Artificial Intelligence In Medicine 147 (2024) 102741

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

MS-CPFI: A model-agnostic Counterfactual Perturbation Feature Importance algorithm for interpreting black-box Multi-State models

Aziliz Cottin [a,b,c,*], Marine Zulian [a], Nicolas Pécuchet [a], Agathe Guilloux [c], Sandrine Katsahian [b,c,d,e]

[a] Healthcare and Life Sciences Research, Dassault Systemes, France
[b] Université Paris Cité, France
[c] HeKa team, INRIA, Paris, France
[d] Medical Informatics, Biostatistics and Public Health Department, Georges Pompidou, Assistance Publique-Hôpitaux de Paris, France
[e] Inserm, Centre d'Investigation Clinique 1418 (CIC1418) Epidémiologie Clinique, Paris, France

ARTICLE INFO

Keywords:
Multi-state model
Deep learning
Interpretability
Feature importance
Medical decision

ABSTRACT

Multi-state processes (Webster, 2019) are commonly used to model the complex clinical evolution of diseases where patients progress through different states. In recent years, machine learning and deep learning algorithms have been proposed to improve the accuracy of these models' predictions (Wang et al., 2019). However, acceptability by patients and clinicians, as well as for regulatory compliance, require interpretability of these algorithms's predictions.

Existing methods, such as the Permutation Feature Importance algorithm, have been adapted for interpreting predictions in black-box models for 2-state processes (corresponding to survival analysis). For generalizing these methods to multi-state models, we introduce a novel model-agnostic interpretability algorithm called *Multi-State Counterfactual Perturbation Feature Importance* (MS-CPFI) that computes feature importance scores for each transition of a general multi-state model, including survival, competing-risks, and illness-death models. MS-CPFI uses a new counterfactual perturbation method that allows interpreting feature effects while capturing the non-linear effects and potentially capturing time-dependent effects.

Experimental results on simulations show that MS-CPFI increases model interpretability in the case of non-linear effects. Additionally, results on a real-world dataset for patients with breast cancer confirm that MS-CPFI can detect clinically important features and provide information on the disease progression by displaying features that are protective factors versus features that are risk factors for each stage of the disease.

Overall, MS-CPFI is a promising model-agnostic interpretability algorithm for multi-state models, which can improve the interpretability of machine learning and deep learning algorithms in healthcare.

1. Introduction

The multi-state approach [1] allow to model within a unique framework the occurrence of different clinical events. For instance, in the case of breast cancer, a three-state model, called illness-death model, with an initial state, an intermediate state of relapse, a final state of death, and three transitions, can help characterize the risks of cancer progression [2]. Similarly, for patients with prostate cancer, a model, called competing risks model, with two final states can analyze the risks of cancer-related deaths and deaths due to other causes [3]. These are specific examples of multi-state models that can be used to model disease progression (see Section 2.2.2).

The Cox proportional hazards (Cox P.H.) model [4], the most popular in survival analysis (i.e. 2 states scenarios), is a popular statistical method designed for modeling transition-specific risks in multi-state models [5], but it has strong assumptions and limitations. To overcome these limitations, an increasing number of machine learning and deep learning algorithms has been proposed in recent years. For example, Lee et al. [6] developed DeepHit, a deep neural network that models

* Corresponding author at: Healthcare and Life Sciences Research, Dassault Systemes, France.
E-mail address: aziliz.cottin@3ds.com (A. Cottin).

https://doi.org/10.1016/j.artmed.2023.102741
Received 31 May 2023; Received in revised form 17 November 2023; Accepted 28 November 2023
Available online 29 November 2023

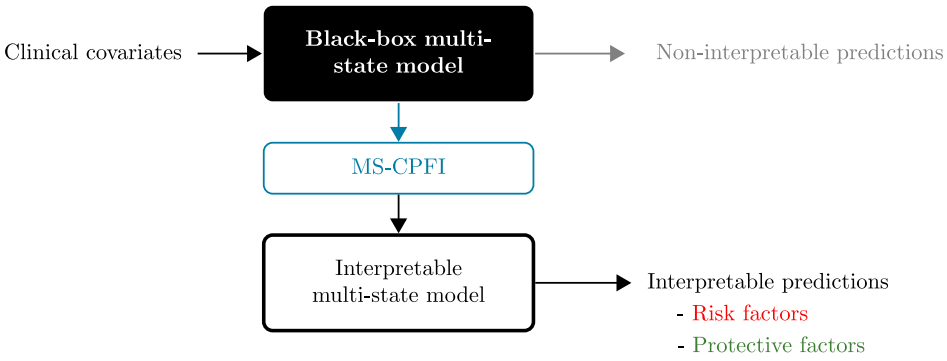


Fig. 1. Interpretable black-box multi-state model.

competing events and handles non-linear relationships between patients' covariates and event risks in a competing-risks design. Similarly, Cottin et al. [7] used deep neural networks to model a three-state illness-death model with IDNetwork. In addition, machine learning [8–10] and deep learning models [11–17] have also been developed for survival analysis.

While machine learning algorithms, especially deep neural networks, have shown promising results in disease progression modeling, interpreting the predictions of these models can be challenging due to their black-box nature. Interpreting the model's predictions is essential to ensure that the predictions are reliable, trustworthy and accurate, making it critical in the context of clinical decision support software. For example, feature selection is a significant concern in machine learning for algorithm trustworthy, and in particular for an application in clinical medicine [18]. Moreover, human understanding and interpretability of this kind of algorithms is a key issue for regulation agencies. In Europe, the AI act requires machine learning-based decision-making softwares to be explainable for their use in practice [19,20]. In the same way, in the USA, the FDA has recently published a new guidance [21] on clinical decision support tools, making necessary for AI-based algorithms to be interpretable and understandable to be certified as medical devices.

The growing concern over the interpretability of AI-based algorithms has prompted significant attention in recent years, resulting in the development of post-hoc interpretability algorithms in various domains. Some of these existing algorithms have been extended for interpreting black-box survival (i.e. two-states) algorithms. In particular, the Permutation Feature Importance (PFI) algorithm [22] stands out as the most intuitive post-hoc interpretability algorithm in the literature. Originally developed for random forests (RFs) [22], PFI computes feature importance scores, for each input feature, by using random permutation and assessing how this feature values perturbation impacts the model error. It has also been extended for application to random survival forests (RSFs) [9] and RSFs for competing risks [23]. However, none of these methods has been extended to interpret a multi-state model.

This paper introduces the Multi-State Counterfactual Perturbation Feature Importance (MS-CPFI) algorithm, aiming to enhance model-agnostic feature importance interpretation for predictions from black-box multi-state algorithms. MS-CPFI computes feature importance for each transition within a multi-state model, making it applicable to diverse models, including survival models, those with competing events, illness-death models, or any multi-state model with irreversible transitions (Fig. 1). Our contributions to the existing PFI algorithm include a model-agnostic, transition-specific feature importance score, a score computed directly on algorithm predictions, enabling interpretable signs, applicability to any time-dependent probability risk-specific distribution, and a novel counterfactual perturbation method, replacing random permutation to identify non-linearities in feature effects.

This paper is organizing as follows. In Section 2, we survey existing strategies to mitigate the 'black box' effect inherent in machine learning algorithms, with a specific focus on model-agnostic approaches (Section 2.1). We present both general approaches and those specifically tailored for interpreting predictions from event history algorithms. Additionally, we delve into the mathematical concepts associated with multi-state models and explore specific instances of multi-state processes (Section 2.2). Moving to Section 3, we detail our contributions, beginning with the adaptation of the Permutation Feature Importance (PFI) algorithm to generalize for any multi-state process (Section 3.1). We introduce our innovative counterfactual perturbation approach (Section 3.3) and provide the mathematical definition of Multi-State Counterfactual Perturbation Feature Importance (MS-CPFI) (Section 3.2). Further, we discuss the computation of confidence intervals for the estimated feature importance score (Section 3.4). In Section 4, we conduct a simulation study to highlight our contributions, showcasing the advantages of MS-CPFI, its robustness, and acknowledging potential limitations. We apply, in Section 5, MS-CPFI to interpret the impact of clinical features on predictions in the METABRIC cohort for patients with breast cancer. We demonstrate how MS-CPFI effectively identifies well-known prognostic factors and provides an intuitive means to interpret a multi-state algorithm.

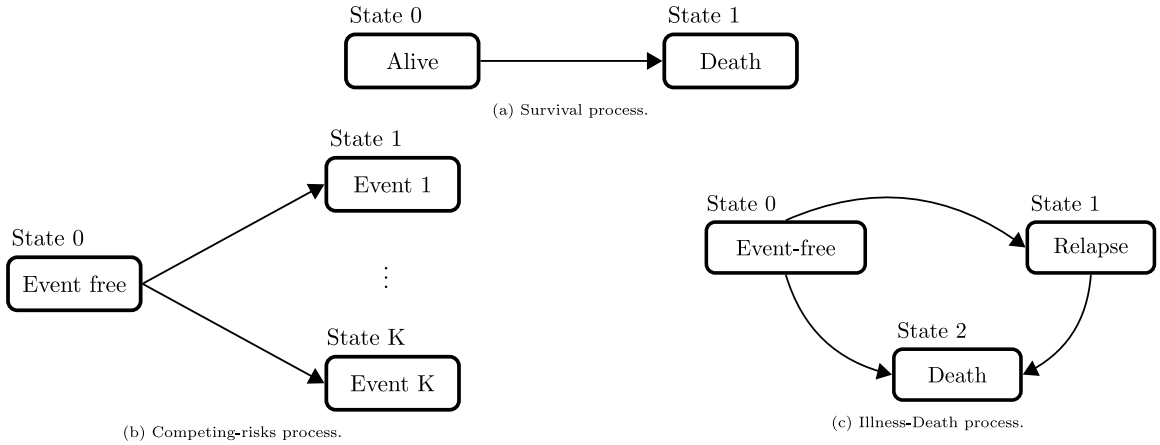
2. Theoretical background

2.1. Related work on interpretability

2.1.1. Interpretability of machine learning models

Interpretability of machine learning algorithms can be categorized into two approaches [24]: model-specific and model-agnostic. Model-specific approaches designate intrinsically interpretable models by their nature. For example, some tree-based algorithms as CART [25] can provide a list of important features that corresponds to the features intrinsically selected to build the trees. Within deep learning approaches, neural additive models [26] are the only ones intrinsically interpretable as they consider additive effects of the features on the prediction. Others models that are not intrinsically interpretables could be interpreted by relying on model-specific knowledge as well. For example, for simple neural network-based algorithms (i.e. with few parameters), the Garson's algorithm [27] provides interpretations but it cannot be applied to large deep learning algorithms [28]. For deep learning algorithms, analyses of the gradients can provide insight on the important features [29], but this kind of approaches is mainly applied for interpreting images classification.

On the other side, model-agnostic methods analyze the predictions after training and apply to any machine or deep learning algorithm. There are two main types of model-agnostic methods. The first uses feature summary visualizations [24] to interpret the nature of the relationship between input features and the outcome. The state-of-the-art methods for this approach are Partial Dependence Plots [30]



(a) Survival process.

(b) Competing-risks process.

(c) Illness-Death process.

Fig. 2. Specific applications of multi-state analysis.

and Accumulated Local Effects (ALE) [31], which describe how a feature influences the model prediction on average according to the distribution of the feature.

The second type of model agnostic approach is based on computing a quantitative importance of each feature through feature summary statistics [24]. The state-of-the-art methods for this approach are PFI [22], cited previously, and Leave-One-Covariate-Out (LOCO) [32].

These previous methods are commonly referred to global methods [33], as they provide insight into the average behavior of an algorithm, as opposed to local methods that focus on interpreting individual predictions. Individual Conditional Expectation (ICE) plots [34], which are the equivalent of the PDP algorithm for measuring individual effects, can be used for this purpose. Local Interpretable model-agnostic Explanations (LIME) [35] explains individual predictions by approximating the predictions using a surrogate model. SHapley Additive exPlanations (SHAP) [36] explains individual predictions as well by computing shapley values. For a more comprehensive review of existing interpretability methods, we recommend reading Christoph Molnar's book [24].

Some of these existing algorithms have been extended for interpreting black-box survival (i.e. two-states) algorithms.

2.1.2. Interpretability of machine learning-based survival models

Within intrinsically interpretable algorithms, Cox-nnet [37] is a deep learning architecture specifically designed to handle high-dimensional gene expression data. The effects of gene expression features are interpreted by analyzing the values of hidden nodes and conducting a gene set enrichment analysis. The subsequent version, Cox-nnet-v2.0 [38], enhances interpretability by integrating feature importance scores. In the same way, Cox-PASNet [39] combines clinical and gene expression features by using a pathway layer with prior biological knowledge, which improves biological interpretability. Similarly, PAGE-Net [40] integrates gene expression and histo-pathological data using specific layers with prior biological knowledge to improve interpretability. However, these approaches are model-specific and primarily designed for interpreting genomic data.

