How to properly evaluate survival models on small cohorts ?

Louis Rebaud

24/11/2021

Table of contents

I. Why training a model to predict the outcome/risk of a patient ?

II. Censored data

III. Available metrics

IV. Literature recommendations to evaluate models

V. Problems associated with cross-validation

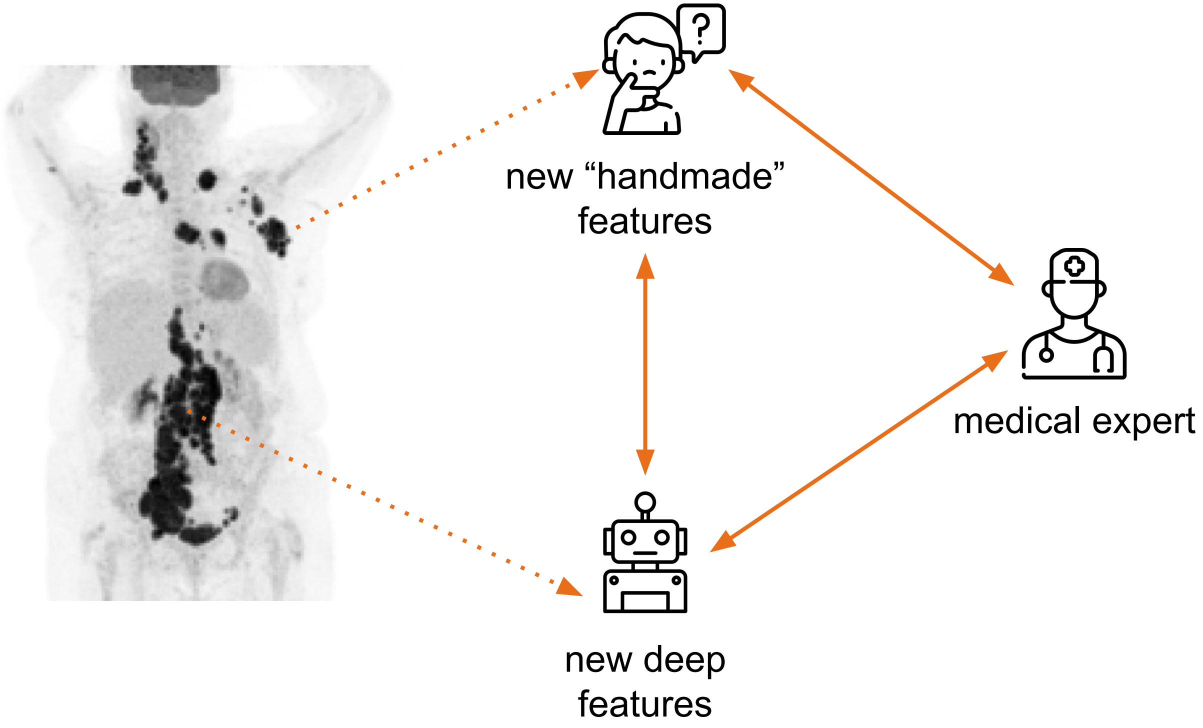
VI. Should we perform hyperparameter optimization ?

VII. Proposed framework

VIII. Conclusion

Why training a model to predict the outcome/risk of a patient ?

- The goal of my PhD is to build new biomarkers for whole-body PET of (lymphoma) patients
- Better understanding of the disease
- Better prediction of outcome and treatment response
- Use of machine learning models to find new relevant features
- To make the model learn new features, I can train it to predict the PFS of a patient from its PET scan
- Then, I need to interpret these models to see what features they build to solve the task



