Counterfactual Perturbation Methods for Interpreting Multi-State Models

Resource Paper: **MS-CPFI: Counterfactual Perturbation Feature Importance for Interpreting Multi-State Models**

Aziliz Cottin, Marine Zulian, Nicolas Pécuchet, Agathe Guilloux, Sandrine Katsahian

Romen Samuel Rodis Wabina, MSc
PhD candidate, Data Science for Healthcare and Clinical Informatics
Data Scientist, Department of Clinical Epidemiology and Biostatistics

MS-CPFI: A Model-Agnostic Counterfactual Perturbation Feature Importance Algorithm for Interpreting black-box Multi-State Models

Model-Agnostic An interpretation method that works with any type of model (e.g., SHAP)

Counterfactual Exploring ***what if*** scenarios,
e.g., what if a particular variable were different?

Perturbation Making ***changes to the input data***
By tweaking one variable at a time, you can observe how each changes impacts the result.

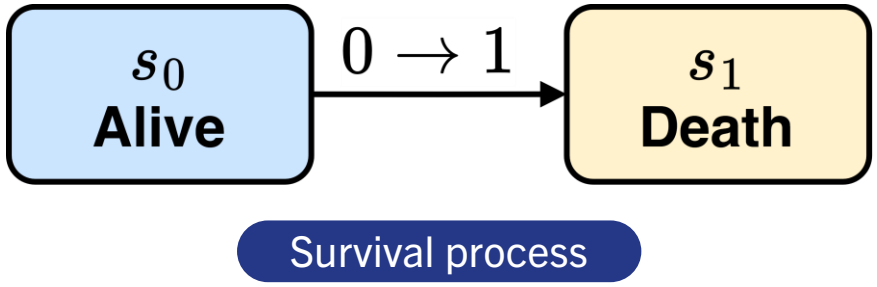
Feature Importance Finding out ***which features have the most effect*** on the model's prediction

Multi-State Models framework used to describe processes where patients move between different states over time

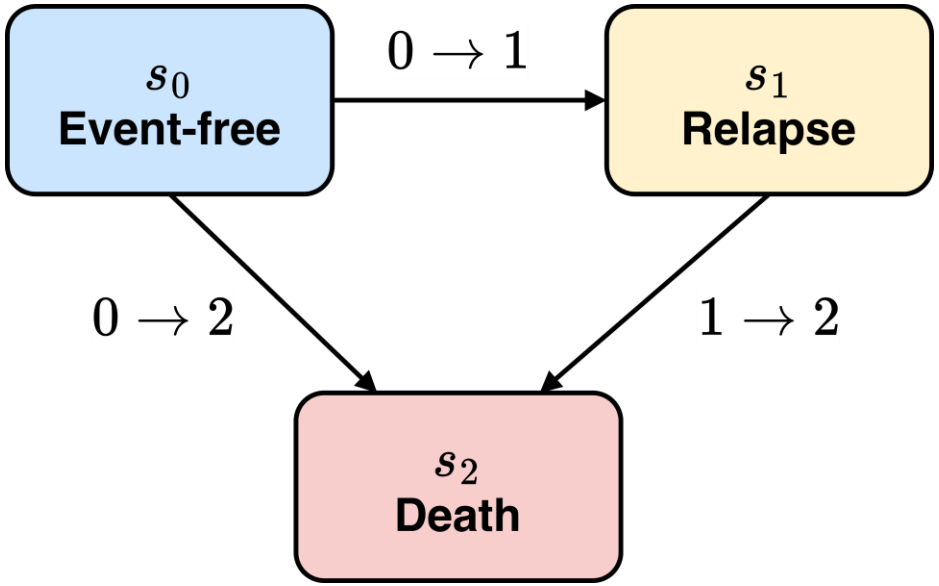
MS-CPFI: A Model-Agnostic Counterfactual Perturbation Feature Importance Algorithm for Interpreting black-box Multi-State Models

- MS-CPFI is an interpretation method that uses ***what if*** scenarios and adjusts the data to figure out which factors are most important in influencing the predictions of any complex model that outputs multiple states.