State-of-the-art model-agnostic interpretability methods have also been adapted for interpreting event history algorithms. As explained previously, PFI have been extended for survival analysis in RSFs [9] in RSfs for competing risks [23]. SurvLIME [41] extended the LIME method. The SHAP method has been extended to interpret survival predictions [42,43]. SurvSHAP(t) [44,45] provides time-dependent explanations for interpreting a black-box survival model. PDP and ICE algorithms have also been implemented for interpreting a survival model [46].

2.2. Overview of multi-state modeling

In this section, we briefly introduce the traditional definition of a multi-state model, referring the reader to Andersen et al. [47] for a more detailed presentation. We focus on progressive multi-state models whose transitions are irreversible, which are widely used in healthcare applications [5]. We then delve into specific types of multi-state models, such as the survival model, competing-risks model, and illness-death model, which are common to model disease progression.

2.2.1. Basic concepts

Multi-state processes extend survival analysis to model successive (≥ 2) events of interest over time. It is a continuous time stochastic process, noted $E(t)$ $(t > 0)$. Let K be the number of states with 0 the initial state, such that $E(t)$ takes values in $\{0, \dots, K-1\}$ indicating the state of the patient at time t. The initial state $E(0) = 0$ is the inclusion time in clinical trials. Intermediate (transient) states indicate the evolution of the disease. The absorbent terminal state is very often death. Disease progression is characterized by transitions between states, represented by the notation $k \to l$ with $(k, l) \in 0, \dots, K-1$, $k < l$.

The instantaneous risk of transitioning from state k to state l after a duration t spent in state k is denoted by α_{kl} for $k < l$ and $(k, l) \in \{0, \dots, K-1\}$ such that

$$\alpha_{kl}(t) = \lim_{h \to 0} \frac{1}{h} \mathbb{P}\left(E(t+h) = l \mid E(t-) = k\right). \tag{1}$$

It is also know as a transition intensity, or transition-specific hazard function. The cumulative transition intensity A_{kl} is defined as

$$A_{kl}(t) = \int_0^t \alpha_{kl}(s) ds. \tag{2}$$

The transition probabilities p_{kl} are defined by

$$p_{kl}(s, t) = \mathbb{P}\big(E(t) = l \mid E(s) = k\big), \forall\ 0 \leq s < t. \tag{3}$$

They can be explicitly linked with the transition intensities α_{kl} by solving Kolmogorov equations [48,49].

2.2.2. Specific cases

Fig. 2 depicts common multi-state models used to represent disease progression. While these models have their own definitions and predictions of interest, their specific quantities can be expressed as functions of the transition intensities (Eq. (1)) or the cumulative transition intensities (Eq. (2)). In this section, we give the mathematical equivalences between these specific quantities and the general quantities used to predict disease progression in multi-state analysis that can be used in the general framework of MS-CPFI (Section 3.1.3).

The survival model. In classical survival analysis, Overall Survival (OS) or RelapseF ree Survival (RFS) are very often the primary endpoints. They can be framed into the multi-state process model by considering two states "Alive" and "Death" (or "Relapse") and one transition $0 \to 1$, as illustrated in Fig. 2(a). In this context, the transition intensity $\alpha_{01}(t)$ represents the instantaneous risk of death at time t, given that the individual has survived up to time t. In the classical framework, the probability of surviving beyond a certain time t is given by the survival function $S(t)$ and is linked to the cumulative transition intensity A_{01} via

$$A_{01}(t) = -\ln(S(t)) \tag{4}$$

The competing-risks model. The competing-risks model applies for competing events, that can prevent the occurrence of the primary event of interest [50]. In a competing-risks model with $K-1$ competing events, the time of occurrence of event k ($1 \le k \le K-1$) is given by the latent random variable $T_k = \inf_{t>0}\{E(t) = k\}$. The observable time is given by the random variable $T = \min\left(T_k\right)_{1\le k\le K-1}$, together with the binary indicator D that indicates the type of event.

The Cumulative Incidence Functions (CIFs) F_k, for $1 \le k \le K-1$, are defined as

$$F_k(t) = \mathbb{P}(T \le t, D = k),\ t > 0,$$

These functions characterize the distribution of T and D.

However, CIFs can sometimes be misinterpreted due to the "reverse effect" [51] because the CIF for one event is conditioned on the CIFs of the competing events. The conditional probabilities (CPs) F_k^c, defined as

$$F_k^c(t) = \frac{F_k(t)}{1 - \sum_{q\neq k} F_q(t)}, \quad t > 0,$$

and which expresses the probability that event k occurs before time t given that no competing event has occurred by time t, was used by Pepe and Mori [51] and others [52–54] to avoid misinterpretation of the CIFs. This approach provides a more accurate and meaningful analysis of competing-risks data by accounting for the possibility of competing events precluding the occurrence of the event of interest.

A competing-risks process can also be represented as a K-state process (see Fig. 2(b)) with $K-1$ irreversible transitions denoted $0 \to k$. The cumulative transition intensity A_{0k} can be rewritten as follows, for $t > 0$,

$$A_{0k}(t) = \int_0^t \frac{F_k(du)}{1 - \sum_{q=1}^K F_q(u)}.$$

These cumulative transition intensities provide a meaningful analysis of competing-risks data as they are not subject to the "reverse effect" issue as well. Then, in the general framework of MS-CPFI (Section 3.1.3), either the CPs or the cumulative transition intensities might be used to provide a right interpretation of disease progression.

The illness-death model. The Fig. 2(c) illustrates an illness-death multi-state process consisting of $K = 3$ states: state 0 is the initial "Event-free" state, state 1 is an intermediate "Relapse" state, and state 2 is an absorbing "Death" state. The process is irreversible and characterized by three transitions: $0 \to 1$, $0 \to 2$, and $1 \to 2$, where transitions from state 0 are competing, and transitions $0 \to 1$ and $1 \to 2$ are successive. We consider here illness-death processes whose transitions times depend only on the duration spent in the current state [47].

An illness-death model can be further characterized by latent random variables (r.v.) T_{kl}, for $(k,l) \in \{(0,1),(0,2),(1,2)\}$. When a subject leaves state 0, it will enter either state 1 at time T_{01} or state 2 at time T_{02}. If a subject is in state 1 at time T_{01}, it will enter state 2 at time $T_{01} + T_{12}$. The observable times can be summarized by T_0, which represents the exit time from state 0

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

together with $D_0 \in \{1,2\}$ that indicates the entered state; and T_2 the entry time to state 2,

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

which is the total survival time.

An illness-death process is conventionally associated with a set of transition intensities α_{kl}, for $(k,l) \in \{(0,1),\ (0,2),(1,2)\}$, as defined in Eq. (1). In parallel, others suggest different modelizations; e.g. Cottin et al. [7] suggest to model the distribution of (T_0, D_0) and T_2 with the cumulative incidence functions F_{kl}, for $(k,l) \in \{(0,1),(0,2),(1,2)\}$, such that, for transitions $0 \to 1$ and $0 \to 2$,

$$F_{0l}(t) = \mathbb{P}\left(T_0 \le t, D_0 = l\right),\ \text{for } l = 1,2,\ t > 0.$$

By following the semi-markovian property, for transition $1 \to 2$, the CIF is defined conditionally to T_0, $D_0 = 1$ and for the duration variable d so that $d = t - T_0$, $d > 0$, as

$$F_{12}\left(d \mid T_0,\ D_0 = 1\right) = \mathbb{P}\left(T_2 - T_0 \le d | T_0,\ D_0 = 1\right).$$

Together with these quantities, the cumulative transition intensities A_{kl} defined in Eq. (2) can be reformulated such that for $l = 1,2,\ t > 0$,

$$A_{0l}(t) = \int_0^t \frac{F_{0l}(ds)}{1 - \left(F_{01}(s) + F_{02}(s)\right)},$$

and, for $d > 0$,

$$A_{12}(d|T_0,\ D_0 = 1) = \int_0^d \frac{F_{12}(ds|T_0,\ D_0 = 1)}{1 - F_{12}(s|T_0,\ D_0 = 1)}.$$

We can also derive the generalization of the conditional probabilities in the illness-death case for the competing transitions $0 \to 1$ and $0 \to 2$ as

$$F_{01}^c(t) = \frac{F_{01}(t)}{1 - F_{02}(t)},\ F_{02}^c(t) = \frac{F_{02}(t)}{1 - F_{01}(t)}.$$

For transition $1 \to 2$, we have the following equivalence: $F_{12}^c(d|T_0,\ D_0 = 1) = F_{12}(d|T_0,\ D_0 = 1)$. In the general framework of MS-CPFI (Section 3.1.3), these cumulative transition intensities and conditional probabilities might be then used provide a meaningful interpretation of disease progression.

In the next section, we develop our algorithm MS-CPFI that can take as input one of these meaningful quantities (i.e. the CPs or the cumulative transition intensities) for each transition. In addition, we consider that we observe a vector of covariates X of dimension P and assume that all the quantities defined below are expressed conditionally to X.

3. Methodology

In this section, we address the critical need to interpret predictions in terms of any of the functions of interest described in Section 2.2. To facilitate the translation of this prediction into meaningful clinical decision support, we present the MS-CPFI algorithm. It serves as a model agnostic interpretation tool specifically designed for multi-state algorithms.

3.1. Main contributions

The concept of feature importance is based on the idea that, if a feature is important, altering its data quality will likely lead to a decline in the accuracy of model predictions. The Permutation Feature Importance (PFI) algorithm [22], as mentioned earlier, is widely employed in the literature for assessing feature importance. In this algorithm, changing the data quality of a feature involves randomly shuffling its values and deeming the feature important if this random perturbation results in a decrease in the model error. This is typically measured through metrics such as discrimination error in classification algorithms. The key advantage is the provision of an intuitive measure of

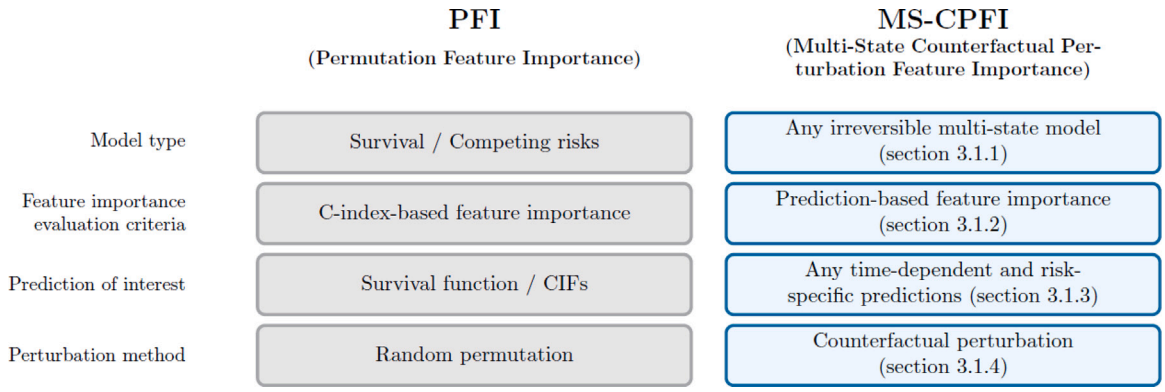


Fig. 3. Illustration of our contributions in MS-CPFI as compared to the existing implemented PFI algorithms in the R package *randomForestSRC* for RSFs [55] and RSFs with competing risks [56].

the features impacting the overall model error. The greater the increase in prediction error when the feature is permuted, the more significant the feature is deemed. Conversely, a decrease or stability in prediction error indicates that the feature is not important for prediction accuracy. More specifically, in Random Survival Forests (RSFs) [9], and likewise in RSFs for competing risks [23], the PFI algorithm analyzes how a random permutation of feature values influences model error, measured by the concordance index or risk-specific concordance indexes. While the PFI algorithm can be extended for multi-state models, certain limitations arise that necessitate further consideration.

To address and overcome these limitations, we present a refined and extended version called Multi-State Counterfactual Perturbation Feature Importance (MS-CPFI). This enhanced algorithm expands upon the PFI methodology in four key dimensions, as illustrated in Fig. 3 and elaborated upon in the subsequent paragraphs.

