Identification of worse prognosis Covid-19 patients

A case study using targeted learning

Vittorio Perduca on behalf of several collaborators

MODCOV-19, 29 June 21

Overview

Our goals:

1. develop a scoring rule to predict, within two days after admission, the outcome of Covid-19 patients
2. define and estimate the importance measures of covariates that could explain the evolution of the disease

A collaboration between medical doctors and mathematicians:

- MAP5: Antoine Chambaz, Etienne Birmelé, VP
- Hôpital Bicêtre, APHP: Benjamin Wyplosz, Antoine Meyer, Catherine Dong

Table of Contents

Table of Contents

Dataset

- data collected retrospectively at Hôpital Bicêtre during the very first weeks of the pandemics
- $n = 209$ patients
- $p \sim 50$ covariates, including
  - demographics: gender, age...
  - comorbidities
  - symptoms: fever, digestive disorders...
  - vital signs: temperature, oxygen saturation and flow, respiration and heart rate...
  - laboratory findings: white blood cell and lymphocytes counts, C-reactive protein...
  - radiological findings: extension of lesions and other results of the annotation by experts
  - likely time since infection
- **repeated measures** for many covariates

Data summary

Characteristic	**Overall**, N = 210[1]	**Good**, N = 153[1]	**Poor**, N = 57[1]	**p-value**[2]
Demographics				
Female gender	94 (45%)	72 (47%)	22 (39%)	0.3
Age, years	62 (51, 77)	60 (48, 70)	78 (65, 88)	<0.001
Smoker	7 (3.3%)	3 (2.0%)	4 (7.0%)	0.088
Time before outcome				
Length of stay, days	3.0 (2.0, 6.0)	4.0 (2.0, 6.0)	2.0 (1.0, 4.0)	
Duration of infection, days	10.0 (7.0, 13.0)	10.0 (8.0, 14.0)	8.0 (6.0, 11.0)	
Vital signs				
Respiration rate	25 (20, 30)	24 (20, 30)	28 (20, 35)	0.2
Heart rate, bpm	92 (82, 103)	92 (81, 102)	92 (86, 107)	0.3
Temperature, C°	37.80 (37.15, 38.60)	37.90 (37.10, 38.55)	37.60 (37.20, 38.70)	0.7
Oxygen saturation	95.0 (93.0, 97.0)	96.0 (94.0, 97.0)	93.0 (88.0, 95.0)	<0.001
Laboratory findings				
C-reactive protein, mg/L	68 (32, 128)	55 (26, 120)	110 (67, 152)	<0.001
Lactate dehydrogenase, U/L	331 (260, 437)	291 (248, 378)	430 (344, 512)	<0.001
Lymphocyte count, x 10^9/L	1.00 (0.70, 1.39)	1.06 (0.79, 1.47)	0.80 (0.57, 1.09)	<0.001

[1] n (%); Median (IQR)

[2] Pearson's Chi-squared test; Wilcoxon rank sum test; Fisher's exact test

Our classification problem

The main goal is to predict the outcome (good or poor) of patients *within two days* after admission:

- **Poor:** death or transfer to ICU
- **Good:** release from the hospital or transfer to day care units

Formalization of the problem:

- given the vector of covariate $x \in \mathcal{X}$ of a patient, we want to predict her outcome y based on a score $S(x) \in [0, 1]$
- the smaller $S(x)$, the less likely y is poor
- S is the output of an algorithm $\widehat{\mathcal{A}}$ trained on the data
- to train and test $\widehat{\mathcal{A}}$ we randomly split the $n = 209$ data points into training ($n_{tr} = 149$) and testing ($n_{te} = 60$) samples

Table of Contents

Super learning I

- instead of arbitrarily choosing a classification algorithm we rely on super learning, $\widehat{\text{SL}}$ [van der Laan 2007]
- $\widehat{\text{SL}}$ evaluates the performance of all algorithms in a user-specified library based on V-fold cross-validation
- $\widehat{\text{SL}}$ either elects one of the base algorithms (discrete SL) or creates a convex combination thereof (continuous SL)
- under mild conditions the discrete SL is asymptotically as accurate as the best base algorithm in the library

Super learning II

- we rely on the `SuperLearner R` package [Polley] tweaked to deal with missing values and allow data enrichment
- we choose the Area Under the Curve (AUC) to assess the performance of each algorithm
- we choose to rely with the discrete SL because
  - our training dataset is limited and we want to compare many algorithms
  - the discrete SL is more interpretable

