Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event

Juliette Murris, Olivier Bouaziz, Michal Jakubczak, Sandrine Katsahian, Audrey Lavenu. Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event. 2024. ffhal-04612431f

Cholatid Ratanatharathorn

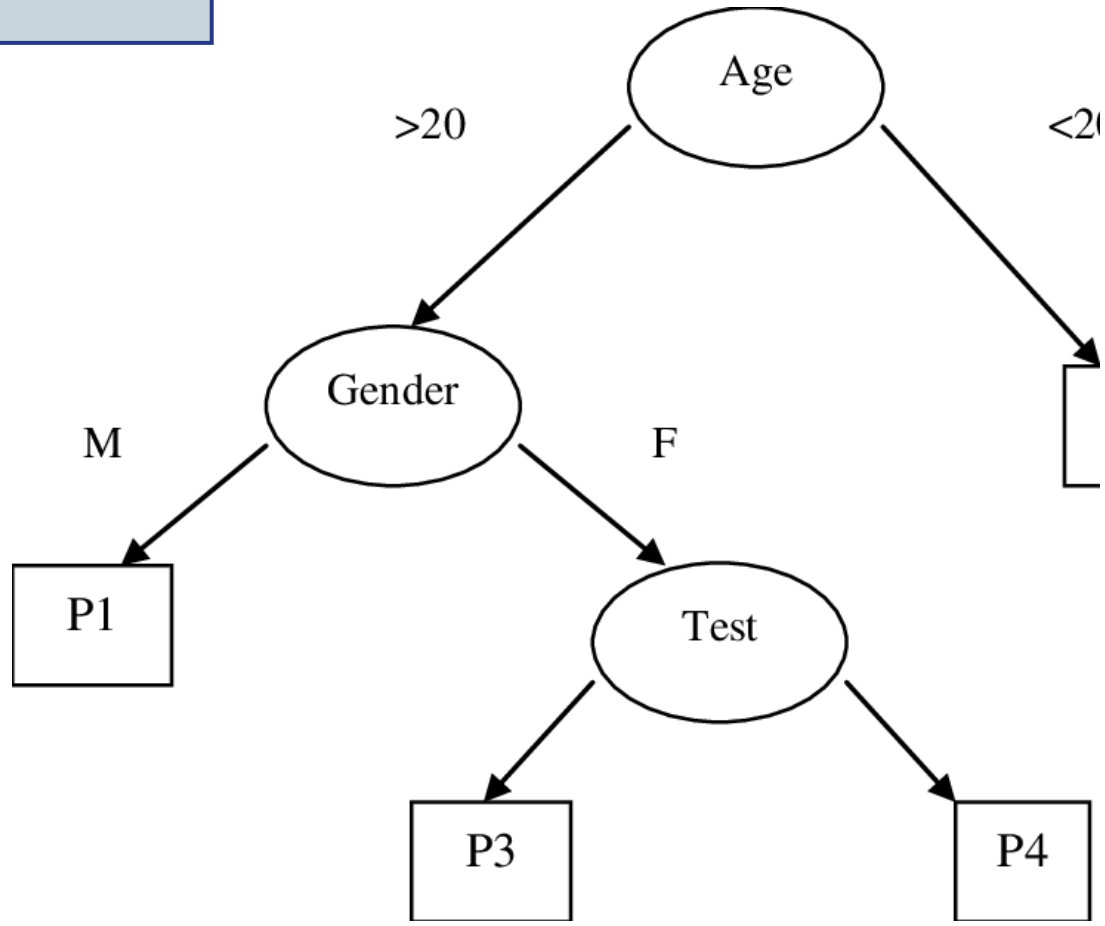
Department of Clinical Epidemiology and Biostatistics
Faculty of Medicine Ramathibodi Hospital, Mahidol University

Steps to recurrent RSF

- Decision tree
- Random forest
- Random survival forest
- Recurrent RSF without terminal event
- Recurrent RSF with terminal event

Decision tree classifier

- Invented since 1986.
- flowchart-like structure where each node represents a feature.
- **Interpretability**: Decision Trees are highly interpretable.
- **Feature Selection**: Decision Trees inherently select features based on their ability to classify the dataset.



