

Machine Learning Methods in Survival Analysis

RETREAT GRBIO 2024

Jordi Cortés Martínez

July 11, 2024



Table of contents I

1 Overview

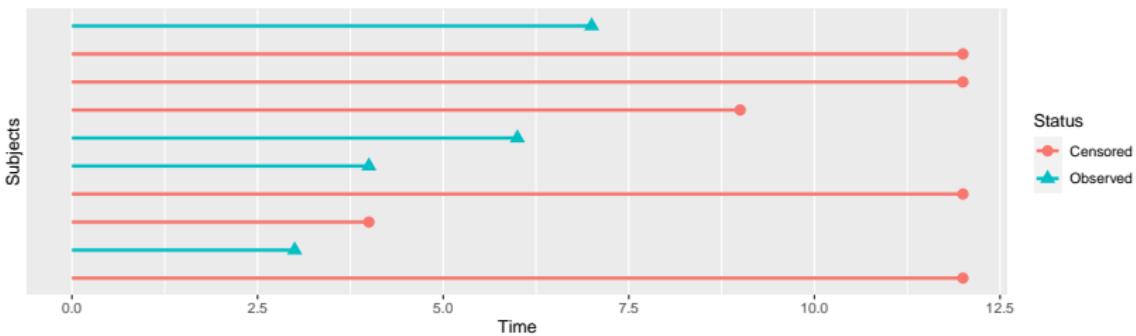
2 ML methods

3 Performance

4 Advices

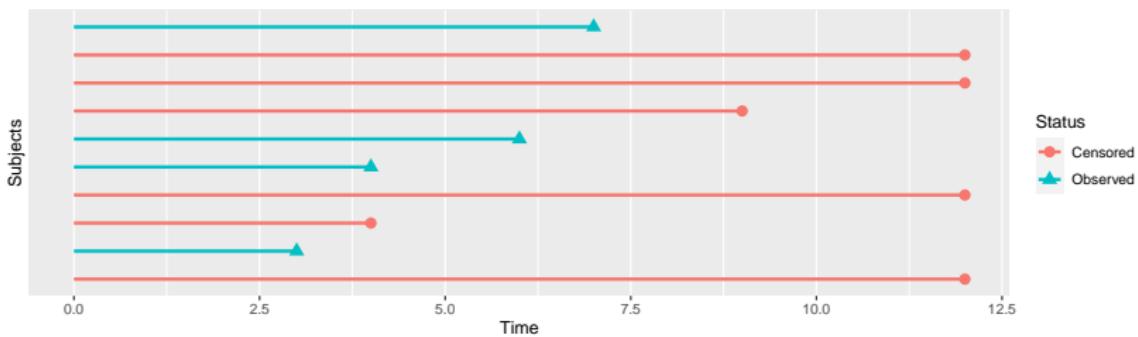
5 R

Type of predictions with time-to-event data



- **Classification problem:** only consider the outcome → 4 out of 10.
[loss of time info]
 - **Regression problem:** only consider the time → Mean survival time
[loss of info of 6 participants & biased results].

Type of predictions with time-to-event data



- **Classification problem:** only consider the outcome → 4 out of 10.
[loss of time info]
 - **Regression problem:** only consider the time → Mean survival time
[loss of info of 6 participants & biased results].

We need to consider the full **time-to-event** nature of the data.

Machine Learning definition (I)

Machine learning

Field of study in **artificial intelligence** concerned with the development and study of **statistical algorithms** that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions.¹

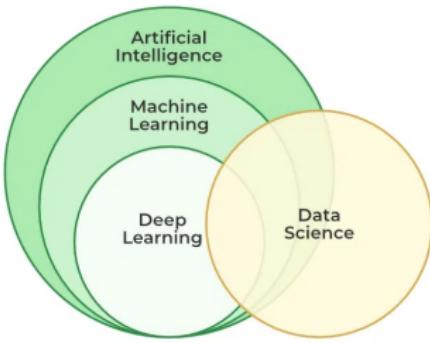
¹ Machine learning. Wikipedia

Figure source: <https://www.geeksforgeeks.org/ml-machine-learning/>

Machine Learning definition (I)

Machine learning

Field of study in **artificial intelligence** concerned with the development and study of **statistical algorithms** that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions.¹



¹Machine learning. Wikipedia

Figure source: <https://www.geeksforgeeks.org/ml-machine-learning/>

Machine Learning definition (II)

Machine Learning and Statistics in Clinical Research Articles—Moving Past the False Dichotomy

Samuel G. Finlayson, MD, PhD^{1,2}; Andrew L. Beam, PhD³; Maarten van Smeden, PhD⁴

» Author Affiliations

JAMA Pediatr. 2023;177(5):448-450. doi:10.1001/jamapediatrics.2023.0034

- Machine learning **depends directly** on statistics.
- Discussing **ML as a alternative to statistics** is equivalent to asking if an automobile is an alternative to its engine

- The **false statistics–ML dichotomy** has negative effects on medical research:
 - Branding an analysis as ML incentivizes some authors/reviewers to favor those methods even if they are not suited for the analysis.

Source: Machine Learning and Statistics in Clinical Research Articles—Moving Past the False Dichotomy.
Finlayson et al.

Machine Learning definition. Exercise (III)

Consider the following prediction models:

- a) A **simple linear regression** model
 - b) A large **regression model** with polynomials and interaction terms
 - c) A small **neural network** with one hidden layer
 - d) A 100-billion parameter **neural network**

Machine Learning definition. Exercise (III)

Consider the following prediction models:

- a) A **simple linear regression** model
 - b) A large **regression model** with polynomials and interaction terms
 - c) A small **neural network** with one hidden layer
 - d) A 100-billion parameter **neural network**

Questions

- ① Which models do you consider as machine learning models?

Machine Learning definition. Exercise (III)

Consider the following prediction models:

- a) A **simple linear regression** model
 - b) A large **regression model** with polynomials and interaction terms
 - c) A small **neural network** with one hidden layer
 - d) A 100-billion parameter **neural network**

Questions

- ① Which models do you consider as machine learning models?
 - ② Which two are more similar?

Machine Learning definition. Exercise (III)

Consider the following prediction models:

- a) A **simple linear regression** model
 - b) A large **regression model** with polynomials and interaction terms
 - c) A small **neural network** with one hidden layer
 - d) A 100-billion parameter **neural network**

Questions

- ① Which models do you consider as machine learning models?
 - ② Which two are more similar?