3.1.1. Model type

First, our MS-CPFI algorithm extends the Permutation Feature Importance (PFI) to accommodate any multi-state algorithm. This broad applicability encompasses survival models, competing risks models, illness-death models, or any other multi-state process. Furthermore, building upon the insights of Ishwaran et al. [23], this extension enables the computation of feature importance scores independently for each transition within a multi-state model. The objective here is to facilitate the measurement of how various clinical factors impact each stage of a disease, providing a more nuanced understanding of their influence across different disease states.

3.1.2. Feature importance evaluation criteria

In the original definition of PFI, feature importance scores are computed based on a feature's contribution to the model error, typically using metrics like risk-specific concordance indexes in a competing-risks model [23]. When transposing this definition to a multi-state model, an equivalent model error could be represented by transition-specific integrated Area Under the Curve (iAUC) and integrated Brier Scores (iBSs), which quantify discrimination and calibration errors, respectively. Refer to the supplementary material B for exact definitions. In this context, a feature importance score would indicate how sensitive the iAUC for a specific transition is to the perturbation of the corresponding feature. While these measures offer insights into how features impact model performance, they do not directly convey how model predictions would change when a feature is perturbed. For clinical interpretation, a more understandable measure of model explainability is necessary. To address this, we implemented a prediction-based feature importance measure that quantifies the average effect on the risk of transition for each feature. A positive (negative) feature importance score indicates an increase (respectively, a decrease) in the risk of transition. Unlike PFI, the sign of this prediction-based feature importance score is interpretable, akin to interpreting log hazard ratios in a Cox Proportional Hazards model [4]. Although we recommend computing prediction-based scores for better interpretability, we recognize that some users may be interested in model-error scores. Therefore, we implemented our algorithm to allow users to choose the type of feature importance (prediction-based versus error-based) through an optional parameter.

3.1.3. Predictions of interest

Our third contribution with MS-CPFI is its ability to accept any predictions of interest, encompassing various cumulative functions of risk derived from a multi-state model. Examples include a survival function from a survival model, risk-specific Cumulative Incidence Functions (CIFs) from a competing-risks model, conditional probabilities, or cumulative transition intensities from an illness-death model (refer to Section 2.2.2 for more details). However, while MS-CPFI offers flexibility in the choice of the function of interest, we provide recommendations as the selection can impact interpretation and lead to biased results in some cases. For instance, in a competing-risks model or any multi-state model with competing transitions, the CIFs are mutually conditioned, as explained in Section 2.2.2. This conditioning can potentially result in misinterpretation of feature importance since the CIF for one event (or transition) is influenced by the CIFs of the competing events (or competing transitions). This issue was not addressed in Ishwaran et al.'s definition of PFI for RSFs with competing risks [23]. To mitigate this bias, we recommend using a quantity of interest that is not conditioned by competing transitions. As discussed in Section 2.2.2, considering either the conditional probabilities or the cumulative hazard functions instead of the CIFs as predictions of interest can address this conditioning issue. Consequently, we implement MS-CPFI by allowing users to specify the type of predictions of interest through an optional parameter.

3.1.4. Perturbation method

In the original definition of PFI, feature importance scores are computed by randomly shuffling the feature values. However, this method has limitations. Firstly, the random shuffle of features can introduce randomness and result in large variance when the permutation is repeated [57]. Secondly, randomly perturbing a feature does not allow for the detection of non-linearities in the feature effect when computing a prediction-based feature importance score (c.f. Section 3.1.2). To address these limitations, we propose a new counterfactual perturbation method, inspired by the notion of counterfactual explanations [58] and the principle of partial dependence plots [30]. The idea is to replace observed values of a feature by the counterfactual ones, all other things being equal. To achieve this, the counterfactual perturbation method changes the feature values with a theoretical one for all the instances before making the predictions. Then, it analyses how this change impacts the predictions, i.e., whether it decreases or increases the prediction.

3.2. Mathematical definition

3.2.1. General formalism

Consider a set of features (or covariates), noted X. Following the approach proposed by Krzyziński et al. [44], we define, for a transition $k \to l$, a prediction of interest A_{kl}, a feature $x \in X$ with a counterfactual scenario x^c, our feature importance score as follows:

$$\mathrm{FI}_{kl}(t, x^c) = \bar{A}_{kl}\left(t|X^{\setminus x}, x^c\right) - \bar{A}_{kl}(t|X), \tag{5}$$

where

$$\bar{A}_{kl}(t) = \frac{1}{n}\sum_{i=1}^{n} A_{kl}\left(t|X_i\right)$$

is the reference cumulative transition intensity averaged over a population of n subjects; and, for $X^{\setminus x}$ the set of covariates excluding the feature x,

$$\bar{A}_{kl}\left(t|X^{\setminus x}, x^c\right) = \frac{1}{n}\sum_{i=1}^{n} A_{kl}\left(t|X_i^{\setminus x}, x^c\right)$$

is the cumulative transition intensity under a counterfactual scenario x^c. The counterfactual perturbation implies then that a feature importance score is computed against the reference (i.e. the average population).

A negative feature importance indicates then that the counterfactual scenario implies a decrease in the prediction (e.g., the cumulative transition intensity) compared to the reference, suggesting a protective factor. Conversely, a positive feature importance indicates that the counterfactual scenario leads to an increase in the prediction compared to the reference, indicating a risk factor.

To measure the overall feature importance of a feature x, we introduce an aggregated feature importance score, which is integrated over a time horizon window $[0, \tau]$, so that

$$\mathrm{FI}_{kl}(x^c) = \int_0^{\tau} \left|\mathrm{FI}_{kl}(t, x^c)\right| dt. \tag{6}$$

If the sign of the difference inside the integral does not change over time, we can remove the absolute value. In that case, a negative feature importance indicates then that the counterfactual scenario implies a decrease in the prediction (e.g., the cumulative transition intensity) compared to the reference, suggesting a protective factor. Conversely, a positive feature importance indicates that the counterfactual scenario leads to an increase in the prediction compared to the reference, indicating a risk factor.

However, this assumption may be too strong and can be relaxed by using the time-dependent feature importance score defined in Eq. (5). In particular, in the case of non-proportional risks with a shifted effect over time, the integration of the feature importance score over time may lead to a biased interpretation.

3.2.2. Illustrative example

Consider the setting of survival analysis and define the cumulative transition intensity (Eq. (4)) or in this case the cumulative hazard ratio (HR) which depends on a feature "age" with a time interaction, such that (Fig. 4(a)):

$$A_{01}(t|\mathrm{age}) = \int_0^t \mathrm{HR}^u(\mathrm{age})\, du = (\mathrm{age} - 50)^2 \times \frac{t^2}{2},$$

where the feature "age" have been simulated for n individuals from the $\mathcal{N}(50, 15)$ distribution, with the hazard ratio

$$\mathrm{HR}^t(\mathrm{age}) = (\mathrm{age} - 50)^2 \times t, \tag{7}$$

for $t = 1, \ldots, \tau$, $\tau = 100$. The reference cumulative transition intensity is then given by (Fig. 4(b)):

$$\bar{A}_{01}(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{age}_i - 50\right)^2 \times \frac{t^2}{2}. \tag{8}$$

Consider the three counterfactual scenarios $\mathrm{age}^c \in \{18, 50, 80\}$. The counterfactual cumulative transition intensity is then given by (Fig. 4(b)):

$$A_{01}(t|\mathrm{age}^c) = (\mathrm{age}^c - 50)^2 \times \frac{t^2}{2}, \tag{9}$$

with age^c the age value in a counterfactual scenario.

With Eq. (5), the time-dependent feature importance of "age" for a counterfactual scenario age^c is computed, for $t \in [0, \tau]$ as follows (Fig. 4(c)):

$$\mathrm{FI}(t, \mathrm{age}^c) = \left((\mathrm{age}^c - 50)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{age}_i - 50\right)^2\right)\right) \times \frac{t^2}{2}. \tag{10}$$

Then, with Eq. (6), the integrated feature importance of "age" for this counterfactual scenario is computed as follows (Fig. 4(d)):

$$\mathrm{FI}(\mathrm{age}^c) = \left((\mathrm{age}^c - 50)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{age}_i - 50\right)^2\right)\right) \times \frac{\tau^3}{6}. \tag{11}$$

As expected (i.e. according to the shape of the HR in Fig. 4(a)), Fig. 4(d) displays that the age values 18 and 80 are risk factors, the age value 50 is a protective factor. From 18 to 50 years old, the FI curve is monotonically increasing; from 50 to 80 years old, the FI curve is monotonically decreasing. Therefore, the shape of the FI curve reflects the shape of the HR and providing these three counterfactual feature importance scores allows understanding the effect of the feature. However, using only the mean as a counterfactual scenario, only a negative importance score could be seen and the true feature effect would have been then misinterpreted. In Section 3.3; we provide some methodology to select the counterfactual scenarios.

We should notice that, in this case, the integrated feature importance is sufficient to provide insight on the feature importance according to the feature effect shape. Previously, we identified a specific case (the case of a shifted effect (SE) of the feature over time) for which the integrated feature importance may lead to a biased interpretation of the feature effect and for which the time-dependent version of our feature importance should be preferred. In the supplementary material C, we give an illustration for this specific case. We illustrate how our time-dependent version of MS-CPFI would be useful to detect a change in the feature effect over time.

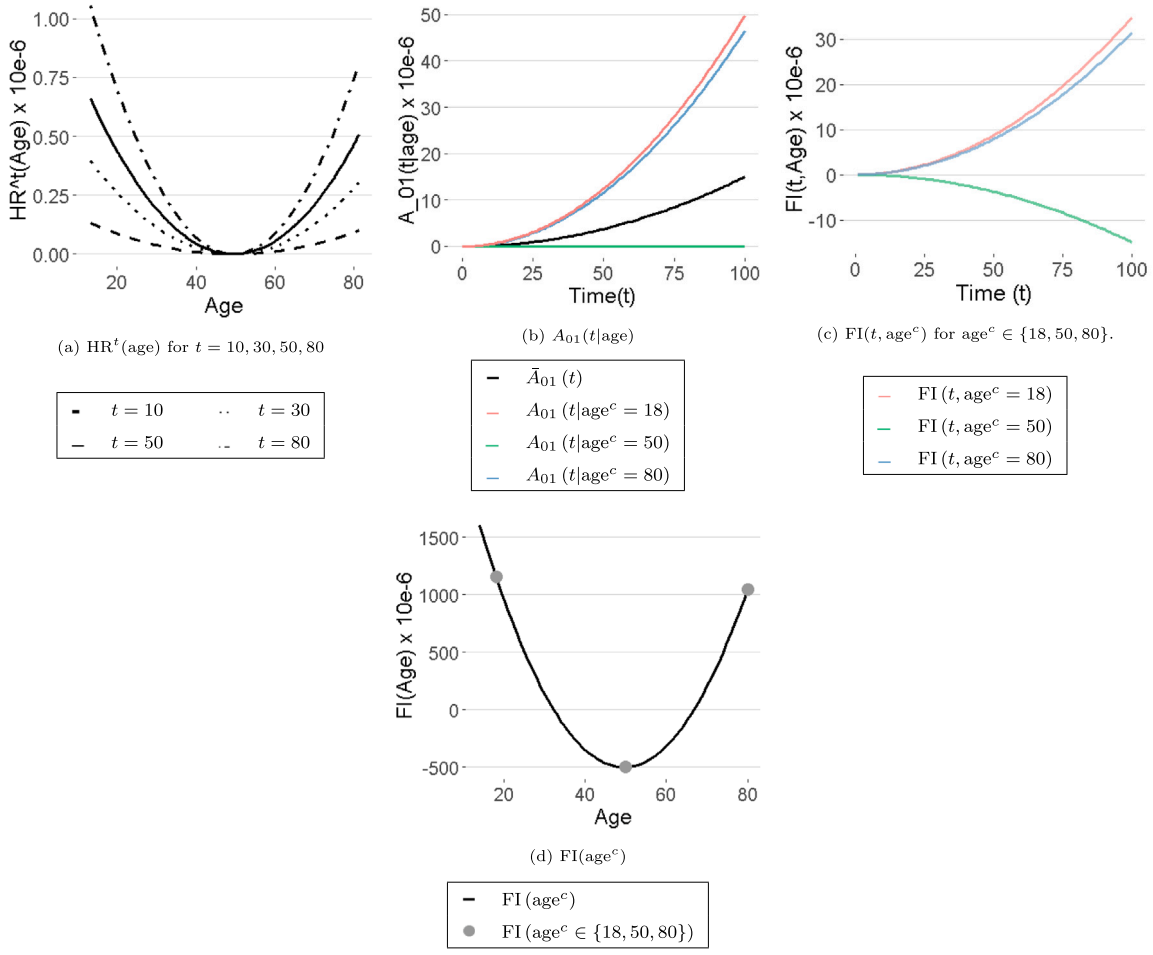
3.3. Selection of the counterfactual scenarios

In the computation of feature importance scores using MS-CPFI for a feature x, the process involves generating counterfactual scenarios by selecting counterfactual values, denoted as x^c. We propose a methodology for choosing these counterfactual scenarios applicable to both numerical and categorical features.