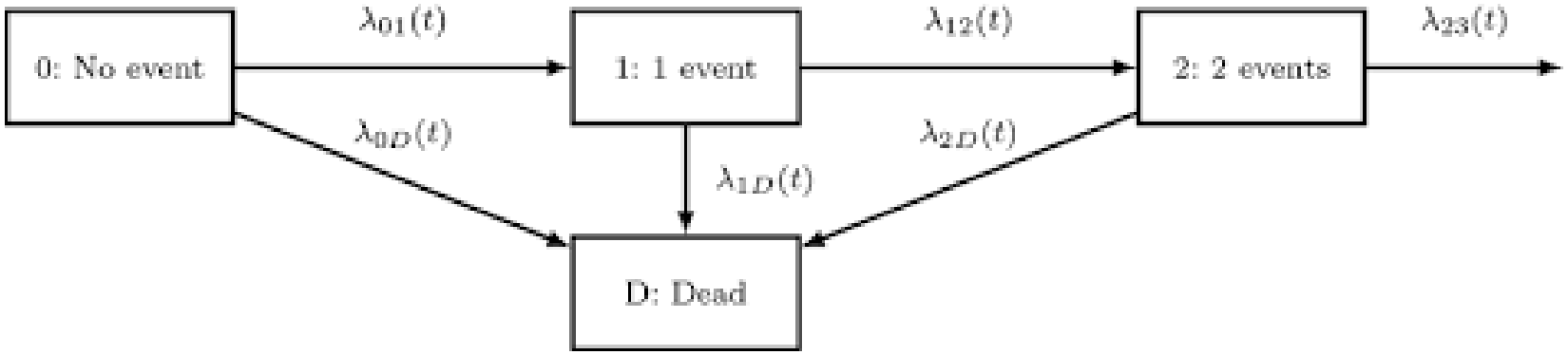
Quinlan, J. R. (1986). "Induction of decision trees". *Machine Learning*, 1(1), 81-106.

Background

- **Recurrent event data**: Many studies track events that happen multiple times to the same individual (e.g., hospital readmissions).
- **Challenges with recurrent events**: Standard survival models often only consider the first event and ignore subsequent occurrences.
- **Limitations of traditional methods**: Cox models and traditional random survival forests aren’t well-equipped to handle the complexity of recurrent events, especially when terminal events are involved.

Recurrent Events and Terminal Events

Recurrent Events: Events that can happen repeatedly to the same individual (e.g., multiple hospitalizations).
Terminal Events: Events that, once they occur, prevent further recurrences (e.g., death).
Impact on Modeling: Terminal events introduce a stopping point, complicating analysis as the model must distinguish between continued risk and final outcomes.



Algorithm 1 Overview of RecForest algorithm

Require: Draw $B > 0$ bootstrap samples from the learning data

for Each node of survival tree b **do**

$mtry$ predictors are randomly selected with $mtry \in \mathbb{N}$, $mtry \leq p$;

A greedy algorithm for optimal threshold research is used to maximize the test statistic;

The tree grows until the stopping rule is met based on the minimal number of events *minsplit* and the minimal number of individuals in terminal nodes *nodesize*;

Estimate $\hat{\mu}_b$ is computed;

end for

Estimate M is computed over the B trees.

Dataset

ID	Time	Age	Gender	Admit
1	20	30	M	1
2	10	62	F	2
3	30	55	F	2
4	15	74	F	3
5	20	15	M	1

Bootstrapped dataset

ID	Time	Age	Gender	Admit
1	20	30	M	1
3	30	55	F	2
4	15	74	F	3
5	20	15	M	1
3	30	55	F	2

Bootstrapped dataset

ID	Time	Age	Admit
1	20	30	1
3	30	55	2
4	15	74	3
5	20	15	1
3	30	55	2

Algorithm 1 Overview of RecForest algorithm

Require: Draw $B > 0$ bootstrap samples from the learning data

for Each node of survival tree b **do**

$mtry$ predictors are randomly selected with $mtry \in \mathbb{N}$, $mtry \leq p$;

A greedy algorithm for optimal threshold research is used to maximize the test statistic;

The tree grows until the stopping rule is met based on the minimal number of events *minsplit* and the minimal number of individuals in terminal nodes *nodesize*;

Estimate $\hat{\mu}_b$ is computed;

end for

Estimate M is computed over the B trees.

Bootstrapped dataset

ID	Time	Age	Gender	Admit
1	20	30	M	1
3	30	55	F	2
4	15	74	F	3
5	20	15	M	1
3	30	55	F	2

Decision on node splitting

Without terminal event

Log rank test

Test each covariate

$$U(t) = \int_0^t \frac{Y_A(u)Y_B(u)}{Y_A(u) + Y_B(u)} (d\hat{\mu}_A(u) - d\hat{\mu}_B(u))$$