Solution: 1) (c) & (d); 2) (b) & (c)

Traditional methods for Survival Analysis

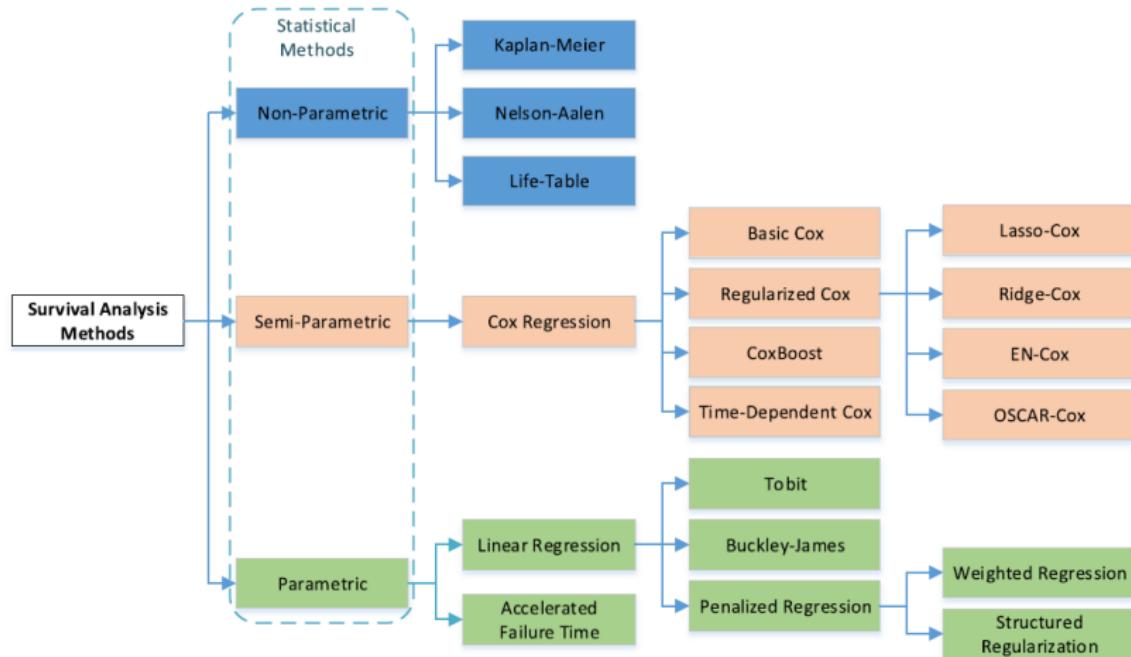


Figure source: Machine Learning for Survival Analysis: A Survey. Wang et al.

ML methods

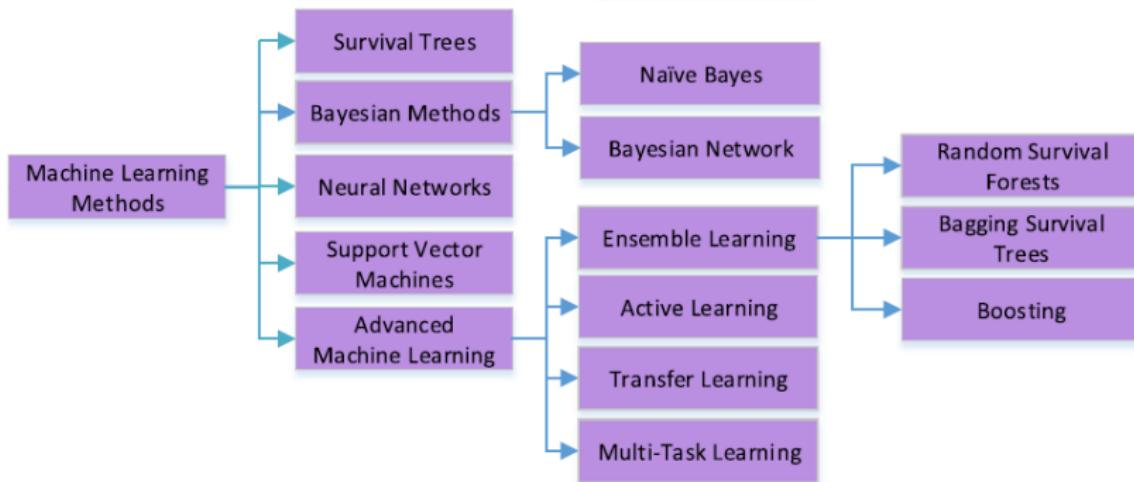


Figure source: Machine Learning for Survival Analysis: A Survey. Wang et al.

Explainability versus prediction

There are two main goals in the biomedical research.

Explainability

We want **to know the cause** of the outcome

Prediction

We want **to anticipate** the outcome

- Each goal leads to **different statistical methodologies** and **distinct models**.
- ML methods are more suitable for the **prediction** goal.

Cox Model: Why not?

- The **Cox model** is the most widely used model in the context of survival analysis.
- However....

...it has some **drawbacks**

- ① It assumes that hazard ratio is **constant** over time
- ② It is not applicable to data with **more features than samples**
- ③ It fails if features are highly **correlated**
- ④ Its decision function is usually **linear** in the covariates
- ⑤ It is not suitable in presence of high order **interactions**

Types of ML methods

- Trees
 - Random Survival Forest (after Survival Trees, Bagging Survival Trees)
- Support Vector Machine (SVM)
 - Ranking
- Deep Learning
 - Deep Surv
- Multi-task learning methods (Advanced ML methods)
 - Multi-Task Logistic Regression

Decision trees (discrete outcome)

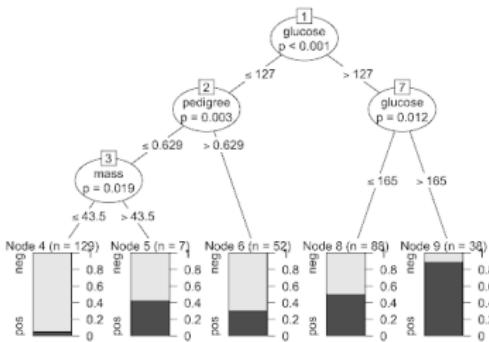
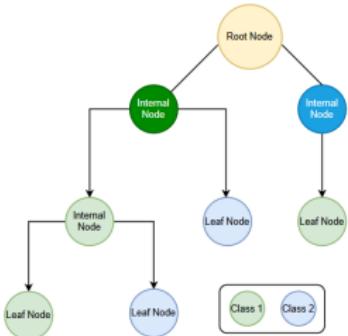
- **Decision tree** algorithm serves to predict a discrete outcome.
- Data is split into **internal/leaf nodes** according to the variables with more predictive capability on the outcome.
- **Split criterion:** maximize a weighted sum of **Gini Indexes** (G_j) of the probabilities (p_i) for the j -node.

$$G_j = 1 - \sum_{i=1}^n (p_{ij})^2$$

Decision trees (discrete outcome)

- **Decision tree** algorithm serves to predict a discrete outcome.
- Data is split into **internal/leaf nodes** according to the variables with more predictive capability on the outcome.
- **Split criterion:** maximize a weighted sum of **Gini Indexes** (G_j) of the probabilities (p_i) for the j -node.

$$G_j = 1 - \sum_{i=1}^n (p_{ij})^2$$



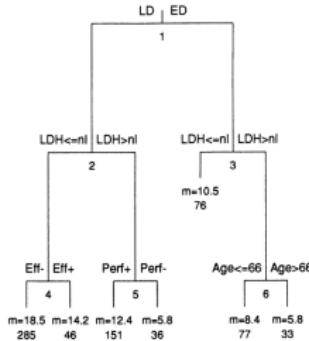
Survival Trees

- **Survival trees** have a survival outcome
- **Split criterion** and the **output** make the difference with the classic decision trees.
 - **Split criterion:** the **logrank test** (most common) or **brier score** (when censoring strongly depends on X).
 - **Output:** Any **survival outcome** (e.g., survival, hazard,...) may be calculated for a specific node participants.

Survival Trees

- **Survival trees** have a survival outcome
- **Split criterion** and the **output** make the difference with the classic decision trees.
 - **Split criterion:** the **logrank test** (most common) or **brier score** (when censoring strongly depends on X).
 - **Output:** Any **survival outcome** (e.g., survival, hazard,...) may be calculated for a specific node participants.

- The figure represents a **survival tree** for patients with lung cancer
- For each leaf node, the **survival curve** and the **median** survival (m) can be calculated.



Bagging Survival Trees

Drawback of survival trees: **variance**

Although they provide **unbiased** estimates, the estimators have **high variances** because of the splitting process. We reduce the problem of large variance by **bagging** several trees.

Bagging Survival Trees

Drawback of survival trees: **variance**

Although they provide **unbiased** estimates, the estimators have **high variances** because of the splitting process. We reduce the problem of large variance by **bagging** several trees.

Procedure:

- ① Draw B **bootstrap** samples from the original data.
- ② Grow a **survival tree** for each bootstrap sample based on **all** features.
- ③ Compute the bootstrap **aggregated survival function** for a new individual with characteristics X_i

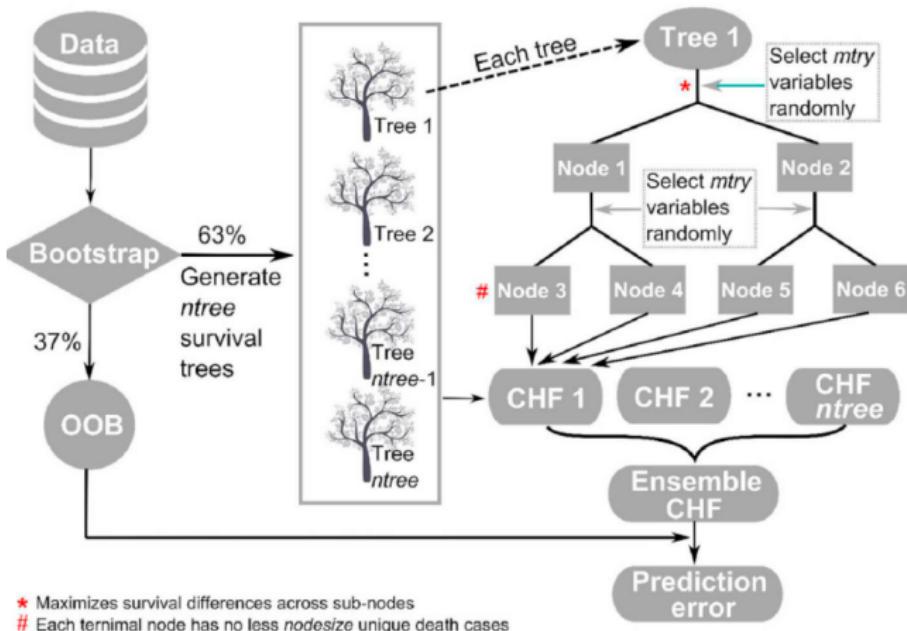
Random Survival Forest: Procedure

Random Survival Forests reduce variance even more than Bagging survival trees.

Procedure:

- ① Draw B **bootstrap** samples from the original data.
 - The part of the sample not included in each tree will be the **Out Of Bag** (OOB) data.
- ② Grow a **survival tree** for each bootstrap sample based on **randomly selected** `mtry` features for each node.
- ③ Grow the tree to **full size**.
- ④ Calculate the **Cumulative Hazard Function (CHF)** for each tree. Average them to obtain the **bootstrap ensemble CHF**.
- ⑤ Calculate the **prediction error** using the **OOB ensemble CHF**.

Random Survival Forest: schema



Source: Implementation of personalized medicine in malignant melanoma patients aided by Al. C. Hernandez.

Random Survival Forest: Pros and cons

- **Motivation:** ensemble learning can be improved further by injecting **randomness** into the learning process.
- **Main idea:** based on effective ***divide and conquer*** principle, i.e., sample fraction of the data; grow a randomized tree predictor on each sub dataset; aggregate the predictions.
- **Main advantage:** **interactions** in classic models must be identified by brute force or must rely on previous knowledge. SRFs automatically handle them in a non-parametric way.
- **Additional pros:** **simple** to use; can be applied to a **wide range** of prediction problems; **accurate** method; can deal with **high-dim data**; easily **parallelizable**.
- **Cons:** having a few tuning parameters with a **black-box** flavor.

Random Survival Forest: CHF

- For each tree and each *leaf* node (h), CHF is obtained via **Nelson-Aalen estimator**:

$$\hat{H}_h(t) = \sum_{t_{l,h} < t} \frac{d_{l,h}}{r_{l,h}}$$

where $t_{l,h}$ is the l^{th} distinct event time of the samples in leaf h . $d_{l,h}$ is the number events at $t_{l,h}$ and $r_{l,h}$ is the number of individuals at risk.

- Bootstrap CHF (**average of all bootstrap samples**):

$$\hat{H}_e^*(t|x_i) = \frac{1}{B} \sum_{b=1}^B \hat{H}_b^*(t|x_i) \quad (\text{prediction future } X)$$

- OOB CHF (**average over samples which $i \in \text{OOB}$**):

$$\hat{H}_e^{**}(t|x_i) = \frac{\sum_{b=1}^B I_{i,b} \hat{H}_b^*(t|x_i)}{\sum_{b=1}^B I_{i,b}} \quad (\text{prediction error})$$

- $\hat{H}_b^*(t|x_i)$ is the CHF of the node in b^{th} bootstrap which x_i belongs to.
- $I_{i,b} = 1$ if i is an OOB case for b .

Random Survival Forest: survival function

Tricky question

Might we estimate the **ensemble survival KM estimator?**

Random Survival Forest: survival function

Tricky question

Might we estimate the **ensemble survival KM estimator**?

- We know:

$$H(t) = -\log(S(t))$$

Random Survival Forest: survival function

Tricky question

Might we estimate the **ensemble survival KM estimator**?

- We know:

$$H(t) = -\log(S(t))$$

- It remains true for each **single tree**:

$$H_b(t) = -\log(S_b(t))$$

Random Survival Forest: survival function

Tricky question

Might we estimate the **ensemble survival KM estimator**?

- We know:

$$H(t) = -\log(S(t))$$

- It remains true for each **single tree**:

$$H_b(t) = -\log(S_b(t))$$

- But we know for **Jensen's inequality**:

$$-\log(E(S_b)) \leq E(-\log(S_b)) = E(H_b) \rightarrow E(H_b) \neq -\log(E(S_b))$$

Random Survival Forest: survival function

Tricky question

Might we estimate the **ensemble survival KM estimator**?

- We know:

$$H(t) = -\log(S(t))$$

- It remains true for each **single tree**:

$$H_b(t) = -\log(S_b(t))$$

- But we know for **Jensen's inequality**:

$$-\log(E(S_b)) \leq E(-\log(S_b)) = E(H_b) \rightarrow E(H_b) \neq -\log(E(S_b))$$

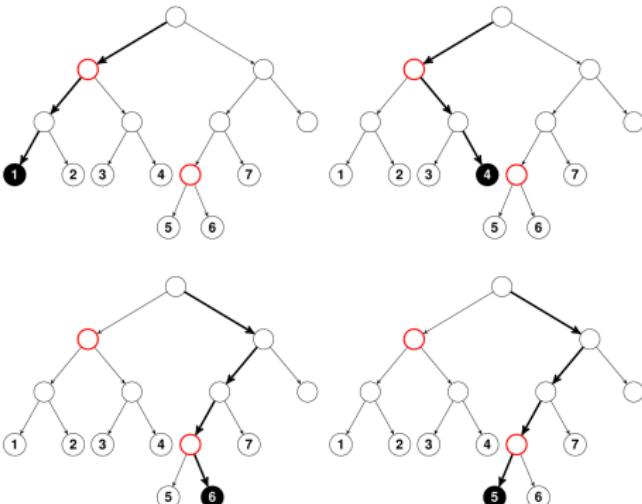
- Therefore, the question is **YES** and we might do it in a similar way to the CHF.

Random Survival Forest: Variable importance (I)

- In the OOB cases for a tree, randomly **permute** all values of the j^{th} variable.
- Put these new covariate values down the tree and compute a **new internal error rate**.
- The **amount by which this new error exceeds** the original OOB error is defined as the **importance** of the j^{th} variable for the tree.
- Averaging over the forest yields the **overall importance**.

Random Survival Forest: Variable importance

- Randomly **permuting** a variable j leads to different assignments.
- Red nodes are nodes that **split** on the target variable j
- Black nodes represent the **assignment** for a new data point.



- The higher variable j splits in the tree, the more an effect permutation has on **prediction error**.

Random Survival Forest: Hyperparameters

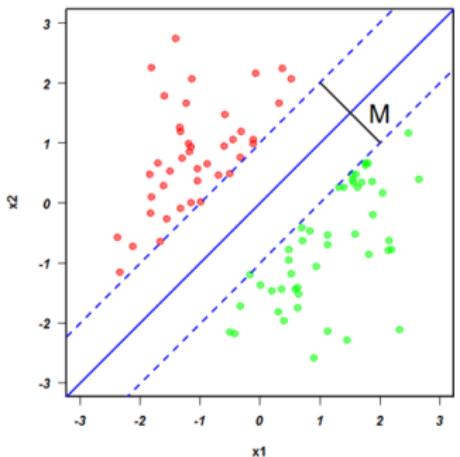
- `ntree` and `mtry` are the most important hyperparameters.
- `ntree` represents the number of trees to be included in the model. **It should not be tuned**: use as many as you can.
- `mtry` represents the number of candidate variables to use in each node.
 - It will largely determine the performance of the model.
 - Some guidelines recommend $p/3$ for continuous outcome and \sqrt{p} for classification. No general advice for survival outcome.
 - Low values favor the inclusion of all variables in each tree.
High values lead to more robust trees.
- Other hyperparameters are less relevant: `nodesize`, `nodedepth`, `splitrule`, `nsplit`, ...

Support Vector Machine: Scenarios (binary outcome)

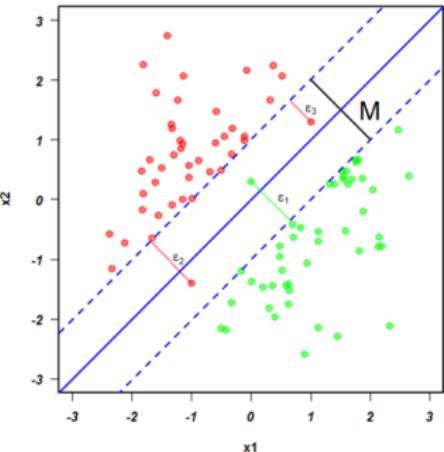
Goal SVM

Given a training set whose classes are known, find a hyperplane that optimally separates the points.

Linearly separable



Soft margins



Support Vector Machine: Scenarios (binary outcome)

• Linear separable scenarios

- There are **infinite hyperplanes** that can properly separate classes.
- We have to choose the one that **maximizes the distance M** .
- The solution is found by a **quadratic optimization problem** with linear constraints.
- It only depends on the **support** points that are on the margin.

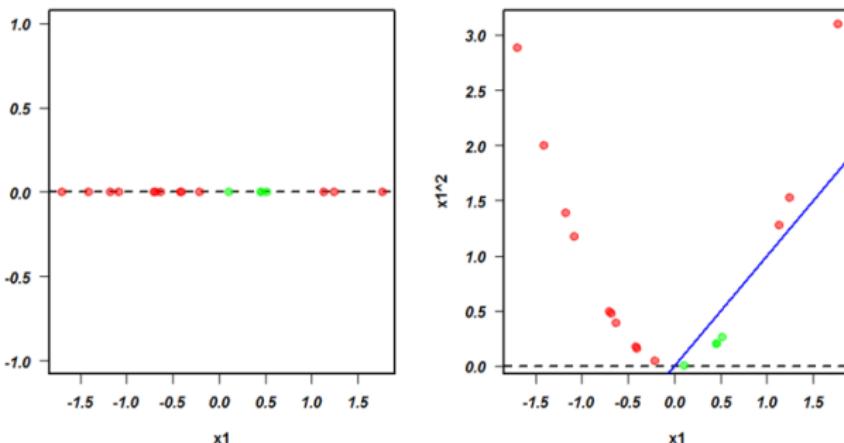
• Non-linear separable scenarios

- The generalization of the previous scenario is one in which certain points are allowed to **violate the margin**
- This is recommended, not only when there is not a separable scenarios, but to **avoid overfitting**
- We have to set the **total allowed amount of overlapping** with margins.
- The solution depends on the **support** points on and within the margin.

Kernel SVM: Mapping in higher dimensions (binary)

- If there is no linearly separable scenario, the dimension of the original space can be increased to get the goal. E.g.:

$$\Phi : R \rightarrow R^2 \quad \text{where} \quad \Phi(x) = (x, x^2)$$



- The mapping is often performed by using Kernel functions.

Survival SVM: Approaches

Three approaches have been proposed for survival SVMs:

- **Ranking**². The output is the **ranking** of a new set of observations regarding their survival time.
- **Regression**³. The output is the **mean survival time**.
- **Hybrid approach**⁴

²Support vector machines for survival analysis. *Van Belle*

³A support vector approach to censored targets. *Shivaswamy*

⁴Support vector methods for survival analysis: comparison between ranking and regression approaches. *Van Belle*

Survival SVM: Ranking (I)

- It aims at **predicting risk ranks** between individuals instead of estimating survival times.
- In training set, only some individuals will be **comparable** (see later)
- **Optimization Problem:**

Objective function to **minimize**

$$f(w) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i,j \in P} \max(0, \xi_{ij})^2$$

where w are the hyperplane coefficients; $\xi_{ij} = 1 - (w^T x_i - w^T x_j)$; x_i, x_j are the features of the individuals i and j ; P is the set on **comparable** pairs (i, j) where we know the survival time is **higher** in i ; and γ is a penalization factor.

Survival SVM: Ranking (II)

$$f(w) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i,j \in P} \max(0, \xi_{ij})^2$$

- The first term accounts for **parsimonious hyperplanes**
- The sum terms penalize **wrong classifications**.
 - Each addend takes a **zero** value when observations are well ranked according to hyperplane
- The concept is similar to **penalized regressions**.
- Each term of the sum involves a **pair of points**.
- New observations x_{new} will be ranked according to **the value of the $x_{new}w$** .

Survival SVM: Ranking (III)

The penalization term:

$$\frac{\gamma}{2} \sum_{i,j \in P} \max(0, \xi_{ij})^2$$

- Only terms where **survival time is higher in i** are added:
 - $P = \{(i, j) | T_i > T_j, \delta_j = 1\}$
- Well classified pairs **contribute nothing** to that term.
 - If $w^T x_i - w^T x_j > 1$, there is no penalization.
- In time-to-event analysis, the SVM algorithm built the hyperplane based on **pairs of comparable points**.

Survival SVM: Remarks

- Use when $p \gg n$. Computation time is $O(n^3)$ and $O(p)$.
- Usually, pure **ranking-based** approaches outperform regression and hybrid models.
- The main drawback of ranking-based techniques is it becomes intractable with large **datasets** (objective function with a quadratic number of terms regarding to the sample size)
- The penalization factor γ is an **overfitting** parameter: the higher the penalization, the greater the chances of overfitting.

Survival SVM: Hyperparameters

- `gamma.mu` and `kernel` are the most important.
 - `gamma.mu` weights the penalization term. It represents a parameter of overfitting.
 - `kernel` is the function to map the observations in higher dimensions.
- Other hyperparameters are less relevant: `opt.meth`

Deep learning: Neural networks (I)

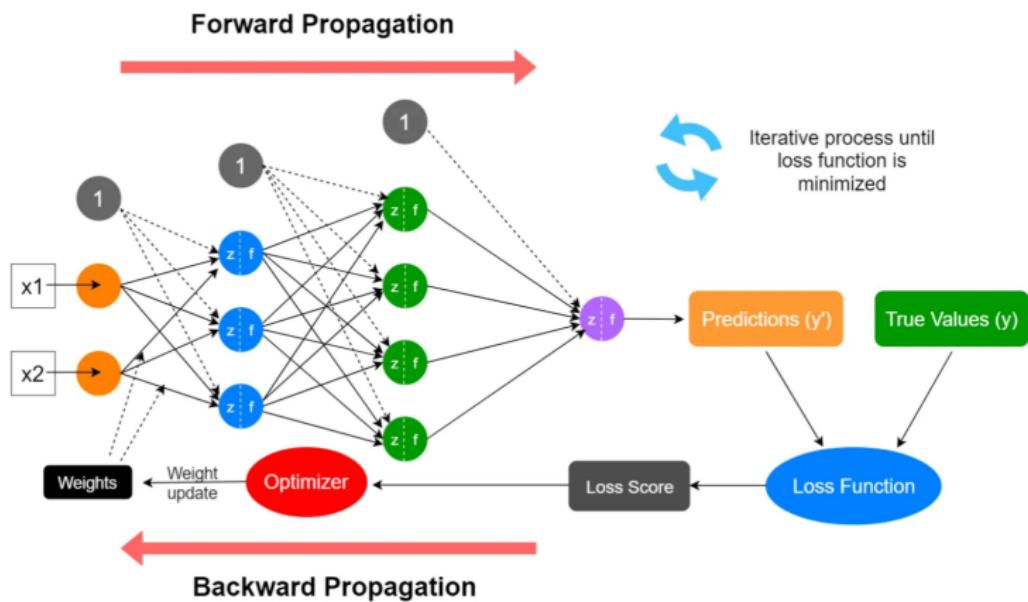


Figure source: Overview of a Neural Network Learning Process. Medium

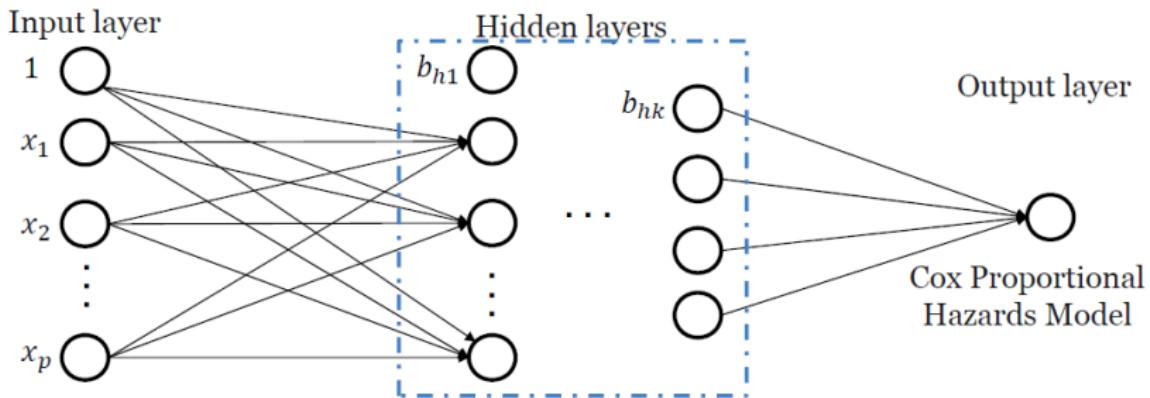
Deep learning: Neural networks (II)

- Neural networks try to imitate human **brain**.
- They have a series of **nodes** (*neurons*) organized in **layers**.
- Each neuron receives a series of computations and produces and output to the next layer.
- In the computations are involved the **bias term** (b); the **weights** (w_i); the **input values** x_i and the non-linear **activation function** $g(\cdot)$:

$$y = g(b + w_i \cdot x_i)$$

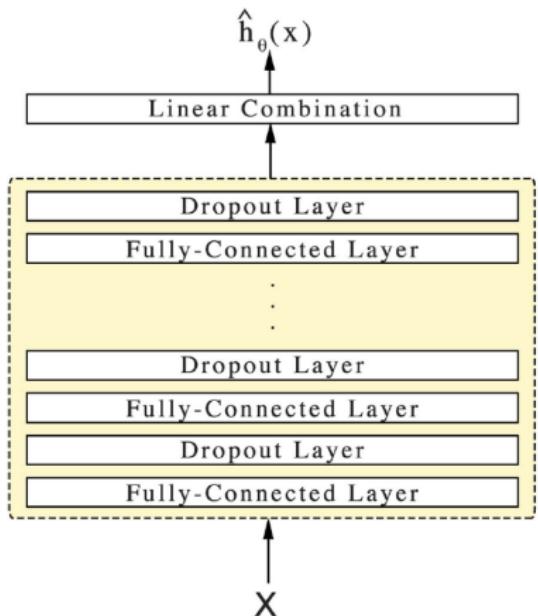
- Three types of layers: **Input**, **Output** and **Hidden** layers.
- There are two stages:
 - **Forward propagation**. To calculate the output.
 - **Backward propagation**. To minimize the error updating the weights.

Deep Surv: Basis



DeepSurv: personalized treatment recommender system using a Cox PH deep neural network. Katzam et al.

Deep Surv: Diagram



- Input → **baseline data** x
- Inputs are propagated by the hidden layers with **weights** θ .
- The hidden layers are fully-connected **nonlinear activation functions** with dropout.
- Last layer performs a linear combination of **hidden features**.
- Output → predicted **log-risk** $\hat{h}_\theta(x_i)$ for each individual i .
- Loss function → negative **log partial-likelihood** (with regularization)
- **Hyperparameters:** no. hidden layers, no. nodes in layers, activation function. They should

Deep Surv vs. Cox Model

• Proportional Cox Model (C)

- Model

$$\lambda_C(t|x_i) = \lambda_0(t) \cdot e^{h_\beta(x_i)}$$

- Partial Likelihood function

$$L_C(\beta) = \prod_{i:E_i=1} \frac{e^{\hat{h}_\beta(x_i)}}{\sum_{j \in \mathbb{R}(T_i)} e^{\hat{h}_\beta(x_j)}}$$

• Deep Surv (D)

- Model

$$\lambda_D(t|x_i) = \lambda_0(t) \cdot e^{h_\theta(x_i)}$$

- Objective function (negative penalized partial log-likelihood):

$$l_D(\theta) := -\frac{1}{N_{E=1}} \sum_{i:E_i=1} \left(\hat{h}_\theta(x_i) - \log \left(\sum_{j \in R(T_i)} e^{\hat{h}_\theta(x_j)} \right) \right) + \lambda \cdot \|\theta\|_2^2$$

where $N_{E=1}$ is the number of patients with an observable event and λ is the regularization

Deep Surv: Hyperparameters

- The **activation** and the **num_nodes** are important hyperparameters.
- The **activation** parameter represents the activation function. The most common are **ReLU** (Rectified linear unit) and **SELU** (Scaled exponential linear unit).
 - ReLU $\rightarrow g(x) = \max(0, x)$
 - SELU $\rightarrow g(x) = \begin{cases} \lambda x & x \geq 0 \\ \lambda\alpha(e^x - 1) & x < 0 \end{cases}$
- The **num_nodes** parameter is the numbers of nodes per layer. It usually goes from a few units (~ 5) to fewer tens (~ 50)

MTLR: Introduction

- Multi Tasks Logistic regression models the survival distribution via a sequence of **dependent regressions**.
- The **logistic regression** can model the probability of surviving more than a given point time t :

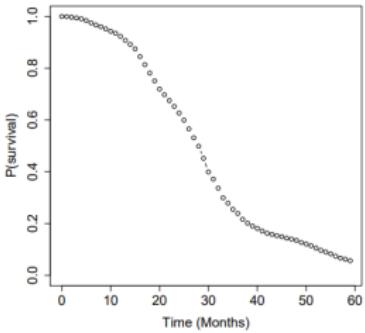
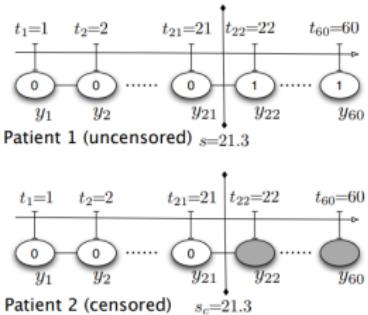
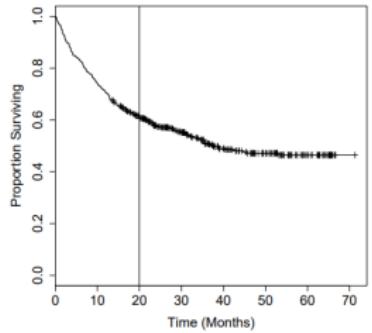
$$P_{\theta}(T \geq t|x) = \frac{1}{e^{\theta x + b}}$$

- This probability can be estimated at **several time points**, $\tau = (t_1, \dots, t_m)$:

$$P_{\theta_i}(T \geq t_i|x) = \frac{1}{e^{\theta_i x + b_i}} \quad 1 \leq i \leq m$$

- The outputs of these models **are not independent**, as an event at or before time t_i implies the event at all subsequent time points t_j for all $j > i$.

MTLR: Schema



Learning patient-specific cancer survival distributions as a sequence of dependent regressors. Yu et al

MTLR: Formulation

- MTLR enforces the **dependency** of the outputs by predicting the survival status of a patient at each of the time snapshots t_i **jointly** instead of independently.

$$P_{\Theta}(Y = (y_1, \dots, y_m) | x) = \frac{e^{\sum_{i=1}^m y_i(\theta_i \cdot x + b_i)}}{\sum_{k=0}^m e^{f_{\theta}(x, k)}}$$

where $\Theta = (\theta_1, \dots, \theta_m)$ and $f_{\theta}(x, k) = \sum_{i=k+1}^m y_i(\theta_i \cdot x + b_i)$

- The **log likelihood** to maximize:

$$\sum_{i=1}^n \left[\sum_{j=1}^m y_j(s_i)(\theta_j \cdot x_i + b_j) - \log \sum_{k=0}^m e^{f_{\Theta}(x_i, k)} \right]$$

- In fact, a **penalization term** is added to this expression to prevent overfitting because we are estimating $2mn$ parameters (n : no. patients, m : no. time points).

MTLR: Advantages

- It models the more intuitive **survival function** rather than the hazard function, avoiding the election of different forms of hazards.
- We can capture the time-varying effects of features and handle censored data easily and naturally.
- This model provides more **accurate predictions** on survival and better calibrated probabilities than other models.

Summary of presented models

Model	Concept	Output
ST	Single Survival tree	Survival
BST	Several survival trees	CHF
RFST	Several random survival trees	CHF
SVMR	Hyperplane to separate by rankings	Rankings
DS	Neural Networks	Hazard function
MTLR	MTLR	Survival

Comparable pairs (I)

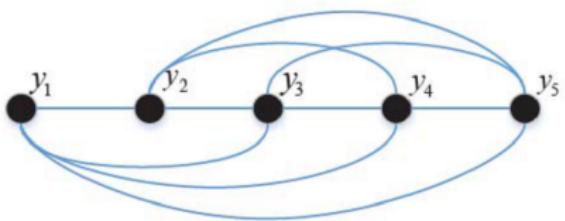
- Two observed times t_i and t_j are **comparable** if we can deduct which event would occur earlier. That is, they are comparable if:
 - Both are uncensored
 - The uncensored time is lower than the censored time

δ_i	δ_j	Times	Class
0	0	$t_i > t_j$	Non Comparable
1	0	$t_i > t_j$	Non Comparable
0	1	$t_i > t_j$	Comparable
1	1	$t_i > t_j$	Comparable

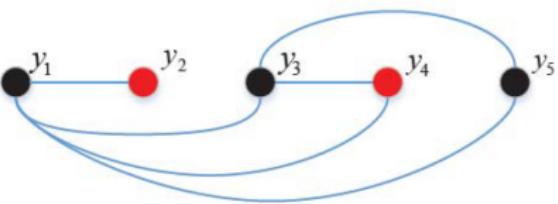
Comparable pairs (II)

Simple rule

If the lower observed time is uncensored, then, the pair is **comparable**



Without censoring
(10 comparable pairs)



With censoring
(6 comparable pairs)

Harrell's Concordance Index (C-Index) (I)

- **Rank order statistic** for comparing predictions against true outcomes.
- **Given a time t** and the comparable instance pair (i, j) with real observed times t_i and t_j , and predicted survival times $\hat{S}_i(t)$ and $\hat{S}_j(t)$, then:
 - The pair (i, j) is **concordant** if $t_i > t_j$ and $\hat{S}_i(t) > \hat{S}_j(t)$
 - The pair (i, j) is **discordant** if $t_i > t_j$ and $\hat{S}_i(t) < \hat{S}_j(t)$
- The **concordance Index** measures the agreement between the rankings of actual and predicted times:
$$C(t) = Pr(\hat{S}_i(t) > \hat{S}_j(t) | t_i > t_j)$$
- It can be measured as the **ratio of the concordant pairs to the total**.
- C-index is equivalent to **AUC** with binary outcomes.

Harrell's Concordance Index (C-Index) (II)

- When the output of the model is the prediction of **survival time** (n_c are the number of comparable pairs):

C-index for survival

$$C(t) = \frac{1}{n_c} \sum_{i:\delta_i=1} \sum_{y_i < y_j} I [S_i(t|X_i) < S_j(t|X_j)]$$

- When the output of the model is the **hazard ratio** provided by the Cox model, the index does not depend on time:

C-index for Cox model

$$C = \frac{1}{n_c} \sum_{i:\delta_i=1} \sum_{y_i < y_j} I [X_i \hat{\beta} > X_j \hat{\beta}]$$

Harrell's Concordance Index (C-Index) (III)

- The C-index during a **time period** $(0, t^*)$ can be calculated aggregating the information for the subintervals contained in the interval:

$$C_{(0,t^*)} = \sum_{t \in T_S} \frac{1}{n_{c_t}} \sum_{i: \delta_i=1} \sum_{y_i < y_j} I [S_i(t|X_i) < S_j(t|X_j)]$$

- It represents a **weighted average** of the time-specific C-index:

$$C_{(0,t^*)} = \sum_{t \in T_S} \frac{1}{n_{c_t}} C(t)$$

- It provides a measure of the **discriminant** capability of our method.

Brier Score (I)

- **Brier score** is used when the outcome is either binary or categorical.
- The **individual contributions** to the empirical Brier score are reweighted based on the censoring information:

Brier Score

$$BS(t) = \frac{1}{N} \sum_{i=1}^N w_i(t) \cdot (I[t_i > t] - \hat{S}(t|x_i))^2$$

$w_i(t)$ denotes the weight for the i^{th} instance.

Brier Score (II)

- The weights can be estimated by considering the Kaplan-Meier estimator of the censoring distribution G on the dataset.

$$w_i(t) = \begin{cases} \delta_i/G(y_i) & y_i \leq t \\ 1/G(y_i) & y_i > t \end{cases}$$

- Weights for censored instances before t will be 0.
- Weights for uncensored instances at t will be greater than 1.
- A useful model will have a Brier score below **0.25**.
- The Brier score is a measure of the overall performance that incorporates both **discrimination** and **calibration**.
- The **Integrated Brier Score** (IBS) is a global measure obtained integratint over all the times.

$$IBS = \int_0^{\tau} BS(t)dt$$

Mean Absolute Error (MAE)

- The mean absolute error (MAE) is the **average of the differences** between the predicted and the actual time values.

MAE

$$MAE = \frac{1}{n} \sum_{i=1}^N (\delta_i \cdot |y_i - \hat{y}_i|)$$

where:

- y_i are the actual observation times.
- \hat{y}_i are the predicted times.

- Condition:** MAE can only be used for the evaluation of survival models which can **provide the event time** as the predicted target value. Only the samples for which the event occurs are considered.

Advices (I)

Some (general) ML advices:⁵

① Before you start to build models

- Do take the time to understand your data. **Garbage in garbage out**
- Do make sure you have enough data. **Signal to noise ratio**

② How to reliably build models

- Don't allow test data to leak into the training process. **E.g., Standardization**
- Do try out a range of different models. **Is this model appropriate for my data?**
- Do optimise your model's hyperparameters. **Optimisation strategy**

⁵How to avoid machine learning pitfalls. *Lones*

Advices (II)

③ How to robustly evaluate models

- Don't do data augmentation before splitting your data.
Balance train set
- Do use a validation set. **If you use the test set to improve your model**
- Don't use accuracy with imbalanced data sets. **Naive accuracy**

④ How to compare models fairly

- Do use statistical tests when comparing models. **McNemar's test**
- Do consider combinations of models **Ensemble learning**

⑤ How to report your results

- Do be transparent. **Reproducibility**
- Do report performance in multiple ways. **False positives or negative**

Pitfalls (I)

- Several systematic reviews authored by **Gary Collins**

Journal of Clinical Epidemiology | Available online 12 March 2017 | In Press, Journal Pre-proof | What's New |

Title: Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review

Authors: Pedro Domingo^{1,2*}, Joaquin A. de la Torre^{1,2}, Constanza L. Andrade-Nunez^{1,2}, Beatrice Sennich^{1,2}, Gianfranco Lovisolo^{1,2}, Daniel M. Gitterman³, Luisa Rofito¹, Steona Kortes¹, Richard D. Meier⁴, Karen Laih-Morris⁵, Steven J. Lofthus⁶, ...

How to cite: <https://doi.org/10.1016/j.jclinepi.2017.03.002> | [Get rights and content](#)

Journal of Clinical Epidemiology | Available online 1 April 2005 | In Press, Journal Pre-proof | What's New |

Title: Systematic review finds "Spin" practices and poor reporting standards in studies on machine learning prediction models

Authors: Constanza L. Andrade-Nunez (Primary author),^{1,2} Jordi Cortés Martínez (Corresponding author),^{1,2} ...

How to cite: <https://doi.org/10.1016/j.jclinepi.2005.01.002> | [Get rights and content](#)

Journal of Clinical Epidemiology | Available online 1 April 2005 | In Press, Journal Pre-proof | What's New |

Title: Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review

Authors: Pedro Domingo^{1,2*}, Joaquin A. de la Torre^{1,2}, Constanza L. Andrade-Nunez^{1,2}, Benjamin Specht^{1,2}, Gaston Bullock¹, ...

How to cite: <https://doi.org/10.1016/j.jclinepi.2005.01.003> | [Get rights and content](#)

BMC Medical Research Methodology | Available online 28 February 2018 | Open Access |

Title: Reporting of prognostic clinical prediction models based on machine learning methods: a critical review to be implemented

Authors: Pedro Domingo^{1,2*}, Joaquin A. de la Torre^{1,2}, Steona Kortes¹, Loris Hooft¹, Richard D. Meier⁴, ...

How to cite: <https://doi.org/10.1186/s12874-018-0477-y> | [Get rights and content](#)

Journal of Clinical Epidemiology | Available online 10 October 2015 | Open Access |

Title: Risk of bias of prognostic models developed using machine learning: a systematic review

Authors: Pedro Domingo^{1,2*}, Joaquin A. de la Torre^{1,2}, Constanza L. Andrade-Nunez^{1,2}, Benjamin Specht^{1,2}, Gaston Bullock¹, ...

How to cite: <https://doi.org/10.1016/j.jclinepi.2015.09.012> | [Get rights and content](#)

Journal of Clinical Epidemiology | Available online 10 October 2015 | Open Access |

Title: Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review

Authors: Constanza L. Andrade-Nunez^{1,2*}, Jordi Cortés Martínez^{1,2}, Steona Kortes¹, Richard D. Meier⁴, ...

How to cite: <https://doi.org/10.1016/j.jclinepi.2015.09.013> | [Get rights and content](#)

BMC Medical Research Methodology | Available online 10 October 2015 | Open Access |

Title: Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review

Authors: Constanza L. Andrade-Nunez^{1,2*}, Jordi Cortés Martínez^{1,2}, Steona Kortes¹, Steven M. Higgins^{1,2}, ...

How to cite: <https://doi.org/10.1186/s12874-015-0170-0> | [Get rights and content](#)

Diagnostic and Prognostic Research | Available online 10 October 2015 | Open Access |

Title: Risk of bias of prognostic models developed using machine learning: a systematic review

Authors: Pedro Domingo^{1,2*}, Joaquin A. de la Torre^{1,2}, Constanza L. Andrade-Nunez^{1,2}, Benjamin Specht^{1,2}, Gaston Bullock¹, ...

How to cite: <https://doi.org/10.1186/s13670-015-0012-z> | [Get rights and content](#)

Pitfalls (II)

- No **sample size** justification (92%)
- High **risk of bias** according to PROBAST scale (87%)
- Comparisons of discriminant measures between **developed models** (78%)
- No model complete **available** (>50%)
- Handled **missing data** inadequately (41%)
- **Categorize** continuous predictors (39%)
- Incorrect assessment of **overfitting** (39%)
- **Inconsistent reporting** between methods and results (27%)

TRIPOD guidelines

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

RESEARCH METHODS AND REPORTING



OPEN ACCESS



Check for updates

TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods

Gary S Collins,¹ Karel G M Moons,² Paula Dhiman,¹ Richard D Riley,^{3,4} Andrew L Beam,⁵ Ben Van Calster,^{6,7} Marzyeh Ghassemi,⁸ Xiaoxuan Liu,^{9,10} Johannes B Reitsma,² Maarten van Smeden,² Anne-Laure Boulesteix,¹¹ Jennifer Catherine Camaradou,^{12,13} Leo Anthony Celi,^{14,15,16} Spiros Denaxas,^{17,18} Alastair K Denniston,^{4,9} Ben Glocker,¹⁹ Robert M Golub,²⁰ Hugh Harvey,²¹ Georg Heinze,²² Michael M Hoffman,^{23,24,25,26} André Pascal Kengne,²⁷ Emily Lam,¹² Naomi Lee,²⁸ Elizabeth W Loder,^{29,30} Lena Maier-Hein,³¹ Bilal A Mateen,^{17,32,33} Melissa D McCradden,^{34,35} Lauren Oakden-Rayner,³⁶ Johan Ordish,³⁷ Richard Parnell,¹² Sherri Rose,³⁸ Karandeep Singh,³⁹ Laure Wynants,⁴⁰ Patricia Logullo¹

Survival analysis reporting guidelines: gap



Survival analysis reporting guidelines: gap



- Is there any guideline specific for reporting ML survival analysis?

Survival analysis reporting guidelines: gap



- Is there any guideline specific for reporting ML survival analysis? **NO**
- Is there any guideline specific for reporting survival analysis?

Survival analysis reporting guidelines: gap



- Is there any guideline specific for reporting ML survival analysis? **NO**
- Is there any guideline specific for reporting survival analysis? **NO** (almost sure)

Survival analysis reporting guidelines: gap



- Is there any guideline specific for reporting ML survival analysis? **NO**
- Is there any guideline specific for reporting survival analysis? **NO** (almost sure)
- **But...There is a draft on how reporting survival analysis**

Survival analysis reporting guidelines: gap



- Is there any guideline specific for reporting ML survival analysis? **NO**
- Is there any guideline specific for reporting survival analysis? **NO** (almost sure)
- **But...There is a draft on how reporting survival analysis**

**Basic Statistical Reporting for
Articles Published in Biomedical Journals:
The “Statistical Analyses and Methods
in the Published Literature” or
The SAMPL Guidelines”**

R packages for survival (I)

Algorithm/Model	Package	Program
Kaplan-Meier	survival	R
Nelson-Aalen	survival	R
Life-Table	survival	R
Basic Cox	survival	R
TD-Cox	survival	R
Lasso-Cox	fastcox	R
Ridge-Cox	fastcox	R
EN-Cox	fastcox	R
Oscar-Cox	RegCox	R
CoxBoost	CoxBoost	R
Tobit	survival	R
BJ	bujar	R
AFT	survival	R

R packages for survival (II)

Algorithm/Model	Package	Program
Bayesian Methods	BMA	R
RSF	randomForestSRC	R
Survival SVM	survivalsvm	R
BST	ipred	R
Boosting	mboost	R
Active Learning	RegCox	R
Deep Surv	survivalmodels	R
Multi-Task Logistic Regression	MTLR	R
Early Prediction	ESP	R
Uncensoring	ESP	R
Calibration	survutils	R
Competing Risks	survival	R
Recurrent Events	survrec	R
survXgboost	survXgboost	R

Figure source: Machine Learning for Survival Analysis: A survey. Wang et al. (modified)

References (I)

- ① E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 1999.
- ② WANG, Ping; LI, Yan; REDDY, Chandan K. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 2019, 51.6: 1-36.
- ③ LeBlanc, M. et al. (1993). Survival Trees by Goodness of Split. *Journal of the American Statistical Association* 88, 457–467.
- ④ H. Ishwaran, U. B. Kogalur, E. H. Blackstone and M. S. Lauer, "Random Survival Forests". *Annals of Applied Statistics*, 2008
- ⑤ P. K. Shivaswamy, W. Chu, and M. Jansche. A support vector approach to censored targets. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 655–660, 2007

References (II)

- ⑥ V. Van Belle, K. Pelckmans, J. Suykens, and S. Van Huffel. Support vector machines for survival analysis. In Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007), pages 1–8, 2007.
- ⑦ V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. K. Suykens. Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artificial Intelligence in Medicine, 53(2):107–118, 2011.
- ⑧ Lones MA. 2021. How to avoid machine learning pitfalls: a guide for academic researchers. Arxiv: 2108.02497.
- ⑨ Katzman, J.L., Shaham, U., Cloninger, A. et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 18, 24 (2018). doi: 10.1186/s12874-018-0482-1

References (III)

- ⑩ Lee C, Zame WR, Yoon J, van der Schaar M. 2018. DeepHit: a deep learning approach to survival analysis with competing risks. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, pp. 2314–21. Palo Alto, CA: AAAI.
- ⑪ C. Yu et al. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. NIPS 2011.
- ⑫ E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine, 1999.