



Inria



Prédiction d'événements récurrents en survie

Développement de méthode d'apprentissage et
application en oncologie

Juliette Murris

Sous la supervision de Pr. S. Katsahian (AP-HP) & A. Lavenu (IRMAR),
En collaboration Cifre avec les laboratoires Pierre Fabre



Outline

1. Motivating clinical data

Motivating clinical context Recurrent events analysis Objectives

2. Combining statistical inference and ensemble machine learning

The RecForest algorithm Performance evaluation Variable importance To wrap-up

3. Transparent use of survival ML algorithms

Rationale Interpretability for survival ML algorithms

4. Conclusion

01

Motivating clinical data





Digestive Cancer in France

Colorectal and
bowel cancer



Gastric and
oesophageal cancer



Hepatobiliary
cancer



Pancreatic cancer



Digestive Cancer in France

Colorectal and
bowel cancer



Gastric and
oesophageal cancer



Hepatobiliary
cancer



Pancreatic cancer



Key Facts

- ▶ Among the most frequent cancers,
affecting over 70,000 patients annually
- ▶ The second leading cause of cancer-related
deaths in France
- ▶ Surgery is the primary treatment strategy

Digestive Cancer in France

Colorectal and
bowel cancer



Gastric and
oesophageal cancer



Hepatobiliary
cancer



Pancreatic cancer



Key Facts

- ▶ Among the most frequent cancers, affecting over 70,000 patients annually
- ▶ The second leading cause of cancer-related deaths in France
- ▶ Surgery is the primary treatment strategy

Public Health Concerns

- ▶ What are the outcomes after the initial cancer surgery?
- ▶ What is the risk of complications or mortality post-surgery?
- ▶ Which factors contribute to readmissions?

Real-World Evidence in Healthcare

Traditional Research Data Sources

- ▶ **Randomized Controlled Trials (RCTs)** are the gold standard for generating healthcare evidence  Hariton & Locascio (2018)
- ▶ **Cohorts and Registries** enable prospective collection of data on specific populations  Porta (2014)

Real-World Data

- ▶ **Electronic Health Records (EHRs)** contain detailed patient histories, diagnoses, treatments, and outcomes  Gunter & Terry (2005)
- ▶ **Claims Databases** provide billing and reimbursement info, including diagnostic codes, procedures, and prescriptions  Cadarette & Wong (2015)

Real-World Evidence in Healthcare

Traditional Research Data Sources

- ▶ **Randomized Controlled Trials (RCTs)** are the gold standard for generating healthcare evidence  Hariton & Locascio (2018)
- ▶ **Cohorts and Registries** enable prospective collection of data on specific populations  Porta (2014)

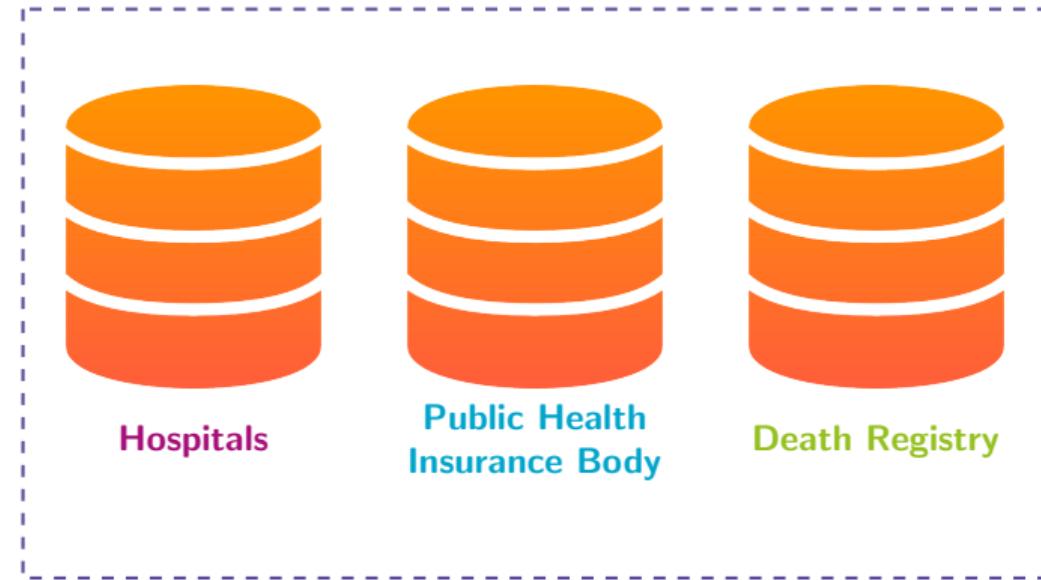
Real-World Data

- ▶ **Electronic Health Records (EHRs)** contain detailed patient histories, diagnoses, treatments, and outcomes  Gunter & Terry (2005)
- ▶ **Claims Databases** provide billing and reimbursement info, including diagnostic codes, procedures, and prescriptions  Cadarette & Wong (2015)