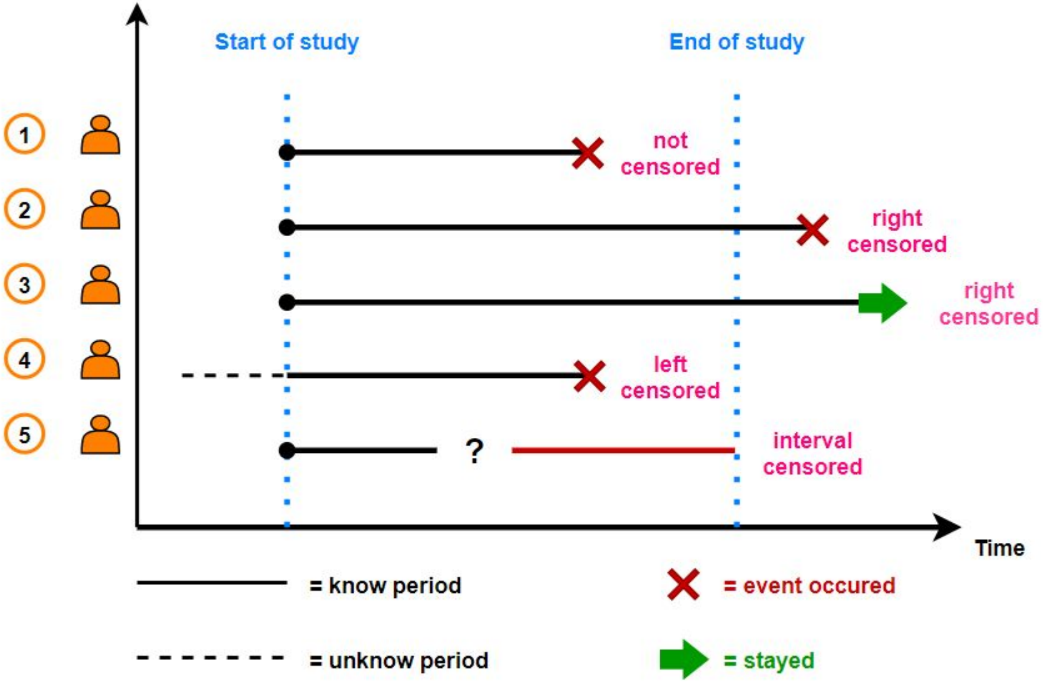
new “handmade” features

medical expert

new deep features

Censored data

- Censored data is a particular type of data where the value of a measurement is only partially known.
- Typical right-censored patient : “the patient was alive until a date D. After this date, we don’t know for how long he survived”.
- A database can contain censored and non-censored patients.
- Survival data almost always have some level of censorship.



Available metrics

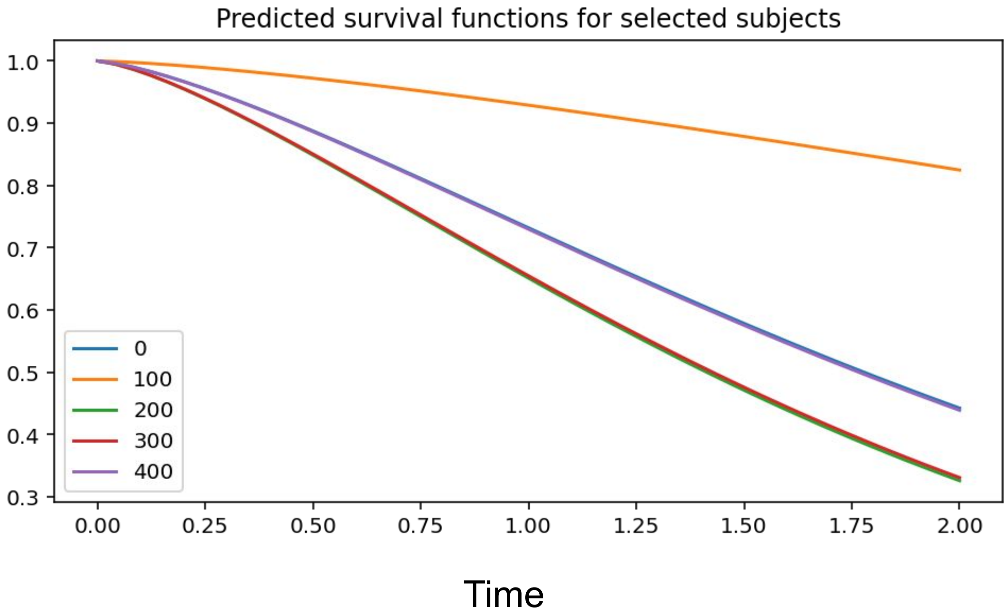
- In survival analysis, when evaluating a model, we are interested in two things:
    - calibration: how close my model prediction is from the true value
    - discrimination: how good is the model at separating a patient from other patients (e.g. predict the correct order of OS, separating patients in low or high risk)

- Traditional metrics cannot be used directly on censored data.