3.3.1. For a numerical feature

The most natural idea for perturbing a numerical feature is to replace feature values with either the feature mean (for a Z-score scaled feature) or the feature range/2 (for a min–max scaled feature), reducing variance by removing randomness and ensuring replication and stability in the estimation of feature importance. However, using a unique value will not capture the shape of the relationships between a feature and the predictions.

While the most straightforward approach is to create counterfactual scenarios for a feature x by using all possible feature values (e.g., for a feature "age" with values in the interval $[18, 90]$, the counterfactual scenarios would be $\mathrm{age}^c \in \{18, 19, \ldots, 90\}$), this can be computationally expensive. Alternatively, selecting equally spaced or quantile values might be considered. However, they may introduce bias in the interpretation, as these values may not capture potential non-linearities in the feature's effect. To address this, we recommend choosing counterfactual scenarios for a feature x using transition-specific univariate Cox Proportional Hazards (Cox P.H.) models [4] with spline-based functions. For a transition $k \to l$, we plot the estimated hazard ratio according to



(a) $\text{HR}^t(\text{age})$ for $t = 10, 30, 50, 80$

(b) $A_{01}(t|\text{age})$

(c) $\text{FI}(t, \text{age}^c)$ for $\text{age}^c \in \{18, 50, 80\}$.

(d) $\text{FI}(\text{age}^c)$

Fig. 4. Illustration of MS-CPFI: example for a feature "age" with the counterfactual scenarios $\text{age}^c \in \{18, 50, 80\}$. Figure (a) plots the time-dependent hazard ratios (Eq. (7)). Figure (b) plots the reference cumulative transition intensity (Eq. (8)) and the cumulative transition intensities for the selected counterfactual values (Eq. (9)). Figure (c) plots the time-dependent feature importance scores (Eq. (10)). Figure (d) plots the integrated feature importance scores (Eq. (11)).

the values of the feature x. The selected counterfactual values include both extreme feature values and the inflection points (which can be chosen graphically) of the hazard ratio curve. In our algorithm, we allow users to choose counterfactual scenarios for a numerical feature, empowering them to specify the counterfactual values for the feature selected upstream.

3.3.2. For a categorical feature

For a categorical feature, the counterfactual scenarios are the feature categories. However, if the feature is dummy-encoded or one-hot-encoded, then the feature-related binary features should be considered as non independent features, and we then need to adjust the counterfactual perturbation.

Indeed, disrupting an encoded feature category independently of the others, as done in state-of-the-art methods, would break the association between categories of the original feature. Therefore, we must take into account the underlying association between encoded features. Additionally, for dummy-encoded features, the reference category leads to a conditional interpretation of the dummy features based on the reference.

To address these issues, we investigated two types of counterfactual perturbation for dummy-encoded features, illustrated in Fig. 5 using the example of a categorical feature BMI (Body Mass Index) with three categories: `Low`, `Normal`, and `High`, which have been dummy-encoded into two binary features: `Low` and `High`, with `Normal` as the reference category.

1. The perturbation method dependent on the reference category (`Normal`) consider each dummy encoded feature (`Low` and `High`) independently and replace their values with 0. This provides a feature importance measure for each dummy encoded feature, which should be interpreted relatively to the reference category. The choice of the reference category does not affect the model results, except for interpretability purposes. Typically, the normative category is chosen as the reference, but in some cases, determining a unique normative category may be challenging.
2. To address this issue, we implemented a second perturbation method that is not conditioned on a reference category. For each feature category, we create a counterfactual scenario where the targeted category is set to 1, while all others are set to 0. These counterfactual scenarios provide a feature importance measure for each feature category, including the reference category, independent of the reference choice.

In our algorithm, we provide users with the option to choose between two perturbation methods for a categorical feature. However, we recommend opting for the second option, i.e., the permutation method that is not conditioned on the reference category.

3.4. Construction of a confidence interval

To provide a confidence interval (CI) for feature importance scores with MS-CPFI, we employ Monte Carlo Simulations (MCS) to validate

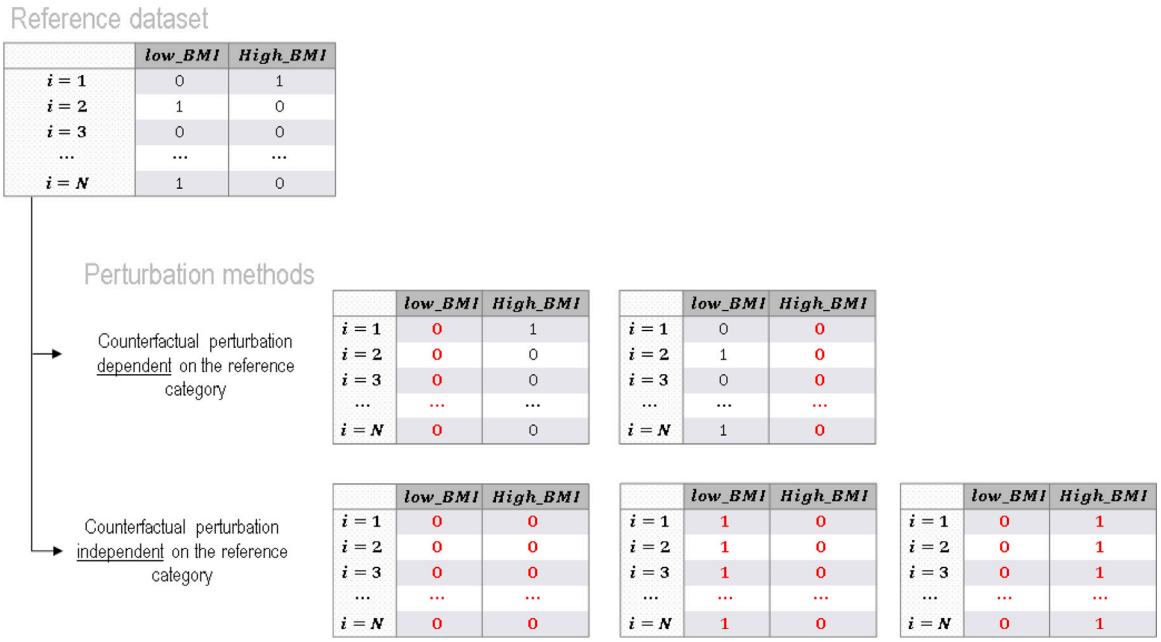


Fig. 5. Counterfactual perturbation for categorical features. Illustration for a categorical feature "BMI" that has been dummy-encoded.

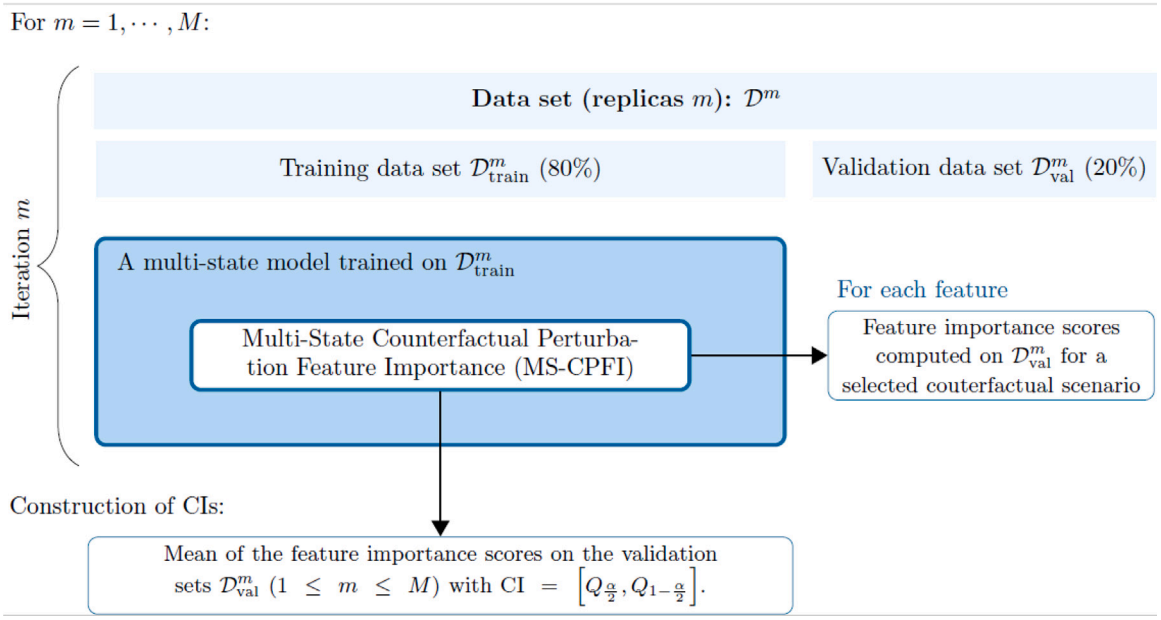


Fig. 6. Construction of confidence intervals (CIs) of feature importance scores in MS-CPFI.

experiments on simulations by generating M data sets. On real data sets, we employ Monte Carlo Cross Validation (MCCV) to provide a CI for feature importance by randomly splitting M times the data set. Let M be the number of iterations ($m = 1, \ldots, M$), where the index m represents either the data set m for simulations, or the data set from split m for a real data set.

An illustration is given in Fig. 6, and a detailed pseudo-code with mathematical notations is given in supplementary material D (Algorithm D.1). This pseudo-code is written for a prediction-based importance score; for computing performance-based, i.e. iAUC or iBS-based, feature importance scores, the pseudo-code is given in the supplementary material E (Algorithm E.1).

4. Experimental results

4.1. Simulation study design

To conduct our evaluation, we used an illness-death model as depicted in Fig. 2(c) and performed Monte Carlo simulations (MCS) by generating M data sets for each simulation model. For each dataset replica m ($m = 1, \ldots, M$), we generated a sample of $n = 5000$ continuous-time illness-death observations and partitioned it into a training set of size $n_{\text{tr}} = 4000$ (80% of the data) and a validation set of size $n_{\text{val}} = 1000$ (20% of the data). We fixed the horizon-time window at $\tau = 100$.

Table 1
Properties of the simulation models.

Model	Covariates generation	Risk functions
A[a]	$X = (X_1, \ldots, X_6)^T$ with $X_p \in \mathbb{R}^n \sim \mathcal{N}(0,1)$, $(1 \le p \le 6)$	$g_{01}(X, \beta_{01}) = 0.5X_1^2 + 0.3X_2^2$ $g_{02}(X, \beta_{02}) = 0.5X_3^2 + 0.3X_4^2$ $g_{12}(X, \beta_{12}) = 0.5X_5^2 + 0.3X_6^2$
B[b]	$X = (X_1, \ldots, X_6)^T$ with $X_p = (X_{p1}, X_{p2})^T \in \mathbb{R}^{n\times 2} \sim \mathcal{N}(0, \Sigma_p)$, $1 \le p \le 6$, with $\Sigma_p \in \mathbb{R}^{2\times 2}$ the matrix of variance–covariance of X_p.	$g_{01}(X, \beta_{01}) = 0.3X_{11}^2 + 0.3X_{21}^2$ $g_{02}(X, \beta_{02}) = 0.3X_{31}^2 + 0.3X_{41}^2$ $g_{12}(X, \beta_{12}) = 0.3X_{51}^2 + 0.3X_{61}^2$

[a] In model A, six covariates were generated, each drawn from a multivariate Gaussian distribution with mean 0 and variance 1.
[b] In model B, six bi-dimensional covariates were generated, each drawn from a multivariate Gaussian distribution with mean 0 and a matrix of variance–covariance:

$$\Sigma_p = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}, (1 \le p \le 6)$$

.

We generate the observation (T_0, D_0) and (T_2) such that the latent illness-death times T_{kl}, for $(k,l) \in \{(01), (02), (12)\}$, are simulated through transition-specific Cox P.H. hazard functions [59], so that, for each observation i $(1 \le i \le n)$,

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

where $g_{kl}(.)$ is a transition-specific risk function, $X_i \in \mathbb{R}^P$ are the individual covariates, $\beta_{kl} \in \mathbb{R}^P$ are fixed effect coefficients, and $\alpha_{kl}^0(.)$ is the baseline hazard function. The three baseline hazard functions are Weibull with scale parameter 0.01, and shape parameter 1.2.

We introduce a censoring process 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.

Supplementary material F provides more details on the statistical methods used for the simulation of the transition times.

