Predicting Recurrent Events in a Survival Framework

Development of a Machine Learning Approach and an Application in Oncology

Juliette Murris

International Day of Women in Statistics and Data Science 2024

01
Motivating clinical data

Digestive Cancer in France

Colorectal and bowel cancer

Hepatobiliary cancer

Gastric and oesophageal cancer

Pancreatic cancer

Digestive Cancer in France

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

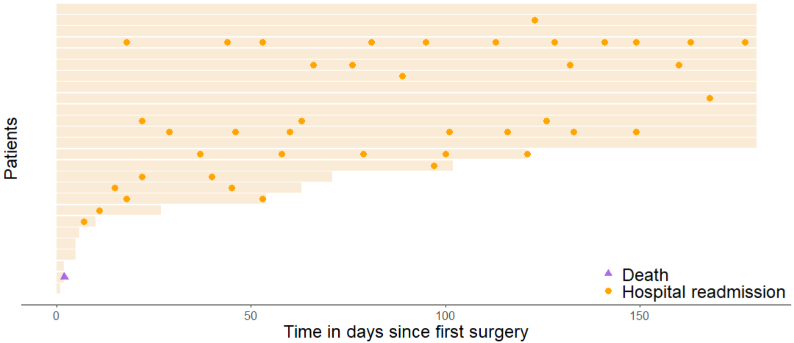
Key Facts

- 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

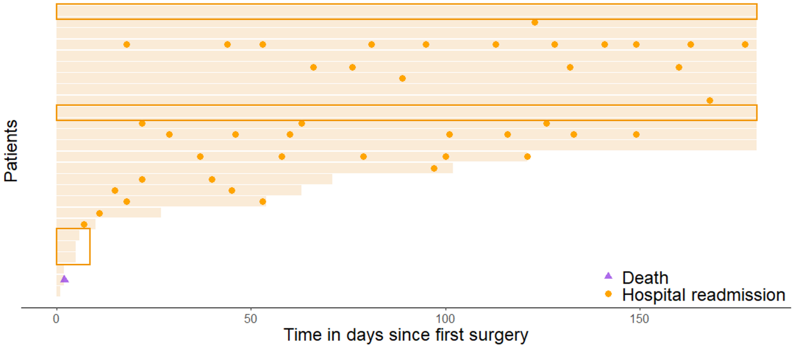
Public Health Concerns

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

What our data are made of

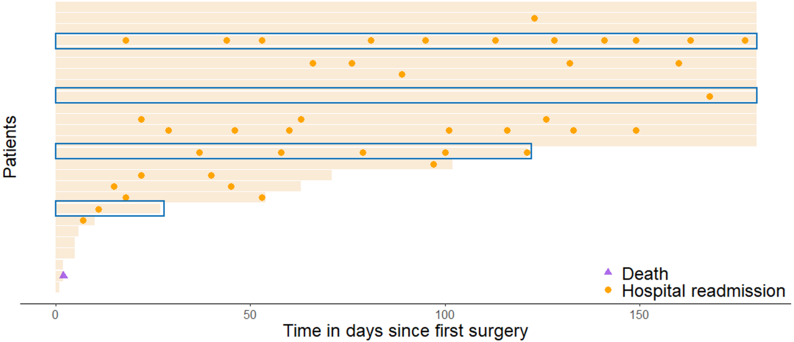


What our data are made of



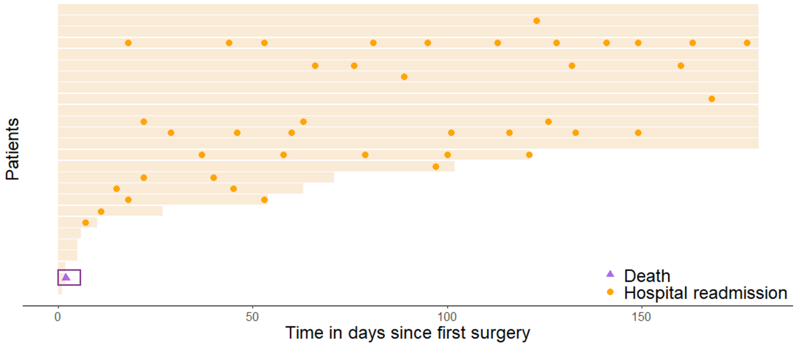
Patients with **no readmissions** over time