With terminal event

Wald Test in Ghosh-Lin (GL) model

$$W = \frac{\hat{\beta}^2}{Var(\hat{\beta})}$$

Without terminal event

$$\hat{\mu}_b(t|x) = \hat{R}_b(t|x) = \int_0^t \frac{N_b(du|x)}{Y_b(du|x)}$$

Nelson-Aalen estimator
for cumulative hazard function

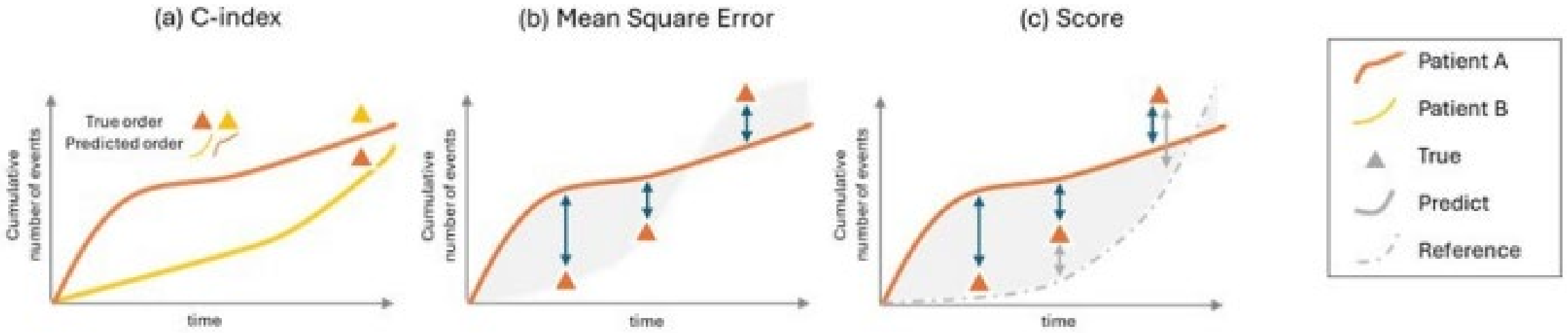
With terminal event

$$\mu_b(t|x) = \int_0^t \hat{S}_b(u|x)\, d\hat{R}_b(u|x)$$

Nelson-Aalen estimator
for cumulative hazard function

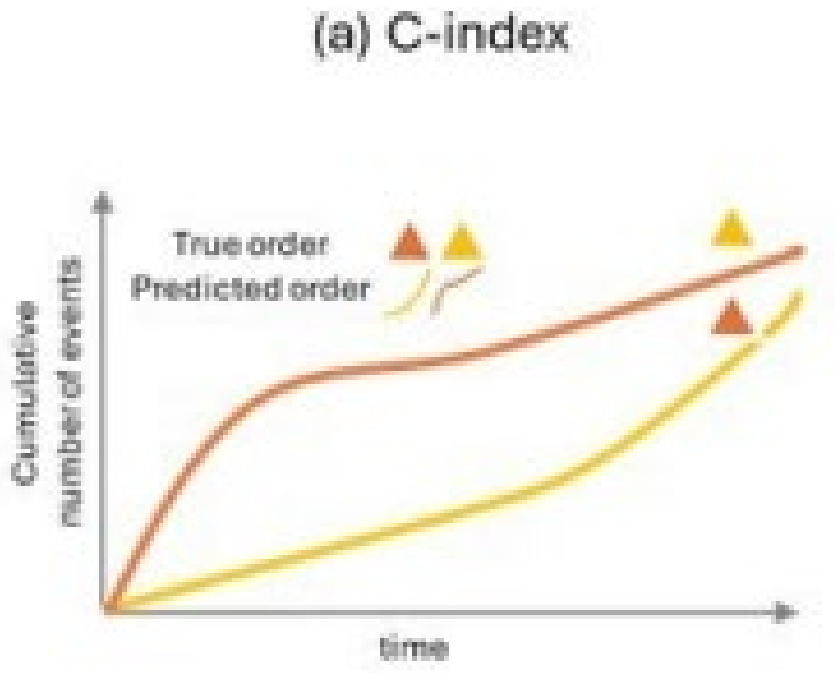
Kaplan Meier Curve
For survival function in terminal event