4.2. Scenario design

To evaluate our interpretability algorithm, we considered two simulation models were consider, in which the covariates generation and the risk functions vary. These simulation models are summarized in Table 1. We then created four distinct scenarios based on these models.

Scenario A.1: Illustration of the reverse effect. Scenario A.1 aims to illustrate the "reverse effect" discussed in Section 2.2.2, which can arise when computing cumulative incidence functions (CIFs) for competing transitions. The data used for this scenario was generated from Model A, which has no correlated features or interaction terms. Two features affect the hazard function for each transition, with no cross-effects through transitions, and risk functions are parameterized with quadratic effects.

Scenario A.2: Illustration of the prediction-based feature importance and of the counterfactual perturbation for numerical features. Scenario A.2 highlights the advantages of the prediction-based MS-CPFI over performance-based (i.e., iAUC-based) feature importance measures. The data used for this scenario were also generated from Model A.

Scenarios B.1 and B.2: Illustration of the counterfactual perturbation for categorical features. Scenarios B.1 and B.2 use data generated from Model B, where continuous features are transformed into discrete features and dummy-encoded by excluding a reference category. These scenarios aim to illustrate our interpretability framework for categorical features. By definition, the choice of reference category has no effect on the results, but it can impact the interpretability of classical methods. Using these scenarios, we aim to show how a counterfactual perturbation for categorical features independent on the reference category can be more easy to interpret than the perturbation method independent on the reference category. Scenarios B.1 and B.2 differ in the choice of the reference category to demonstrate this advantage.

4.3. Results

To conduct our simulations and estimate the predictions of interest, we use IDNetwork [7] that is a deep learning architecture designed to model an illness-death process and to provide transition-specific and time-dependent predictions (see the supplementary material A for a brief description). Output of IDNetwork are the density probability functions. Using the equivalence defined in Section 2.2, we can compute one of the predictions functions as input to MS-CPFI.

Main results on scenarios A.1 and A.2, B.1 and B.2 are displayed in the next sections. Additional results are given subsequently. In our results, we focus on the integrated feature importance (with no absolute value).

4.3.1. Scenario A.1: Illustration of the reverse effect

To illustrate how the reverse effect constraint between competing transitions can be overcome, we compare the CIFs with conditional probabilities (CP) and cumulative hazard functions in Fig. 7, using a performance-based (i.e. iAUC-based) feature importance that show the impact of model performance (see the supplementary material E). Specifically, we focus on the competing transitions $0 \to 1$ and $0 \to 2$, and illustrate the reverse effect in Fig. 7(a) when CIFs are used as the functions of interest. For transition $0 \to 1$ (respectively transition $0 \to 2$), features X_1 and X_2 (respectively features X_3 and X_4) have a positive feature importance. However, features X_3 and X_4 (which actually have an effect on the competing transition $0 \to 2$) display a negative feature importance for transition $0 \to 1$ (respectively, features X_1 and X_2 display a negative feature importance for transition $0 \to 2$). Therefore, computing feature importance with CIFs can lead to a biased interpretation of competing transitions.

To address this issue, we use CPs (Fig. 7(b)) or Hs (Fig. 7(c)) to overcome the reverse effect. For the subsequent results, we use Hs to illustrate our findings.

4.3.2. Scenario A.2: Illustration of the prediction-based feature importance and of the counterfactual perturbation for numerical features

In scenario A.1, we demonstrated the reverse effect that can occur in the presence of competing transitions with an iAUC-based feature performance. However, an iAUC-based feature importance does not provide information of the sign of the feature effect; in that case a prediction-based feature importance is more helpful. In scenario A.2, we illustrate advantages of the prediction-based feature importance and of the counterfactual perturbation for numerical features.

As explained in Section 3.1, to compute a prediction-based MS-CPFI, counterfactual scenarios need to be created for each numerical feature by selecting the relevant feature values characterizing changes in the risk of the transition. We illustrate how we choose these values for feature X_1 in Fig. 9 where we plot the hazard ratio estimated

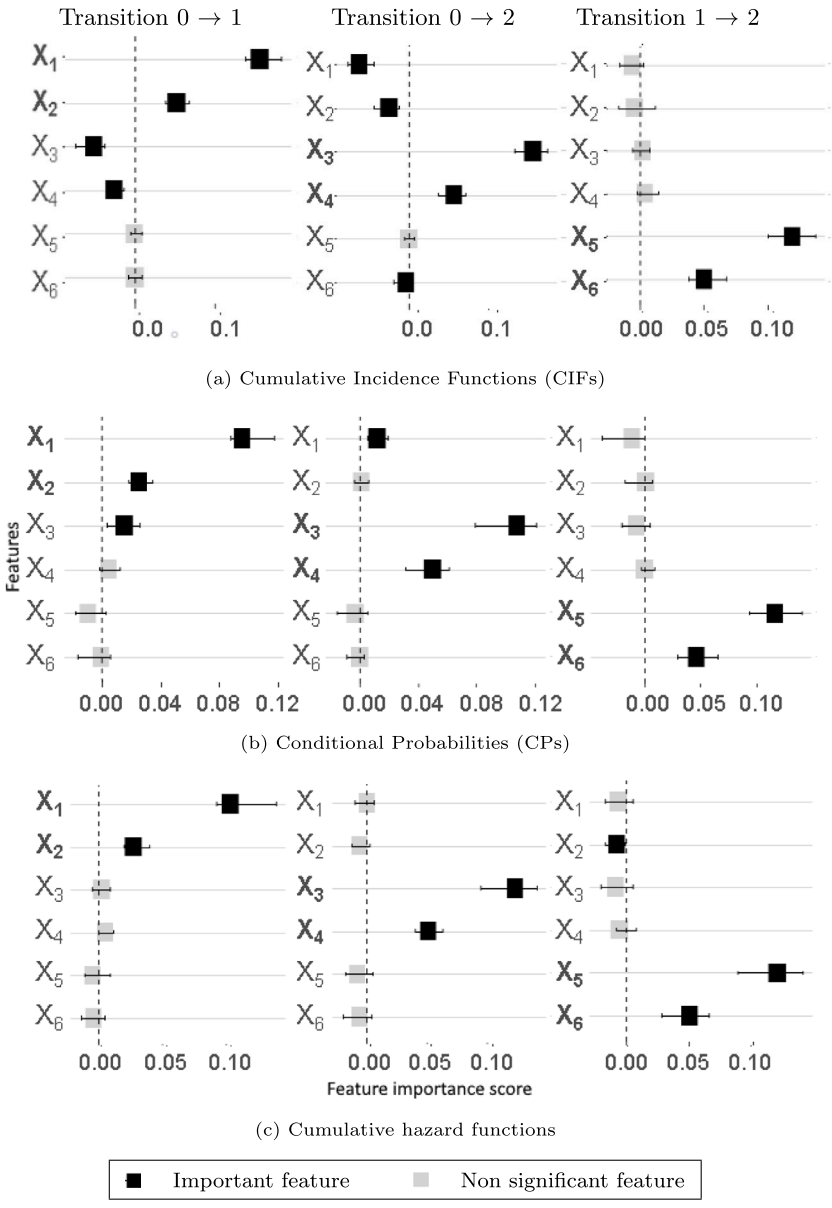


Fig. 7. Illustration of the reverse effect mentioned in Section 2.2 (Scenario A.1). Feature importance score computed with an iAUC-based MS-CPFI versus the type of prediction: (a) CIFs versus (b) CPs or (c) cumulative hazard functions. Features written in bold are those which have an effect in the simulation.

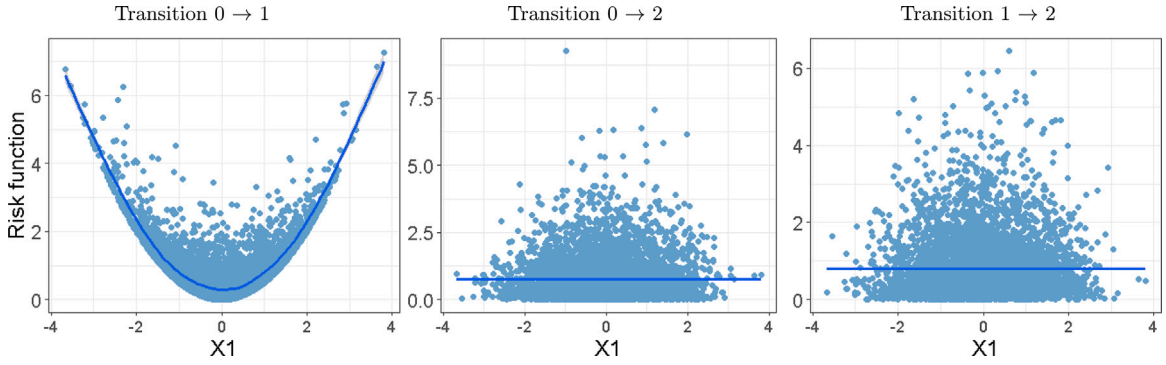


Fig. 8. Visualization of the true risk functions (Scenario A.2). True transition-specific risk functions,

$\left\{g_{kl}\left(X^i,\beta_{kl}\right),X_1^i\right\}_{1\le i\le n},\ \left((k,l)\in\{(0,1),(0,2),(1,2)\}\right),$

versus values of X_1. Blue points represent individual risks; the blue line represents a linear smooth of the individual points.

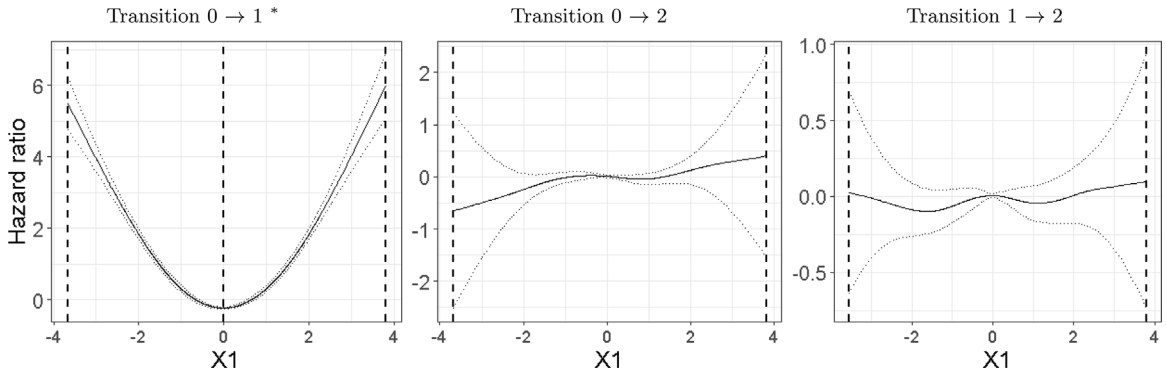


Fig. 9. Selection of the counterfactual values (Scenario A.2). Transition-specific hazard ratios $\widehat{\text{HR}}_{kl}^{i}$ versus the values X_1^i estimated with spline-based transition-specific univariate Cox models [60]. We tested the significance of the non-linear part for each model with a threshold of 0.05; * indicates a significant p-value.

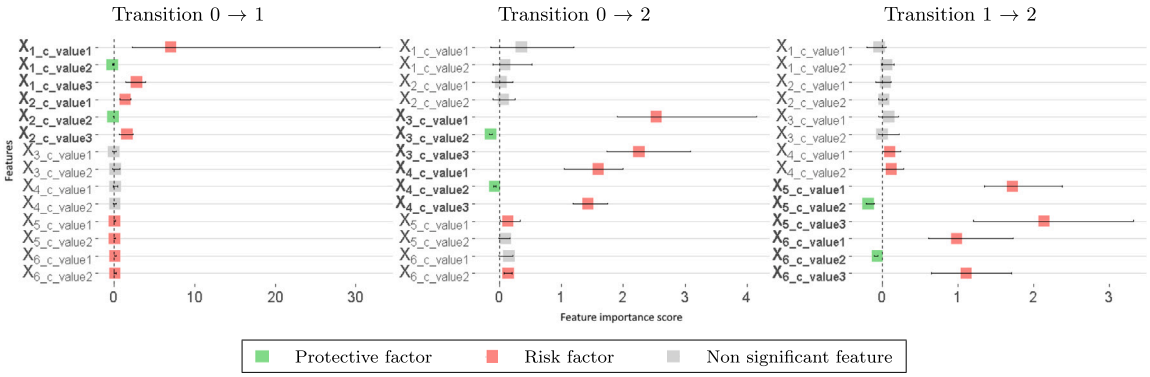


Fig. 10. Illustration of the prediction-based feature importance with numerical features (Scenario A.2). Feature importance scores computed with a prediction-based MS-CPFI and the cumulative hazard functions as the predictions of interest. Features written in bold are those which have an effect in the simulation. Features names following by the suffix "c_value" designate the counterfactual values selected in Fig. 9.

with spline-based transition-specific marginal Cox models [60]. Fig. 8 displays the true simulated hazard ratios. For each transition, we select graphically the values of X_1 that correspond to the effect changes in the estimated hazard ratio curves. Dashed lines represent the selected values. For transition 0 → 1, selected values are $\{\min(X_1), 0, \max(X_1)\}$; for transitions 0 → 2 and 1 → 2, selected values are $\{\min(X_1), \max(X_1)\}$. We repeat this process for each feature.