- Metrics that can be used with censored data:
    - Time-dependent Brier Score
    - Time-dependent Area under the ROC curve
    - Concordance index

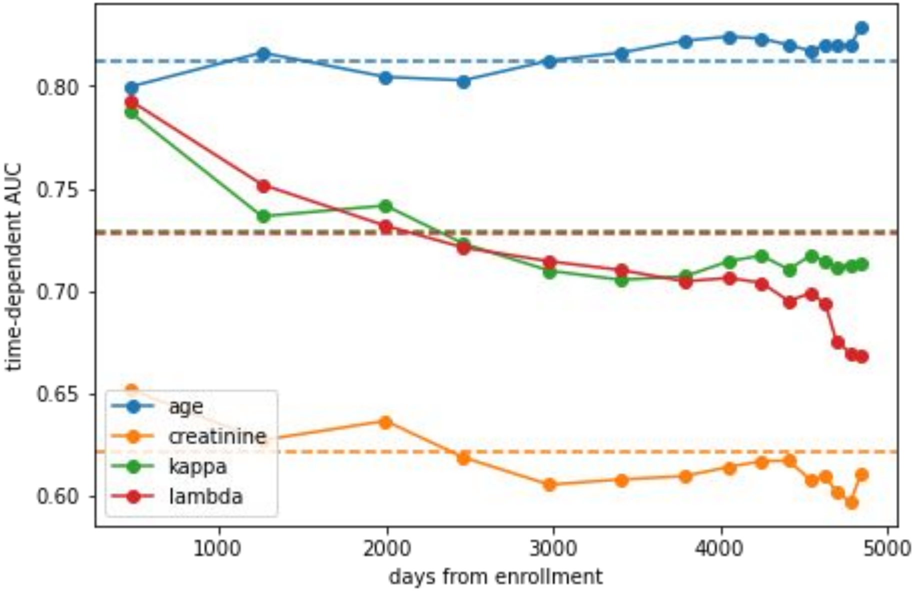
Time-dependent Brier Score

- It is an extension of the mean squared error to right censored data.
- Only applicable for models that estimate a survival function.
- Evaluate calibration and discrimination



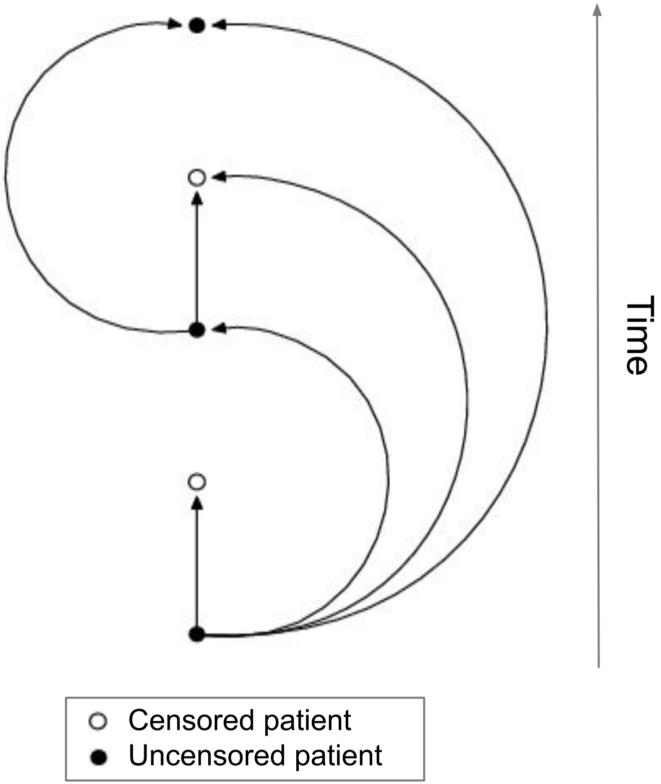
Time-dependent Area under the ROC curve

- With censored data, we can compute the ROC curve at a fixed time point
- This metric evaluates the AUC at multiple time points.
- We can use the mean to have a unique value
- Evaluate discrimination but not calibration.



Concordance index

- Censored patients can be used when compared to uncensored patients.
- This metric counts the proportion of correctly ordered pairs.
- Evaluate discrimination but not calibration.