Model Performance



C-index

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

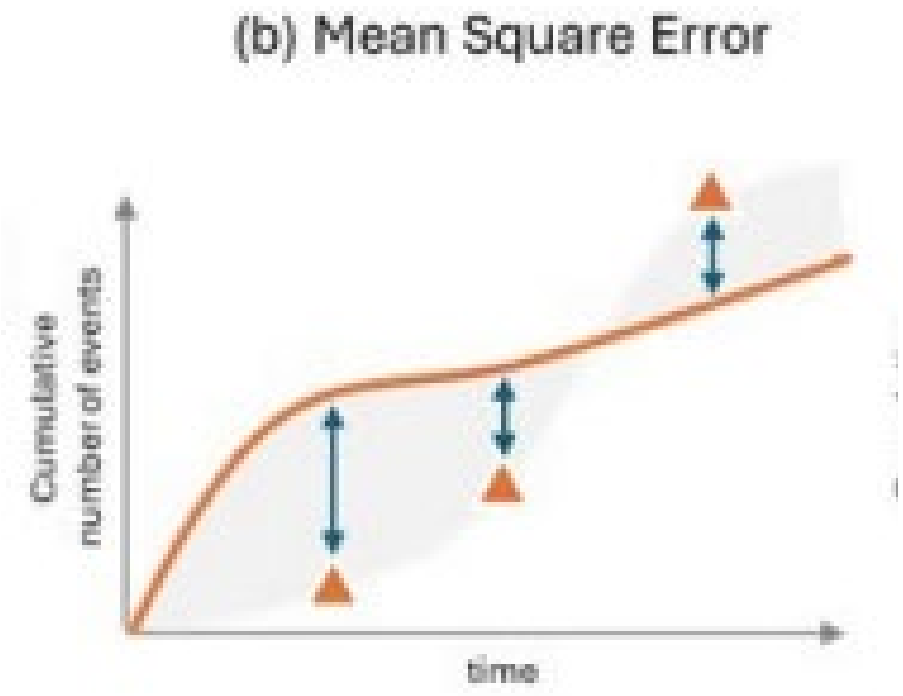


Patient	Observed Cumulative Events	Predicted Cumulative Events
1	2	1.5
2	4	3.8
3	1	0.9

Patient 1 and Patient 2
Observed: Patient 2 has 4 events (higher) and Patient 1 has 2 events.
Predicted: Patient 2 has 3.8 (higher) and Patient 1 has 1.5.
Concordant Pair.

$$\text{C-index} = \frac{\text{Number of Concordant Pairs}}{\text{Total Number of Comparable Pairs}}$$

Integrated Mean Squared Error (IMSE)



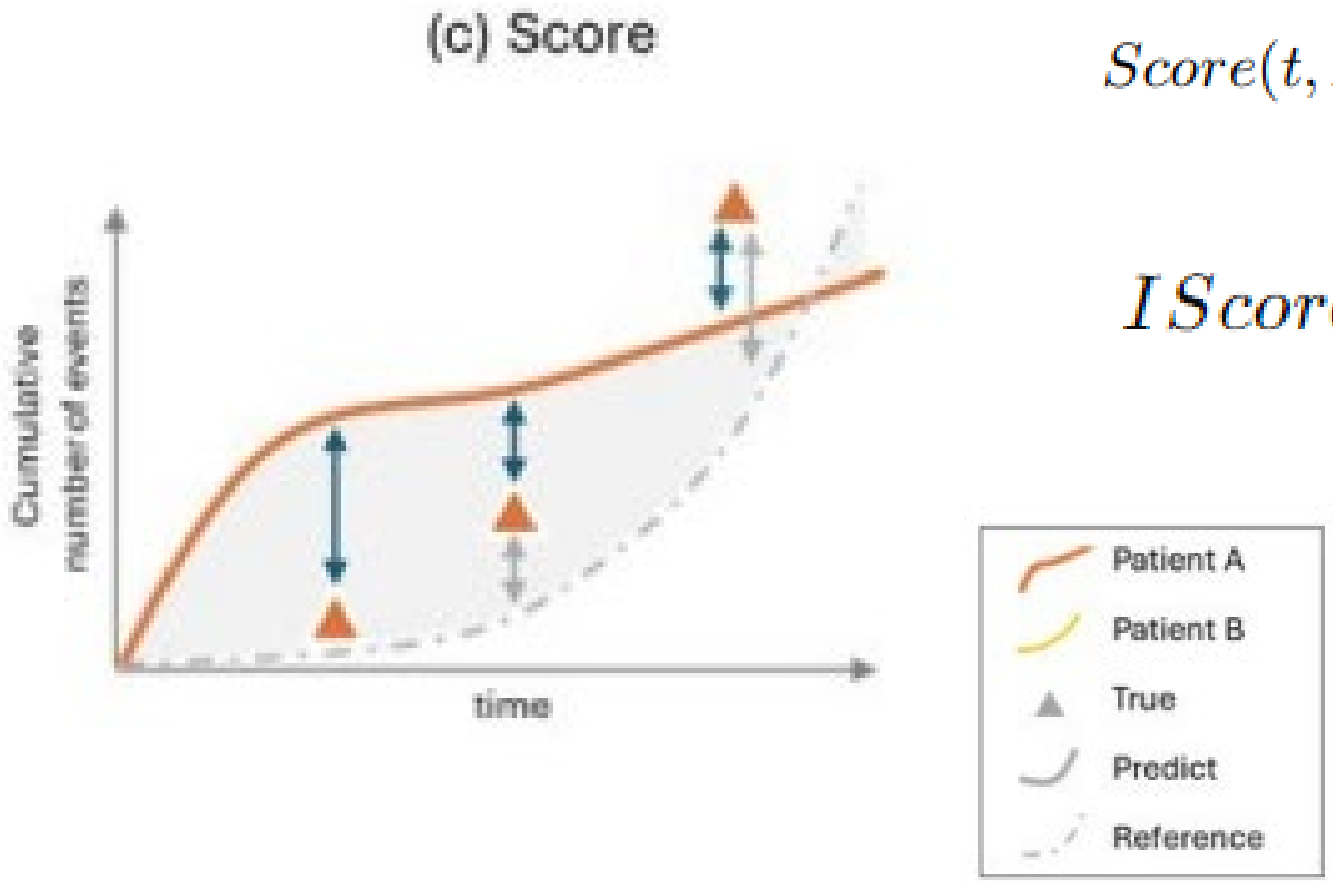
$$IMSE = \frac{1}{T}\int_0^T MSE(t)\,dt$$