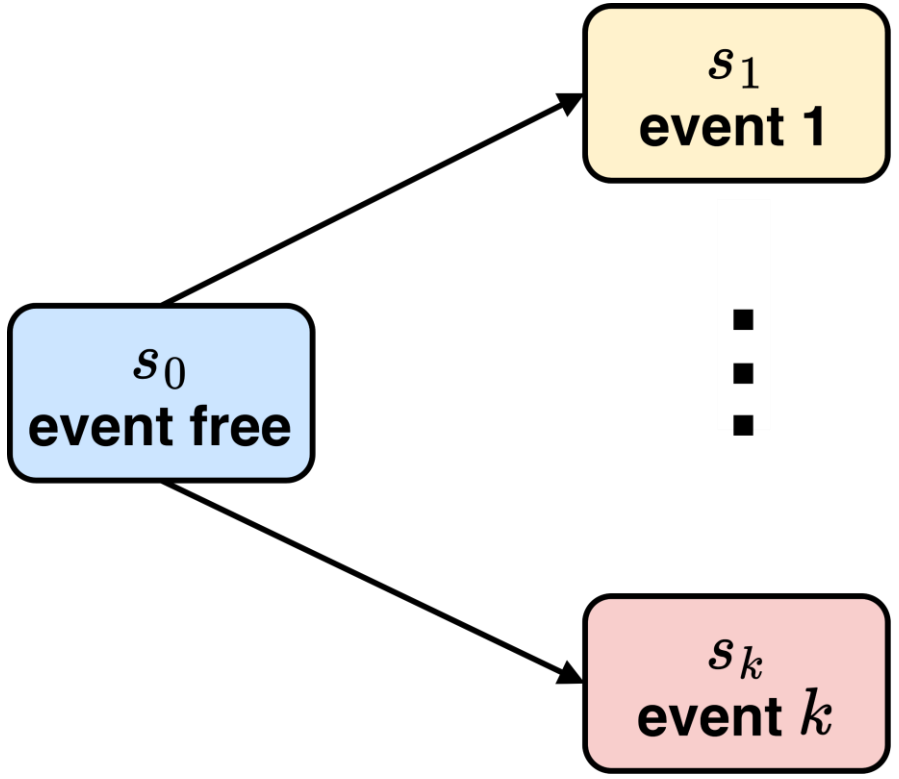
Multi-State Models



Survival process

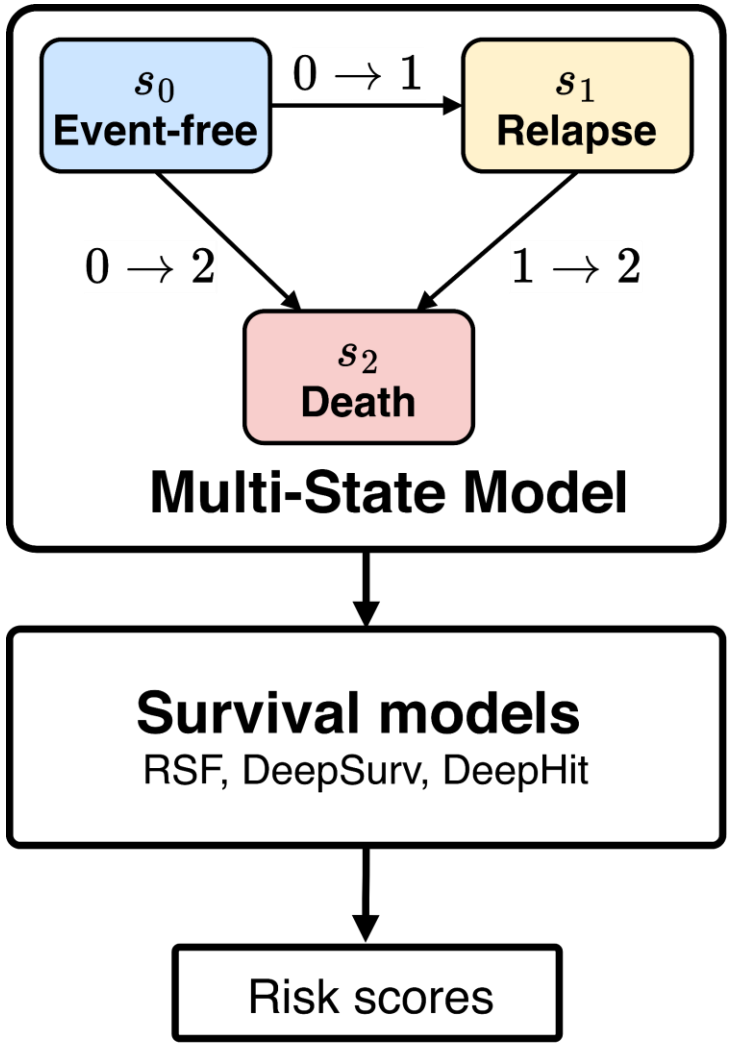


Illness-death model

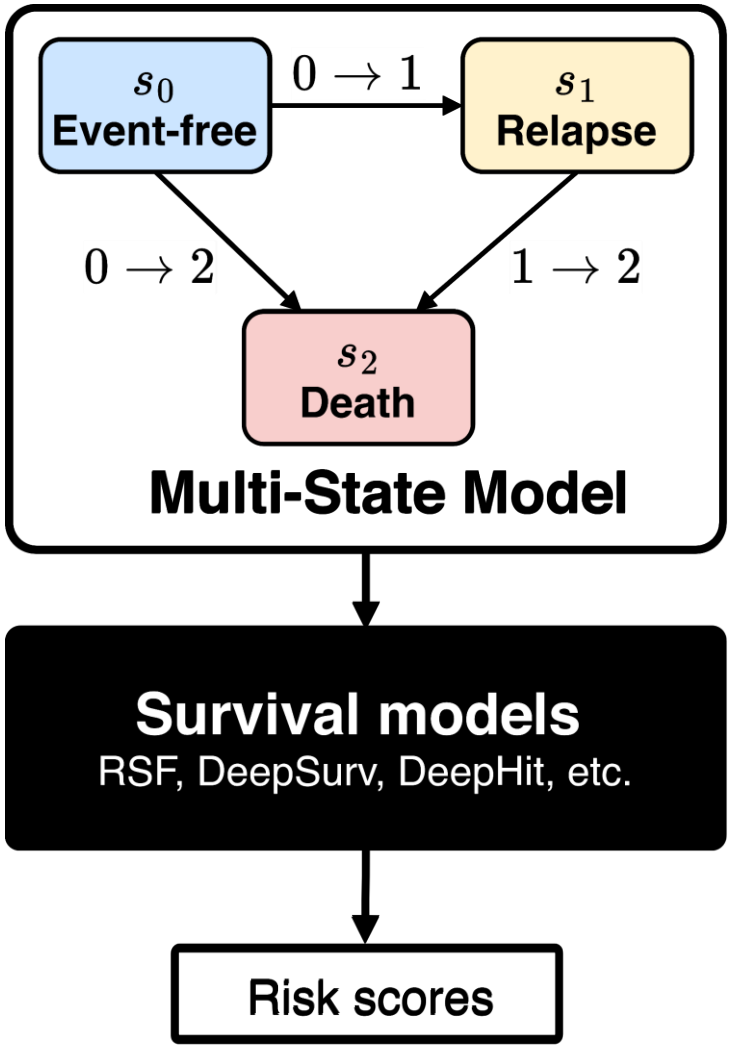


Competing risks

Survival models



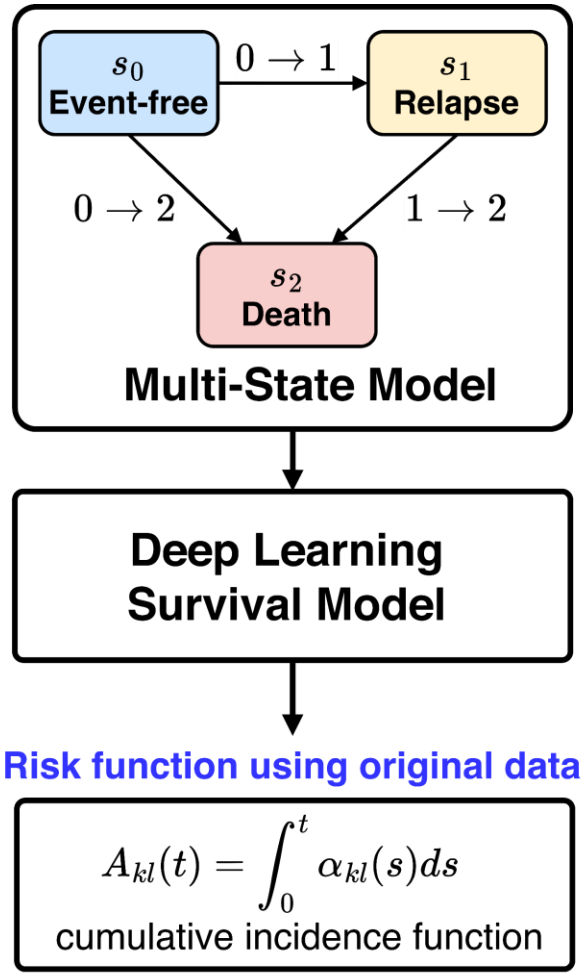
Survival models



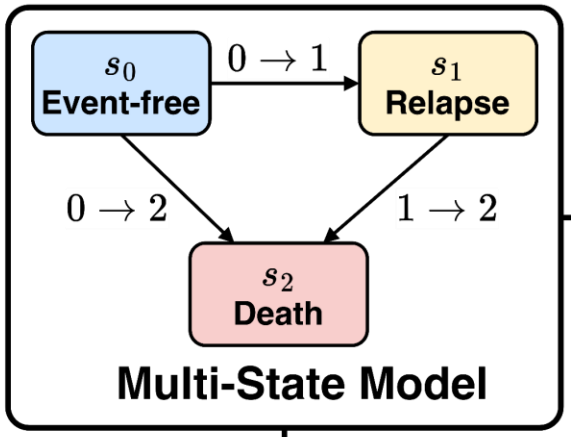
- Black-box models provide non-interpretable results
- Model interpretability is essential to ensure that the predictions are **reliable**, **trustworthy**, and **accurate**.

- Human understanding is a key issue
  - AI Act (Europe) requires ML-based softwares to be explainable.
  - Research guidelines on publishing clinical AI articles and clinical decision support tools.

