. *Breakthroughs in Statistics*. New York, NY: Springer; 1992:196-202.
43. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med*. 2005;24(11):1713-1723.
44. Alberts SR, Sargent DJ, Nair S, et al. Effect of oxaliplatin, fluorouracil, and leucovorin with or without cetuximab on survival among patients with resected stage III colon cancer: a randomized trial. *JAMA*. 2012;307(13):1383-1393.
45. André T, Boni C, Mounedji-Boudiaf L, et al. Oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment for colon cancer. *N Engl J Med*. 2004;350(23):2343-2351.
46. Curtis C, Shah SP, Chin S-F, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346-352.
47. Van Wieringen WN, Kun D, Hampel R, Boulesteix A-L. Survival prediction using gene expression data: a review and comparison. *Comput Stat Data Anal*. 2009;53(5):1590-1603.
48. Yousefi S, Amrollahi F, Amgad M, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci Rep*. 2017;7(1):1-11.
49. Ching T, Zhu X, Garmire LX. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput Biol*. 2018;14(4):e1006076.
50. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44(W1):W90-W97.
51. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. *Cell Syst*. 2015;1(6):417-425.
52. Wu Y, Sarkissyan M, Vadgama JV. Epithelial-mesenchymal transition and breast cancer. *J Clin Med*. 2016;5(2):13.
53. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*. 2010;137(2):263-272.
54. Sparano JA, Crager MR, Tang G, Gray RJ, Stemmer SM, Shak S. Development and validation of a tool integrating the 21-gene recurrence score and clinical-pathological features to individualize prognosis and prediction of chemotherapy benefit in early breast cancer. *J Clin Oncol*. 2021;39(6):557-564.
55. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for deep neural networks; 2017. arXiv preprint arXiv:1711.06104.
56. Jackson C. *Multi-state Modelling with R: The MSM Package*. Cambridge, UK: Cambridge University Press; 2007:1-53.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Cottin A, Pecuchet N, Zulian M, Guilloux A, Katsahian S. IDNetwork: A deep illness-death network based on multi-state event history process for disease prognostication. *Statistics in Medicine*. 2022;41(9):1573-1598. doi: 10.1002/sim.9310