Claims Databases in France

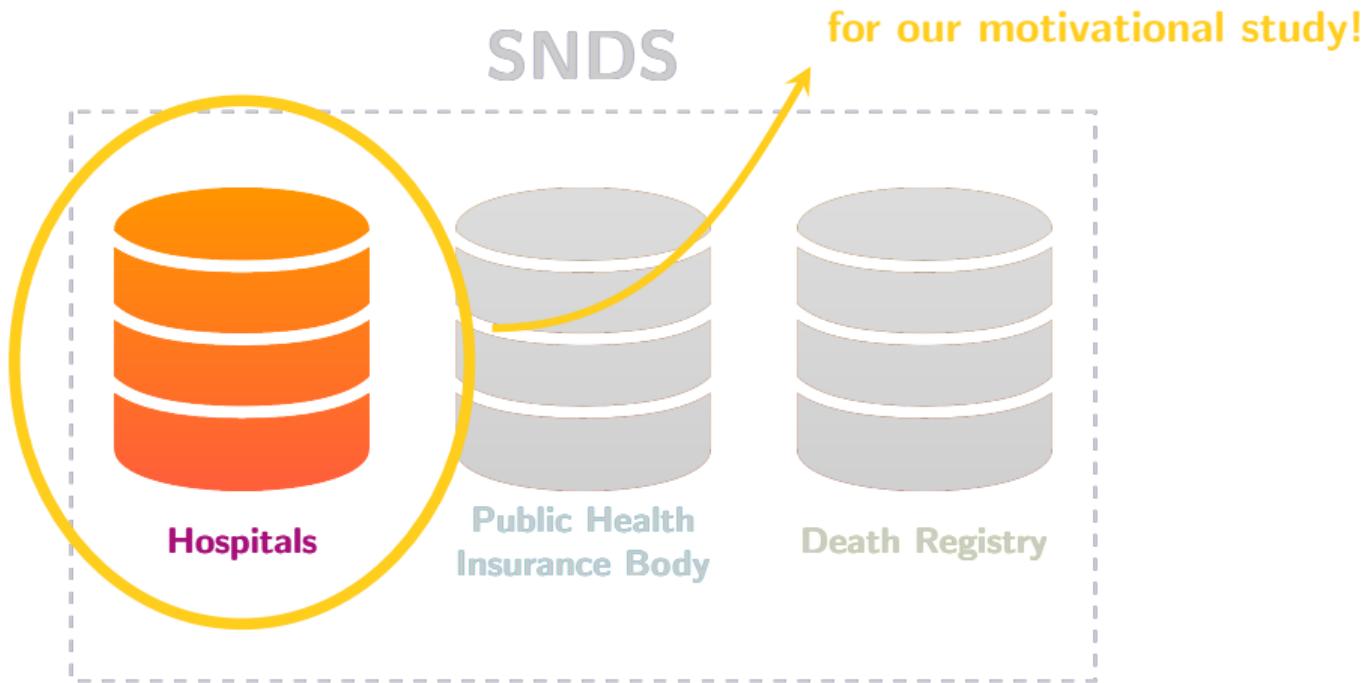
SNDS



SNDS: Système national des données de santé



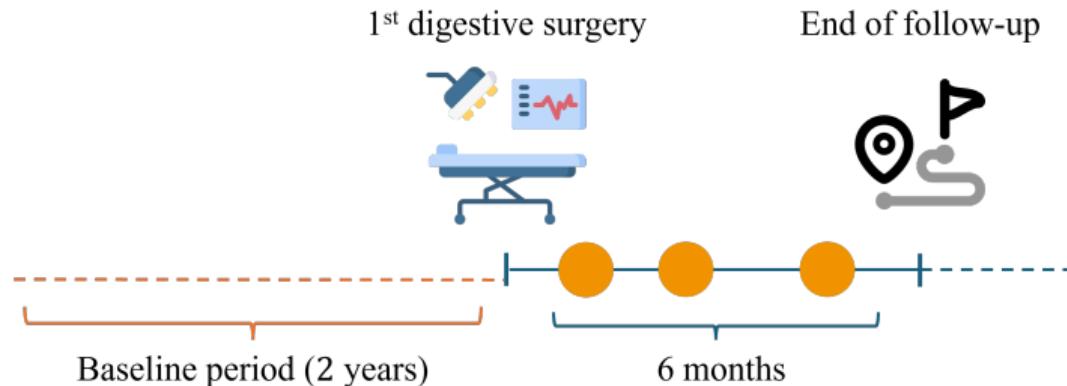
Claims Databases in France



SNDS: Système national des données de santé



Claims Databases in France – our motivational study

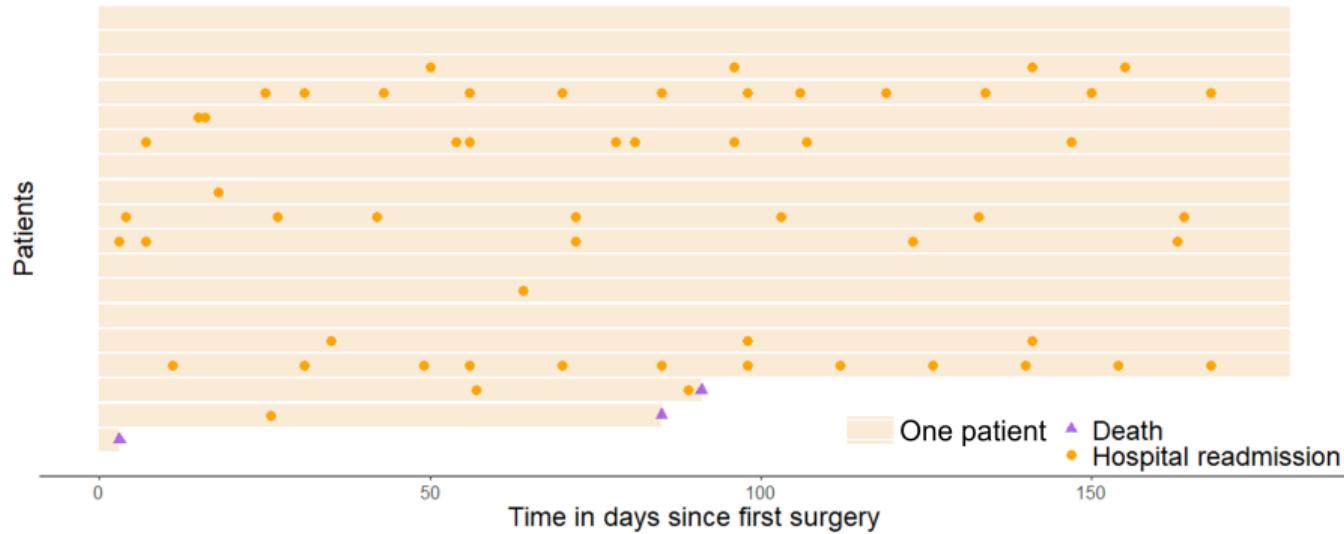


- ▶ Population – adult patients who have undergone digestive surgery (colorectal surgery, small bowel surgery, hepatobiliary surgery, pancreatic surgery, oesogastric surgery)
- ▶ Intervention – first digestive surgery between January 2020 and December 2022
- ▶ Comparator – Not for our study
- ▶ Outcome – Cumulative number of hospital readmissions over time in a 6-month window



What our data is made of

Subsample of 17/255,732 patients extracted





What our data is made of

Subsample of 17/255,732 patients extracted

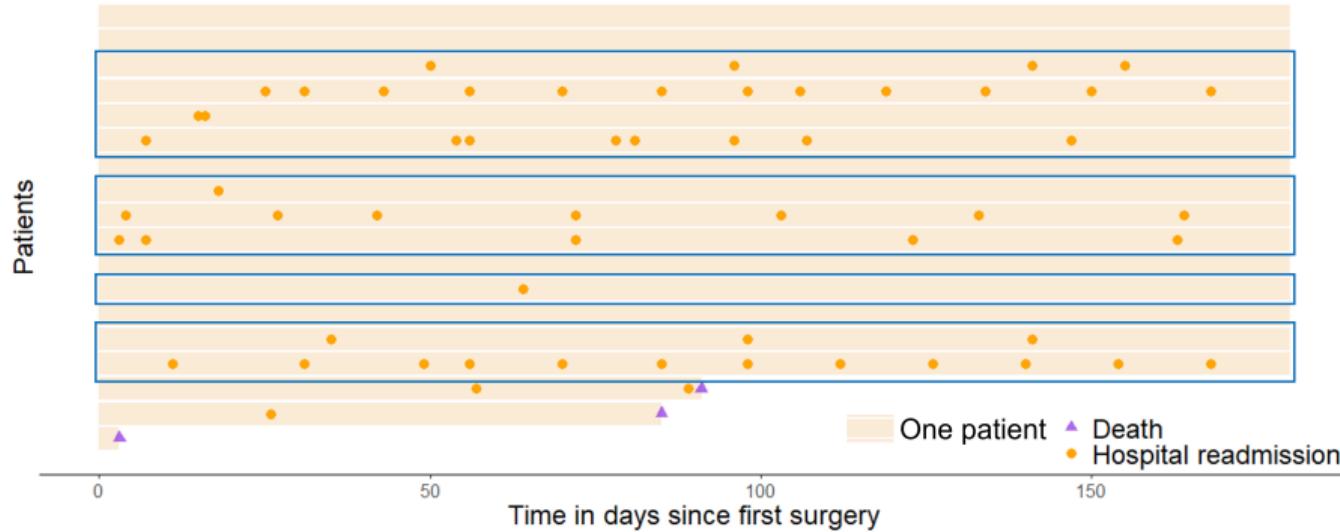


Patients with **no readmissions** over time



What our data is made of

Subsample of 17/255,732 patients extracted

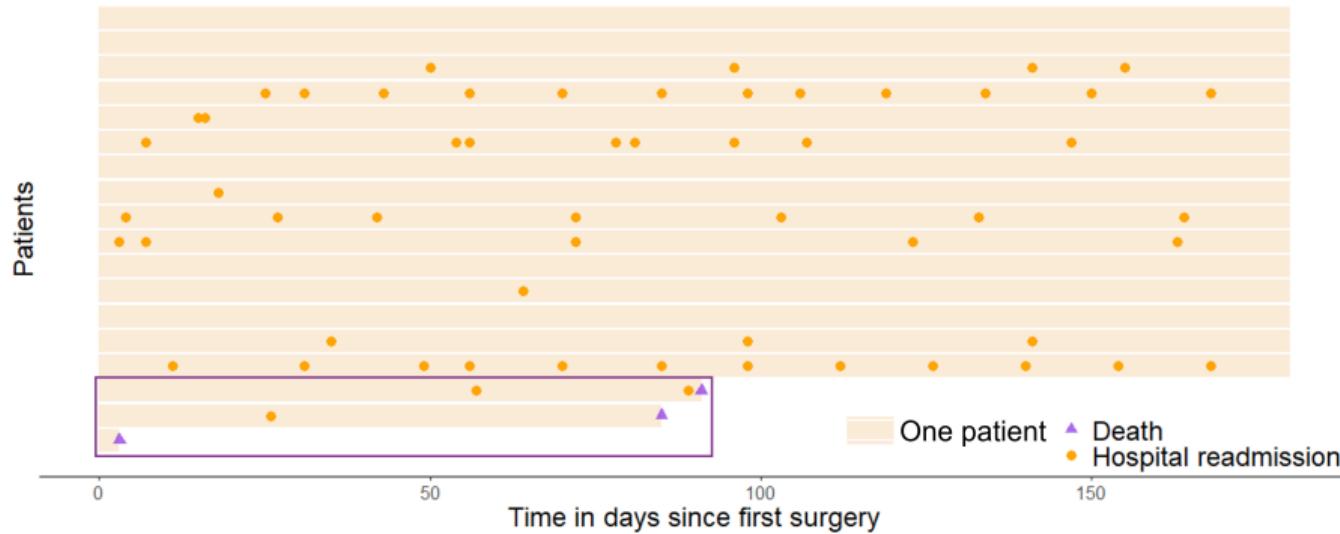


Patients with **one or more readmissions** over time



What our data is made of

Subsample of 17/255,732 patients extracted

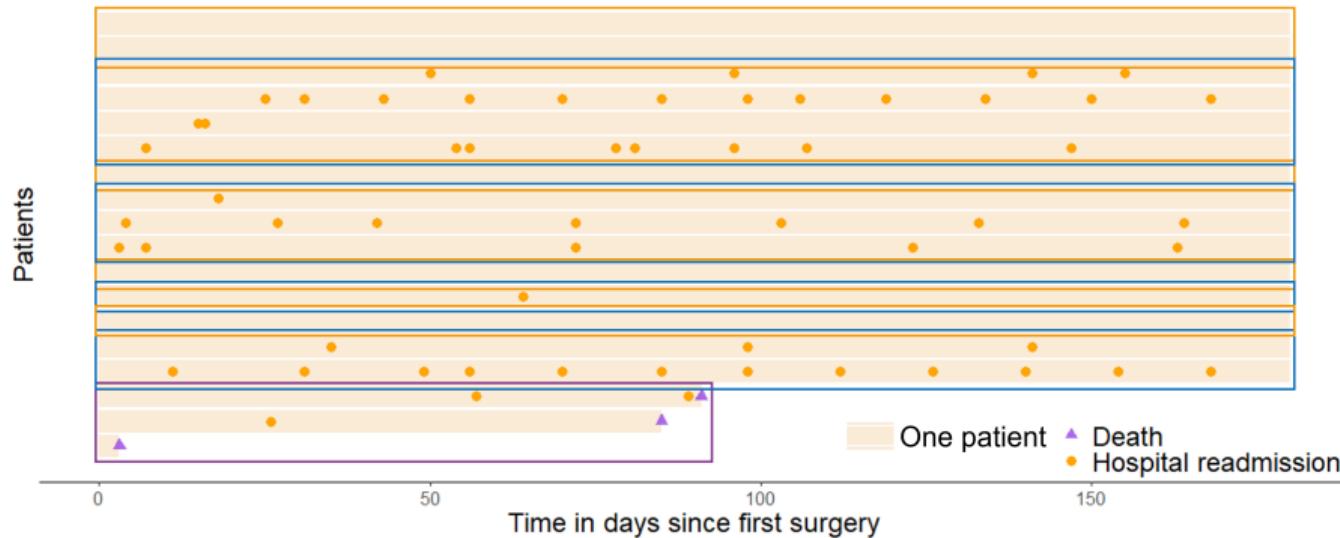


Patients who died during follow-up



What our data is made of

Subsample of 17/255,732 patients extracted



How to analyze multiple hospital readmissions over time for each patient?