Feature Importance



Feature Importance



Feature Importance
Permutation-based

Randomly shuffles a feature's values to see how it affects model performance

Deep Learning Survival Model

Deep Learning Survival Model

Risk function using original data

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

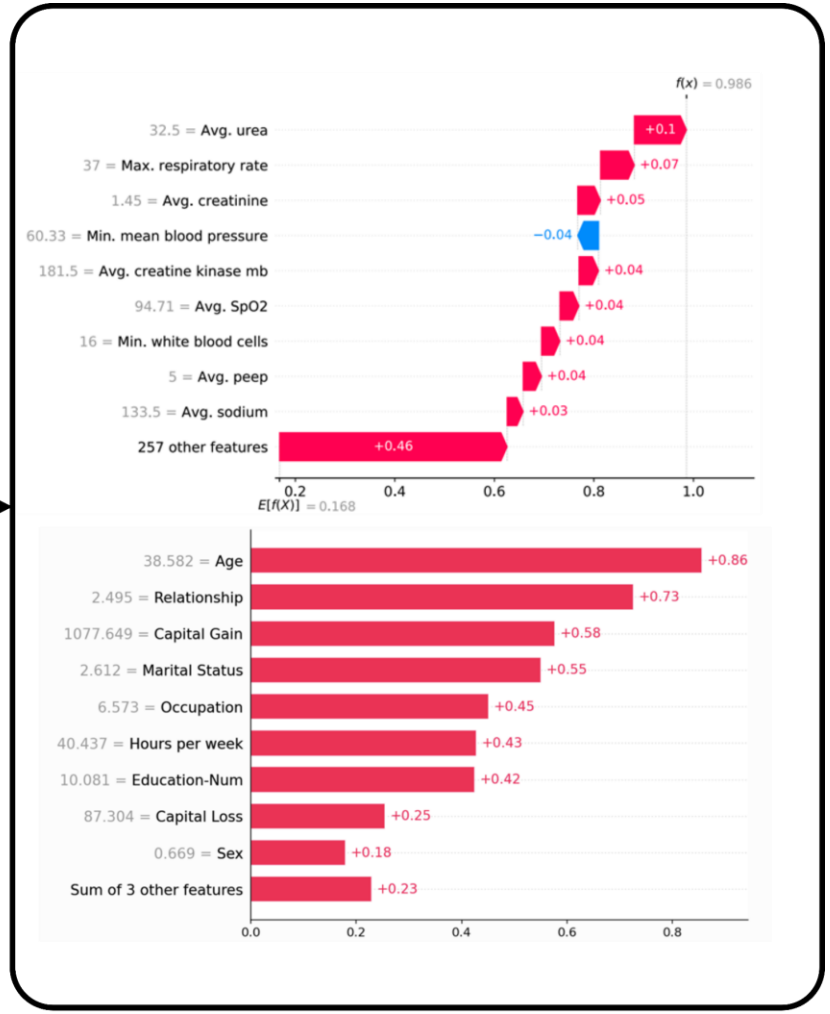
cumulative incidence function

versus

Risk function using permutated data

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

cumulative incidence function



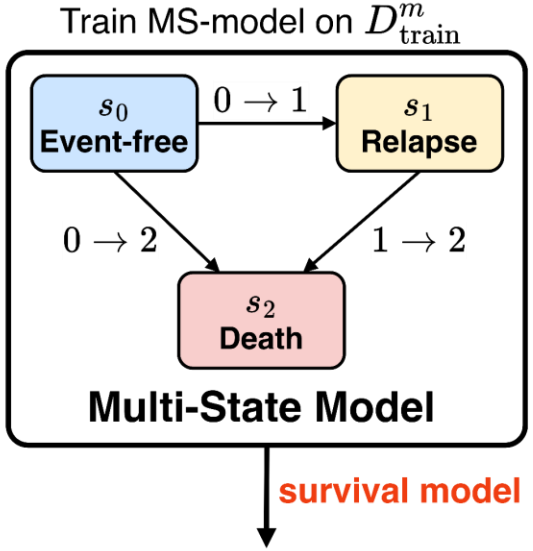
Random permutation vs counterfactual perturbation

- Traditional approach: permutation feature importance (PFI)
  - Randomly shuffles a feature's values to see how it affects model performance
  - Purely random swaps can inject noise, fail to isolate non-linear effects of a feature value
  - Yield high-variance estimates

- Drawbacks of random permutation:
  - Unstable: re-running permutation multiple times produce fluctuating importance scores
  - A random shuffle doesn't reveal specific value

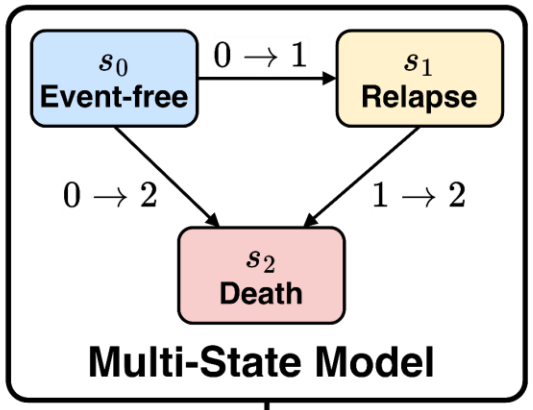
- Counterfactual perturbation
  - Replace observed values of a feature by the counterfactual ones, all other things being equal
    - Evaluate how much the predicted risk changes under that specific new value—revealing protective vs. risk-prone states more directly.

MS-CPFI algorithm



MS-CPFI algorithm

Train MS-model on D^m_{train}



survival model

Compute the predictions D^m_{val}

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

cumulative transition intensity

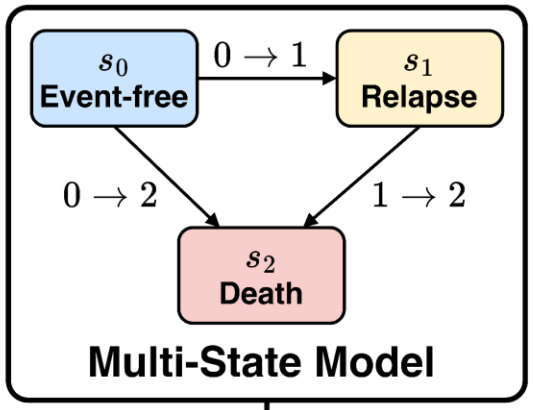
Compute the average reference transition-specific predictions

$$\overline{A}^m_{kl}(t) = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} A_{kl}(t \mid X^i)$$

risk scores using original data

MS-CPFI algorithm

Train MS-model on D^m_{train}



survival model

Compute the predictions D^m_{val}

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

cumulative transition intensity

Compute the average reference transition-specific predictions

$$\overline{A}^m_{kl}(t) = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} A_{kl}(t \mid X^i)$$

risk scores using original data

Consider a feature X_j $(1 \leq j \leq P)$ to create counterfactual scenarios

For a selected counterfactual scenario, perturb values on the feature j in D^m_{val} for each observation i

original data

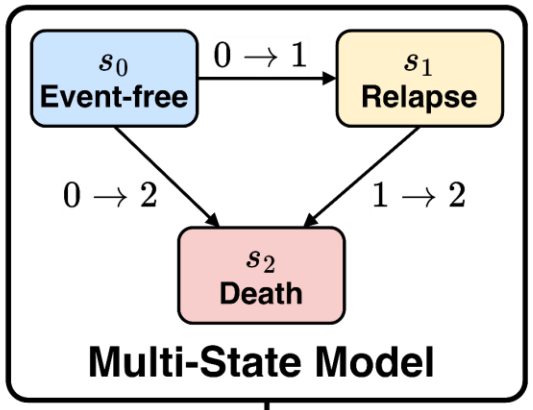
Patient	BMI
1	28.5
2	26.3
3	31.2
4	29.7
5	27.8