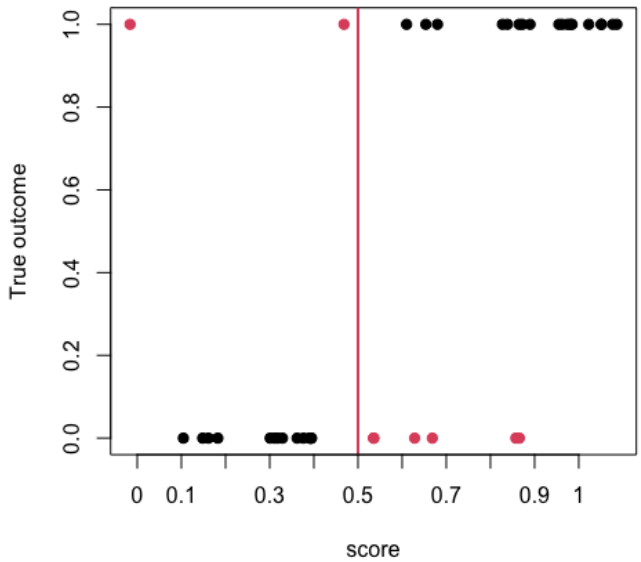
Performance measures

Reminders I

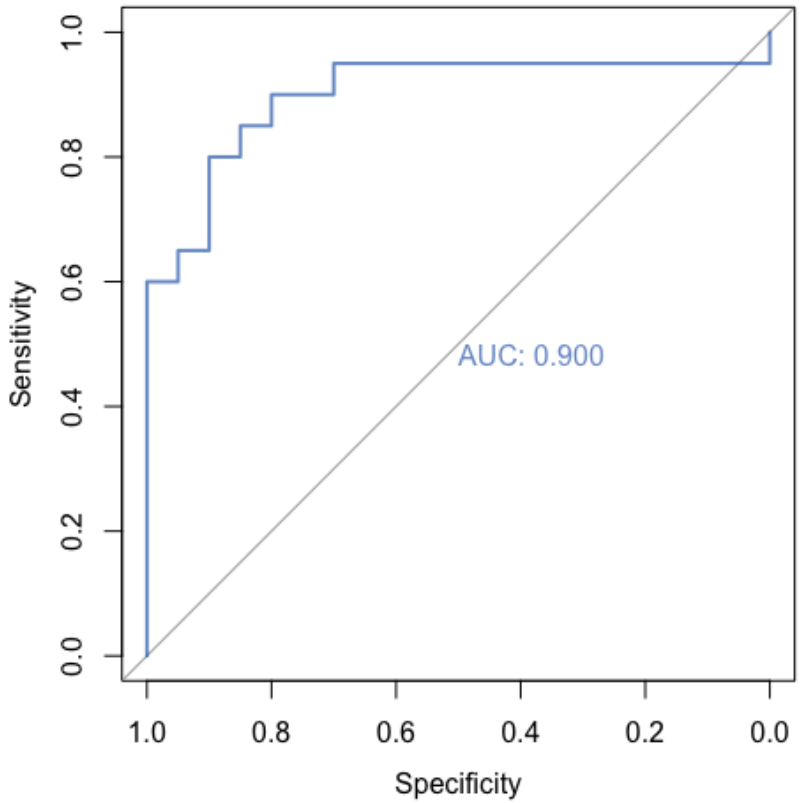
- classification algorithms generally produce a mapping $x \mapsto S(x)$ from $\mathcal{X}$ to $[0, 1]$
- classification boundary between classes is determined by a threshold T
- for each T we can tabulate a confusion matrix from labeled data:

		True classes	
		good	poor
Predicted classes	good	n_{00}	n_{01}
	poor	n_{10}	n_{11}

- sensitivity $= \frac{n_{11}}{n_{01}+n_{11}}$ (*true positive rate*, where here positive is poor)
- specificity $= \frac{n_{00}}{n_{00}+n_{10}}$ (*true negative rate*)
- positive predictive value $= \frac{n_{11}}{n_{10}+n_{11}}$
- negative predictive value $= \frac{n_{00}}{n_{00}+n_{01}}$



- ROC curve:



- AUC = area under the ROC curve = probability that $S(x_p) > S(x_n)$ with x_p and x_n random true positive and negative subjects

Building our library of learners

- we build a library of base algorithms of the type

$$\widehat{\mathcal{A}}_k \circ \widehat{\mathcal{S}}_\ell \circ \widehat{\mathcal{E}} \circ \widehat{\mathcal{I}}$$

  where
  - $\widehat{\mathcal{I}}$ is an imputation algorithm to deal with missing data
  - $\widehat{\mathcal{E}}$ is a data enrichment algorithm
  - $\widehat{\mathcal{S}}_\ell$ is a screening algorithm that filters out some covariates, $\ell = 1, \ldots, 7$
  - $\widehat{\mathcal{A}}_k$ is a main classification algorithm, $k = 1, \ldots, 25$

- In total, we consider $127 = (1 + 1 + 6) \times 1 + (1 + 4 + 4 + 8) \times 7$

Imputation algorithm $\widehat{\mathcal{I}}$

cautious imputation algorithm:

- for each variable Z, missing values are replaced with the mean of z in the group of patients characterised by a similar number of days since the probable date of infection
- **example:** if the cardiac frequency of a patient 5 days after the probable date of infection is missing, then we impute the mean of all the available cardiac frequencies of patients who were probably infected between 3 and 6 days earlier
- binned number of days since infections are

  (0-3, 4-6, 7-9, 10-12, 13-15, 16-20, 21-24, more than 24 days)

Screening algorithms $\widehat{\mathcal{S}}_\ell$

- reduce overfitting and increase robustness
- improve interpretability
- we consider only *expert* procedures (not data-driven)
- $\widehat{\mathcal{S}}_1$ keeps the covariates deemed relevant by [Wynant 20] (11 variables)
- the other algorithms keep different subsets of covariates then $\widehat{\mathcal{S}}_\ell$ ($\leq$ 30 variables)

Main classifiers $\widehat{\mathcal{A}}_k$

```
> sl.library
 [1] "SL.mean"       "SL.mean_by_1"  "SL.mean_by_2"  "SL.mean_by_3"  "SL.mean_by_4"
 [6] "SL.mean_by_5"  "SL.mean_by_6"  "SL.Wuhan_tree" "SL.glmnet"     "SL.rpart_1"
[11] "SL.rpart_2"    "SL.rpart_3"    "SL.rpart_4"    "SL.ranger_1"   "SL.ranger_2"
[16] "SL.ranger_3"   "SL.ranger_4"   "SL.xgboost_1"  "SL.xgboost_2"  "SL.xgboost_3"
[21] "SL.xgboost_4"  "SL.xgboost_5"  "SL.xgboost_6"  "SL.xgboost_7"  "SL.xgboost_8"
```