What options do we have?

- ▶ Focus on the presence of at least one readmission?
- ▶ Focus on the number of readmissions at 6 months?
- ▶ Focus on time to first hospital readmission?



What options do we have?

- ▶ Focus on the presence of at least one readmission?
 - **Classification** problem, Solution: classifier, **No consideration of multiple events**
- ▶ Focus on the number of readmissions at 6 months?
- ▶ Focus on time to first hospital readmission?



What options do we have?

- ▶ Focus on the presence of at least one readmission?
 - **Classification** problem, Solution: classifier, **No consideration of multiple events**
- ▶ Focus on the number of readmissions at 6 months?
 - **Regression** problem, Solution: regressor, **No consideration of time**
- ▶ Focus on time to first hospital readmission?



What options do we have?

- ▶ Focus on the presence of at least one readmission?
 - **Classification** problem, Solution: classifier, **No consideration of multiple events**
- ▶ Focus on the number of readmissions at 6 months?
 - **Regression** problem, Solution: regressor, **No consideration of time**
- ▶ Focus on time to first hospital readmission?
 - **Survival** problem, Solution: Survival analysis, **No consideration of subsequent events**



What options do we have?

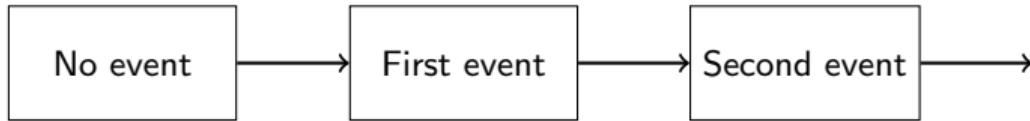
- ▶ Focus on the presence of at least one readmission?
 - **Classification** problem, Solution: classifier, **No consideration of multiple events**
- ▶ Focus on the number of readmissions at 6 months?
 - **Regression** problem, Solution: regressor, **No consideration of time**
- ▶ Focus on time to first hospital readmission?
 - **Survival** problem, Solution: Survival analysis, **No consideration of subsequent events**
- ▶ **Focus on time to recurrent readmission**
 - **Survival** problem, Solution: Survival analysis for **recurrent events**



Recurrent events

Definition

Stochastic processes that generate events of the same type repeatedly over time.

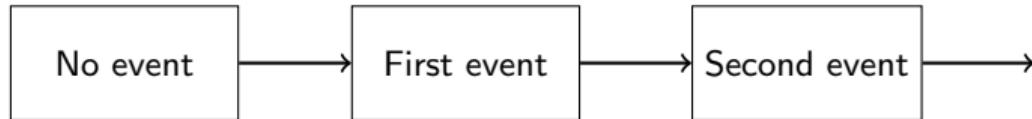




Recurrent events

Definition

Stochastic processes that generate events of the same type repeatedly over time.



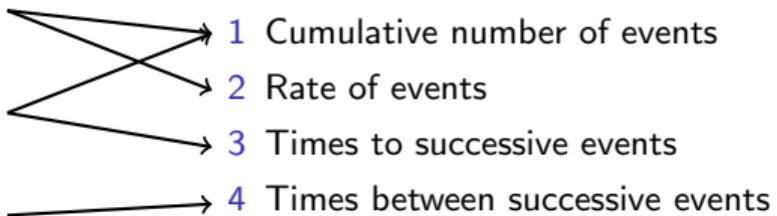
Censoring

When the **exact** time of an event is **not fully observed** for some subjects **within the study period**



Scientific questions for appropriate endpoints

- A Does the intervention decrease the event number over the study period?
- B How many events does the intervention prevent, on average?
- C What is the intervention effect on the number of subsequent events amongst patients with a preceding event?



ICH E9 (2019), Schmidli (2023), Wei (2023)



Non-parametric approach

The **Mean Cumulative Function** is the marginal expected number of events in $[0, t]$:

$$\mu(t) = \mathbb{E}[N(t)]$$



Non-parametric approach

The **Mean Cumulative Function** is the marginal expected number of events in $[0, t]$:

$$\mu(t) = \mathbb{E}[N(t)]$$

MCF Estimator:

$$\hat{\mu}(t) = \int_0^t d\hat{\mu}(u) du = \int_0^t \frac{\sum_{i=1}^n Y_i(t) dN_i(t)}{\sum_{i=1}^n Y_i(t)}$$

total number of events observed
over $[t, t + \Delta t)$

total number at risk
over $[t, t + \Delta t)$

 Cook & Lawless (1997)



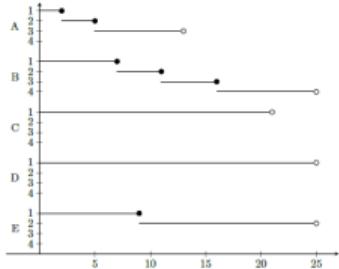
Modeling strategies

Modeling strategies

Conditional models

Andersen & Gill (1982), Prentice, Williams & Peterson (1981)

- ▶ Focus: Intensity – instantaneous probability of observing any event in a small time period $[t; t+]$
- ▶ Time scale: counting process

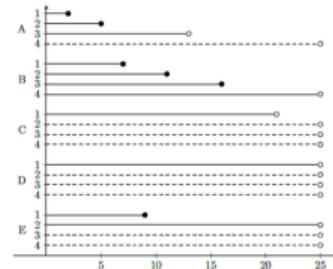


- ▶ Dependence structure between recurrent events by **full specification** of the recurrent event process

Marginal models

Wei, Lin & Weissfeld (1989), Lee, Wei & Amato (1992)

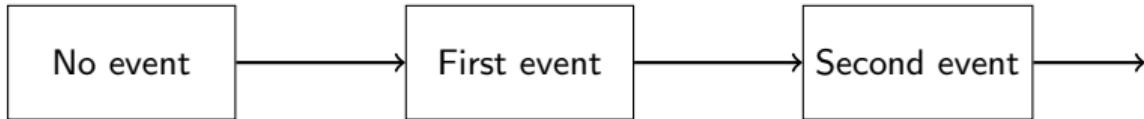
- ▶ Focus: Marginal features – marginal distribution of times to the first, second, third, ... event
- ▶ Time scale: total time



- ▶ Dependence structure between successive events may remain unspecified

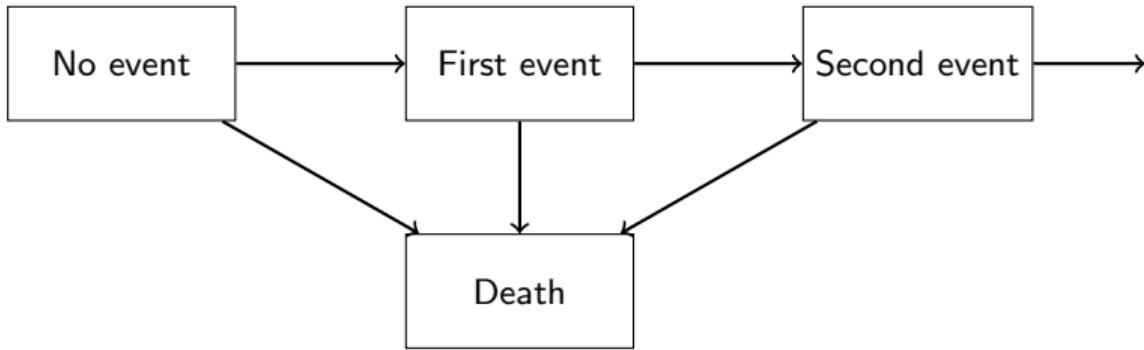


Non-informative censoring ?



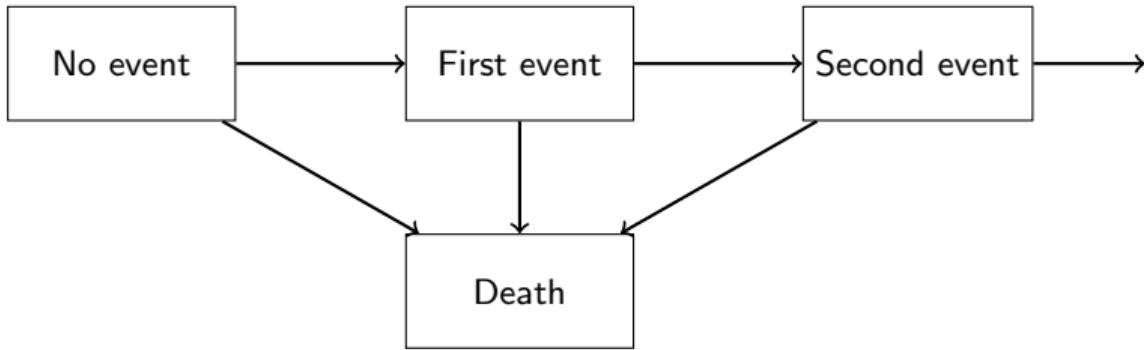


Non-informative censoring ?





Non-informative censoring ?



= with a **terminal** event



With a Terminal Event

MCF Non-parametric Estimator:

$$\hat{\mu}(t) = \int_0^t \hat{S}(u-) \frac{\sum_i Y_i(u) dN_i(u)}{\sum_i Y_i(u)}$$

↑
Kaplan-Meier estimator
of survival just before u

increment
at time u

Modeling:

$$\mu(t|X) = \begin{cases} \mu_0(t) \cdot \exp(\beta^T X) & \text{if } X \text{ is time-independent} \\ \int_0^t \exp(\beta^T X(s)) d\mu_0(s) & \text{if } X \text{ is time-dependent} \end{cases}$$

Ghosh & Lin (2000, 2002)

Raising questions – from the statistician's perspective

Key challenges

- ▶ How to manage situations with **high-dimensional data**?
- ▶ How to handle **multicollinearity** amongst variables?
- ▶ How to avoid **overfitting** and ensure reliable **generalization** to new data?

