Prediction models for survival analysis

M2 Données massives en santé

Juliette Murris

Januray 2025

Outline

Introduction

Prediction models

There are thousands of prediction models in the medical literature

Use "baseline" variables to predict an outcome (most of time occurrence of an event): diagnosis or prognosis

- Binary (dead/alive): logistic regression
- Survival: Cox regression
- Continuous (rare): linear regression

But any other approach could be used (e.g. random forests, neural networks, SVM, etc.)

Examples

Short-term outcome (binary)

- Hospital mortality – SAPS-II (Simplified Acute Physiology Score), APACHE II/III

Longer-term outcome

- 10-year cardiovascular disease risk – Framingham risk score, QRISK2
- 2-year non-relapse mortality – HCT-CI (comorbidity index)
- 12-year recurrence after radical prostatectomy

What is a good prediction model?

"All models are wrong, but some are useful" Box (1976)

- **Multiple definitions and perspectives:**
  - A model is *"good"* if it is **useful** – but how do we define usefulness?
  - The ultimate goal: positively impacting **patient outcomes** (gold standard, but rare).
- **Steps before assessing therapeutic or clinical impact:**
  - Focus first on evaluating how well the model **performs**.
- **Key Aspects of model performance:**
  1. **Calibration:** How closely do predicted probabilities align with observed outcomes?
  2. **Discrimination:** How effectively does the model differentiate between cases (e.g., diseased) and controls (e.g., non-diseased)?

Evaluating performance

Concordance Index for Survival Analysis

What is the Concordance Index?

- The most common method for evaluating survival models is based on the relative risk of an event rather than the absolute survival times.
- This is done by calculating the concordance probability or the concordance index (C-index).

$$\mathbb{C} = \mathbb{P}(\eta_i > \eta_j | T_i < T_j), \tag{1}$$

where η_i is the risk score for individual i and T_i is the observed survival time.

Key References: Harrell (1982), Uno (2011), Gerds (2013)

C-index by Harrell (1982)

$$\hat{\mathbb{C}}_{\mathsf{H}} = \frac{\sum_{i \neq j} 1(\hat{\eta}_i > \hat{\eta}_j) 1(T_i < T_j, \delta_i = 1)}{\sum_{i \neq j} 1\left(T_i < T_j, \delta_i = 1\right)}, \tag{2}$$

where $\hat{\eta}_i$ is the predicted risk score, δ_i is an indicator for censoring, and T_i, T_j are the survival times.

- The numerator counts the concordant pairs, i.e., pairs where the model correctly predicts the order of events.
- The denominator normalizes by the total number of comparable pairs.

C-index by Uno (2011) with IPCW

- Uno et al. (2011) extended the C-index by introducing inverse probability of censoring weighting (IPCW)
- This method accounts for the censoring mechanism in survival data, improving the accuracy of the C-index estimation

$$\hat{\mathbb{C}}_{\mathsf{Uno}} = \frac{\sum_{i \neq j} 1(\hat{\eta}_i > \hat{\eta}_j) 1(T_i < T_j, \delta_i = 1) \hat{w}_i(T_i)}{\sum_{i \neq j} 1(T_i < T_j, \delta_i = 1) \hat{w}_i(T_i)}, \tag{3}$$

where

$$\hat{w}_i(T_i) = \begin{cases} \frac{\delta_i}{\hat{G}(T_i)} & \text{if } T_i \leq t, \\ \frac{1}{\hat{G}(T_i)} & \text{if } T_i > t, \end{cases} \tag{4}$$

with $\hat{G}(T_i)$ is the estimated Kaplan-Meier estimate of the censoring distribution.

C-index for Specific Follow-Up Period

Time-Dependent C-index: To evaluate performance over a fixed follow-up period $[0, t^*]$, Heagerty (2005) defined the time-dependent C-index.
The time-dependent AUC at a given time t is calculated as:

$$\mathrm{AUC}(t) = \mathbb{P}(\eta_i < \eta_j | T_i < t, T_j > t), \tag{5}$$

and the time-dependent C-index is:

$$\hat{\mathbb{C}}_{t^*} = \sum_t \widehat{\mathrm{AUC}}(t) \cdot \mathrm{num}(t), \tag{6}$$

where $\widehat{\mathrm{AUC}}(t)$ is the estimated AUC at time t.

Brier Score for Survival Models

- Initially developed for weather forecasting (Brier 1950), the Brier score assesses the accuracy of probabilistic predictions.
- For binary outcomes, it is equivalent to the mean squared error.

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{y}_i - y_i\right]^2 , \tag{7}$$

where $\hat{y}_i$ is the predicted probability and y_i is the actual outcome.

Brier Score for Survival with Censoring

- The Brier score is extended for survival analysis by adjusting for censoring (Graf 1999)
- The individual contributions are weighted according to the censoring distribution.

Weighted Brier Score Formula:

$$\mathrm{BS}(t) = \frac{1}{\sum_{i=1}^{N} Y_i(t)} \sum_{i=1}^{N} \hat{w}_i(t) \left[\hat{S}_i(t) - 1(T_i > t) \right]^2, \tag{8}$$

where $\hat{w}_i(t)$ is the weight for individual i, and $\hat{S}_i(t)$ is the predicted survival probability.

Integrated Brier Score (IBS)

- The Integrated Brier Score (IBS) evaluates the overall performance of a model over a period of time.
- It is defined as the integral of the Brier score over the time period $[\tau_1, \tau_2]$.

$$\mathrm{IBS} = \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} \mathrm{BS}(t)\, dt. \tag{9}$$

- $\tau_1 = 0$ and τ_2 is typically the maximum observed follow-up time.
- The IBS provides an overall assessment of model calibration over time.