- `SL.mean*`: x is associated to the proportion of good prognosis according to her subgroup in terms of age and/or comorbidities
- `SL.Wuhan_tree`: decision tree built from Wuhan data and based on crp, ldh and lymphocyte proportion in [Yan 2020]
- `SL.glmnet`: default elastic net
- `SL.rpart_*`: decision trees with 4 different maximum depths
- `SL.ranger_*`: random forests with 4 different number of trees
- `SL.xgboost_*`: gradient boosting with 8 different configurations

How we deal with repeated measures

- goal is to identify, within two days ($t = 1$) after hospital admission ($t = 0$), patients who have poor prognosis
- instead of training only on data collected at $t = 0, 1$ we exploit all data, even those collected at $t > 1$
  however we test on data collected at $t = 0, 1$
- each patient contributes 1 to 8 *blocks* of data $(x_t, y) \in \mathcal{X} \times \{0, 1\}$ where
  - x_t = vector of covariates at admission or $2 \times t$ days thereafter, $0 \leq t \leq 7$
  - y = prognosis
- overall the 209 subjects contribute 570 blocks of data

number of blocks of data	1	2	3	4	5	6	7	8
number of patients	43	76	36	28	9	10	3	4

Training the super learner

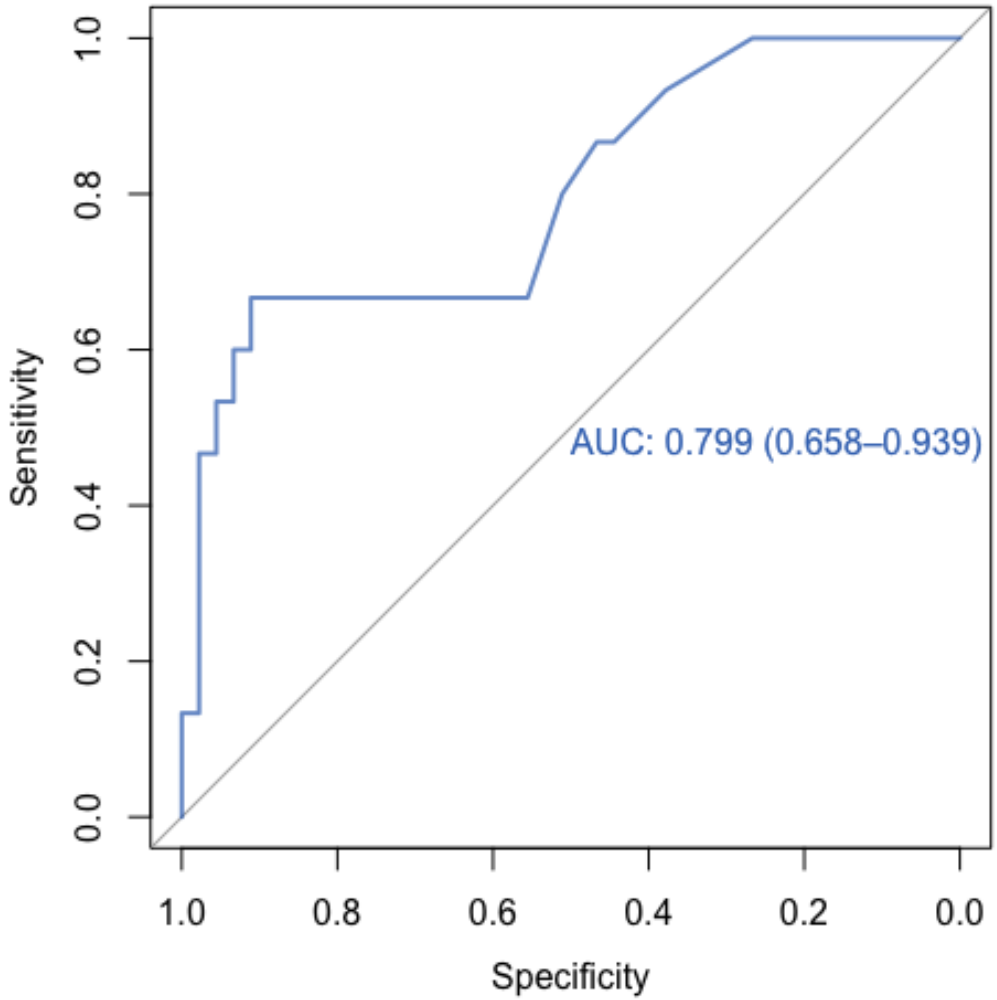
- training $\widehat{\text{SL}}$ requires training and testing all the base learners in our library following a V-fold cross validation scheme
- we use $V = 5$ folds and measure the performance of base algorithms with AUC
- we make sure that for each patient, all her blocks fall in the same fold so as to ensure the mutual independence between the data folds
- the performance of the discrete SL is learnt by nested cross-validation

Cross-validated AUC

- cross-validated AUC: 0.83 (95% CI [0.71; 0.94])
- the discrete SL is a random forest (50 or 200 trees) in 3 out of 5 folds:

```
> sl$whichDiscreteSL
[[1]]
[1] "SL.xgboost_1_screening.expert_one_sl"

[[2]]
[1] "SL.ranger_1_screening.expert_four_sl"

[[3]]
[1] "SL.xgboost_1_screening.expert_five_sl"

[[4]]
[1] "SL.ranger_1_screening.expert_two_sl"

[[5]]
[1] "SL.ranger_4_screening.expert_six_sl"
```