Raising questions – from the statistician's perspective

Key challenges

- ▶ How to manage situations with **high-dimensional data**?
- ▶ How to handle **multicollinearity** amongst variables?
- ▶ How to avoid **overfitting** and ensure reliable **generalization** to new data?

Current insights

Machine learning (ML) and survival counterparts

However, no ML algorithm *specifically designed* for recurrent events in a survival framework

 Murris (2023)



Raising questions – from the user's perspective

"Machine learning is frequently referred to as a black box – data goes in, decisions come out, but the processes between input and output are opaque."  [The Lancet editorial \(2018\)](#)

Clinician



Why a treatment is recommended for the patient at hand?

Researcher



Data-induced hypothesis
Discovery!

Patient



Informed consent
Life-style changes



Objectives

- 1 Sharpen recurrent events modeling with machine learning
- 2 Explore conditions for understanding survival machine learning

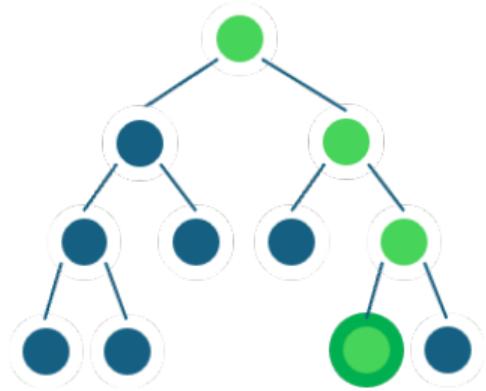
02

Combining statistical inference and ensemble machine learning





Growing Trees

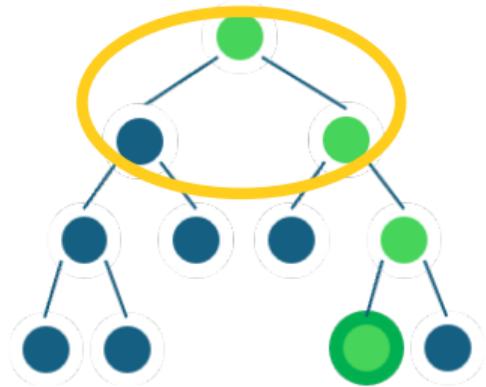


Key Components

Breiman (1996)



Growing Trees



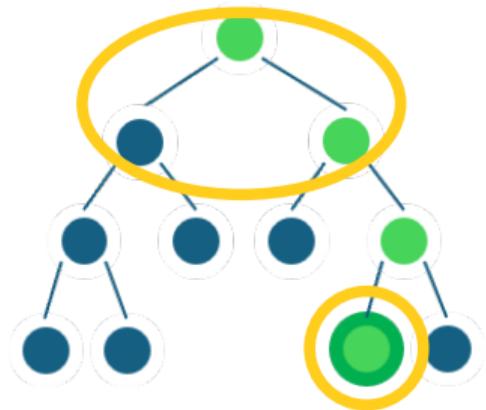
Key Components

- ▶ **Splitting Rule:** Identifies the optimal way to partition data at each node.

Breiman (1996)



Growing Trees



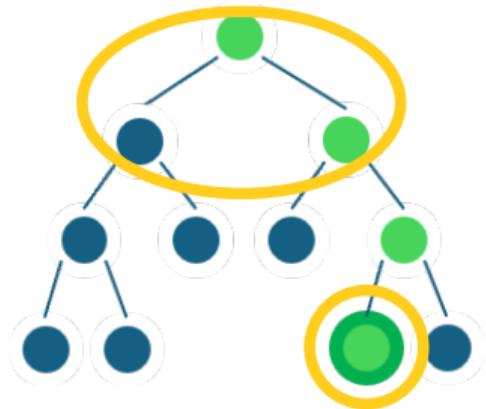
Key Components

- ▶ **Splitting Rule:** Identifies the optimal way to partition data at each node.
- ▶ **Terminal Node Estimator:** Selects the most suitable estimator to summarize final nodes.

Breiman (1996)



Growing Trees



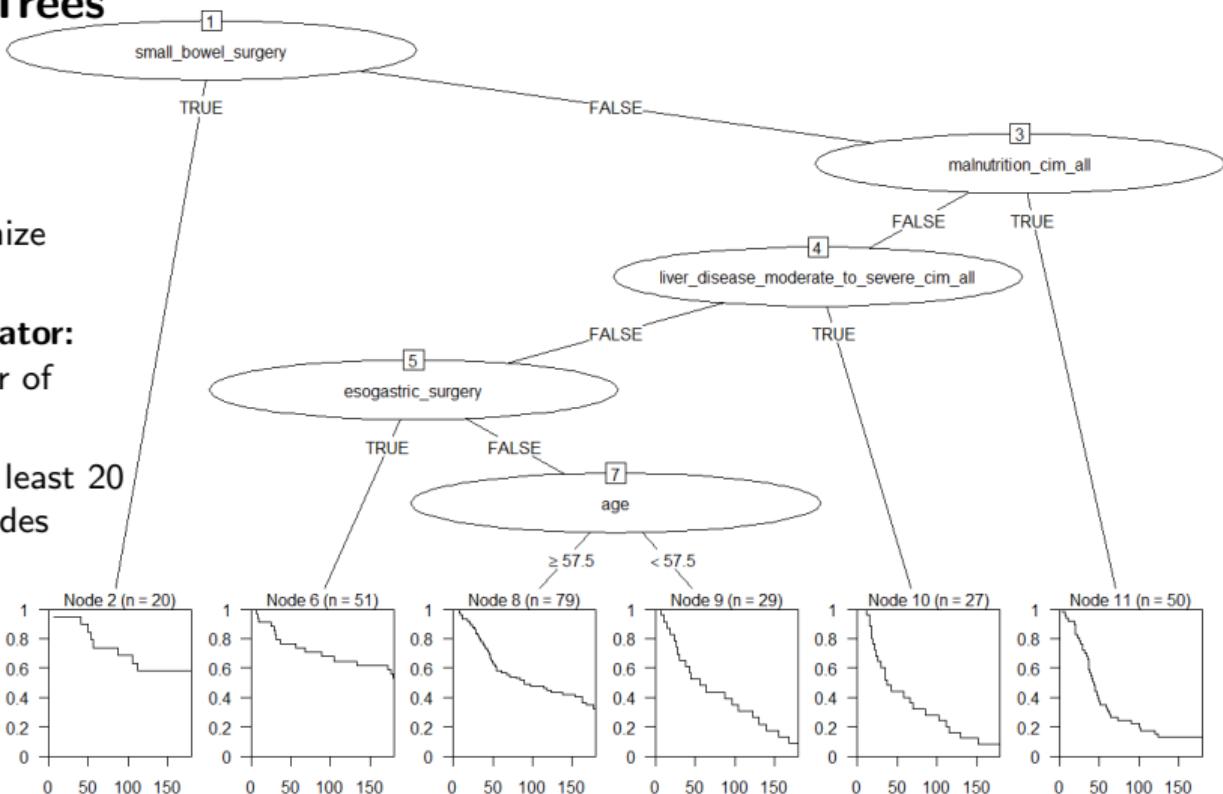
Key Components

- ▶ **Splitting Rule:** Identifies the optimal way to partition data at each node.
- ▶ **Terminal Node Estimator:** Selects the most suitable estimator to summarize final nodes.
- ▶ **Pruning Strategy:** Applies techniques to refine and simplify the tree structure.

Breiman (1996)

Growing Survival Trees

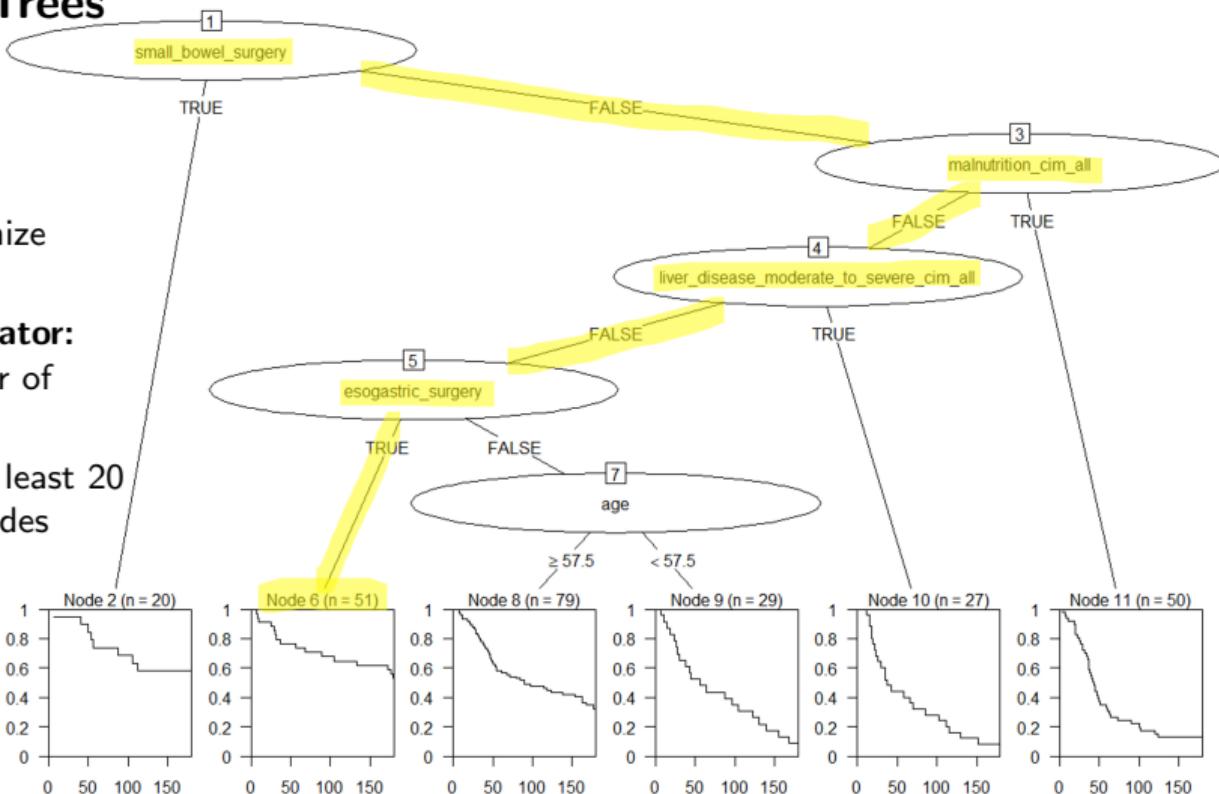
- ▶ **Splitting Rule:** Maximize Logrank test statistic
- ▶ **Terminal Node Estimator:** Kaplan-Meier estimator of survival function
- ▶ **Pruning Strategy:** At least 20 subjects in terminal nodes