What our data are made of



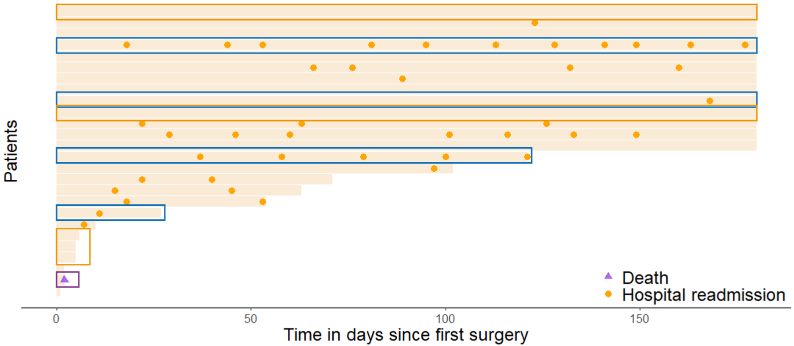
Patients with **one or more readmissions** over time

What our data are made of



Patients who died during follow-up

What our data are made of



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: logistic regressions, **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: logistic regressions, **No consideration of multiple events**

- Focus on the number of readmissions at 6 months?
  - **Regression** problem, Solution: linear regressions, **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: logistic regressions, **No consideration of multiple events**

- Focus on the number of readmissions at 6 months?
  - **Regression** problem, Solution: linear regressions, **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: logistic regressions, **No consideration of multiple events**

- Focus on the number of readmissions at 6 months?
  - **Regression** problem, Solution: linear regressions, **No consideration of time**

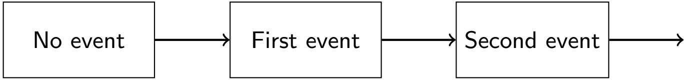
- Focus on time to first hospital readmission?
  - **Survival** problem, Solution: Survival analysis, **No consideration of subsequent events**

- **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.

No event $\rightarrow$ First event $\rightarrow$ Second event $\rightarrow$

Censoring

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

State-of-the-art – Non-parametric approach

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

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

State-of-the-art – Non-parametric approach

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

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

Nelson-Aalen Estimator:

$dN(t) = \sum_{i=1}^{n} Y_i(t) dN_i(t)$
total number of events observed over $[t, t + \Delta t)$

$$\hat{\mu}(t) = \sum_{\{h | t_{(h)} \leq t\}} \frac{dN(t_{(h)})}{Y(t_{(h)})}$$

H distinct event times across all n patients

$Y(t) = \sum_{i=1}^{n} Y_i(t)$
total number at risk over $[t, t + \Delta t)$

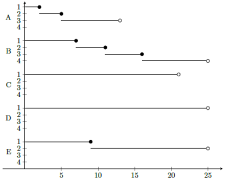
Cook & Lawless (1997)

State-of-the-art – 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**

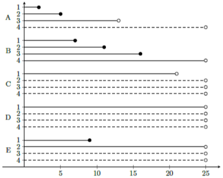


- Dependence structure between recurrent events by **full specification** of the recurrent event process

Marginal models

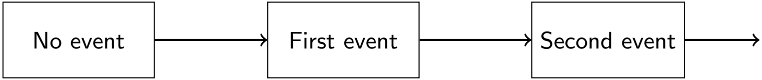
Wei, Lin & Weissfeld (1989), Lee, Wei & Amato (1992)