Patient	Observed Cumulative Events (6 Months)	Predicted Cumulative Events (6 Months)	Squared Error
1	2	1.8	0.04
2	3	2.5	0.25
3	4	3.9	0.01

MSE (6) = 1/3 * (0.04 + 0.25 +0.01)

Calculate for all time point

IScore

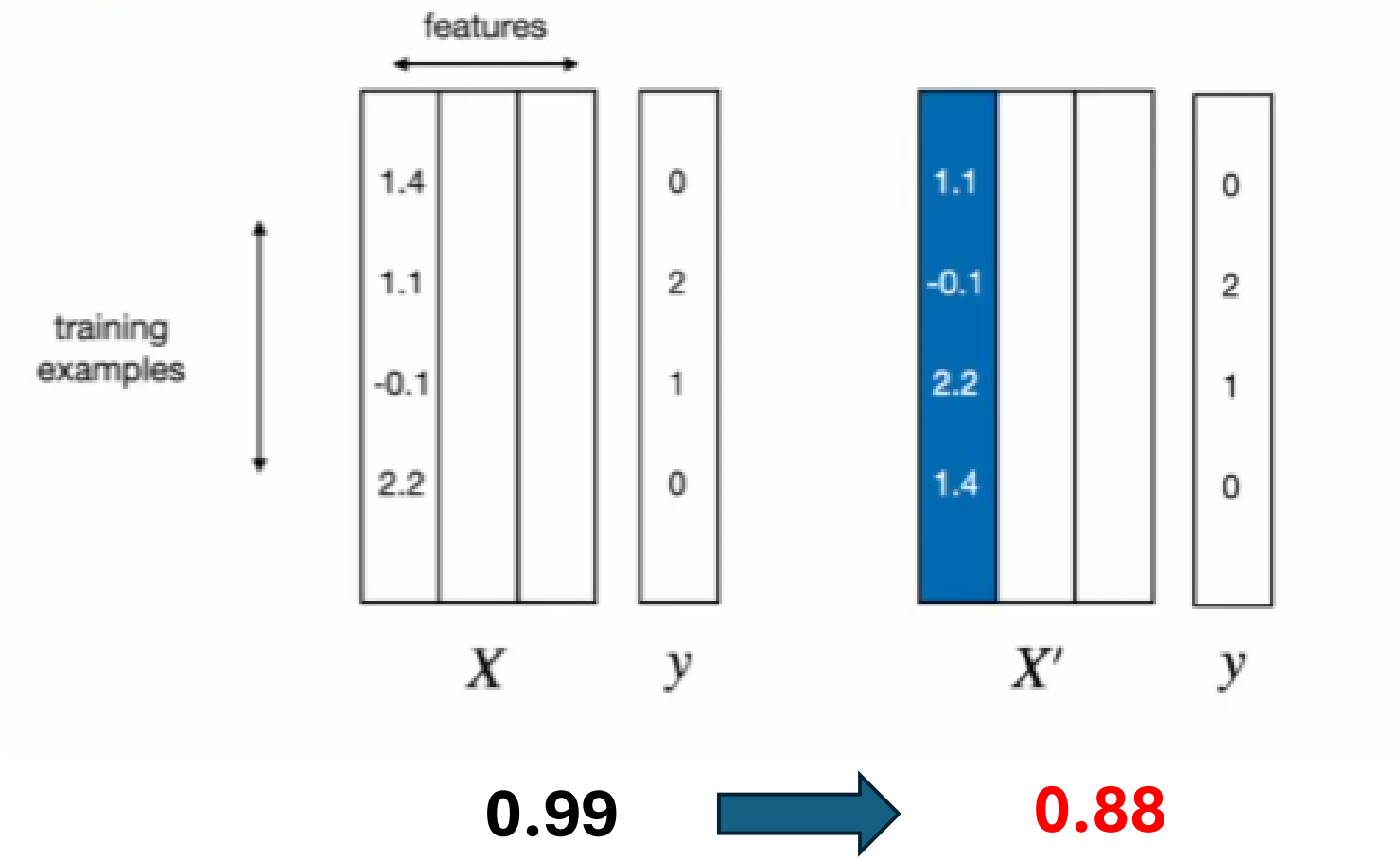


$$Score(t, \hat{M}) = \frac{1}{B} \sum_{b=1}^{B} Score_b(t, \hat{\mu}_b, \hat{\mu}_{b,0})$$

$$IScore = \frac{1}{T} \int_0^T \mathrm{IScore}(t)\, dt$$

Positive = better than reference model
Negative = worse than reference model

Feature importance



Performance measurement

C-index
Integrated Mean Squared Error (IMSE)