In Fig. 10, we plots results of feature importance for each transition and for each selected values of the features. A significant positive prediction-based feature importance for a counterfactual feature value indicates that this feature value is a deleterious risk factor for the transition. Respectively, a significant negative prediction-based feature importance for a feature value indicates that the feature value is a protective factor for the transition.

For transition 0 → 1, for feature X_1, according to the simulated effects (Fig. 8), we expected a null feature importance for the value2, a positive feature importance for value1 and value3; idem for feature X_2; for other features, we expected a null feature importance. First plot of Fig. 10 displays this expectation. For transitions 0 → 2 and 1 → 2, the second and the third plots display a null feature importance for the three counterfactual values, for these two features, as expected. For the other features, conclusions are those expected (see plots of the true feature importance scores in the supplementary materials G, Figure G.1). We can also compare these results with the plots in Fig. 7(c), computed with an iAUC-based feature importance. An iAUC-based feature importance does not provide an explanation on the sign of the feature effect and could not detect the variations in the effects when the feature value changes. Rather, the prediction-based feature importance with the counterfactual perturbation can identify associations between values of a feature and risks of transition.

4.3.3. Scenarios B.1 and B.2: Illustration of the counterfactual perturbation for categorical features

We used scenarios B.1 and B.2 to demonstrate the advantages of the counterfactual perturbation method for dummy-encoded categorical features. The purpose was to show how a feature importance score that is independent of the reference category can provide the correct interpretation regardless of the chosen reference category, as compared to a feature importance score that is dependent on the reference category.

To generate our results, we employed equal-frequency discretization and created four categories (cut1, cut2, cut3, cut4) from the initial numerical features in model B. Subsequently, each discretized feature was transformed using dummy encoding, excluding a reference category. We present our results specifically for the transition 0 → 1, as the conclusions for other transitions are equivalent. Based on the way the features were discretized (refer to Fig. 11), we expect the following feature importance score patterns. For features X_{11} and X_{21}, we anticipate a positive feature importance score for cut1 and cut4, and a negative feature importance score for cut2 and cut3. For the remaining features, we expect non-significant importance scores across all dummy features.

In scenario B.1, the reference category is cut1; in scenario B.2, the reference category is cut2. In Fig. 12, features written in bold are those which have an effect in the simulation. Graphs have been zoomed only for the four first features; the others features have no effect on the transition (complete plots are given in the supplementary material H).

The counterfactual perturbation method that is independent of the reference category provides consistent interpretation regardless of the chosen reference category (Fig. 12(b)). On the other hand, a counterfactual perturbation method that is dependent on the reference category (Fig. 12(a)) yields different interpretations when the reference category changes. Consequently, it may lead to a misinterpretation of the true effect of the feature categories.

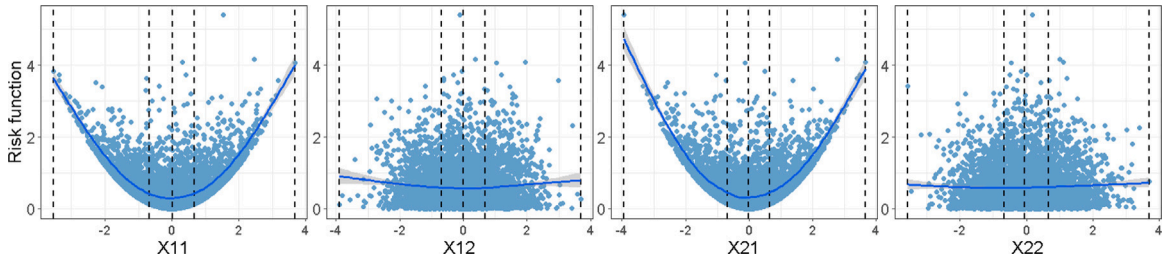
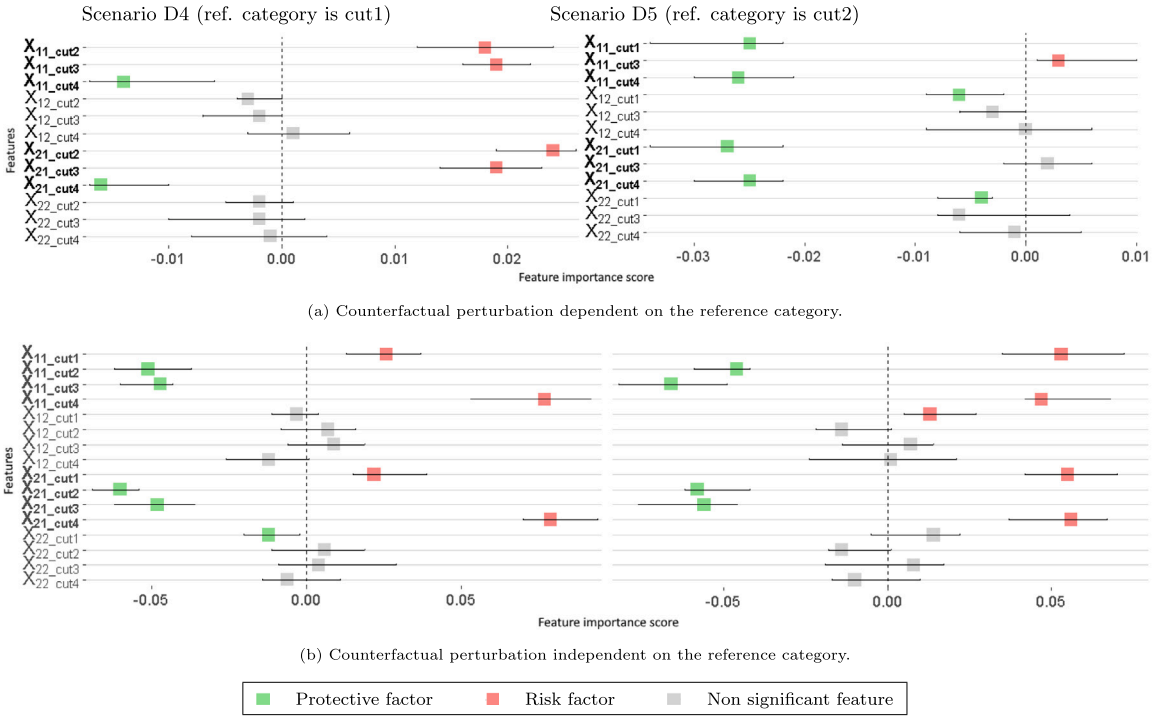


Fig. 11. Transformation of numerical features into categorical features (Scenarios B.1 and B.2). True simulated transition-specific risk functions for transition 0 → 1, $\left\{g_{01}\left(X^i,\beta_{01}\right),X^i_{p1}\right\}_{1\le i\le n}$, $\left\{g_{01}\left(X^i,\beta_{01}\right),X^i_{p2}\right\}_{1\le i\le n}$, versus values of the feature for features $\left(X_{p1},X_{p2}\right)^T$, $p=1,2$. Dashed lines display bounds of the created categories from the quantile values 0.25, 0.5, 0.75, such that cut1 = $\left[Q_{0\%},Q_{25\%}\right]$, cut2 = $\left]Q_{25\%},Q_{50\%}\right]$, cut3 = $\left]Q_{50\%},Q_{75\%}\right]$, cut4 = $\left]Q_{75\%},Q_{100\%}\right]$. Blue points represent individual risks; the blue line represents a linear smooth of the individual points. We did not put graphs for features $X_{31},\dots,X_{62}$; graphs are similar to those of features X_{12} and X_{22} as they have no effect on this transition either.



(a) Counterfactual perturbation dependent on the reference category.

(b) Counterfactual perturbation independent on the reference category.

Fig. 12. Illustration of the prediction-based feature importance on categorical features (Scenarios B.1 and B.2). Feature importance scores for transition 0 → 1, with a prediction-based MS-CPFI and the cumulative hazard functions as the predictions of interest, versus the type of counterfactual perturbation: (a) dependent versus (b) independent on the reference category.

4.3.4. Additional experiments

We conducted additional experiments on MS-CPFI using three additional scenarios. The simulation schemes for these scenarios are provided in the supplementary material I.1.

In scenario C, the advantage of the prediction-based MS-CPFI is illustrated in the case of exponential effects of the features on the risks of transition. Detailed results can be found in the supplementary material I.2, which show that the prediction-based MS-CPFI algorithm accurately estimates feature importance in the case of exponential effects and is capable of detecting non-linear risk functions.

In scenario D the robustness of MS-CPFI is illustrated in the presence of cross effects (i.e. shared effects of the features between the transitions). Detailed results are presented in the supplementary material I.3, confirming the robustness of MS-CPFI in detecting important features even when cross effects are present.

Lastly, scenario E was designed to highlight the effects of correlated features. In the computation of feature importance, only the marginal effect of each feature is measured, which can lead to biased interpretation in the presence of strongly correlated features [61]. Detailed results are provided in the supplementary material I.4, they illustrate the biased interpretation that may happen when computing feature importance in the presence of strongly correlated features. To address this issue, we suggest two alternative methods: either preprocessing the dataset by keeping only one of the correlated features or computing conditional feature importance by perturbing the correlated features simultaneously [57].

5. Results on the METABRIC data set

5.1. Data description

The data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort, as described in the study by Curtis et al. (2012) [62][1] were used. Our analysis focuses on a sample

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

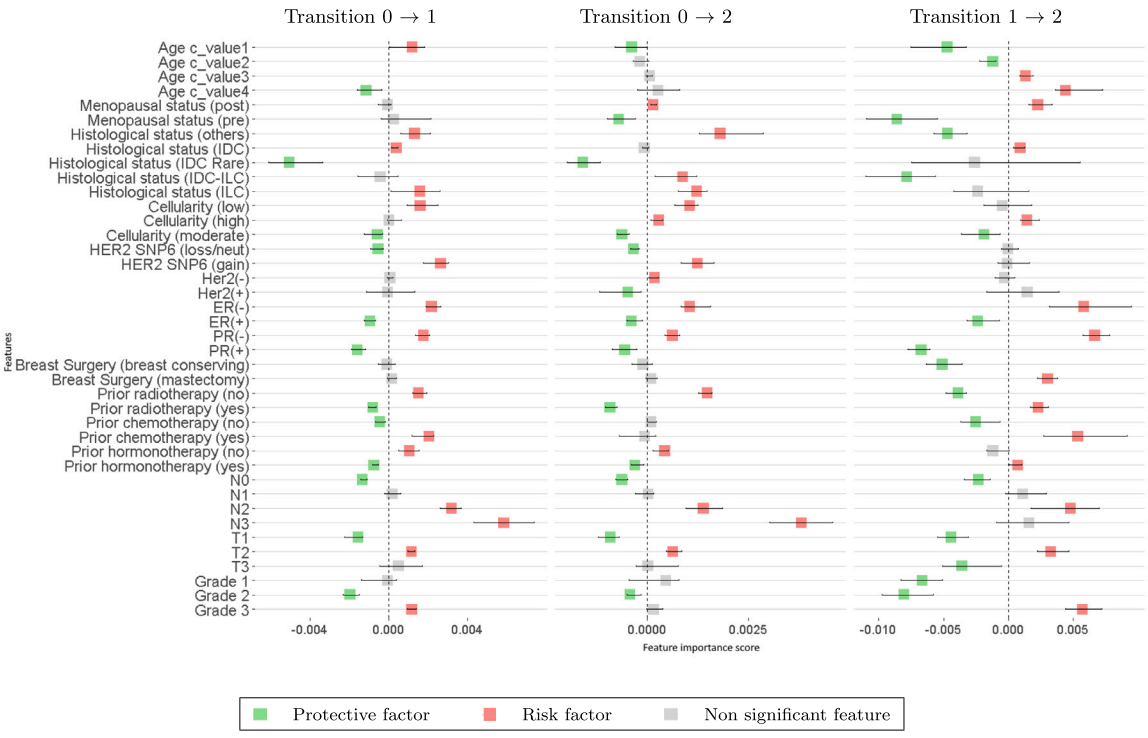


Fig. 13. Estimated important prognostic factors in the METABRIC data set on patients with breast cancer. Feature importance with a prediction-based MS-CPFI and the cumulative hazard functions as the predictions of interest. The features "Age" following by the suffix "c_value" designate the selected counterfactual values illustrated in Fig. 14.

Table 2
Number of (No.) observed event and right-censored patients (METABRIC data set on patients with breast cancer).

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

[a] Among patients at risk.

of 1903 patients who were followed up for a period of 360 months (30 years). To model the progression of breast cancer in this dataset, a three-states illness-death model was used. The initial state is denoted by 0 and corresponds to "Inclusion in the cohort", while state 1 represents an intermediate stage of "Relapse", and state 2 is an absorbing state that denotes "Death". Descriptive statistics for these events are presented in Table 2.

The METABRIC dataset comprises clinical, histo pathological, gene copy number, and gene expression features used to determine breast cancer subgroups. Based on the literature, we selected the most relevant clinical and histo pathological features, excluding those that exhibited high correlation, as confirmed by the results presented in the supplementary material (Appendix A).

In addition, we transformed some of the initial numerical features into categorical features based on clinical classification, ultimately retaining one numerical and fourteen categorical features for analysis.

To address missing values, we imputed the median value for the numerical feature and the mode for categorical features. We applied dummy-encoding to the categorical features and standardized the numerical feature using a min–max scaler. A description of these features and their transformations is presented in Table J.1.1 of the supplementary material (Appendix J.1).

To generate the necessary predictions for applying MS-CPFI to the METABRIC dataset, we employ the deep learning-based prediction algorithm IDNetwork for illness-death model [7], as detailed in the supplementary material (Appendix A). The outputs of IDNetwork correspond to density probability functions, which we convert into cumulative hazard functions as the predictions of interest, as per the equivalences defined in Section 2.2.

5.2. Results

In this section, we demonstrate the effectiveness of the prediction-based MS-CPFI algorithm in identifying clinically important features using real data. Here, we focus on the integrated feature importance (with no absolute value). Fig. 13 displays results of feature importance for each transition. We analyzed these interpretability results by checking in the literature if they are clinically valid. In the next paragraphs, we provide an analyze of the main important risk factors; additional analyses are given in Section 5.3 and in the supplementary material J.2.

To determine the counterfactual values of interest for the numerical feature "Age" corresponding to changes in the risk of transition for each state, we utilize spline-based univariate transition-specific Cox P.H. models, as shown in Fig. 14. For transition 0 → 1, we can consider that between the extreme values of age the effect is monotone, and choose the counterfactual values as the extreme values, i.e. {22, 96}. For transition 0 → 2 the selected age values are {22, 45, 70, 96}; for transition 1 → 2 the selected age values are {22, 50, 70, 96}. Between each selected value, we consider a monotonic effect of the risk.

Cancer grade [63]. Tumor grade is a prognostic classification based on the proliferation rate of cancer cells. A grade 1 tumor grows slowly and is associated with a low likelihood of metastasis, while a grade 3 tumor grows rapidly and is associated with a high likelihood of metastasis. Grade 2 tumors grow faster than grade 1 tumors but slower than grade 3 tumors and have an intermediate probability of metastasis. Therefore, patients with grade 1 and grade 2 tumors generally have a better prognosis than those with grade 3 tumors. Our analysis confirms these findings, as we observe that grade 3 is a risk factor for relapse (i.e., transition 0 → 1), while grade 2 is a protective factor. In addition,

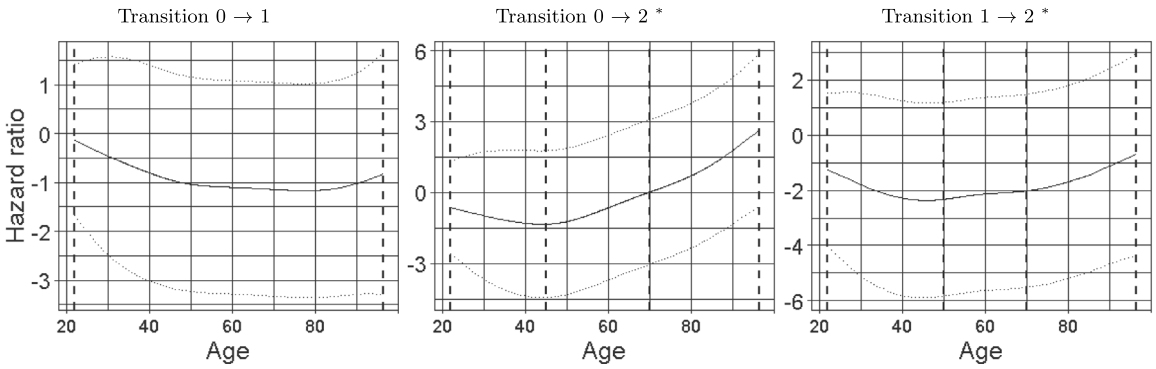


Fig. 14. Selection of the counterfactual values for the feature "Age" (METABRIC data set on patients with breast cancer). Estimated transition-specific hazard ratios for feature "age" as functions of "age". Dashed lines represent the selected counterfactual values. We tested the significativity of the non-linear part for each model with a threshold of 0.05; * indicates a significant *p*-value.

for the risk of death after a relapse (i.e., transition 1 → 2), grades 1 and 2 are protective factors, while grade 3 is a risk factor. These findings are consistent with previous research in the field.

TNM classification [64]. The TNM (Tumor size, Nodes involvement, presence of Metastasis) classification from the American Joint Committee on Cancer (AJCC) is used to establish tumor stage of patients with breast cancer. The tumor size (T) is a classification from $T0$ (no evidence of primary tumor) to $T3$ (tumor size ≥ 5 cm), and $T4$ (tumor size growing into the chest wall and/or skin). The number of invaded lymph nodes (N) is a classification from $N0$ (cancer has not spread to nearby lymph nodes), to $N3$ (cancer has spread to ≥ 10 auxiliary lymph nodes). There features are well-known risk factors of breast cancer progression. Our results reveal that for transitions 0 → 1 and 0 → 2, a $T1$ cancer is a protective factor and $T2$ is risk factor. $T3$ is not significant, that is probably due to the fact that the frequency of observations of $T3$ is low against frequencies of $T1$ and $T2$. For all three transitions, $N0$ is a protective factor, $N1$ is not significant, $N2$ and $N3$ are risk factors (except for transition 1 → 2 where $N3$ is not significant). These findings align with the existing literature, indicating that a T0 (or N0) cancer is a protective factor, while a higher T (respectively a higher N) increases the risks of relapse and death.

Hormone receptor status [65]. Breast cancers can be classified in two groups: hormone receptor-positive versus hormone receptor-negative cancers. In hormone receptor-positive breast cancers, female sex hormones (estrogen - ER - and/or progesterone - PR) stimulate tumor growth. These cancers usually grow slower than hormone receptor-negative cancers and have a better short-term prognosis, but may have a higher risk of late recurrence. In hormone receptor-negative breast cancers, female sex hormones do not affect cancer cells growth. They have a greater risk of relapse in the first years after the end of treatment.

In our results, we found that hormone receptor-positive (ER+ and/or PR+) cancers are favorable prognostic factors, while hormone receptor-negative (ER- and/or PR-) cancers are unfavorable prognostic factors for all three transitions. This means that hormone receptor-positive cancers have a lower risk of relapse or death than hormone receptor-negative cancers; these conclusions are consistent with the literature.

Her2 status [65]. Breast cancers can also be classified as Her2-positive (noted Her2+) or Her2-negative (noted Her2-) cancers. Her2+ breast cancers have a higher level of the protein Her2 (Human epidermal growth factor receptor 2). This protein increases the growth of cancer cells, which makes the cancer more aggressive than Her2- cancers. However, Her2-targeted treatments are very effective, that makes Her2+ cancers having a very good prognosis for relapse and death. In ours results, we can see that, for transition 0 → 2, Her2- is a risk factor, Her2+ is a protective factor; for transitions 0 → 1 and 1 → 2, there is no significant effect of the protein Her2.

Age at diagnosis [66]. Age is a known risk factor of beast cancers; incidence increases with age. However, patients diagnosed at a young age (less than 40 years old) have a poorer prognosis than patients aged from 40 to 60 years old. They present a higher probability to have a triple-negative cancer (i.e. ER-, PR- and Her2-) because they are about 2 to 3 times more likely to have the BRCA1 mutation that is linked to the development of aggressive breast cancer as the triple-negative breast cancer. These cancers have a poorer remission prognosis than other sub-types of breast cancer. Consequently, this age group has a higher recurrence rate than the others age groups. Patients over 70 years old have the lowest survival; their survival is affected by their age and their risk of developing medical comorbidities. However, cancers of elderly patients are mainly hormone receptor-positive cancers; therefore they present a low risk of cancer-related death.

In our results, for transition 0 → 1 (i.e. the risk of relapse), the age value 22 (Age c-value1) is a risk factor, the age value 96 (Age c-value4) is a protective factor. These results are consistent with the literature; younger patients have a higher probability of relapse; older patients have a good prognosis of cancer recurrence as they can be treated effectively with hormonotherapy. For transition 0 → 2 (i.e. the risk of death with no cancer relapse), the age value 22 (Age c-value1) is a protective factor, the age values 45, 70 and 96 (Age c-value2, Age c-value3 and Age c-value4) have no significant effect. These results are consistent with the literature; younger patients have a low probability of death with no cancer relapse because the occurrence of this transition is strongly linked to non-cancer deaths and comorbidities-related deaths which rather affect older patients. For transition 1 → 2 (i.e. the risk of death after a cancer relapse), the age values 22 and 50 (Age c-value1 and Age c-value2) are protective factors, the age values 70 and 96 (Age value3" and "Age c-value4) are risk factors. We can also see that patients aged of 96 years old have a higher risk for this transition than patients aged of 70 years old. These results are consistent as the risk of death for patients over 70 years old is strongly correlated to comorbidities, and this risk increases when age increases.

In parallel, effects of the age values are ordered. For transition 0 → 1, the risk of relapse decreases when age increases; for transition 0 → 2, the risk of non-cancer deaths decreases when age decreases; for transition 1 → 2, the risk death after a cancer relapse decreases when age decreases. This allows to characterize the shape of the feature effect for each transition (see supplementary material J.2 - Figure J.2.1 for an illustration).

5.3. Additional results

In the supplementary material J.3, we compare our results with the interpretability results from transition-specific Cox P.H. models; we conclude that the state-of-the-art statistical method does not allow to detect all the known prognostic factors, and is limited in the presence of non-linear effects of the features.

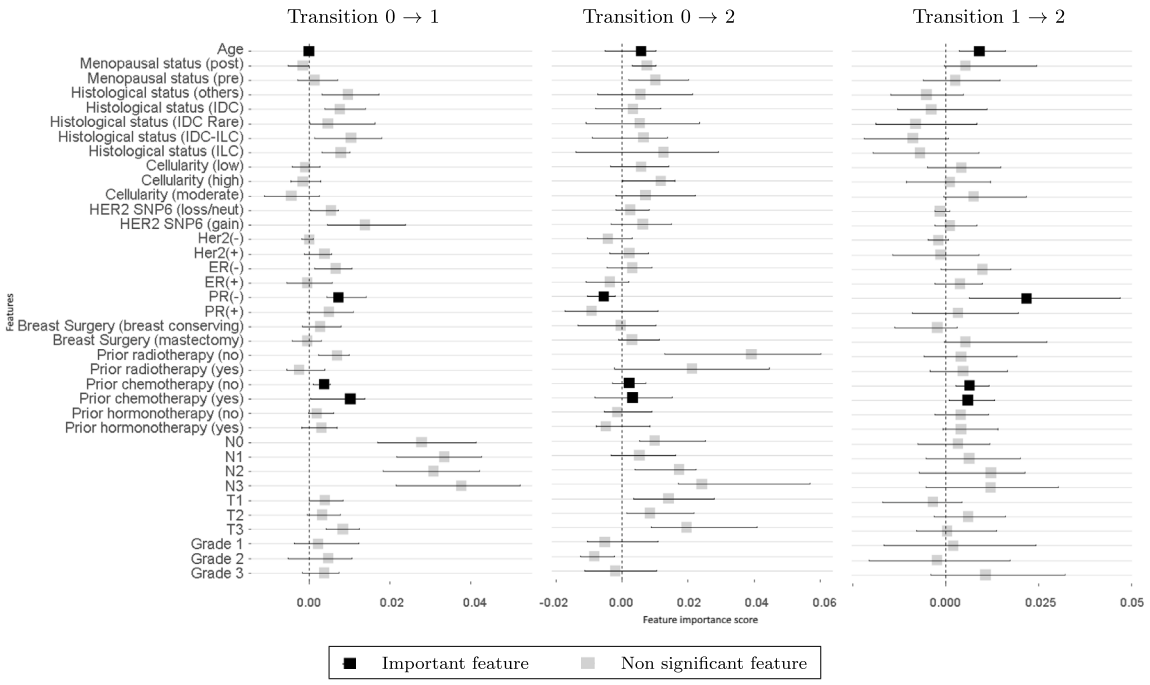


Fig. 15. iAUC-based feature importance (METABRIC data set on patients with breast cancer). Feature importance with an iAUC-based MS-CPFI and the cumulative hazard functions as the predictions of interest.

We also compared the prediction-based feature importance, as shown in Fig. 13, with an iAUC-based feature importance, illustrated in Fig. 15. Our goal was to highlight the limitations of using a performance criterion to compute feature importance scores for clinical interpretation.

We observed that the sign of the feature importance, i.e. whether it is positive or negative, is not interpretable when computing an iAUC-based feature importance. A positive feature importance indicates that perturbing the feature decreases the iAUC of the model, implying that the feature is important for model training. On the other hand, a negative feature importance is not interpretable.

We found that some well-known important prognostic factors for breast cancer, such as the hormone-receptor status, are non-significant when using the iAUC-based feature importance method. Particularly for transition $1 \rightarrow 2$, we observed several important features, including the T and N classifications, cancer grade, hormone-receptor status, and histological classification, which were not significant. Additionally, as the sign of an iAUC-based feature importance score is not interpretable, it does not provide any information on the sign of the feature effect. Therefore, we concluded that an iAUC-based feature importance method does not provide the expected interpretation of the features in contrast to our prediction-based MS-CPFI feature importance scores.

We also analyzed the results of an iBS-based MS-CPFI, as described in the supplementary material J.4, and reached the same conclusions. Thus, we validated that a prediction-based feature importance method is more relevant for providing a real clinical sense of the interpretations.

6. Conclusion

We have developed MS-CPFI, a new model-agnostic algorithm for interpreting any irreversible multi-state model and understanding the effects of covariates on disease progression. This is the first interpretability algorithm of its kind developed in the context of multi-state analysis. The use of interpretability methods is crucial for integrating black-box-based disease progression predictions into a clinical decision support system. MS-CPFI is particularly useful in personalized medicine, providing information on the clinical factors influencing the evolution of a specific disease.

MS-CPFI extends the class of feature importance algorithms to a larger class of disease models dealing with time-to-event data, including the successive occurrence of multiple clinical events. We have derived the definition of the state-of-the-art PFI algorithm used in RSFs, with major improvements to better interpret covariate effects and address limitations of the PFI algorithm in this context. By design, the input of MS-CPFI are time-dependent and transition-specific risk predictions and the output is statistical summaries of feature importance for each transition. By using a new counterfactual perturbation method and using risk predictions directly instead of model performance, the MS-CPFI algorithm provides an ordered list of features that are risk factors or protective factors for each stage of a disease (or transition of a multi-state process).

We have tested the robustness of MS-CPFI in the case of non-linear effects of the features on the risks of transition through experiments on simulated data sets. We have also evaluated MS-CPFI in interpreting an illness-death black-box model trained on a real dataset on patients with breast cancer, showing that MS-CPFI can detect clinically important features for evaluating prognosis in such patients.

In summary, MS-CPFI's strength lies in its algorithm agnosticism, allowing interpretation across various multi-state models, and its clinical interpretability, providing insights into risk factors and protective factors with confidence intervals. However, the current implementation has limitations that warrant improvement. One limitation involves the choice of counterfactual scenarios for interpreting the effect of a numerical feature, which should be integrated and automated in the next version of MS-CPFI. Another limitation is the management of correlated data, which may introduce bias, and we provide methodologies to address this in the supplementary material I.4 (Figure I.4.2) Additionally, MS-CPFI is currently designed to interpret multi-state models using tabular data, but considerations for other data types, such as imaging, anatomo-pathological, or connected device data, are essential for comprehensive patient prognosis evaluation. The algorithm also focuses on static transition-specific predictions, limiting its applicability. Future iterations should explore integrating time-varying features and

dynamic interpretations based on the temporal evolution of patients' characteristics, enhancing the prediction of patients' prognosis and adapting MS-CPFI for this purpose. As a future work and inspired by SurvShap(t) [44] and SurvLIME [41], we also aim to combine MS-CPFI with a local interpretability functionality to provide individual feature importance scores in a multi-state model.

CRediT authorship contribution statement

Aziliz Cottin: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Marine Zulian:** Supervision, Validation, Writing – original draft, Writing – review & editing. **Nicolas Pécuchet:** Supervision, Validation. **Agathe Guilloux:** Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Sandrine Katsahian:** Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare no potential conflict of interests.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artmed.2023.102741.

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] Hajihosseini M, Faradmal J, Sadighi-Pashaki A. Survival analysis of breast cancer patients after surgery with an intermediate event: Application of illness-death model. Iran J Public Health 2015;44(12):1677.
[3] Dignam JJ, Zhang Q, Kocherginsky M. The use and interpretation of competing risks regression ModelsModeling with competing risks. Clin Cancer Res 2012;18(8):2301–8.
[4] Cox DR. Regression models and life-tables. J R Stat Soc Ser B Stat Methodol 1972;34(2):187–202.
[5] 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 Programs Biomed 2010;99(3):261–74.
[6] Lee C, Zame WR, Yoon J, van der Schaar M. Deephit: A deep learning approach to survival analysis with competing risks. In: Thirty-second AAAI conference on artificial intelligence. 2018.
[7] 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. Stat Med 2022;41(9):1573–98.
[8] Wang P, Li Y, Reddy CK. Machine learning for survival analysis: A survey. ACM Comput Surv 2019;51(6):1–36.
[9] Ishwaran H, Kogalur UB. Random survival forests for R. R News 2007;7(2):25–31.
[10] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. 2016.
[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, van der Schaar M. Rnn-surv: A deep recurrent model for survival analysis. In: International conference on artificial neural networks. Springer; 2018, p. 23–32.
[16] Ren K, Qin J, Zheng L, Yang Z, Zhang W, Qiu L, et al. Deep recurrent survival analysis. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 33. 2019, p. 4798–805.
[17] Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. PeerJ 2019;7:e6257.
[18] Pfeifer B, Holzinger A, Schimek MG. Robust random forest-based all-relevant feature ranks for trustworthy ai. Stud Health Technol Inform 2022;294:137–8.
[19] Panigutti C, Hamon R, Hupont I, Fernandez Llorca D, Fano Yela D, Junklewitz H, et al. The role of explainable AI in the context of the AI act. In: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency. 2023, p. 1139–50.
[20] Müller H, Holzinger A, Plass M, Brcic L, Stumptner C, Zatloukal K. Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European in vitro diagnostic regulation. New Biotechnol 2022;70:67–72.
[21] Geller J. Food and drug administration published final guidance on clinical decision support software. J Clin Eng 2023;48(1):3–7.
[22] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[23] Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. Biostatistics 2014;15(4):757–73.
[24] Molnar C. Interpretable machine learning. Lulu. com; 2020.
[25] Lewis RJ. An introduction to classification and regression tree (CART) analysis. In: Annual meeting of the society for academic emergency medicine in san francisco, california. Vol. 14. Citeseer; 2000.
[26] Agarwal R, Melnick L, Frosst N, Zhang X, Lengerich B, Caruana R, et al. Neural additive models: Interpretable machine learning with neural nets. In: Advances in neural information processing systems. Vol. 34. 2021, p. 4699–711.
[27] Olden JD, Jackson DA. Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. Ecol Model 2002;154(1–2):135–50.
[28] 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.
[29] Ancona M, Ceolini E, Öztireli C, Gross M. Gradient-based attribution methods. Explainable AI: Interpret Explain Visual Deep Learn 2019;169–91.
[30] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat 2001;1189–232.
[31] Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Ser B Stat Methodol 2020;82(4):1059–86.
[32] Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. J Amer Statist Assoc 2018;113(523):1094–111.
[33] Zhang Y, Tiňo P, Leonardis A, Tang K. A survey on neural network interpretability. 2020, arXiv preprint arXiv:2012.14261.
[34] Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 2015;24(1):44–65.
[35] Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016, p. 1135–44.
[36] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. Vol. 30. 2017.
[37] 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.
[38] Wang D, He K, Garmire LX. Cox-nnet v2. 0: Improved neural-network based survival prediction extended to large-scale EMR dataset. 2020, arXiv preprint arXiv:2009.04412.
[39] Hao J, Kim Y, Mallavarapu T, Oh JH, Kang M. Cox-PASNet: Pathway-based sparse deep neural network for survival analysis. In: 2018 IEEE international conference on bioinformatics and biomedicine. IEEE; 2018, p. 381–6.
[40] Hao J. Biologically interpretable, integrative deep learning for cancer survival analysis. 2019.
[41] Kovalev MS, Utkin LV, Kasimov EM. SurvLIME: A method for explaining machine learning survival models. Knowl-Based Syst 2020;203:106164.
[42] Li R, Shinde A, Liu A, Glaser S, Lyou Y, Yuh B, et al. Machine learning–based interpretation and visualization of nonlinear interactions in prostate cancer survival. JCO Clin Cancer Inform 2020;4:637–46.
[43] Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. 2018, arXiv preprint arXiv:1802.03888.
[44] Krzyziński M, Spytek M, Baniecki H, Biecek P. SurvSHAP (t): Time-dependent explanations of machine learning survival models. Knowl-Based Syst 2022;110234.
[45] Spytek M, Krzyziński M, Baniecki H, Biecek P. survex: Model-agnostic explainability for survival analysis.
[46] Ehrlinger J. GgRandomForests: Exploring random forest survival. 2016, arXiv preprint arXiv:1612.08974.
[47] Andersen PK, Borgan O, Gill RD, Keiding N. Statistical models based on counting processes. Springer Science & Business Media; 2012.
[48] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: Competing risks and multi-state models. Stat Med 2007;26(11):2389–430.
[49] Heggland T. Estimating transition probabilities for the illness-death model. 2015.
[50] Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. Vol. 360. John Wiley & Sons; 2011.

[51] Pepe MS, Mori M. Kaplan—Meier, marginal or conditional probability curves in summarizing competing risks failure time data? Stat Med 1993;12(8):737–51.
[52] Pintilie M. Competing risks: a practical perspective. Vol. 58. John Wiley & Sons; 2006.
[53] Cabarrou B, Dalenc F, Leconte E, Boher J-M, Filleron T. Focus on an infrequently used quantity in the context of competing risks: The conditional probability function. Comput Biol Med 2018;101:70–81.
[54] Zhang M-J, Fine J. Summarizing differences in cumulative incidence functions. Stat Med 2008;27(24):4939–49.
[55] Ishwaran H, Lauer MS, Blackstone EH, Lu M, Kogalur UB. Randomforestsrc: Random survival forests Vignette. 2021.
[56] Ishwaran H, Gerds TA, Lau BM, Lu M, Kogalur UB. randomForestSRC: Competing risks Vignette. 2021.
[57] Molnar C, König G, Herbinger J, Freiesleben T, Dandl S, Scholbeck CA, et al. General pitfalls of model-agnostic interpretation methods for machine learning models. In: XxAI-beyond explainable AI: international workshop, held in conjunction with ICML 2020. Springer; 2022, p. 39–68.
[58] Kovalev M, Utkin L, Coolen F, Konstantinov A. Counterfactual explanation of machine learning survival models. Informatica 2021;32(4):817–47.
[59] Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Stat Med 2005;24(11):1713–23.
[60] Therneau T. Spline terms in a Cox model. 2017, Self-published at: https://cran.r-project.org/web/packages/survival/vignettes/splines.pdf.
[61] Molnar C, Gruber S, Kopper P. Limitations of interpretable machine learning methods. 2020.
[62] Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486(7403):346–52.
[63] Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, et al. Breast cancer prognostic classification in the molecular era: The role of histological grade. Breast Cancer Res 2010;12(4):1–12.
[64] Giuliano AE, Connolly JL, Edge SB, Mittendorf EA, Rugo HS, Solin LJ, et al. Breast cancer—major changes in the American joint committee on cancer eighth edition cancer staging manual. CA: A Cancer J Clin 2017;67(4):290–303.
[65] Allison KH, Hammond MEH, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, et al. Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update. 2020.
[66] McGuire A, Brown JA, Malone C, McLaughlin R, Kerin MJ. Effects of age on the detection and management of breast cancer. Cancers 2015;7(2):908–29.