Performance on validation sample I

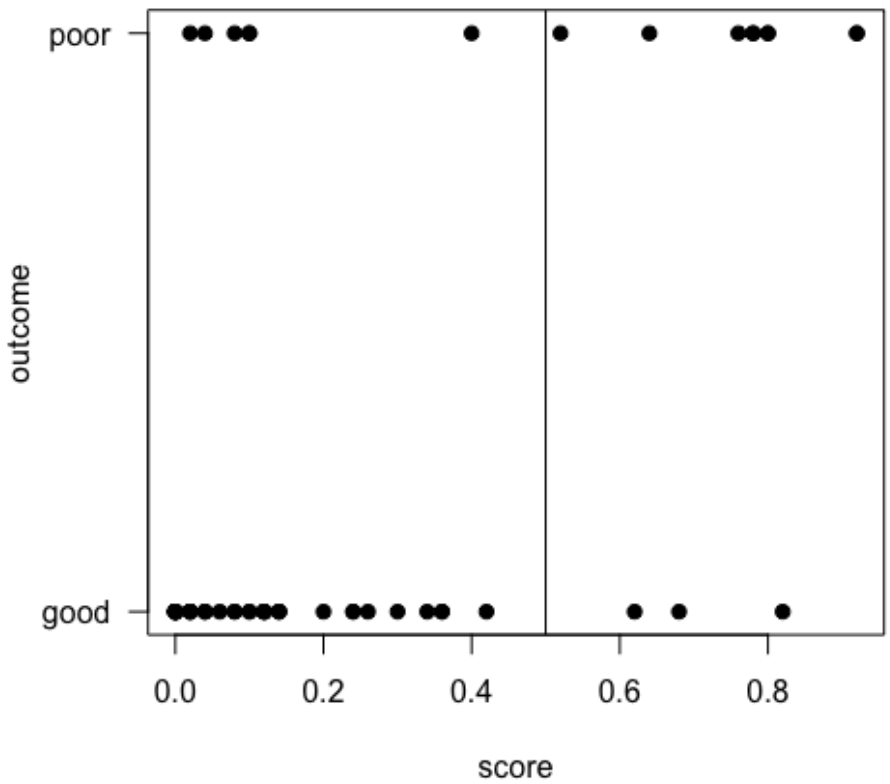


Performance on validation sample II

Threshold: 0.5

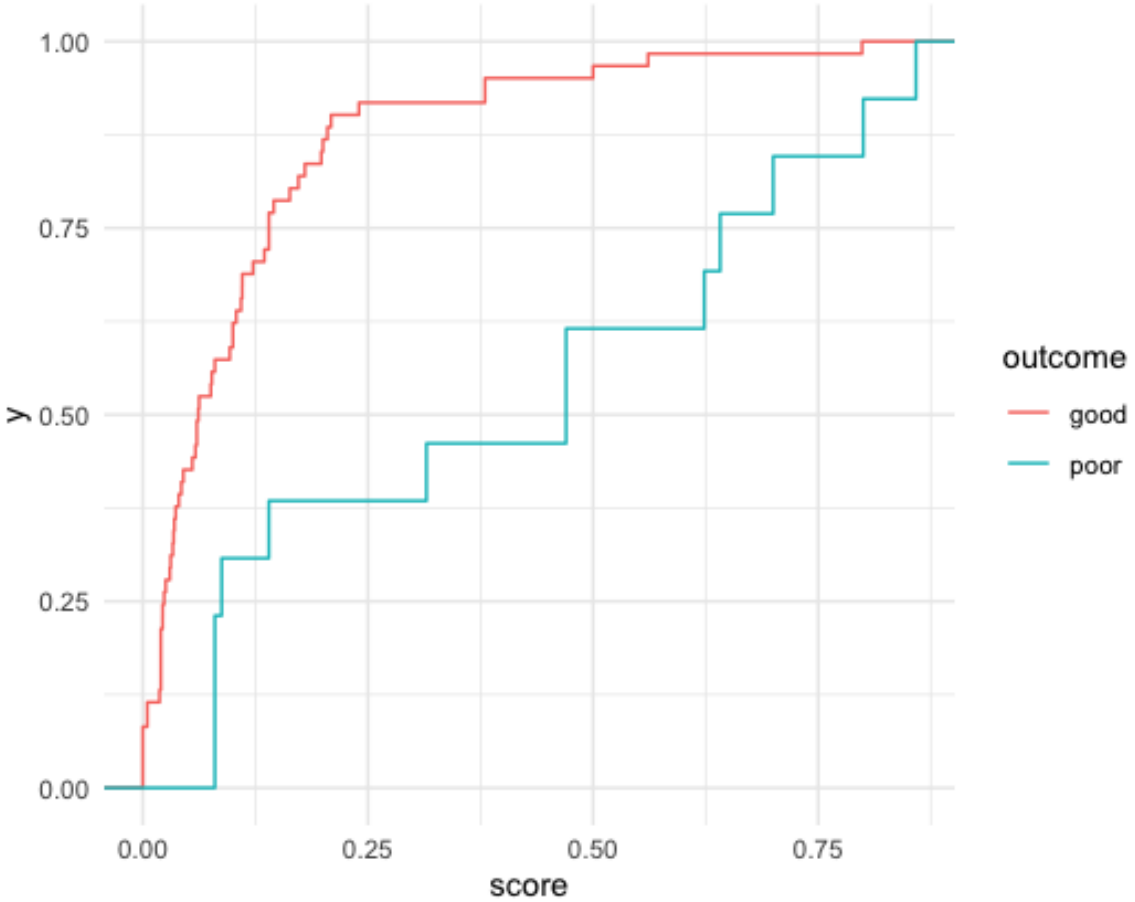
		Truth	
		good	poor
Prediction	good	42	6
	poor	3	9

- sensitivity: 0.60
- specificity: 0.93
- positive predictive value: 0.75
- negative predictive value: 0.87



Empirical cumulative distribution of the score I

Cross-validated ECD



Empirical cumulative distribution of the score II

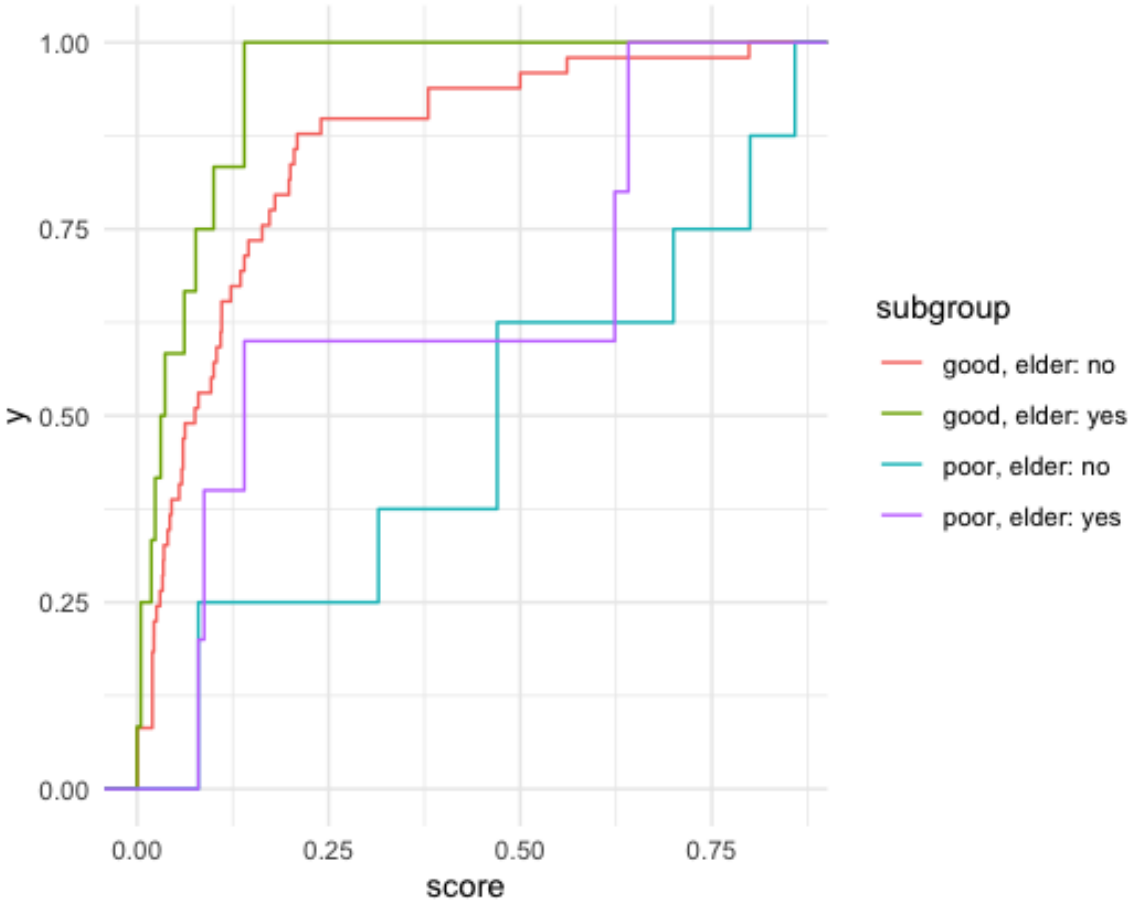


Table of Contents

From predicting to explaining

- our second aim is to
  - define the importance of putatively relevant covariates as model-agnostic statistical parameters
  - calculate point estimates and confidence intervals for such parameters

- we do so by applying the targeted learning framework [van der Laan and Rose 2011, 2018]

Targeted learning (TL)

- traditionally a parametric model is chosen based on the outcome type and then each covariate-specific coefficient is interpreted as the covariate's measure of importance
- this approach is clearly flawed under model misspecification
- it makes more sense to first define the covariate's measure of importance as a statistical parameter (the *target quantity*) and then to develop specifically an estimator thereof
- TL makes it possible to use machine learning methods to obtain estimates that are closer to the target quantities than those that would be obtained using classic parametric models
- under empirical processes conditions, the TL estimators are asymptotically Gaussian thus allowing quantification of uncertainty (standard errors, confidence intervals and p-values)

Non-parametric variable importance (NPVI)

- let P be the law of (X, Y)
- let $X_j \subset X$ be a real-valued covariate
- let x_j^{ref} be a reference value for X_j s.t. $P(X_j = x_j^{\text{ref}} | X \setminus X_j) > 0$
- let $\theta_P(x) = P(Y = 1 | X = x)$
- let $\theta_P(x_j^{\text{ref}}, X \setminus X_j) = P(Y = 1 | X_j = x_j^{\text{ref}}, X \setminus X_j = x_{-j})$
- introduced by [Chambaz et al 2012],

$$\Psi_j(P) := \frac{\mathrm{E}_P \left[(X_j - x_j^{\text{ref}}) \left(\theta_P(X) - \theta_P(x_j^{\text{ref}}, X \setminus X_j) \right) \right]}{\mathrm{E}_P \left[\left(X_j - x_j^{\text{ref}} \right)^2 \right]} \in \mathbb{R}$$

is a relevant NPVI of X_j on Y accounting for X_j

NPVI interpretation

- $\Psi_j(P)$ is the slope coefficient in the linear regression of the risk difference