The original Concordance index proposed by Harrell is optimistically biased when the level of censorship is high.

The corrected version proposed by Uno must be used instead. It fixes this issue by weighting the patients according to the censorship level.

Literature recommendations to evaluate models

- There is some consensus on how to properly evaluate models on small datasets
- The recommendations favour “accuracy like” metrics.
- What about other metrics (AUC, C-index) that compare patients to other patients to evaluate a model ?
- Can we use the same strategies with the censored metrics ?

Performance estimation

- (Repeated) k-fold cross-validation **without** independent test set
- Leave-one-out cross-validation **without** independent test set
- Confidence interval via 0.632(+) bootstrap

Hyperparameter selection

- (Repeated) k-fold cross-validation **with** independent test set
- Leave-one-out cross-validation **with** independent test set

Model comparison

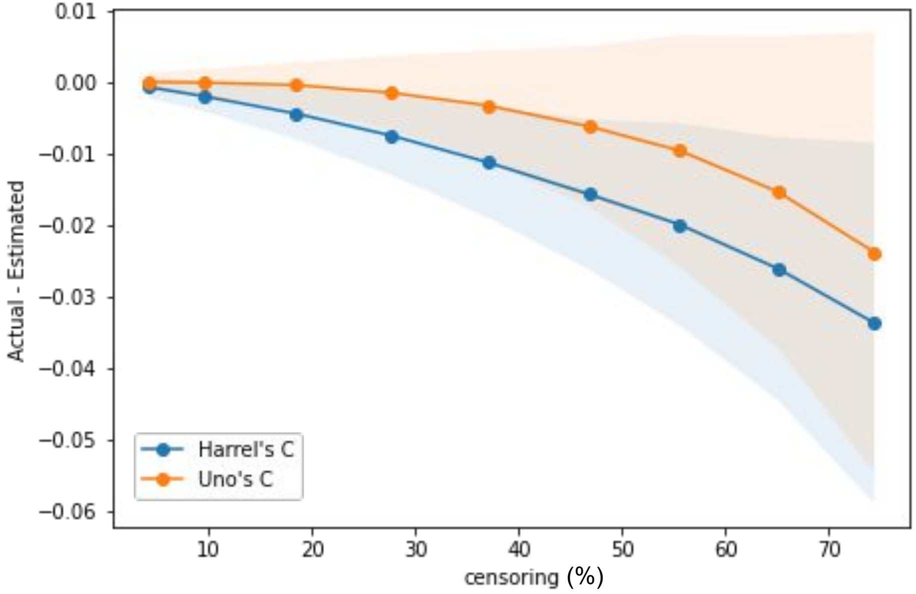
- Combined 5x2cv F test
- Nested cross-validation

The problems associated with cross-validation

- To answer this question, I performed several tests on synthetic data
- I measured:
  - the “actual” metric (no censoring)
  - the “estimated” measure of the metric with censoring
  - the “CV” cross-validated version of the estimated metric
- I used a sample size of 300 (same as the REMARC cohort).
- I can also control the strength of the evaluated feature (hazard ratio). To mitigate its effect, it was set to a random value at each iteration.
- I repeated each measure 10 000 times and measured the mean and standard deviation