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

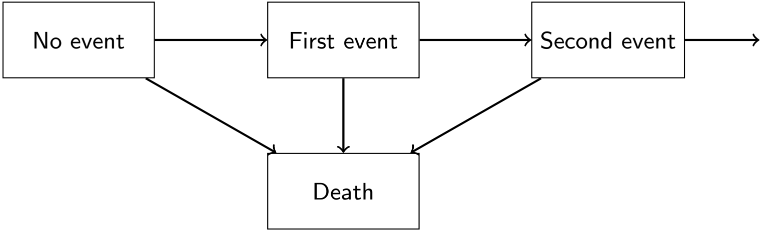


- Dependence structure between successive events may **remain unspecified**

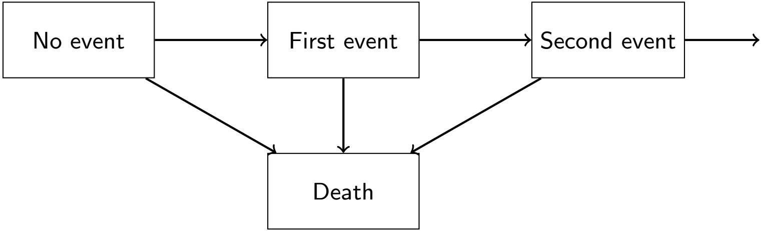
Non-informative censoring ?



Non-informative censoring ?



Non-informative censoring ?



= with a **terminal** event

State-of-the-art – With a Terminal Event

MCF Non-parametric Estimator:

increment
at time u

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

Kaplan-Meier estimator
of survival just before u

Modeling:

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

Ghosh & Lin (2000, 2002)

Raising questions – from the statistician's perspective

Key challenges

- How to manage situations with **high-dimensional data**?
- How to select **independent variables** when dealing with recurrent events?
- 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 select **independent variables** when dealing with recurrent events?
- How to avoid **overfitting** and ensure reliable **generalization** to new data?

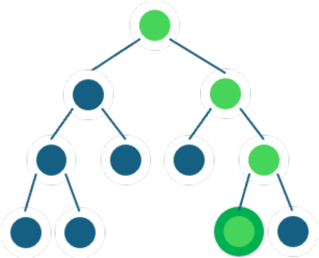
Current insights

Machine learning (ML) and survival counterparts
However, no ML algorithm *specifically designed* for recurrent events in a survival framework

Murris (2023)

02
Combining statistical inference and ensemble machine learning

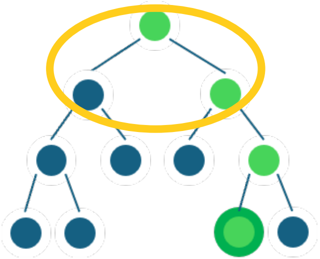
Growing Trees



Key Components

Breiman (1996)

Growing Trees

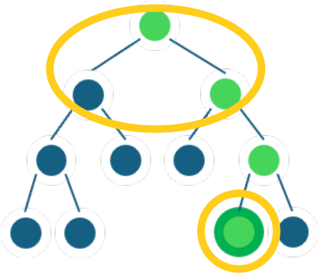


Key Components

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

Breiman (1996)

Growing Trees

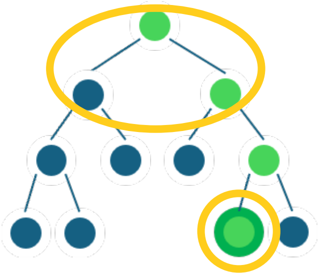


Key Components

- **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

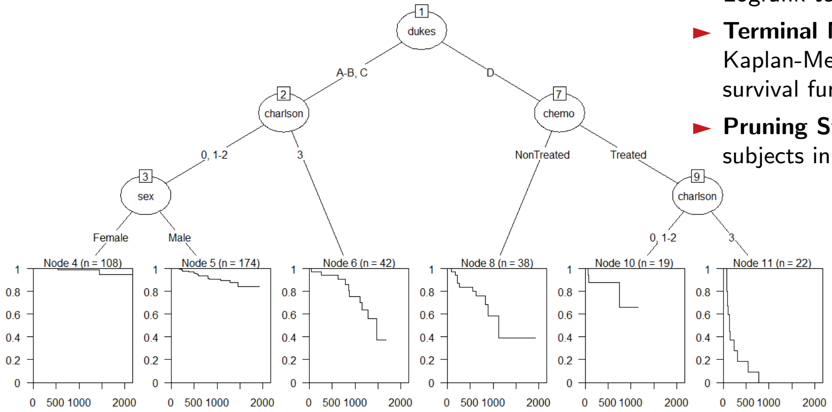


Key Components

- **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 15 subjects in terminal nodes