Feature importance calculate by **average difference** of original model performance and permuted model performance

Rehospitalization colorectal cancer

rehospitalization times after surgery in patients diagnosed with colorectal cancer

https://github.com/cran/frailtypack/blob/master/data/readmission.rda

id	enum	t.start	t.stop	time	event	chemo	sex	dukes	charlson	death
1	1	0	24	24	1	Treated	Female	D	3	0
1	2	24	457	433	1	Treated	Female	D	0	0
1	3	457	1037	580	0	Treated	Female	D	0	0
2	1	0	489	489	1	NonTreated	Male	C	0	0
2	2	489	1182	693	0	NonTreated	Male	C	0	0
3	1	0	15	15	1	NonTreated	Male	C	3	0
3	2	15	783	768	0	NonTreated	Male	C	3	1
4	1	0	163	163	1	Treated	Female	A-B	0	0
4	2	163	288	125	1	Treated	Female	A-B	0	0
4	3	288	638	350	1	Treated	Female	A-B	0	0
4	4	638	686	48	1	Treated	Female	A-B	0	0
4	5	686	2048	1362	0	Treated	Female	A-B	0	0
5	1	0	1134	1134	1	NonTreated	Female	C	0	0
5	2	1134	1144	10	0	NonTreated	Female	C	3	0
6	1	0	627	627	1	Treated	Male	A-B	0	0
6	2	627	1190	563	1	Treated	Male	A-B	0	0
6	3	1190	1406	216	1	Treated	Male	A-B	0	0
6	4	1406	1407	1	0	Treated	Male	A-B	0	0

Ghosh-Lin (GL) model

- statistical approach designed to handle **recurrent event data with terminal events**.
- Semi-parametric model.
- Joint model **WLW Model for Recurrent Event Data**

$$\lambda_{ik}(t) = \exp\{\beta_k z_i\}\,\lambda_{0k}(t), \qquad k = 1, \ldots, K.$$

Cox proportional-hazards model for **terminal events**

$$\lambda_i(t) = \exp\{\beta z_i\}\,\lambda_0(t),$$

TABLE 6

Means and standard deviations over the 10-fold cross-validation for readmission dataset

Metric\Model	Np	GL1	GL2	GL3	GL4	RecForest	GL*
C-index ↑	0.58	0.53	0.48	0.48	0.45	0.80	0.60
	(0.05)	(0.08)	(0.08)	(0.07)	(0.05)	(0.04)	(0.06)
IMSE ↓	7 883.50	7 843.99	8 361.16	8 229.08	9 981.50	706.02	7 934.28
	(6 229.47)	(6 106.36)	(6 292.29)	(6 478.35)	(6 064.23)	(508.96)	(6 606.23)
IScore ↑	ref.	39.41	-477.67	-345.62	-2 098.44	188.22	51.33
	ref.	(230.6)	(348.48)	(432.6)	(541.59)	(89.00)	(142.63)

Np = non-parametric estimator; GL1 = Gosh-Lin model with sex; GL2 = Gosh-Lin with sex and chemotherapy; GL3 = Gosh-Lin model with sex, chemotherapy and Dukes' tumoral stage; GL4 = Gosh-Lin model with sex, chemotherapy and Dukes' tumoral stage and Charlson's index; GL* = Ghosh-Lin model with best variables from RecForest. Arrows indicate whether higher are lower scores lead to best performances.