counterfactual perturbation

Patient	BMI
1	22
2	22
3	22
4	22
5	22

MS-CPFI algorithm

Train MS-model on D^m_{train}



survival model

Compute the predictions D^m_{val}

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

cumulative transition intensity

Compute the average reference transition-specific predictions

$$\overline{A}^m_{kl}(t) = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} A_{kl}(t \mid X^i)$$

risk scores using original data

Consider a feature X_j $(1 \leq j \leq P)$ to create counterfactual scenarios

For a selected counterfactual scenario, perturb values on the feature j in D^m_{val} for each observation i

original data

Patient	BMI
1	28.5
2	26.3
3	31.2
4	29.7
5	27.8

→

counterfactual perturbation

Patient	BMI
1	22
2	22
3	22
4	22
5	22

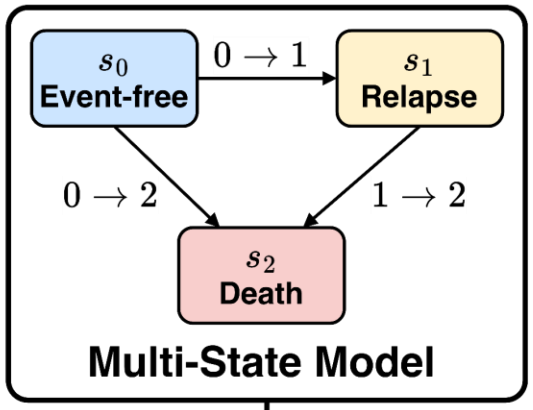
Compute the average transition-specific prediction in the counterfactual scenario as:

$$\overline{A}_{kl}(t; \overline{x}_j) = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} A_{kl}(t \mid X^i_j, \overline{x}_j)$$

risk scores using perturbed data

MS-CPFI algorithm

Train MS-model on D^m_{train}



survival model

Compute the predictions D^m_{val}

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

cumulative transition intensity

Compute the average reference transition-specific predictions

$$\overline{A}^m_{kl}(t) = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} A_{kl}\left(t \mid X^i\right)$$

risk scores using original data

Consider a feature X_j $(1 \leq j \leq P)$ to create counterfactual scenarios

For a selected counterfactual scenario, perturb values on the feature j in D^m_{val} for each observation i

original data

Patient	BMI
1	28.5
2	26.3
3	31.2
4	29.7
5	27.8

→

counterfactual perturbation

Patient	BMI
1	22
2	22
3	22
4	22
5	22

Compute the average transition-specific prediction in the counterfactual scenario as:

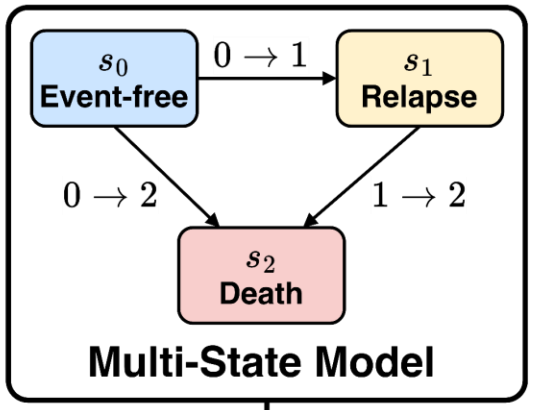
$$\overline{A}_{kl}(t; \overline{x}_j) = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} A_{kl}\left(t \mid X^i_j, \overline{x}_j\right)$$ risk scores using perturbed data

Compute the time-dependent feature importance using risk scores from original and perturbed data

$$\text{FI}^m_{kl}(t, \overline{x}_j) = \overline{A}^m_{kl}(t; \overline{x}_j) - \overline{A}^m_{kl}(t) \quad \text{OR} \quad \text{FI}^m_{kl}(t, \overline{x}_j) = \int_0^\tau \left(\overline{A}^m_{kl}(t; \overline{x}_j) - \overline{A}^m_{kl}(t)\right) dt$$

MS-CPFI algorithm

Train MS-model on D^m_{train}



survival model

Compute the predictions D^m_{val}

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

cumulative transition intensity

Compute the average reference transition-specific predictions

$$\overline{A}^m_{kl}(t) = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} A_{kl}(t \mid X^i)$$

risk scores using original data

Consider a feature X_j $(1 \le j \le P)$ to create counterfactual scenarios

For a selected counterfactual scenario, perturb values on the feature j in D^m_{val} for each observation i

original data

Patient	BMI
1	28.5
2	26.3
3	31.2
4	29.7
5	27.8

→

counterfactual perturbation

Patient	BMI
1	22
2	22
3	22
4	22
5	22

Compute the average transition-specific prediction in the counterfactual scenario as:

$$\overline{A}_{kl}(t; \overline{x}_j) = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} A_{kl}\left(t \mid X^i_j, \overline{x}_j\right)$$ risk scores using perturbed data

Compute the time-dependent feature importance using risk scores from original and perturbed data

$$\text{FI}^m_{kl}(t, \overline{x}_j) = \overline{A}^m_{kl}(t; \overline{x}_j) - \overline{A}^m_{kl}(t) \quad \text{OR} \quad \text{FI}^m_{kl}(t, \overline{x}_j) = \int_0^\tau \left(\overline{A}^m_{kl}(t; \overline{x}_j) - \overline{A}^m_{kl}(t)\right) dt$$

Output: Feature importance score for each feature x

$$\text{FI}_{kl}(\overline{x}_j) = \frac{1}{M} \text{FI}^m_{kl}(\overline{x}_j)$$

MS-CPFI Interpretation

$$\mathrm{FI}_{kl}^{m}(t, \overline{x}_j) = \overline{A}_{kl}^{m}(t; \overline{x}_j) - \overline{A}_{kl}^{m}(t)$$

$\overline{A}_{kl}^{m}(t; \overline{x}_j) < \overline{A}_{kl}^{m}(t)$

negative feature importance score

Feature is associated with a lower likelihood of the event, i.e., **protective effect**

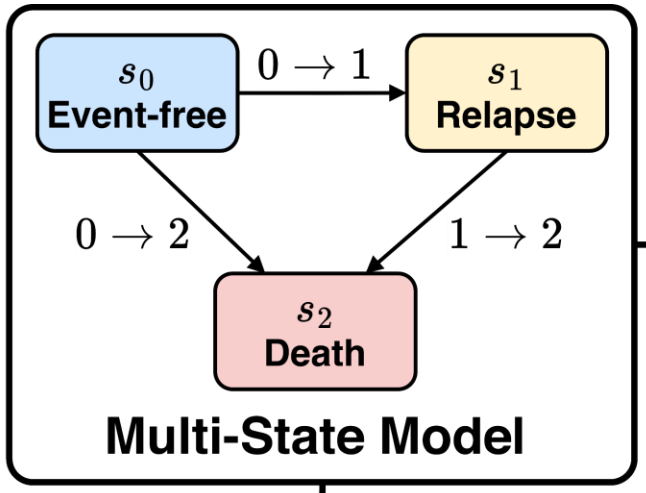
$\overline{A}_{kl}^{m}(t; \overline{x}_j) > \overline{A}_{kl}^{m}(t)$

positive feature importance score

Feature is associated with a higher likelihood of the event, i.e., **risk factor**

Suppose we perturbed age with a higher age value.