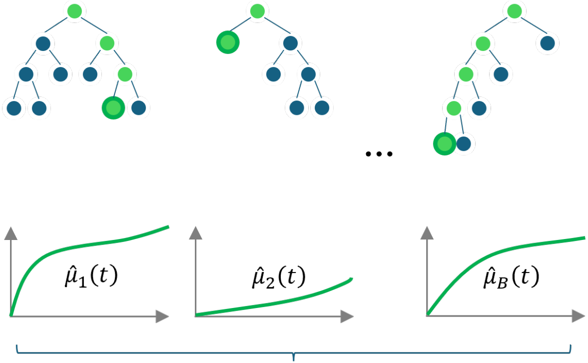
Ishwaran (2008)

Growing Survival Trees with Recurrent Events

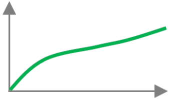
	Without a Terminal Event	**With** a Terminal Event
Splitting Rule	**Maximize the Test Statistic**	
At each node, $m \in \mathbb{N}$ predictors are randomly selected	Pseudo-score test	Wald test from Ghosh-Lin model
Terminal Node Estimator	**MCF Estimator** $\hat{\mu}_b(t\|\mathbf{x})$	
For tree b	$\int_0^t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

Murris (2024)

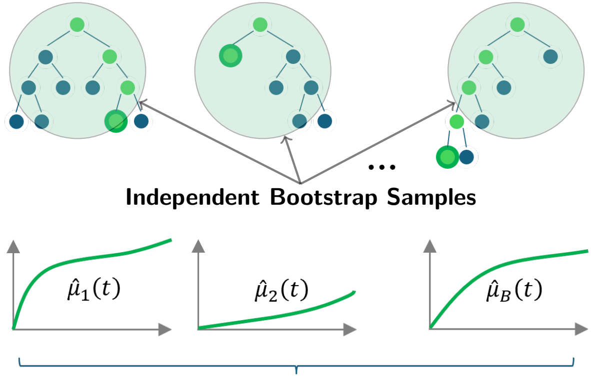
Aggregating to build random forests



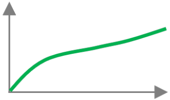
$$\widehat{\mathrm{M}}(t|\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{\mu}_b(t|\mathbf{x})$$



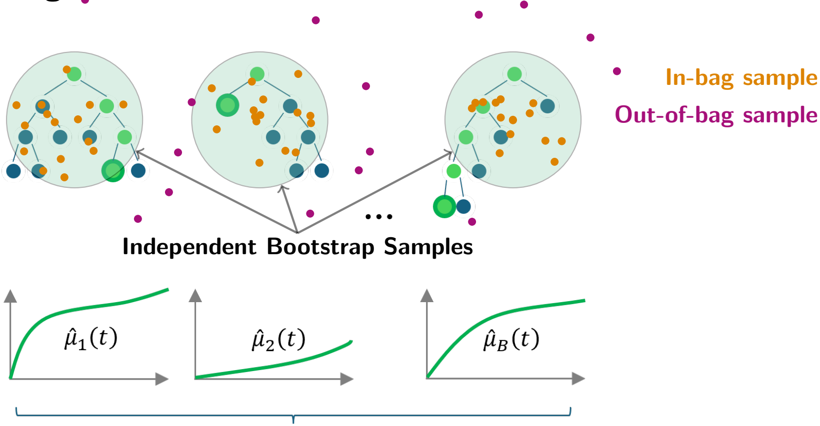
Aggregating to build random forests



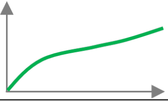
$$\widehat{\mathrm{M}}(t|\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_b(t|\mathbf{x})$$



Aggregating to build random forests

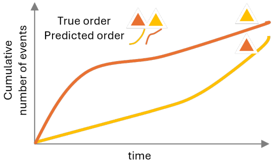


$$\widehat{\mathrm{M}}(t|\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_b(t|\mathbf{x})$$



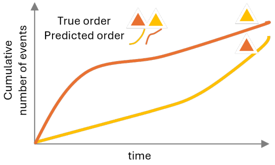
Performance evaluation – (a) The concordance index

- 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)



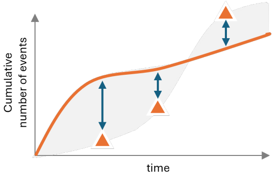
New C-index based on event **occurrence rate** Murris (2024)

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

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.