Mean Absolute Error (MAE) for Survival

- The Mean Absolute Error (MAE) measures the average absolute difference between predicted and observed survival times.
- For survival analysis, MAE is only calculated for uncensored observations.

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} \delta_i |T_i - \hat{T}_i|, \tag{10}$$

where δ_i is an indicator for event occurrence, and T_i and $\hat{T}_i$ are the observed and predicted survival times, respectively.

Addressing Censoring in MAE

Challenges with Censoring:

- A naïve approach excludes censored subjects from MAE calculation, but this may introduce bias, especially with high censoring rates.

Advanced Approaches:

- Using inverse probability censoring weighting to account for censored observations (Haider 2020)
- **MAE-margin:** Assigning a "best guess" margin time to censored subjects based on Kaplan-Meier estimators (Qin 2023)

Building prediction models

The aim of the model

- **To be considered at the very beginning:**
  - **What does the model aim to clinically achieve?**
  - **When and where will it be used?**
    - Will it be implemented in a computer or app?
- **What resources are needed?**
  - Does it rely on **readily available information**, or
  - Are **additional data** (e.g., lab values) required?
- **What is the desired "final product"?**
  - A calculator?
  - A tool that is quick and easy to use?
  - A **simplified model** for specific contexts?

Which predictors?

- **Common temptation:** Include everything
  - Risk of **overfitting**.
- **Use subject matter (clinical) expertise:**
  - Knowledge almost always exists – leverage expertise.
  - While medicine has many unknowns, there is considerable knowledge on **pathophysiology**.
  - Avoid including **irrelevant predictors**, especially when data availability is limited.
- **Consider sample size:**
  - No explicit sample size formula, but follow the rule of thumb:
    - **10 to 20 events per variable (EPV).**
  - Small EPV increases the risk of **overfitting**.

Overfitting and Underfitting

Definition: Overfitting and underfitting describe how a model generalizes from training data (Bishop 2006)

Key Concepts:

- **Overfitting:** Model fits training data too well, capturing noise.
  - Symptoms: High complexity, poor test performance.
- **Underfitting:** Model is too simple to capture data patterns.
  - Symptoms: Poor performance on both training and test data.

Goal: Achieve a balance between complexity and generalization (Hasite 2009)

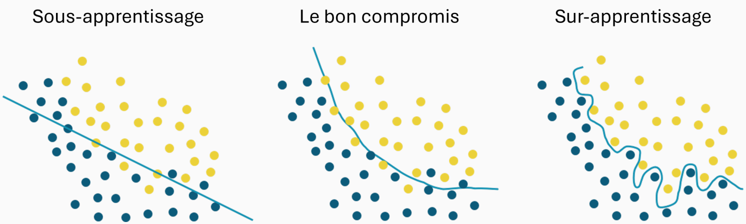
Bias-Variance Trade-off

Definitions:

- **Bias:** Error due to overly simplistic assumptions.
- **Variance:** Error due to sensitivity to data fluctuations.

Trade-off:

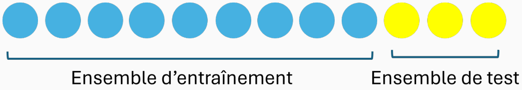
- Increasing model complexity reduces bias but increases variance.
- Objective: Find the optimal balance for generalization.



Model Evaluation and Validation

Objective: Evaluate the model's ability to generalize to unseen data (Hastie 2009)

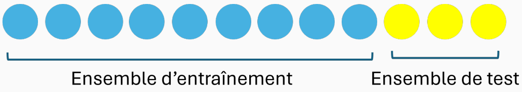
Train-test split (*e.g.*, 80%-20%)



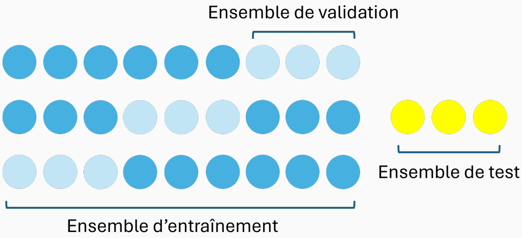
Model Evaluation and Validation

Objective: Evaluate the model's ability to generalize to unseen data (Hastie 2009)

Train-test split (*e.g.*, 80%-20%)



Cross-validation for robust evaluation



Hyperparameter Optimization

Definition: Hyperparameters influence model behavior and must be tuned for optimal performance.

Key Techniques:

- Grid search: Systematic search over a parameter grid.
- Random search: Randomized sampling of hyperparameter space.
- Bayesian optimization: Probabilistic model for optimization.

Warning: Avoid overfitting the validation set; use separate test data.

External validation

Rationale

We have seen before the development of the model with *internal validation*.

How well does the model perform on new unseen data?

- On another center,
- Data collected in another time batch,
- ...

Existing (yet imperfect) solutions

Challenges in External Validation:

- Data from external sources may differ in terms of **quality**, **availability**, and **measurement techniques**.
- Requires careful consideration of **heterogeneity** between training and validation datasets.

Best Practices for External Validation:

- Ensure **similarity in data structure** (e.g., outcome definitions, predictor variables).
- Perform **stratified analysis** to assess performance across different subgroups.
- Report **performance metrics** for external datasets.

Handling different data distributions (Andrew Ng's reco)

1. **Mix a small portion of external data into the training set:**
    - Take a small subset of the external validation data and combine it with the original training data.
    - This helps to expose the model to the new distribution, reducing the risk of poor performance due to distributional differences.
2. **Fine-tune the model on the augmented training set:**
    - After mixing the data, fine-tune the model on the combined training data to allow the model to adapt to the new distribution.
    - Fine-tuning ensures that the model can generalize better to both the original and the new distribution.

TP1

TP1