$$\theta_P(X) - \theta_P(x_j^{\text{ref}}, X \setminus X_j)$$

against

$$X_j - x_j^{\text{ref}}$$

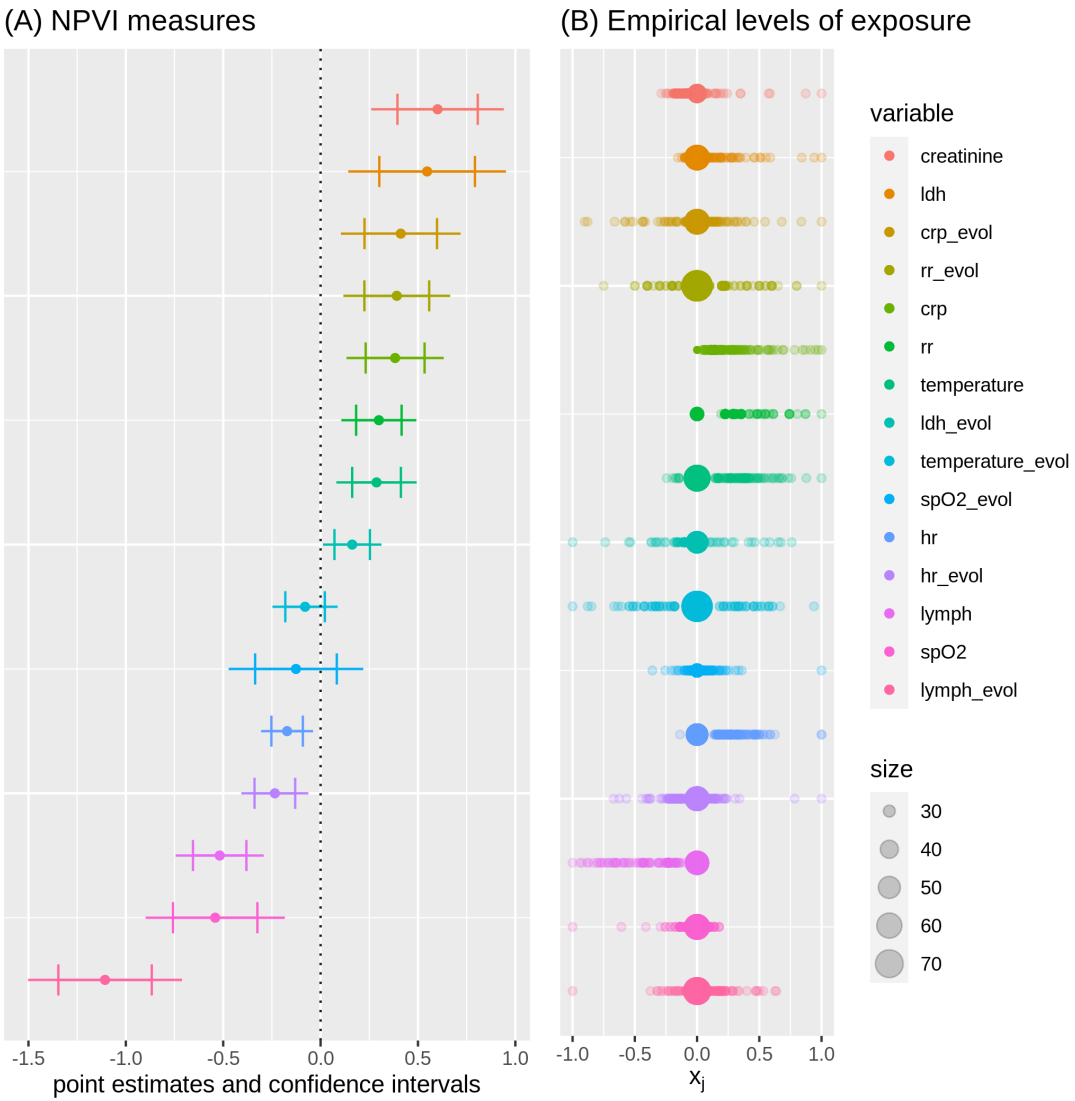
$\Rightarrow \Psi_j(P)$ measures how a deviation of X_j from x_j^{ref} correlates with the change in risk going from X to $(x_j^{\text{ref}}, X \setminus X_j)$

- the definition of $\Psi_j(P)$ makes it possible to integrate the fact that our covariates come with reference values established in the medical practice

Inference

- an estimator based on the targeted minimum loss estimation (TMLE) methodology is introduced and studied in [Chambaz et al 2012]
- the estimation strategy relies, among other things, on estimating relevant features of the law P such as $P(X_j = x_j^{\text{ref}} | X \setminus X_j)$ and $E_P[X_j | X_j \neq x_j^{\text{ref}}, X \setminus X_j]$
- under mild assumptions, the point estimate of $\Psi_j(P)$ can be accompanied with an asymptotic confidence interval
- the validity of the confidence interval is guaranteed provided that the relevant features of P are sufficiently well estimated
- those features are estimated with super learning
- we rely on a customized version of the `tlme.npvi` package [Chambaz, Neuvial 2015]

NPVI estimates



- variables have been normalized: one unit increase means the same across all of them
- NPVI > 0: an increase of X_j is correlated to an increase of the risk $P(Y = 1|X) - P(Y = 1|X_j = x_j^{\text{ref}}, X \setminus X_j)$
- 95% global CI
- In (B) the size of the disk represents the number of patients in the reference interval, and dots represent values outside of it

Table of Contents

Discussion et conclusions

- similar works:

  [Yan 20]:

  - reported performance is very good
  - subsequent studies reported limited applicability.
  - the approximation given by the decision tree showed in the published article works less better than other algorithms on our data

  [Lassau 21]:

  - development of a severity score from biological variables enriched with a score obtained with a deep neuronal network trained on chest CT scan images
  - the gain in AUC obtained by deep learning is 0.03
- our algorithm has good specificity but not so good sensitivity
- as such, it can be used to identify patients who deserve in-depth follow-up
- we have developed an interactive application implementing our classification algorithm using the `shiny` package

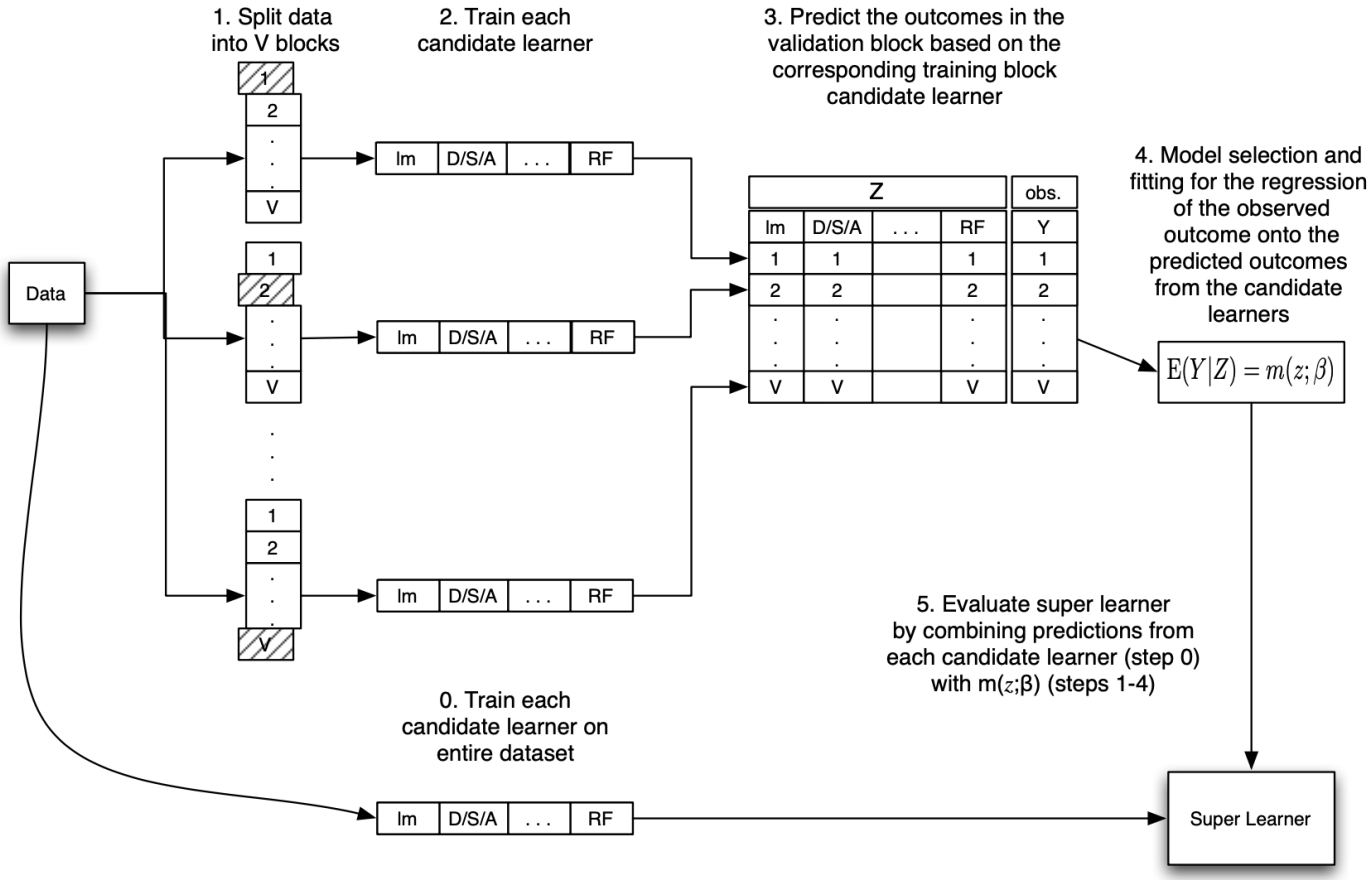
Some references

- van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. "Super learner." Statistical applications in genetics and molecular biology 6.1 (2007).
- Yan, Li, et al. "An interpretable mortality prediction model for COVID-19 patients." Nature machine intelligence 2.5 (2020): 283-288.
- Wynants, Laure, et al. "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal." BMJ 369 (2020).
- Van der Laan, Mark J., and Sherri Rose. Targeted learning in data science. Cham: Springer International Publishing, 2018.
- Williamson, Brian D., et al. "Nonparametric variable importance assessment using machine learning techniques." Biometrics 77.1 (2021): 9-22.
- Chambaz, Antoine, Pierre Neuvial, and Mark J. van der Laan. "Estimation of a non-parametric variable importance measure of a continuous exposure." Electronic journal of statistics 6 (2012): 1059.
- Chambaz, Antoine, and Pierre Neuvial. "tmle. npvi: targeted, integrative search of associations between DNA copy number and gene expression, accounting for DNA methylation." Bioinformatics 31.18 (2015): 3054-3056.
- Lassau, Nathalie, et al. "Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients." Nature communications 12.1 (2021): 1-11.

Appendix

Supplemental slides

Super learning III



[van der Laan 2007]

$\widehat{\mathcal{E}}$: two-step super learning for data enrichment

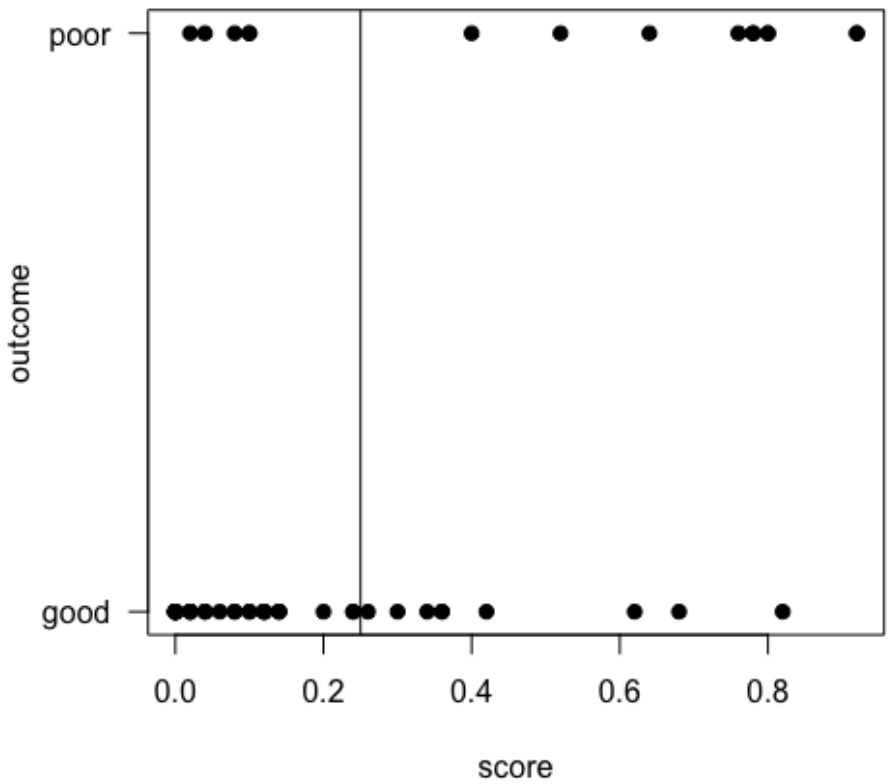
0. we split the initial training data ($n_{tr} = 149$) into two subsamples $\mathcal{F}_1$ and $\mathcal{F}_2$ of sizes $0.25 \times n_{tr}$ and $0.75 \times n_{tr}$
1. first step:
   a. we train a SL on on $\mathcal{F}_1$ using only data at $t = 0, 1$ and a reduced library of base algorithms (random forests)
   b. we predict the score for the observations in $\mathcal{F}_2$
2. second step: the main SL described above is trained on $\mathcal{F}_2$

Performance on validation sample III

Threshold: 0.25

		Truth	
		good	poor
Predicton	good	36	5
	poor	9	10

- sensitivity: 0.67
- specificity: 0.80
- positive predictive value: 0.53
- negative predictive value: 0.88



Another definition of variable importance

- let X_j be a subset of the vector of covariates X for which we want to define an importance measure on the outcome Y
- let P be the law of (X, Y) and $\theta_P(x) = P(Y = 1|X = x)$
- let $\theta_{P,-j}(x_{-j}) = P(Y = 1|X \setminus X_j = x_{-j})$
- the importance measure of X_j can be defined according to [Williamson et al 2021] as

$$\frac{\mathrm{E}_P\left[(Y - \theta_P(X))^2\right]}{\mathrm{Var}_P(Y)} - \frac{\mathrm{E}_P\left[(Y - \theta_{P,-j}(X \setminus X_j))^2\right]}{\mathrm{Var}_P(Y)} \in [0, 1]$$

- this parameter can be interpreted as the additional proportion of variability in the outcome explained by including X_j
- extension of the standard R^2 coefficient

Prepocessing for NPVI estimation

variable		centering ($x_j^{\text{ref}} = 0$)	units
creatinine	at day 0	if $\lvert x - 95\rvert \leq 15$ then 0 else $x - 95$	mg/L
C-reactive protein	at day 2	if $x \leq 35$ then 0 else $x - 25$	mg/L
C-reactive protein	day 2 - day 0	if $\lvert x\rvert \leq 5$ then 0 else x	mg/L
heart rate	at day 2	if $\lvert x - 70\rvert \leq 10$ then 0 else $x - 70$	beats pm
heart rate	day 2 - day 0	if $\lvert x\rvert \leq 5$ then 0 else x	beats pm
lactate dehydrogenase	at day 2	if $\lvert x - 275\rvert \leq 50$ then 0 else $x - 275$	U/L
lactate dehydrogenase	day 2 - day 0	if $\lvert x\rvert \leq 5$ then 0 else x	U/L
blood lymphocyte count	at day 2	if $x \geq 1.25$ then 0 else $x - 1.25$	$\times 10^3/\text{mm}^3$
blood lymphocyte count	day 2 - day 0	if $\lvert x\rvert \leq 0.1$ then 0 else x	$\times 10^3/\text{mm}^3$
oxygen saturation	at day 2	if $\lvert x - 95.5\rvert \leq 0.5$ then 0 else $x - 95.5$	%
oxygen saturation	day 2 - day 0	if $\lvert x\rvert \leq 0.5$ then 0 else x	%
respiratory rate	at day 2	if $\lvert x - 17\rvert \leq 5$ then 0 else $x - 17$	breaths pm
respiratory rate	day 2 - day 0	if $\lvert x\rvert \leq 1$ then 0 else x	breaths pm
temperature	at day 2	if $\lvert x - 37\rvert \leq 0.5$ then 0 else $x - 37$	$^\circ$C
temperature	day 2 - day 0	if $\lvert x\rvert \leq 0.5$ then 0 else x	$^\circ$C