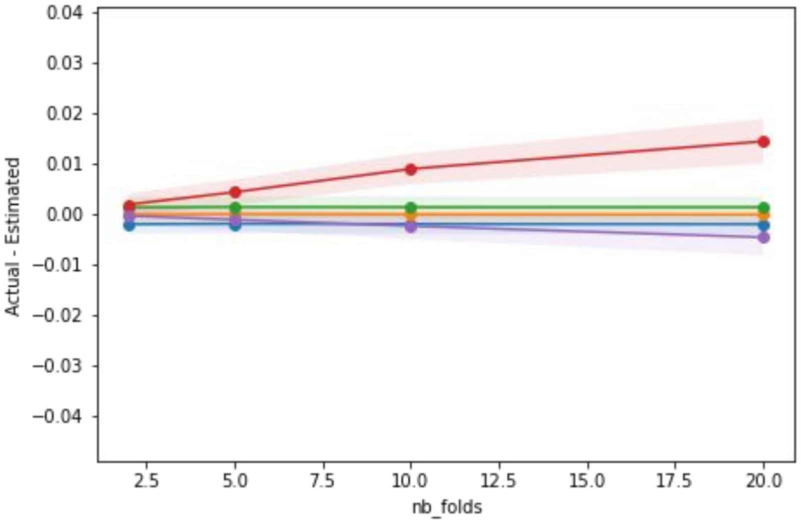
The problems with cross-validation

- As sanity check for the synthetic data, I tested if the Uno’s C-index was indeed better than Harrell’s C-index
- Observations:
  - The optimistic bias increases with the censorship level
  - Uno’s C-index indeed reduces the optimistic bias.
  - But it does not remove it completely.
- The synthetic data seems coherent on this point. So I used it to test the effect of the cross-validation on the AUC and the C-index

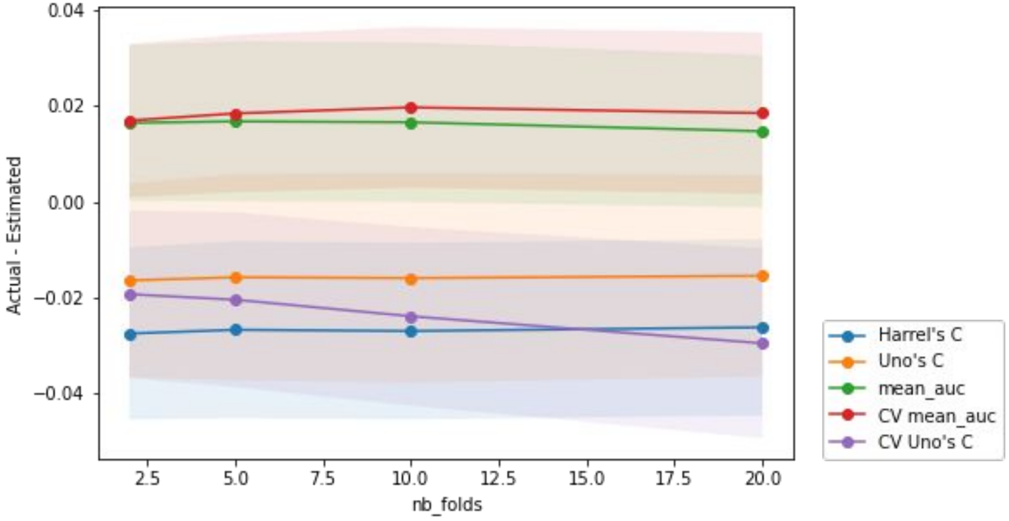


The problems with cross-validation

Low censoring (~10%)



High censoring (~70%)



- When the censoring is **low**, cross-validated **C-index** seems to be more accurate than cross-validated **AUC**
- When the censoring is **high**, cross-validated **AUC** seems to be more accurate than cross-validated **C-index**

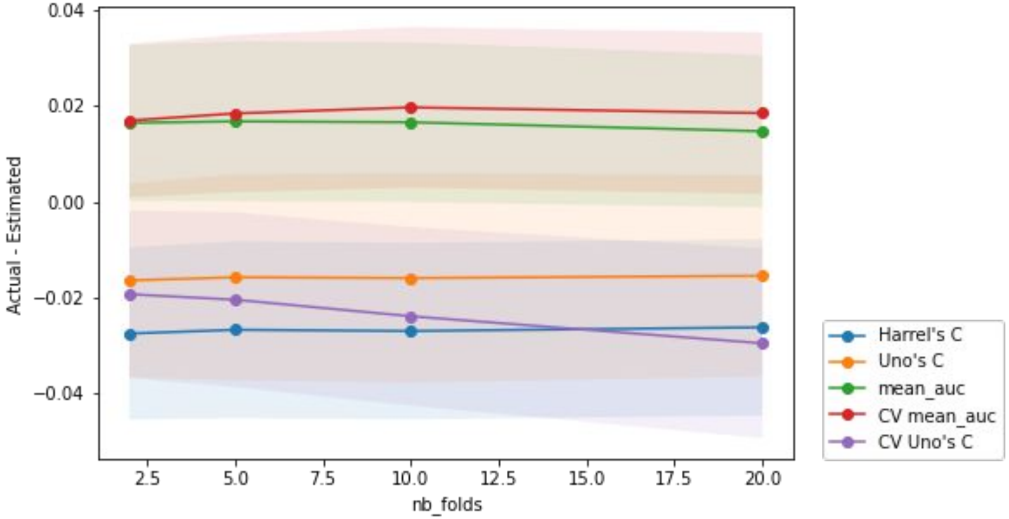
- Let’s focus on the case where the censoring is high :