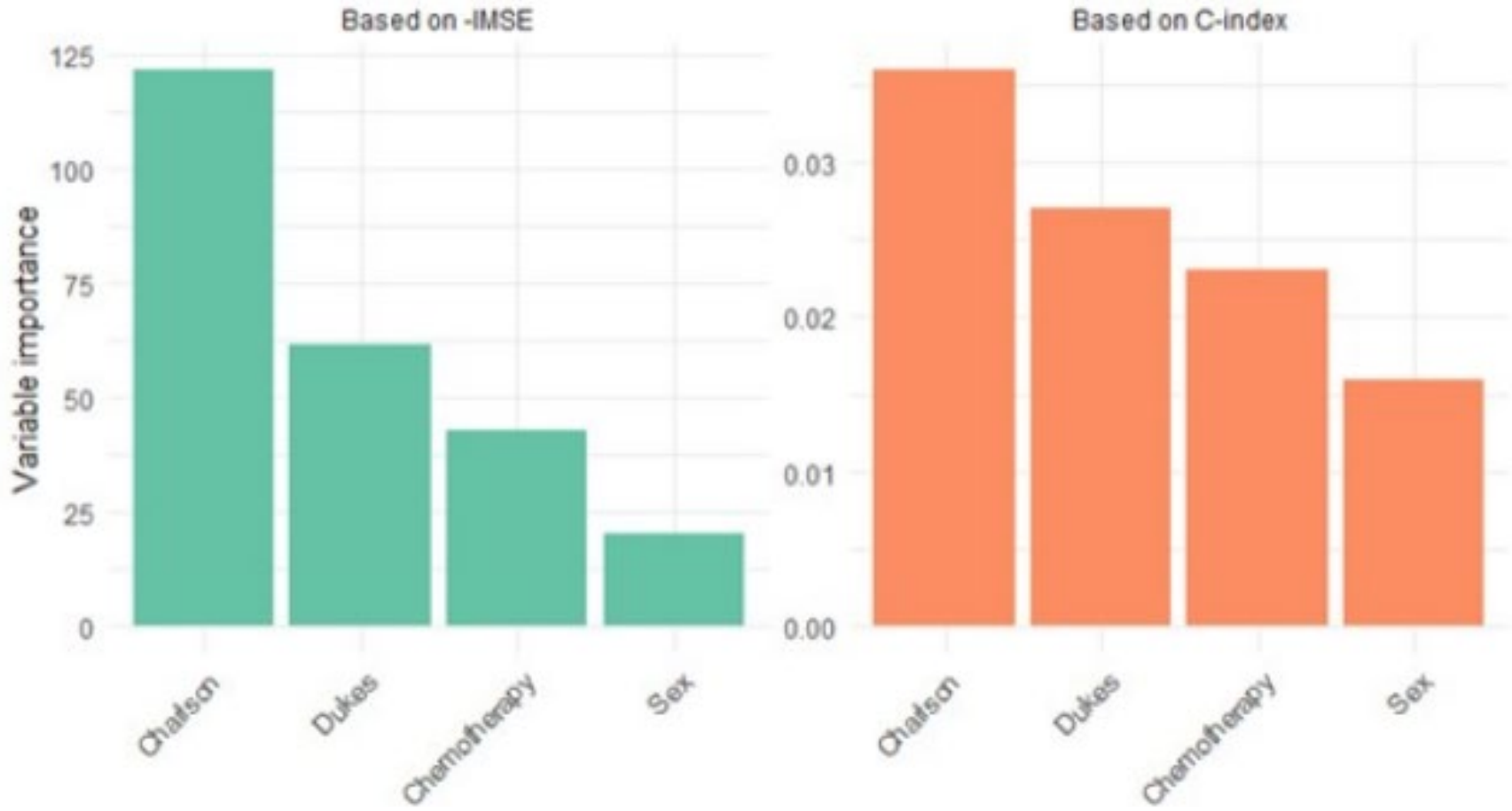


FIG 4. *Variable importance of RecForest computed on the C-index and the opposite of the integrated MSE. Charlson refers to Charlson comorbidity index, Dukes refers to tumoral Dukes stage.*