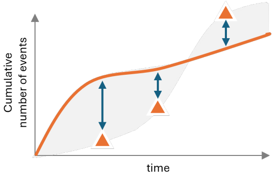
Performance evaluation – (b) The mean square error

- No MSE metric for recurrent events until very lately Bouaziz (2023)
- We adapted it for an ensemble framework



Performance evaluation – (b) The mean square error

- No MSE metric for recurrent events until very lately Bouaziz (2023)
- We adapted it for an ensemble framework



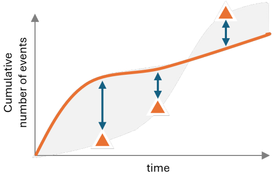
For each tree b,

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

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 (2023)
- 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:

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

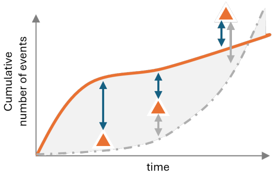
Performance evaluation – (c) The score

But ✋

Performance evaluation – (c) The score

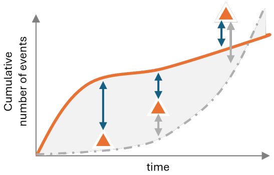
But ✋

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.



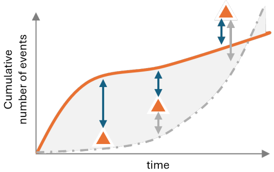
Need for a score to represent the prediction gain compared to a reference estimator $\hat{\mu}_0$ and we define for each tree b

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

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:

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

Importance of Variables

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

1. Estimate the original model error err_{OOB} from a chosen evaluation metric
2. For each feature $j \in \{1, \ldots, p\}$ do:

- Generate feature matrix Z^{perm} by permuting feature j in the data Z
- Estimate error $\widehat{err}_{OOB}^{Z^{\text{perm}}}$ based on the predictions of the permuted data
- Calculate permutation variable importance over B trees as:

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

This breaks the association between j and y

Output: Importance scores for all variables

Application to French Digestive Cancer Data

Table: RecForest performances

C-index ↑	IMSE ↓	IScore ↑
0.72	1,398.04	409.32

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**: $\%V_{\text{imp}} < 1\%$

Murris (2025?)

03
To wrap-up

Multiple Medical Applications of `RecForest`

RecForest

- **Oncology**
  Analysis of cancer re-lapses, progressions
- **Cardiovascular Events**
  Heart failures, strokes, angina
- **Chronic Diseases**
  Hospital readmissions, chronic episodes, flare-ups
- **Psychiatric Disorders**
  MDD relapses, schizophrenia
- **Infectious Diseases**
  Recurrent infections (e.g., HIV, Hepatitis)
- **Orthopedic Surgery**
  Pain recurrence, complications

To Wrap-Up – Key Takeaways

RecForest

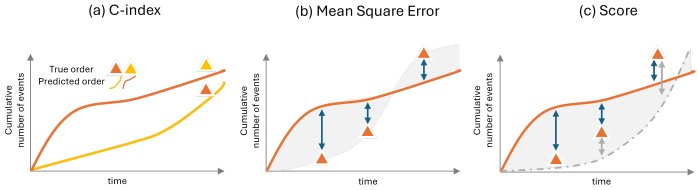
- **Non-Parametric** when no terminal event
- **High-Dimensional Data**
- **Robust to Multicollinearity**
- **Variable Importance**

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

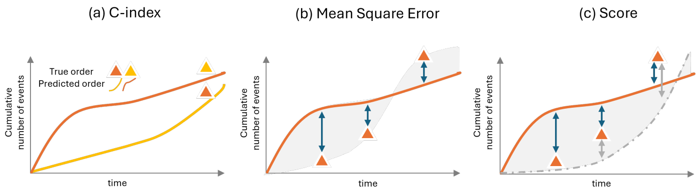


To Wrap-Up – Key Takeaways

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

References I

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.

Krzyziński, 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.