The problems with cross-validation

When the censoring is high:

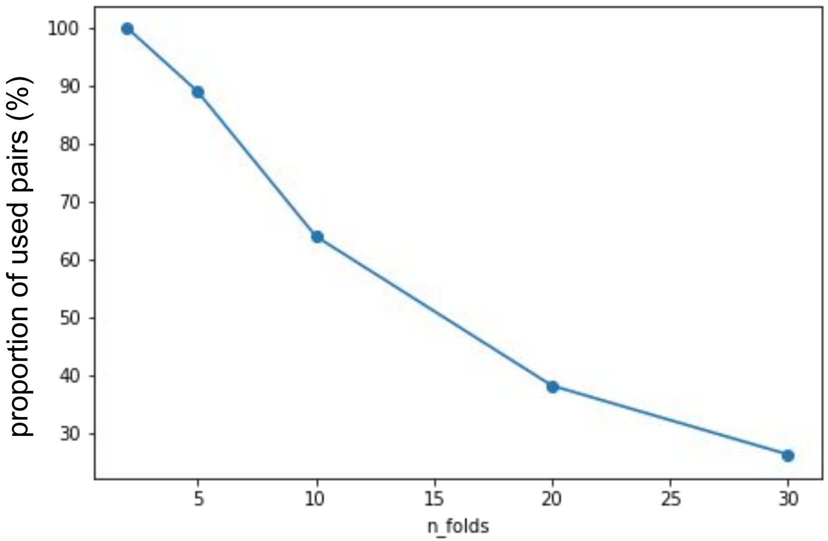
- The C-index is always optimistically biased.
- The time-dependent AUC is always pessimistically biased.
- Increasing the number of folds seems to make the C-index more optimistic.
- The Uno's C-index can get as bad as the Harrell's C-index if n_folds ≥ 10.
- The time dependent AUC seems less affected.
- Why does the C-index get worse when we increase the number of folds ?

High censoring (~70%)