Scenario questions



MS-CPFI algorithm

Scenario 2: How to select counterfactual values for numerical and categorical features?

Risk function using original data

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

cumulative incidence function

Risk function using perturbed data

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

cumulative incidence function

Scenario 1: What should be the function that serves as the reference for MS-CPFI?

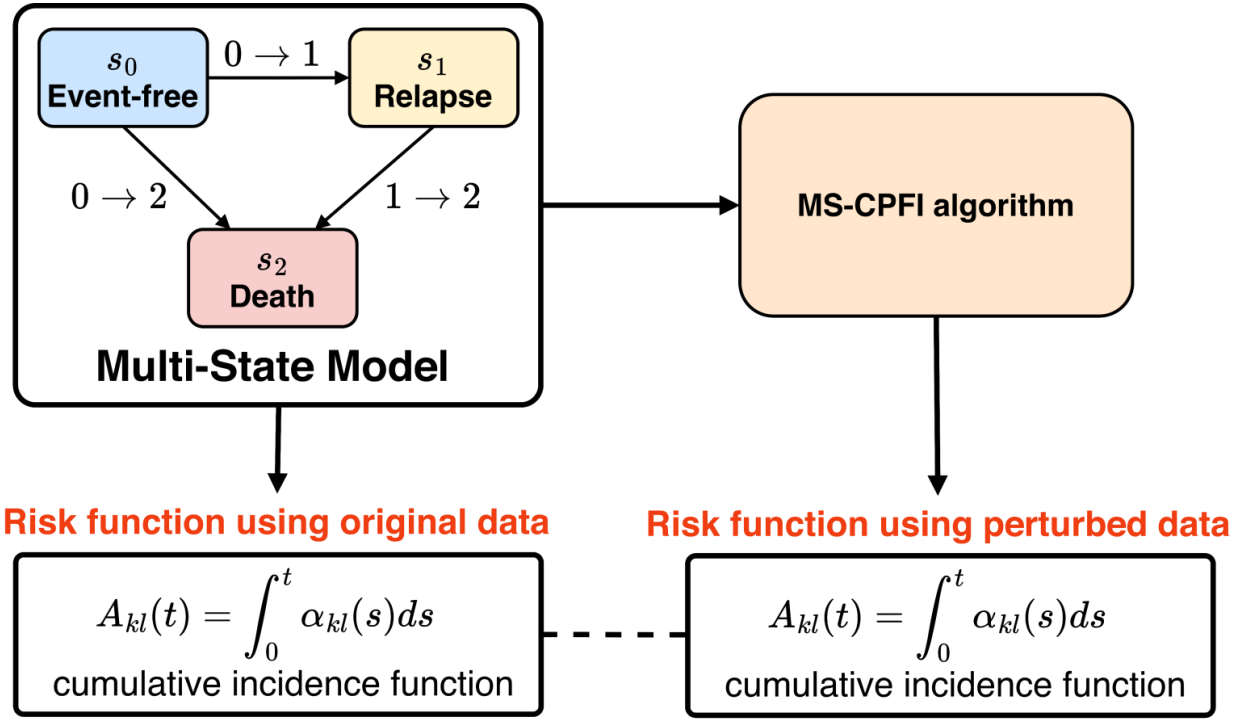
S1: Reference to MS-CPFI algorithm

- Multi-state models (especially competing risks) often use cumulative incidence functions (CIFs) to analyze how features influence the probability of each event.

Problem using CIFs: **reverse effect**

- A feature that ***increases*** risk of transition $0 \rightarrow 2$ can falsely appear to ***decrease*** the risk of $0 \rightarrow 1$
- The rise in one event's incidence necessarily lowers the apparent incidence of the competing event.

- Use conditional probabilities and cumulative hazard functions (CFHs) as reference.



S2: Counterfactual values for numerical features

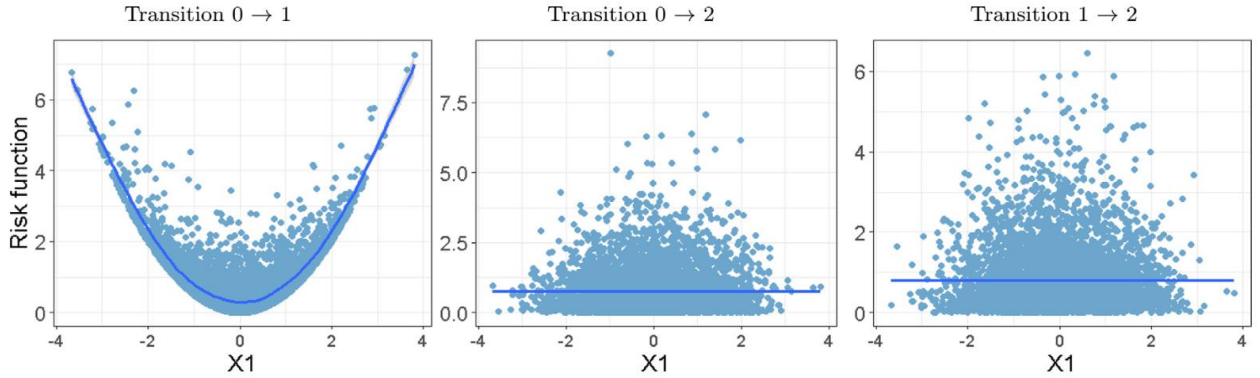
- Traditional approaches are exhaustive
  - all possible feature values
  - quantile values

- Choose counterfactual values using transition-specific univariate CoxPH with spline-based functions.

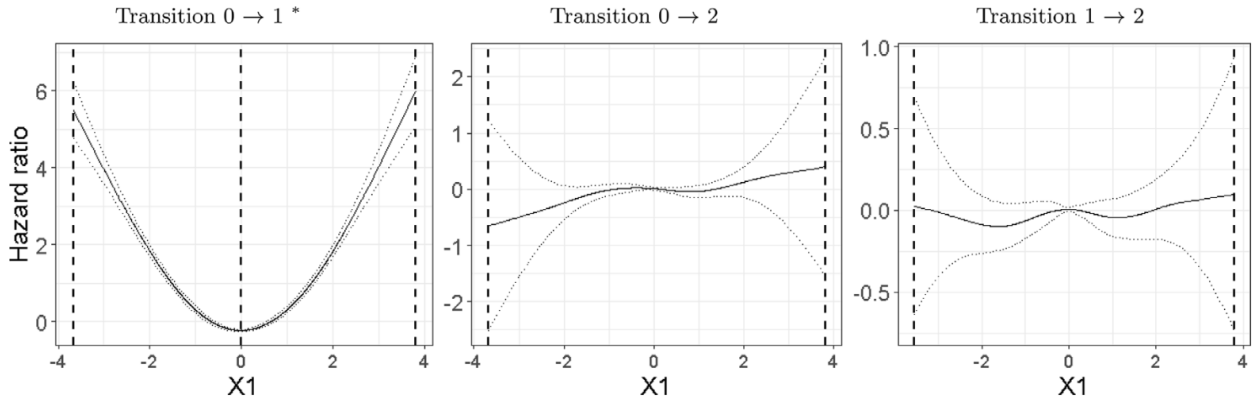
- Spline functions can capture non-linear relationships between the feature and the hazard.

- Plot the estimated hazard ratio as a function of the feature values.