Ishwaran (2008)

Growing Survival Trees

- ▶ **Splitting Rule:** Maximize Logrank test statistic
- ▶ **Terminal Node Estimator:** Kaplan-Meier estimator of survival function
- ▶ **Pruning Strategy:** At least 20 subjects in terminal nodes



Ishwaran (2008)



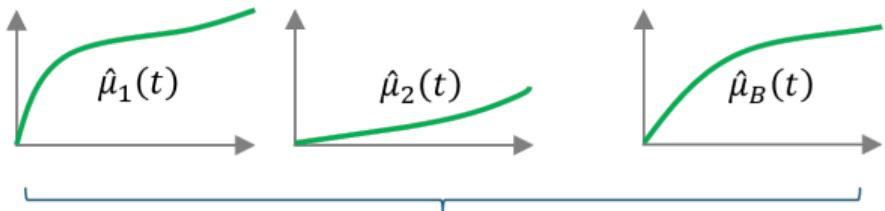
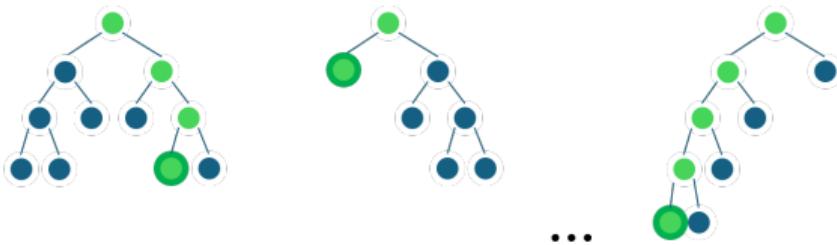
Growing Survival Trees with Recurrent Events

	Without a Terminal Event	With a Terminal Event
Splitting Rule	Maximize the Test Statistic	
At each node, $m \in \mathbb{N}$ predictors are randomly selected	Pseudo-score test	Wald test from Ghosh-Lin model
Terminal Node Estimator	MCF Estimator $\hat{\mu}_b(t \mathbf{x})$	
For tree b	$\int_0^t \frac{dN_b(u)}{Y_b(u)}$	$\int_0^t \hat{S}_b(u) \frac{\sum_i Y_{b,i}(u) dN_{b,i}(u)}{\sum_i Y_{b,i}(u)}$
Pruning Strategy	A Minimal Number of Events and/or Individuals	

Murris (2024)



Aggregating to build random forests – RecForest

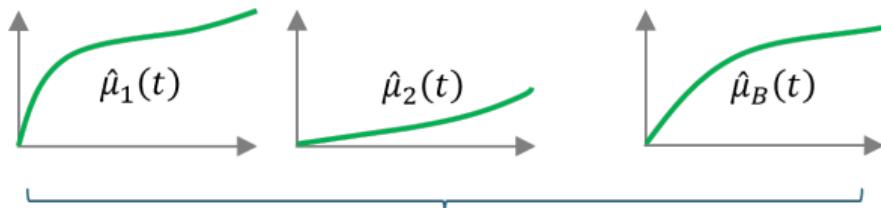
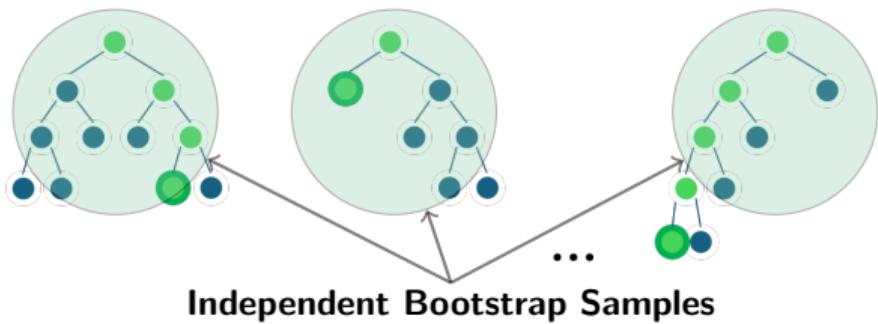


$$\hat{M}(t|\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{\mu}_b(t|\mathbf{x})$$





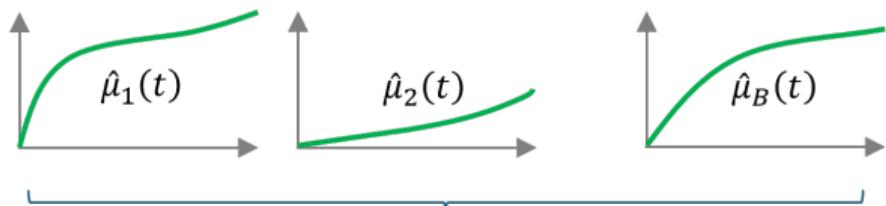
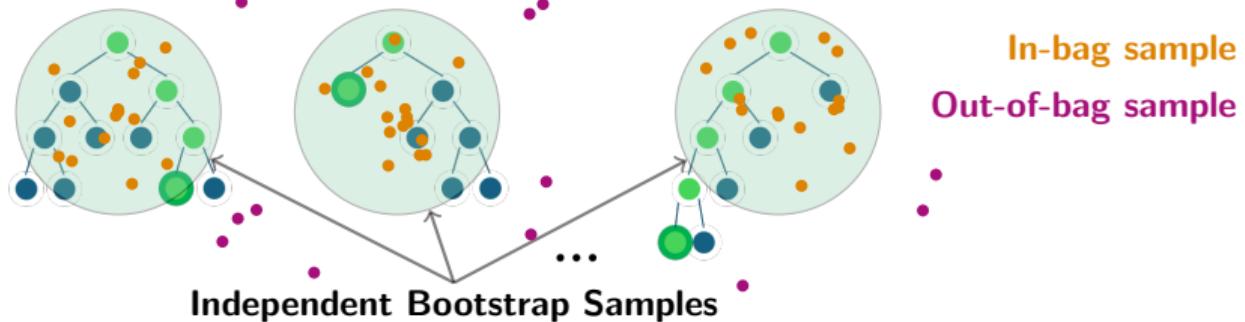
Aggregating to build random forests – RecForest



$$\hat{M}(t|\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{\mu}_b(t|\mathbf{x})$$



Aggregating to build random forests – RecForest



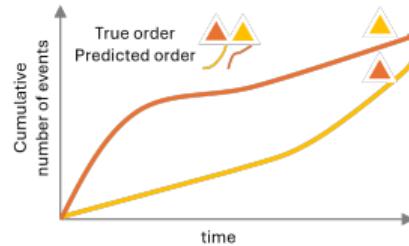
$$\hat{M}(t|\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{\mu}_b(t|\mathbf{x})$$





Performance evaluation – (a) The concordance index

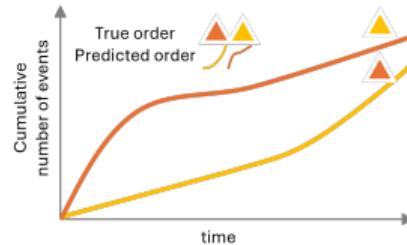
- ▶ C-index widely used as a performance metric
 Harrell (1982)
- ▶ Extension needed to take into account subsequent event occurrences  Kim (2018)





Performance evaluation – (a) The concordance index

- ▶ C-index widely used as a performance metric
☞ Harrell (1982)
- ▶ Extension needed to take into account subsequent event occurrences ☞ Kim (2018)



New C-index based on event occurrence rate ☞ Murris (2024)

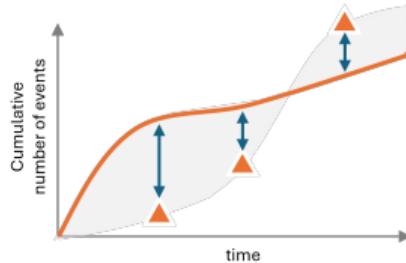
$$\hat{C}_{\text{rec}} = \frac{\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{r_i > r_j} \times \mathbb{1}_{\hat{r}_i > \hat{r}_j}}{\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{r_i > r_j}}$$

with $r_i = \frac{N_i(T_i)}{T_i}$ and $\hat{r}_i = \frac{\hat{\mu}(T_i | \mathbf{x}_i)}{T_i}$ the observed and predicted event occurrence rates, respectively.



Performance evaluation – (b) The mean square error

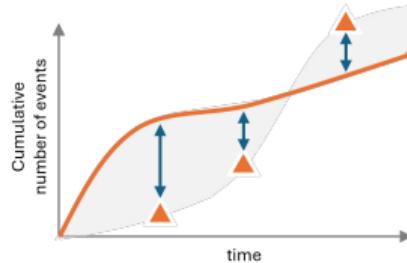
- ▶ No MSE metric for recurrent events until very lately [Bouaziz \(2024\)](#)
- ▶ We adapted it for an ensemble framework





Performance evaluation – (b) The mean square error

- ▶ No MSE metric for recurrent events until very lately [Bouaziz \(2024\)](#)
- ▶ We adapted it for an ensemble framework



For each tree b ,

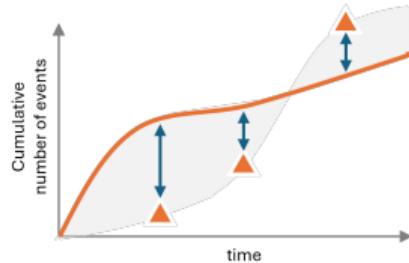
$$\widehat{MSE}_b(t, \hat{\mu}_b) = \frac{1}{n} \sum_{i=1}^n \left(\int_0^t \frac{dN_i(u)}{\hat{G}_c(u|\mathbf{x})} - \hat{\mu}_b(t|\mathbf{x}) \right)^2$$

Where $\hat{G}_c(u|\mathbf{x}) = 1 - \hat{G}(u - |\mathbf{x})$ is an estimator of $G_c(u|\mathbf{x}) = 1 - G(u - |\mathbf{x})$, the conditional cumulative distribution function of the censoring variable C given \mathbf{x} .



Performance evaluation – (b) The mean square error

- ▶ No MSE metric for recurrent events until very lately [Bouaziz \(2024\)](#)
- ▶ We adapted it for an ensemble framework



For each tree b ,

$$\widehat{MSE}_b(t, \hat{\mu}_b) = \frac{1}{n} \sum_{i=1}^n \left(\int_0^t \frac{dN_i(u)}{\hat{G}_c(u|\mathbf{x})} - \hat{\mu}_b(t|\mathbf{x}) \right)^2$$

Where $\hat{G}_c(u|\mathbf{x}) = 1 - \hat{G}(u - |\mathbf{x})$ is an estimator of $G_c(u|\mathbf{x}) = 1 - G(u - |\mathbf{x})$, the conditional cumulative distribution function of the censoring variable C given \mathbf{x} .

Therefore:

$$\widehat{MSE}(t, \hat{M}) = \frac{1}{B} \sum_{b=1}^B \widehat{MSE}_b(t, \hat{\mu}_b)$$



Performance evaluation – (c) The score

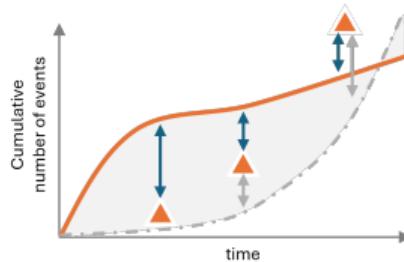
But 



Performance evaluation – (c) The score

But 

Two different models may lead to similar MSE values over time.

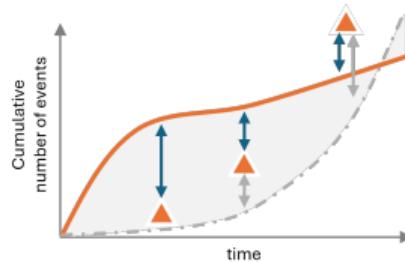




Performance evaluation – (c) The score

But 

Two different models may lead to similar MSE values over time.



Need for a score to represent the prediction gain compared to a reference estimator $\hat{\mu}_0$ and we define for each tree b

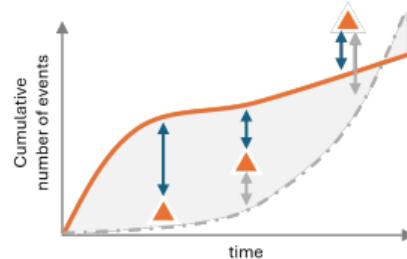
$$Score_b(t, \hat{\mu}_b, \hat{\mu}_{b,0}) = \widehat{MSE}_b(t, \hat{\mu}_{b,0}) - \widehat{MSE}_b(t, \hat{\mu}_b)$$



Performance evaluation – (c) The score

But

Two different models may lead to similar MSE values over time.



Need for a score to represent the prediction gain compared to a reference estimator $\hat{\mu}_0$ and we define for each tree b

$$Score_b(t, \hat{\mu}_b, \hat{\mu}_{b,0}) = \widehat{MSE}_b(t, \hat{\mu}_{b,0}) - \widehat{MSE}_b(t, \hat{\mu}_b)$$

Therefore:

$$Score(t, \hat{M}) = \frac{1}{B} \sum_{b=1}^B Score_b(t, \hat{\mu}_b, \hat{\mu}_{b,0})$$



Performance evaluation – Integrated counterparts

But 



Performance evaluation – Integrated counterparts

But 

There is a need for the estimation of the expectation of single-time MSE and derived score over time (e.g. hyperparameter tuning, generalized metric, etc.)



Performance evaluation – Integrated counterparts

But 

There is a need for the estimation of the expectation of single-time MSE and derived score over time (e.g. hyperparameter tuning, generalized metric, etc.)

$$\begin{cases} \widehat{IMSE}(\tau_1, \tau_2, \hat{M}) &= \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} \widehat{MSE}(t, \hat{M}) dt \\ \widehat{IScore}(\tau_1, \tau_2, \hat{M}) &= \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} Score(t, \hat{M}) dt \end{cases}$$

With $\tau_1 = 0$ and τ_2 the maximum event time on the original sample.



Importance of Variables

Input: Trained model \hat{f} , variable matrix X , target vector y

1. Estimate the original model error err_{OOB} from a chosen evaluation metric
2. For each feature $j \in \{1, \dots, p\}$ do:
 - Generate feature matrix \hat{X}^{perm} by permuting feature j in the data X 
 - Estimate error $\widehat{\text{err}}_{OOB}^{X^{\text{perm}}}$ based on the predictions of the permuted data
 - Calculate permutation variable importance over B trees as:

*This breaks the association
between j and y*

$$VImp(j) = \frac{1}{B} \sum_{b=1}^B (\widehat{\text{err}}_{OOB}^{X^{\text{perm}}} - \text{err}_{OOB})$$

Output: Importance scores for all variables

Application to French Digestive Cancer Data

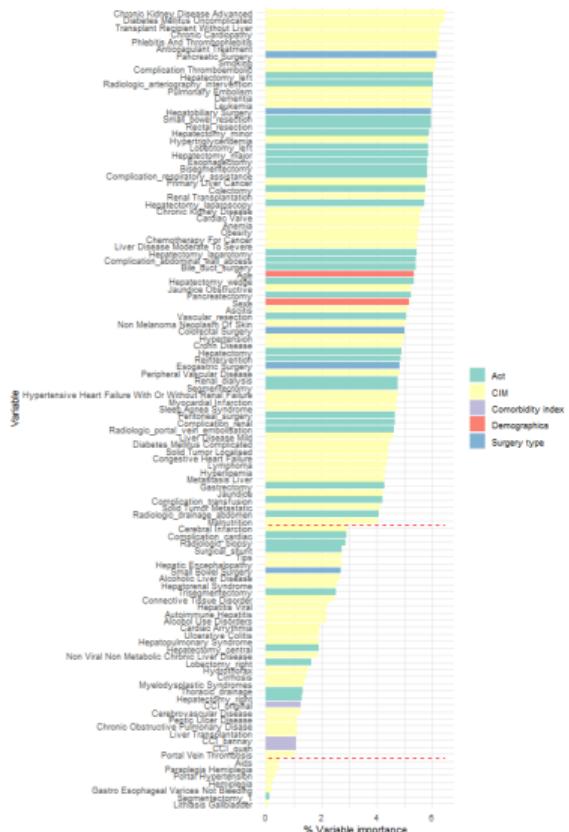
Table: Performances

	C-index ↑	IMSE ↓	IScore ↑
RecForest	0.72	1,398.04	409.32
Np estimator	0.52	3,773.21	ref.

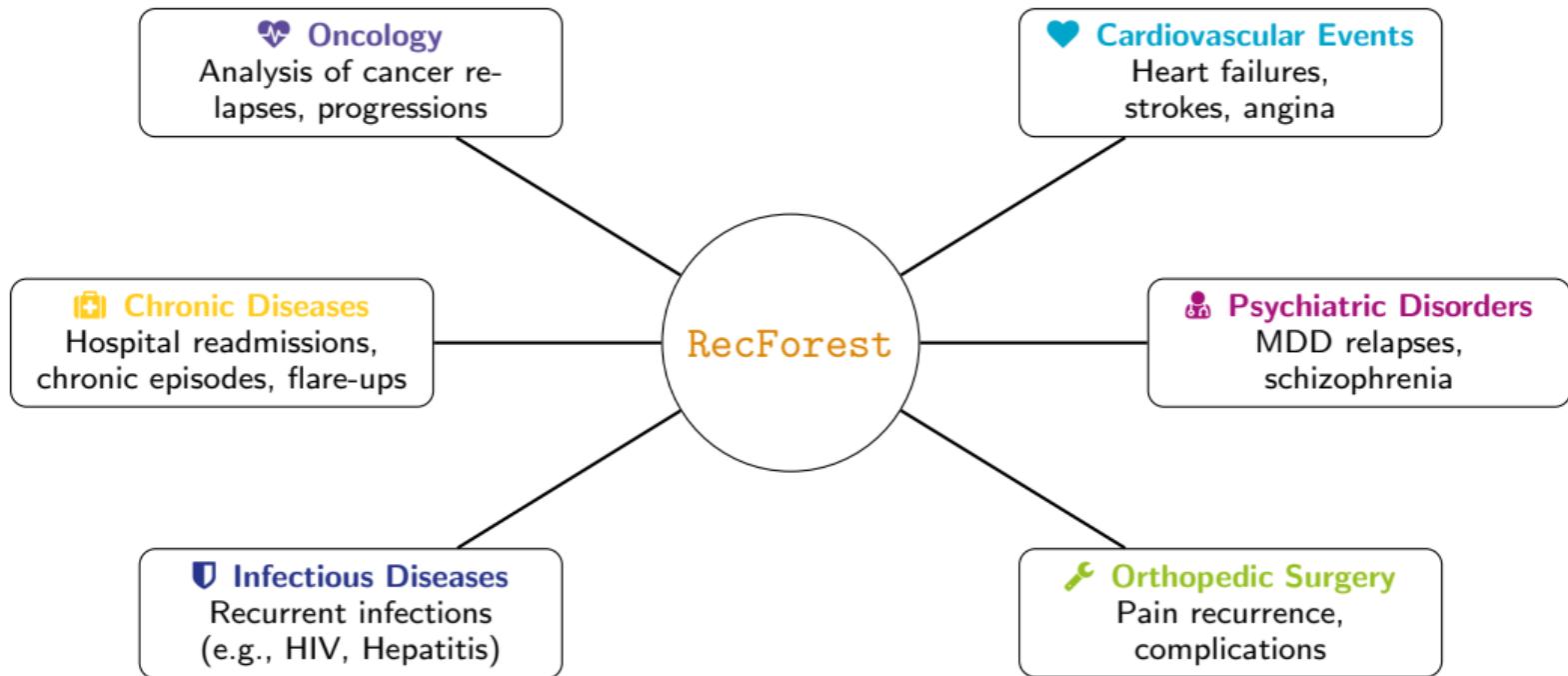
Importance of Variables

Demographics, ICD-10 codes, Procedures,
Comorbidity indices, Surgery types

- **Most important:** $\% V_{\text{imp}} \geq 4\%$
- **Moderately important:** $1\% \leq \% V_{\text{imp}} < 4\%$
- **Least important:** $\% V_{\text{imp}} < 1\%$



Multiple Medical Applications of RecForest





To Wrap-Up – Key Takeaways

RecForest

- ✓ Non-Parametric when no terminal event
- ✓ High-Dimensional Data
- ✓ Robust to Multicollinearity
- ✓ Variable Importance

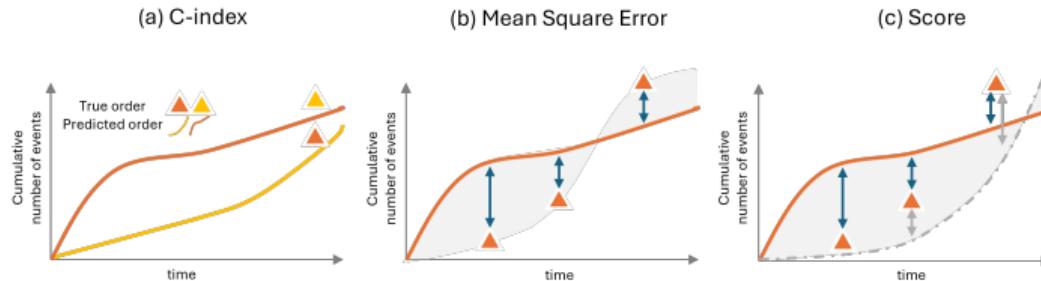
To Wrap-Up – Key Takeaways

RecForest

- ✓ Non-Parametric when no terminal event
- ✓ High-Dimensional Data

- ✓ Robust to Multicollinearity
- ✓ Variable Importance

3 metrics for performance evaluation

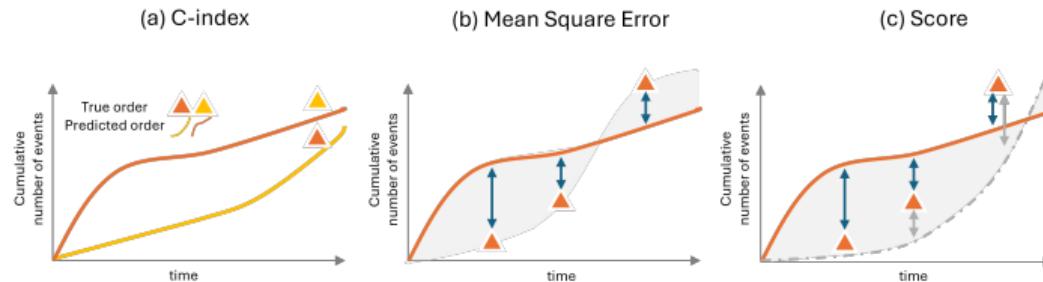


To Wrap-Up – Key Takeaways

RecForest

- ✓ Non-Parametric when no terminal event
- ✓ High-Dimensional Data
- ✓ Robust to Multicollinearity
- ✓ Variable Importance

3 metrics for performance evaluation



- A powerful and flexible tool for recurrent events analysis in many medical fields
- Allows for potential extensions, e.g. tree-based boosting techniques

03

Transparent use of survival ML algorithms



Why?



Users need

- ▶ **Transparency:** Better understand *how* the model makes predictions
- ▶ **AI-based decision-making risk understanding:** *Assessment, management and quantification*
- ▶ **Bias handling:** Ensure the model doesn't learn *unintended biases*
- ▶ **Scientific discovery:** Gain *insights* and uncover *new knowledge* from the model

✉ Liao (2020), Markus (2021), Farah & Murris (2023)



What?

Model explainability

- ▶ Intrinsically understandable model
- OR
- ▶ Non-understandable model complemented with understandable and faithful explanations

Explanation interpretability

- ▶ Unambiguous explanation
- AND
- ▶ Avoid cognitive overload to foster understanding

Markus (2021)



What?

Model explainability

- ▶ Intrinsically understandable model

OR

- ▶ Non-understandable model complemented with understandable and faithful explanations

Explanation interpretability

- ▶ Unambiguous explanation

AND

- ▶ Avoid cognitive overload to foster understanding

Interpretability methods

✉ Markus (2021)



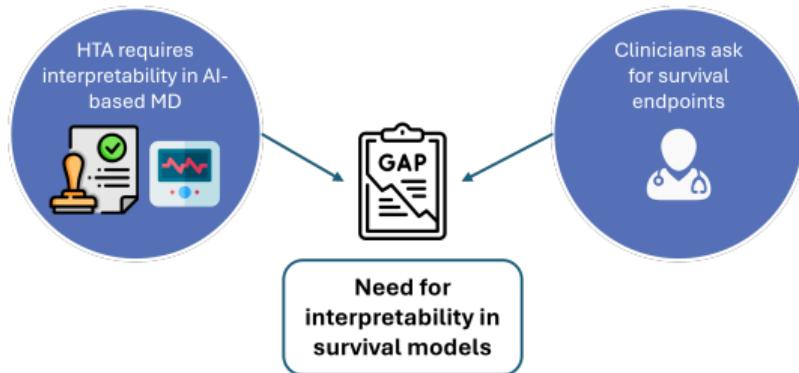
How?

Interpretability methods from the literature

- ▶ **Global feature importance:** Identifies impactful features at the population scale
- ▶ **Local feature importance:** Focuses on features at the most granular scale
- ▶ **Model-agnostic method:** Applicable regardless of the assessed model
- ▶ **Post-hoc method:** Applied on top of model inference

 Guidotti (2019), Miller (2019), Ali (2023)

Gap identified in survival problems



HTA: Health Technology Assessment; MD: Medical Device.

Widely adopted interpretability methods:

- ▶ LIME and SHAP
- ▶ Over 6,000 citations in PubMed.

However, for survival problems:

- ▶ SurvLIME and SurvSHAP are natural extensions
- ▶ Only 4 citations in PubMed*.

*While '*survival machine learning*' gets 10,745 hits since 2020

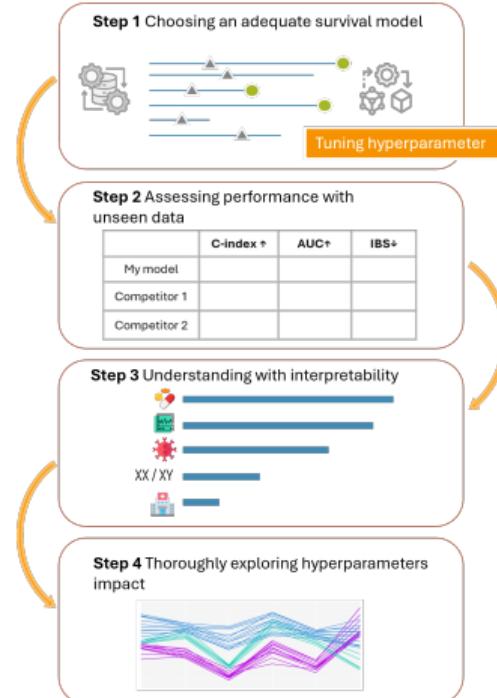
❑ Ribeiro (2016), Lundberg & Lee (2017), Kovalev (2020), Krzyński (2023)



A Comprehensive Tutorial for Interpretability in Survival ML

Our 4-Step Approach

- 1. Model Selection and Tuning:** Choose and fine-tune a survival model adapted to data characteristics
- 2. Model Evaluation:** Ensure accuracy and reliability with metrics such as C-index, AUC, and IBS
- 3. Interpretability Methods:** Explore advanced tools like **SurvLIME** and **SurvSHAP**
- 4. Impact of Hyperparameters:** Demonstrate how hyperparameters influence model efficiency and complexity



Murris (2024b)

04 Conclusion





To wrap-up

- 1 Sharpen recurrent events modeling with machine learning
- 2 Explore conditions for understanding survival machine learning



To wrap-up

1 Sharpen recurrent events modeling with machine learning

- ▶ Identified a gap in handling both recurrent event data and statistical learning Murris (2023)
- ▶ Developed **RecForest**: An extension of the Random Survival Forests algorithm to handle recurrent events, *with or without a terminal event* Murris (2024)
 - Refine the splitting rule and terminal node estimation
 - Provide appropriate metrics and error evaluation methods
 - Adjust variable importance calculations accordingly

2 Explore conditions for understanding survival machine learning



To wrap-up

1 Sharpen recurrent events modeling with machine learning

- ▶ Identified a gap in handling both recurrent event data and statistical learning Murris (2023)
- ▶ Developed **RecForest**: An extension of the Random Survival Forests algorithm to handle recurrent events, *with or without a terminal event* Murris (2024)
 - Refine the splitting rule and terminal node estimation
 - Provide appropriate metrics and error evaluation methods
 - Adjust variable importance calculations accordingly

2 Explore conditions for understanding survival machine learning

- ▶ Assessed *evaluation criteria of ML algorithms* by Health Technology Assessment bodies Farah & Murris (2024)
- ▶ Developed **ready-to-use tools** to consider interpretability methods for survival problems Murris (2024b)



Take home messages

From the statistician's perspective

From the user's perspective



Take home messages

From the statistician's perspective

From the user's perspective





Take home messages

From the statistician's perspective

From the user's perspective



What is the **research question**?



Take home messages

From the statistician's perspective



From the user's perspective

What is the **research question**?

Give attention to **recurrent events** analysis!



Take home messages

From the statistician's perspective



From the user's perspective

What is the **research question**?

Give attention to **recurrent events** analysis!

Care about **transparency, explainability** and **interpretability**



Next Steps

RecForest development:

R package in progress!

Made possible by *Guillaume D (Pierre Fabre)*

reforest 1.0.0 Articles ▾ Reference Changelog

reforest



{reforest} offers a flexible solution for analyzing recurrent events in survival data, outperforming traditional methods like the Cox model, which struggle with repeated events (e.g., hospital readmissions) and terminal events like death. By leveraging machine learning (Random Survival Forests), RecForest models both the timing and frequency of events, even with right-censored data, leading to more accurate predictions and insights, ultimately aiding in better decision-making and patient care.

The methodology is fully described in [Murris J., Bouaziz O., Jakubczak M., Katsahian S., & Lavenu A. \(2024\).](#)

Application on French Hospital Data:

Our study on post-operative readmissions in digestive cancer is on track

In collaboration with *Stylianos T (AP-HP)* and *Pierre Fabre*

Tree-based model-specific interpretability method for survival outcomes:

Project shaping up, submitted for ENSAE work

Kudos to *Lucas D (Inria)* for handing over



References |

- Ali, S., Abuhmed, T., El-Sappagh, S., ... & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion*, 99, 101805.
- Bezin, J., Duong, M., Lassalle, R., Droz, C., ... & Moore, N. (2017). The national healthcare system claims databases in France, SNIIRAM and EGB: powerful tools for pharmacoepidemiology. *Pharmacoepidemiol. Drug Saf.*, 26(8), 954-962.
- Breiman, L. (1996). Bagging predictors. *Mach. Learn.*, 24, 123-140.
- Bouaziz, O. (2024). Assessing model prediction performance for the expected cumulative number of recurrent events. *Lifetime Data Anal.*, 30(1), 262-289.
- Cadarette, S. M., & Wong, L. (2015). An introduction to health care administrative data. *Can. J. Hosp. Pharm.*, 68(3), 232.
- Cook, R. J., & Lawless, J. F. (1997). Marginal analysis of recurrent events and a terminating event. *Stat. Med.*, 16(8), 911-924.
- Farah, L., Murris, J. M., Borget, I., Guilloux, A., Martelli, N. M., & Katsahian, S. I. (2023). Assessment of performance, interpretability, and explainability in artificial intelligence-based health technologies: what healthcare stakeholders need to know. *Mayo Clin. Proc. Digit. Health*, 1(2), 120-138.
- GUIDELINE, I. H. Estimands and Sensitivity Analysis in Clinical Trials.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), 1-42.



References II

- Gunter, T. D., & Terry, N. P. (2005). The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. *J. Med. Internet Res.*, 7(1), e383.
- Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG*, 125(13), 1716.
- Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, 247(18), 2543-2546.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests.
- Kim, S., Schaubel, D. E., & McCullough, K. P. (2018). A C-index for recurrent event data: application to hospitalizations among dialysis patients. *Biometrics*, 74(2), 734-743.
- Kovalev, M. (2020). On the Complexity of Modern Machine Learning Algorithms. *J. Comput. Sci.*, 48, 101234.
- Krzyżysiński, P. (2023). Advances in Explainable Artificial Intelligence: A Survey of Methods and Applications. *Artif. Intell. Rev.*, 56(3), 1457-1481.
- Liao, Q. V., Gruen, D., & Miller, S. (2020, April). Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Proc. NeurIPS*, 31, 4765-4774.



References III

- Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care. *J. Biomed. Inform.*, 113, 103655.
- Medicine, T. L. R. (2018). Opening the black box of machine learning. *Lancet Respir. Med.*, 6(11), 801.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267, 1-38.
- Murris, J., Charles-Nelson, A., Tadmouri Sellier, A., Lavenu, A., & Katsahian, S. (2023). Towards filling the gaps around recurrent events in high dimensional framework: a systematic literature review and application. *Biostat. Epidemiol.*, 7(1), e2283650.
- Murris, J., Bouaziz, O., Jakubczak, M., Katsahian, S., & Lavenu, A. (2024). Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event.
- Porta, M. S., Greenland, S., Hernán, M., dos Santos Silva, I., & Last, J. M. (Eds.). (2014). A dictionary of epidemiology. *Oxford Univ. Press*.
- Ribeiro, M. T. et al. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proc. ACM SIGKDD*, 1135-1144.
- Schmidli, H., Roger, J. H., & Akacha, M. (2023). Estimands for recurrent event endpoints in the presence of a terminal event. *Stat. Biopharm. Res.*, 15(2), 238-248.
- Wei, J., Mütze, T., Jahn-Eimermacher, A., & Roger, J. (2023). Properties of two while-alive estimands for recurrent events and their potential estimators. *Stat. Biopharm. Res.*, 15(2), 257-267.

Merci





Appendix I – Modeling recurrent events, conditional models

- ▶ **Poisson models:**

$$\alpha_{j(j+1)}(t) = \alpha_0(t) \cdot r(\beta, X(t)) \quad \text{or} \quad \alpha_{j(j+1)}(t) = \alpha_0 \cdot r(\beta, X)$$

with β as the regression coefficient and $r(\beta, X)$ the relative risk function.

- ▶ **AG model:**

$$\alpha_{j(j+1)}(t) = \alpha_0(t) \cdot r(\beta, X(t))$$

- ▶ **PWP-CP model:**

$$\alpha_{j(j+1)}(t) = \alpha_0(t) \cdot r(\beta, X(t))$$

- ▶ **PWP-GT model:**

$$\alpha_{j(j+1)}(t) = \alpha_0(t) \cdot r(t - T_{N(t-)}, \beta, X(t))$$

- ▶ **NB models:**

$$\alpha_{j(j+1)}(t | U) = U \cdot \alpha_0(t) \cdot r(\beta, X(t)) \quad \text{or} \quad \alpha_{j(j+1)}(t | U) = U \cdot \alpha_0 \cdot r(\beta, X)$$

where U is a gamma-distributed random effect.



Appendix II – Theoretical MSE Criterion with Known Censoring Distribution

We introduce a theoretical criterion that would be available if the censoring distribution was known. For some function $\mu \in \mathcal{M}$, let:

$$MSE(t, \mu) = \mathbb{E} \left[\left(\int_0^t \frac{dN(u)}{G_c(u | \bar{X}(u))} - \mu(t | \bar{X}(t)) \right)^2 \right] \quad (1)$$

The crucial idea behind this comes from the fact that:

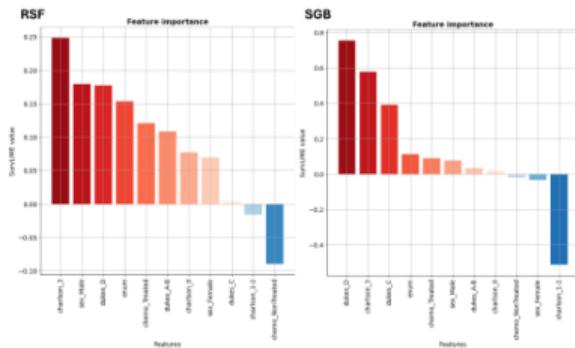
$$\mathbb{E} \left[\int_0^t \frac{dN(u)}{G_c(u | \bar{X}(u))} \right] = \mathbb{E}[\mu^*(t | \bar{X}(t))] \quad (2)$$

demonstrated in Bouaziz (2024)

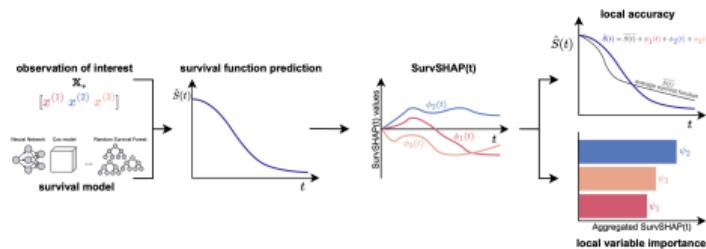
Appendix III – SurvLIME and SurvSHAP

SurvLIME

- ▶ Local explanation method for survival models
 - ▶ Perturbs the input data and fits a simple interpretable model (e.g., linear regression) to approximate the model's behavior around a specific instance.



SurvSHAP



- ▶ Global and local interpretability method based on Shapley values
 - ▶ Calculates the contribution of each feature to the prediction by considering all possible feature combinations