The problems with cross-validation

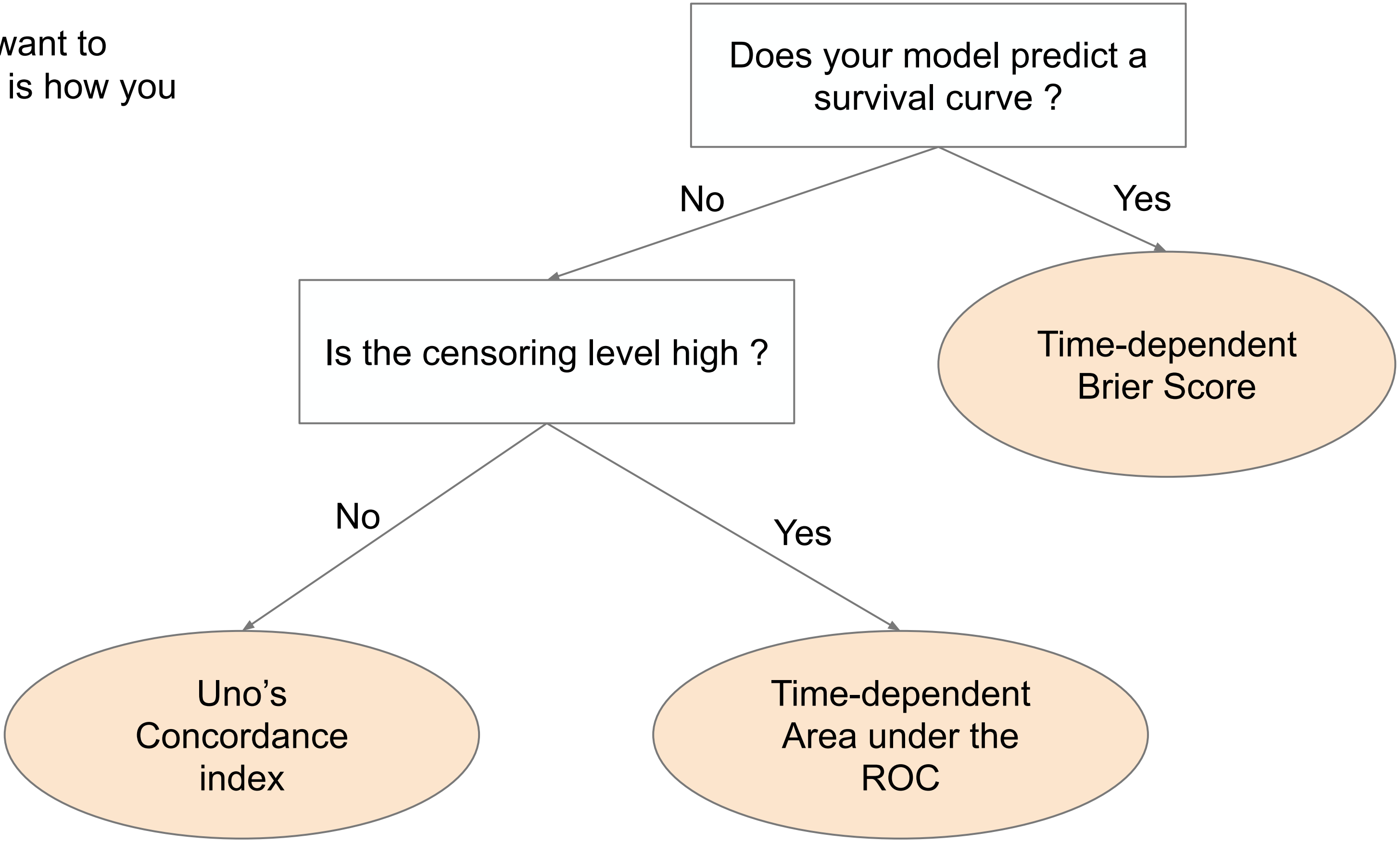
- When performing patient-based cross validation, we reduce the number of pairs evaluated.
- The Concordance Index is therefore evaluated on a fraction of the dataset.
- With datasets such as REMARC, cross-validated C-index should be avoided for performance estimation :
  - when n_fold is low: the model is trained on too few data (pessimistic)
  - when n_fold is high: too few pairs are used for evaluation (optimistic)



- It can still be used for model comparison if we accept the hypothesis that the bias is the same for every models (can be discussed)
- From this experiment, if the censoring is high, it seems that we can use the recommended methods if we use the time-dependent AUC to evaluate our survival models

The problems with cross-validation

- From this experiment, if you want to perform cross-validation, this is how you should select your metric:



Does your model predict a survival curve ?

No

Yes

Time-dependent Brier Score

Is the censoring level high ?

No

Yes

Uno's Concordance index

Time-dependent Area under the ROC

Should we perform hyperparameter optimization ?

- Hyperparameter optimization improves the performance of an algorithm by tuning it for the specific task.
- For a fair and realistic evaluation of algorithms, each one should be tuned for the problem.
- Because of cross validation, hyperparameter optimization is expensive on small dataset:

  number of models to train =
  number of repetitions
  × number of outer folds
  × (number of trials × number of inner folds + 1)