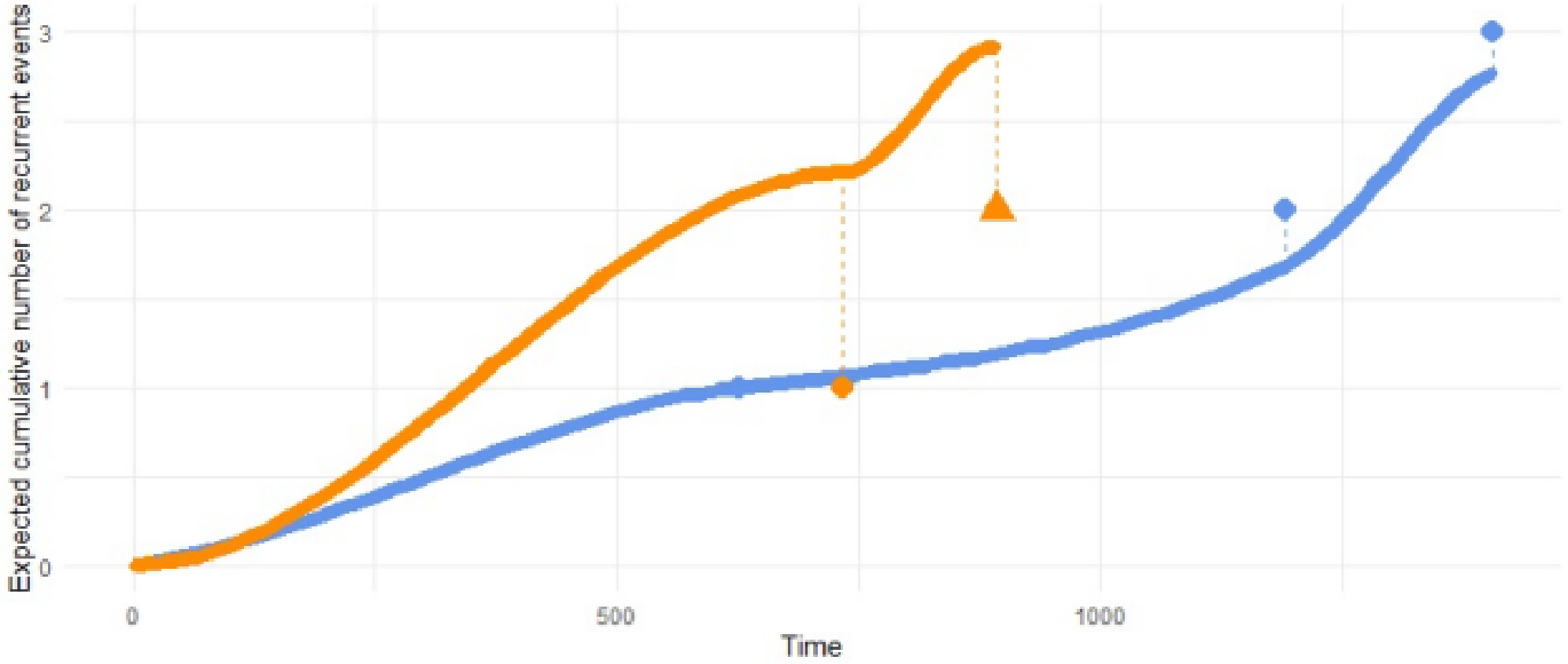
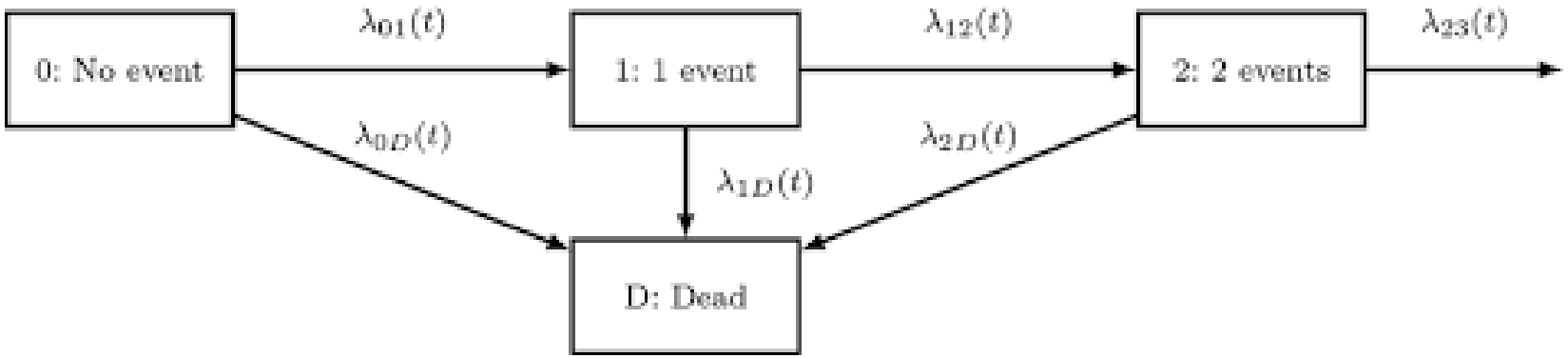


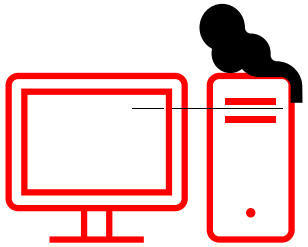
FIG 5. *Expected cumulative number of recurrent events with RecForest for two patients, one in orange with the highest Charlson comorbidity score, and the other in blue with the lowest. Data points outside the prediction curves are observed data. Triangle indicates the patient died.*

Advantages of RecForest

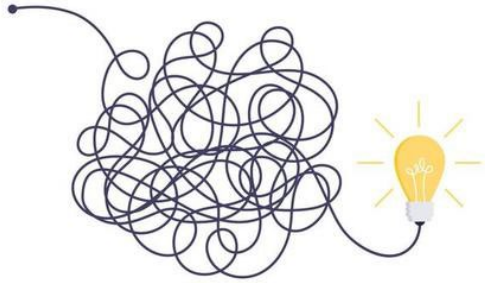
- **Handles High-Dimensional Data**
- **Adaptable to Time-Varying Covariates**: Allows dynamic risk adjustment.
- **Accommodates Terminal Events**: Provides reliable estimates even with censoring.



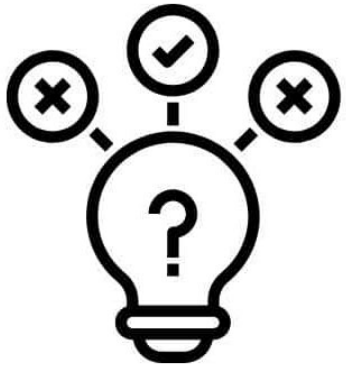
Limitations of RecForest



•**Computational Intensity**: Ensemble learning increases computational demands.



•**Potential Interpretability**: Complexity may limit direct interpretability.



•**Assumptions in GL Model**: Relies on assumptions like the proportional hazard.