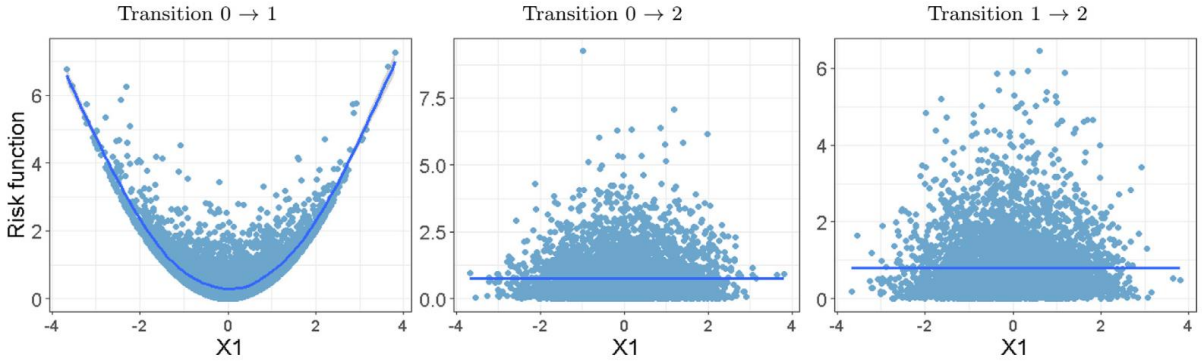
- Counterfactuals for numerical features:
  - extreme values (min or max)
  - inflection points

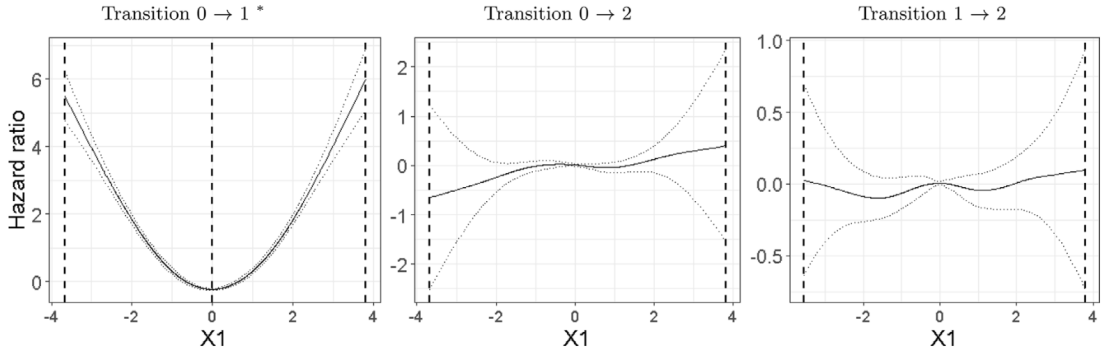


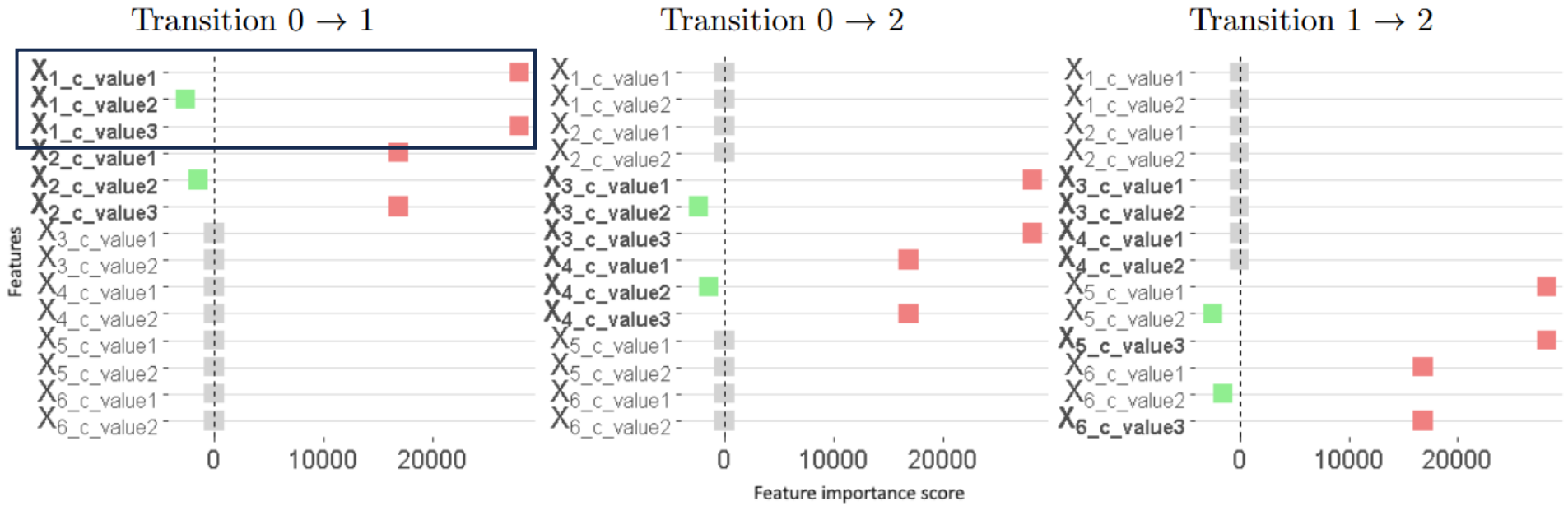
Counterfactuals (broken lines) for numerical feature x_1



S2: Counterfactual values for numerical features







S2: Counterfactual values for categorical features

- Demonstrates counterfactual perturbation for dummy-encoded categorical features.
  - MS-CPFI treats each category as a separate ***what if*** (counterfactual) scenario.

- Using BMI as categorical feature
  - dummy-encoded into binary features

Original data (dummy encoded BMI)

	low_BMI	*High_BMI*
$i = 1$	0	1
$i = 2$	1	0
$i = 3$	0	0
...	...	...
$i = N$	1	0

Counterfactual Perturbation

	low_BMI	*High_BMI*
$i = 1$	0	0
$i = 2$	0	0
$i = 3$	0	0
...	...	...
$i = N$	0	0

	low_BMI	*High_BMI*
$i = 1$	1	0
$i = 2$	1	0
$i = 3$	1	0
...	...	...
$i = N$	1	0

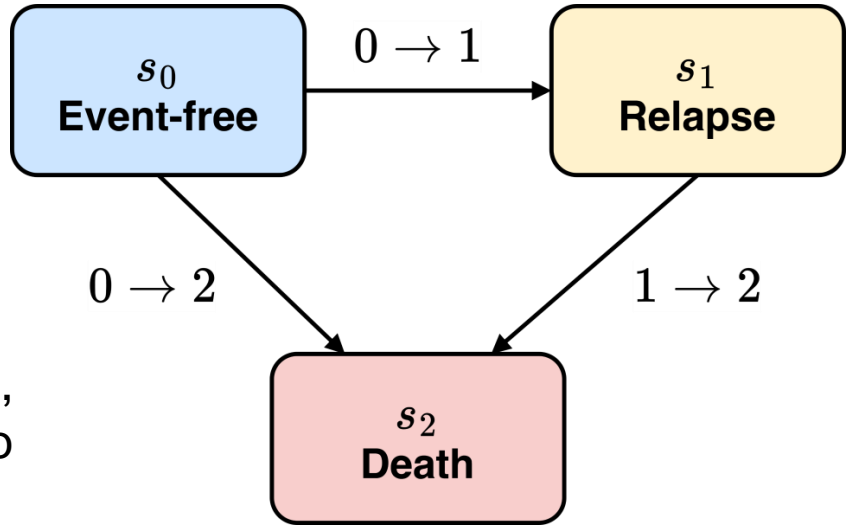
	low_BMI	*High_BMI*
$i = 1$	0	1
$i = 2$	0	1
$i = 3$	0	1
...	...	...
$i = N$	0	1

Each BMI category would be considered a counterfactual value.

METABRIC dataset

- Breast cancer progression using illness-death framework.
  - Utilized 1,903 patient data; followed-up for 30 years

$0 \rightarrow 1$	$0 \rightarrow 2$	$0 \rightarrow$ censored	$1 \rightarrow 2$	$1 \rightarrow$ censored	Total
677 (36%)	509 (27%)	717 (38%)	593 (88%)	84 (12%)	1903



- Relapse and death are the key transitions being analyzed, with a proportion of patients transitioning from relapse to death (88%).

- Transformed numerical features into categorical
  - 1 numerical; 14 categorical features

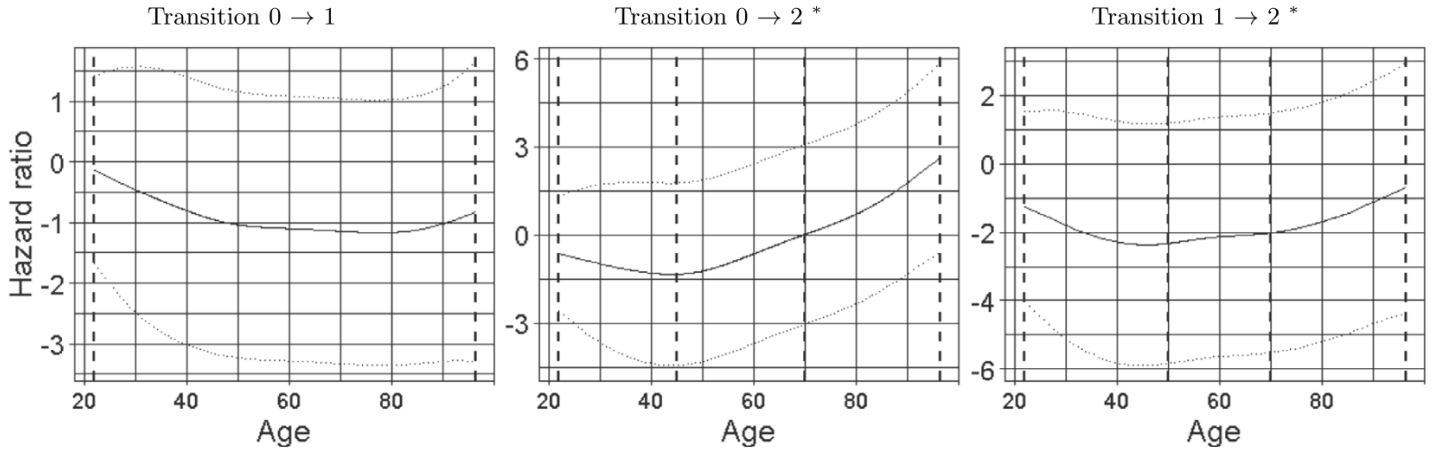
- Missing data imputation:
  - median (numerical); mode (categorical)

METABRIC dataset

Features	Modalities
Age at diagnosis	Numerical feature in years.
Inference on the menopausal status	post-menopausal, pre-menopausal.
Cancer histological type	others, IDC, IDC rare, IDC+ILC, ILC.
Cancer cellularity	low, high, moderate.
Density of receptor Her2 (human epidermal growth factor receptor 2) measured with SNP6-based measures	Neutral/Loss, Gain.
Her2 expression status	negative (Her2-), positive (Her2+).
Estrogen receptor status	negative (ER-), positive (ER+).
Progesterone receptor status	negative (PR-), positive (PR+).
Presence of a breast surgery	breast conserving, mastectomy.
Prior radiotherapy	No, Yes.
prior chemotherapy	No, Yes.
Prior hormonotherapy	No, Yes.
Nodes involvement classification (TNM)	N0: 0, N1: [1-3], N2: [4-9], N3: $\geq$10.
Tumor size classification (TNM)	T1:]0-20], T2:]20-50], T3: >50.
Cancer grade	Grade 1, Grade 2, Grade 3.

METABRIC: Counterfactual Perturbation on age

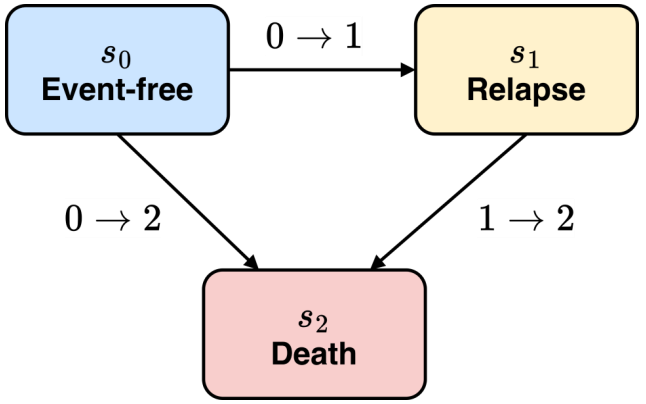
- **Age** is a crucial factor in cancer progression
- Determine the **counterfactual values in numerical features**
  - Used **spline-based univariate transition-specific Cox models**



Counterfactual values for every transition:

- Transition 0 → 1: Ages 22 and 96
- Transition 0 → 2: Ages 22, 45, 70, and 96
- Transition 1 → 2: Ages 22, 50, 70, and 96

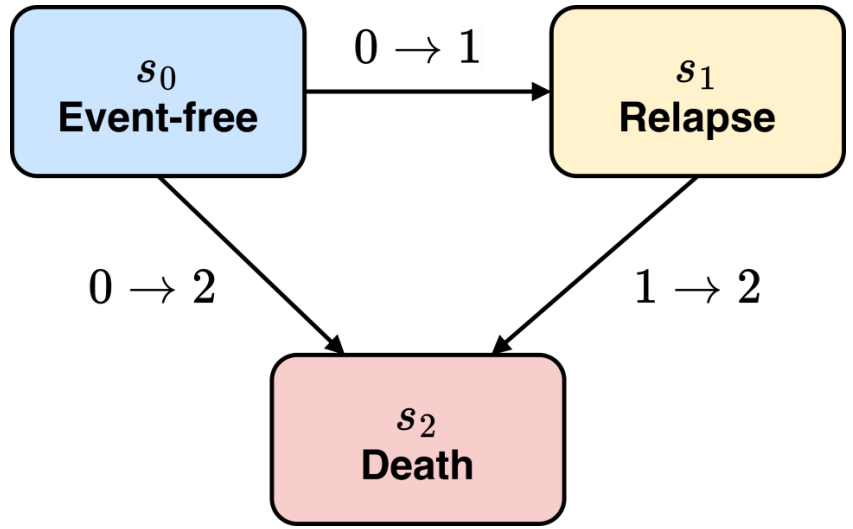
		original	counterfactual	
Patient ID	Transition	Tumor	c(T2)	c(T3)
1	0 → 1	45	22	96
1	0 → 2	45	22	96
1	1 → 2	45	22	96
2	0 → 1	60	22	96
2	0 → 2	60	22	96
2	1 → 2	60	22	96

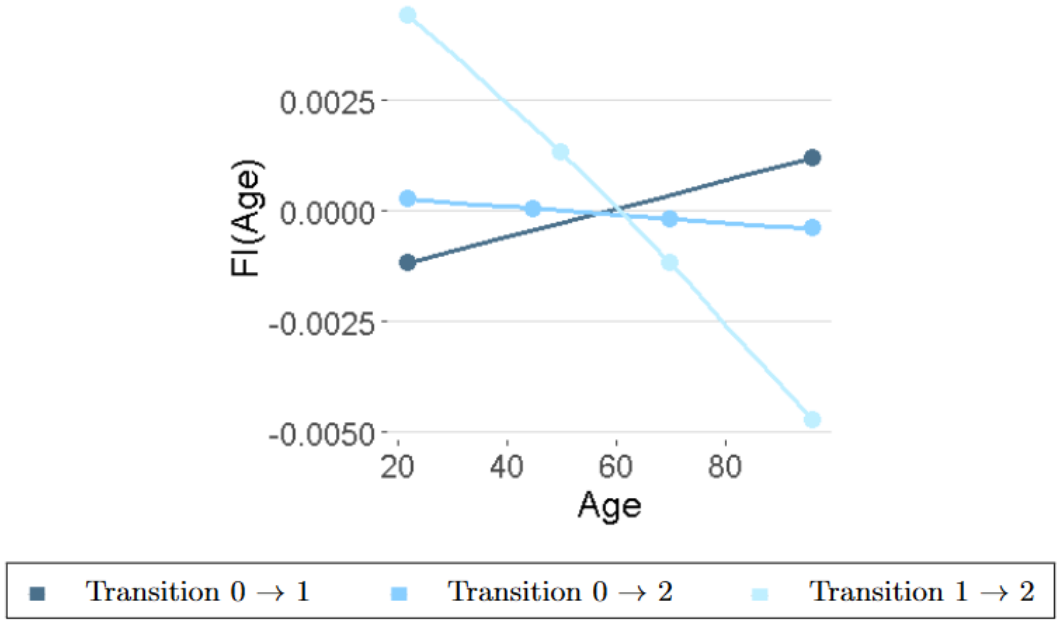


METABRIC: Importance of age

- Feature importance for age varies across transitions, showing different effects depending on the specific disease progression state.

- Transition 0 → 1
  - Older age increases the risk of relapse after being event-free.
- Transition 0 → 2
  - As age increases, the risk of death from non-cancer-related causes decreases.
- Transition 0 → 2
  - Older patients have an increased risk of death after relapse.

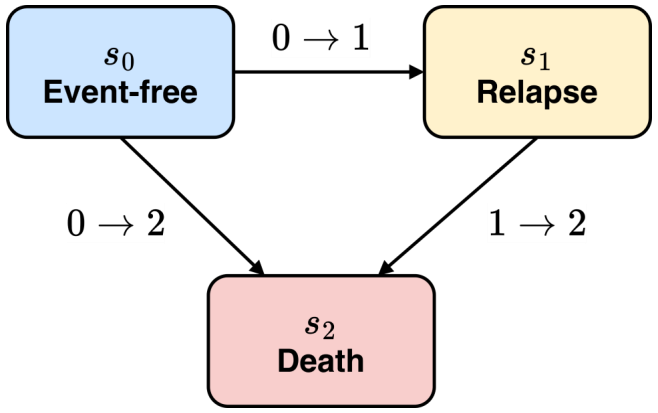




METABRIC: Counterfactual Perturbation on TNM

- TNM: Tumor size (T), nodes (N), and metastasis (M) are key features influencing progression.
- Perturb these features (T and N) to simulate disease progression and understand their effects.

- Counterfactual scenarios:
  - Each feature is altered to represent a hypothetical progression of the disease
    - e.g., changing tumor size from T1 to T3 (study did not show T4)
  - Tumor size and lymph node involvement are perturbed across different **counterfactual values**.



original data | counterfactual values

Patient ID	Transition	Tumor	Node	c(T2)	c(T3)	c(N1)	c(N2)	c(N3)
1	$0 \to 1$	T1	N0	T2	T3	N1	N2	N3
1	$0 \to 2$	T1	N0	T2	T3	N1	N2	N3
1	$1 \to 2$	T1	N0	T2	T3	N1	N2	N3
2	$0 \to 1$	T2	N1	T1	T3	N0	N2	N3
2	$0 \to 2$	T2	N1	T1	T3	N0	N2	N3
2	$1 \to 2$	T2	N1	T1	T3	N0	N2	N3
3	$0 \to 1$	T3	N2	T1	T2	N0	N1	N3
3	$0 \to 2$	T3	N2	T1	T2	N0	N1	N3
3	$1 \to 2$	T3	N2	T1	T2	N0	N1	N3

METABRIC: Importance of TNM classification

- MS-CPFI aligns with clinical understanding, confirming that perturbing TNM features offers meaningful insights into the likelihood of cancer progression.
  - Smaller tumor sizes (T1) and no lymph node involvement (N0) have a protective effect across all disease transitions.
  - Larger tumor sizes (T2, T3) and higher lymph node involvement (N2, N3) are risk factors, indicating higher probabilities of relapse and death.

Feature	Transition		
	0 → 1	**0 → 2**	**1 → 2**
Tumor size			
T1	Protective	Protective	Protective
T2	Risk	Risk	Risk
T3	Not Significant	Not Significant	Not Significant
Lymph node			
N0	Protective	Protective	Protective
N1	Not Significant	Not Significant	Not Significant
N2	Risk	Risk	Not Significant
N3	Risk	Risk	Not Significant

Comparison with CoxPH

- Compared the performance of MS-CPFI with CoxPH
- Both CoxPH and MS-CPFI consistently identify lymph node involvement and tumor size as Risk factors across all disease transitions.
- Hormone therapy is a protective factor for relapse, non-cancer death, and death after relapse.
- MS-CPFI can handle non-linear effects, but Cox is more constrained with linear assumptions, potentially oversimplifying relationship between features and transitions.

	CoxPH	MS-CPFI
Transition 0 → 1		
Histological Status	Risk factor	Risk factor
N Classification	Risk factor	Risk factor
T Classification	Risk factor	Risk factor
Hormone Therapy	Protective	Protective
Transition 0 → 2		
N Classification	Risk factor	Risk factor
T Classification	Risk factor	Risk factor
Her2-SNP6	Risk factor	Risk factor
Hormone Therapy	Protective	Protective
Transition 1 → 2		
N Classification	Risk factor	Risk factor
Hormone Therapy	Protective	Protective

Discussion

- First interpretability algorithm in the context of multi-state analysis
- 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

- By using a new counterfactual perturbation method and using risk predictions directly instead of model performance, the MSCPFI 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).

- Limitations
  - Choice of counterfactual scenarios for interpreting the effect of a numerical feature
  - Management of correlated data, which may introduce bias
  - MS-CPFI is currently designed to interpret multi-state models using tabular data
  - **No comparison of MS-CPFI to other explainability approaches (e.g., SHAP, LIME, or SurvLIME)**

- Future works
  - combine MS-CPFI with a local interpretability functionality to provide individual feature importance scores in a multi-state model.