- With typical values : 5 × 5 × (100 × 5 + 1) = 12 525 models

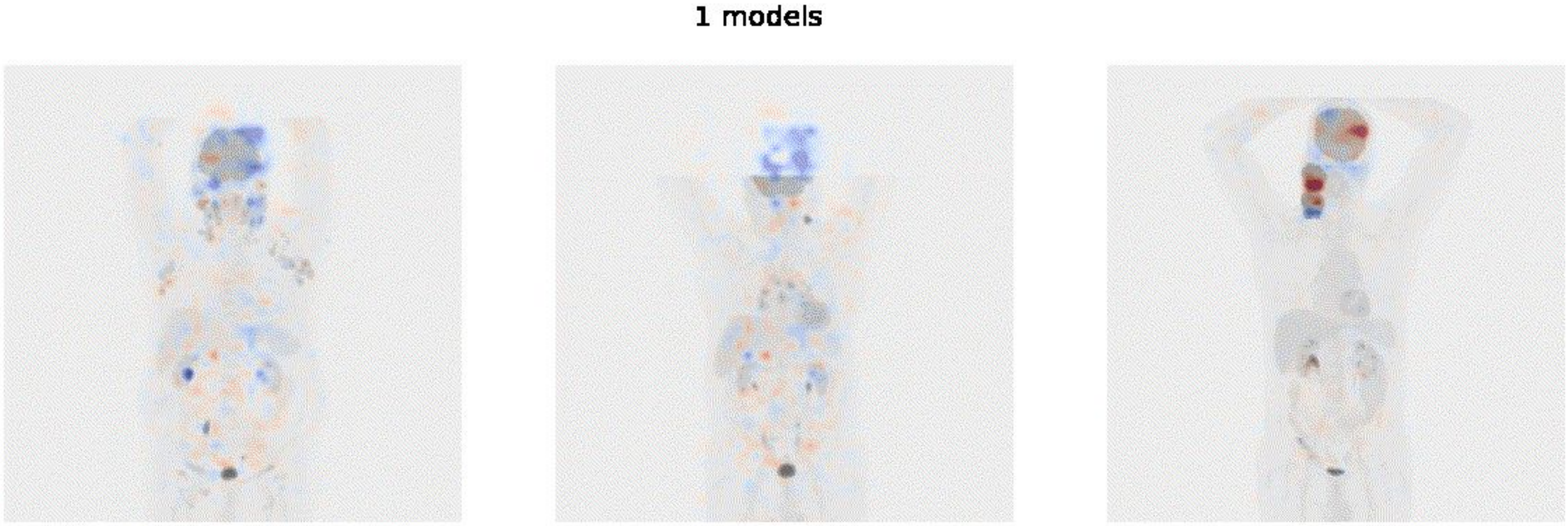
Should we perform hyperparameter optimization ?

- While being expensive, this approach also suffers from the small size of the dataset.
- The dataset is probably too small for a realistic hyperparameter search. The selected hyperparameters will have a high probability of being overfitted on the training set of the inner fold.
- This will result in pessimistic and noisy estimations of the performance of the model.
- Proposed methods: do not evaluate the trials and use all the models trained with different hyperparameters to perform an ensembled prediction on the test set (e.g. average of predictions)
- From my experiments, it provides better and more stable performance (I will also test that on synthetic data)
- It is also cheaper since we remove the inner cross-validation:

  with previous numbers: 12 525 models down to 2 500 models

- Ensembling models could also provide a less noisy interpretation of models by averaging individual interpretation

Should we perform hyperparameter optimization ?



Patient 1
27 months PFS (censored)
prediction: 1.4

Patient 2
38 months PFS (censored)
prediction: 2.2

Patient 3
19 months PFS (censored)
prediction: 1.3

Proposed framework

- Repeat N times a stratified k-fold cross validation
- For each fold:
  - Train M models with random hyperparameters on the training set
  - Average the predictions on the test set of all the models
  - Evaluate the averaged prediction on the test set with the selected metric
- Report the mean and standard deviation of the evaluation across all the folds

- The training set can be resampled differently for each model to have a better ensembling

Conclusion

- We saw what are the challenges of working with censored data
- If we want to use cross-validation, we should be careful when choosing the metric.
- Uno correction to the C-index does not seem to completely fix the optimistic bias.
- We saw a framework that is aligned with state-of-the-art methods to evaluate models while being adapted to our small datasets.
- Ensembling models is interesting on small datasets to reduce the variance, the computational cost and to increase the interpretability of the models

Bibliography

- Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808.
- Nadeau, C., Bengio, Y. Inference for the Generalization Error. Machine Learning 52, 239–281 (2003). https://doi.org/10.1023/A:1024068626366
- Thomas G. Dietterich; Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput 1998; 10 (7): 1895–1923.
- Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011 May 10;30(10):1105-17. doi: 10.1002/sim.4154. Epub 2011 Jan 13. PMID: 21484848; PMCID: PMC3079915.
- Blanche P, Kattan MW, Gerds TA. The c-index is not proper for the evaluation of t-year predicted risks. Biostatistics. 2019 Apr 1;20(2):347-357. doi: 10.1093/biostatistics/kxy006. PMID: 29462286.
- S. Pölsterl, “scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn,” Journal of Machine Learning Research, vol. 21, no. 212, pp. 1–6, 2020.
- I got the reference to some of this articles from this blog post (recommended by Thibault):

  https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms/