Derivation of a Cost-Sensitive COVID-19 Mortality Risk Indicator Using a Multistart Framework

Rubén Armañanzas, PhD
rarmananzas@unav.es

DATAI - The Institute for Data Science and Artificial Intelligence

February 23rd, 2022

Outline

- AI Schools of Thought
- Introduction
- Data & Data Sets
- Feature Filtering & Selection
- Derivation of the Mortality Risk Indicator
- Indicator Performance
- Time Horizon
- Closing
- Acknowledgements

Artificial Intelligence Today

Model-based AI, aka *represent-and-reason*

* Represents knowledge about [physical] entities and involves reasoning with such knowledge.

Main tool: *Logic and probability*
Exemplar: *Ontologies*

Function-based AI, aka *curve-fitting*

* Formulates a task as a function-fitting problem, with function inputs coming directly from the raw data and outputs corresponding to the high-level recognitions.

Main tool: *Artificial neural networks*
Exemplar: *Deep learning*

Strong AI vs. Weak AI

Adnan Darwiche, UCLA CS Chair (CACM 2018, 61(10), 56-67)

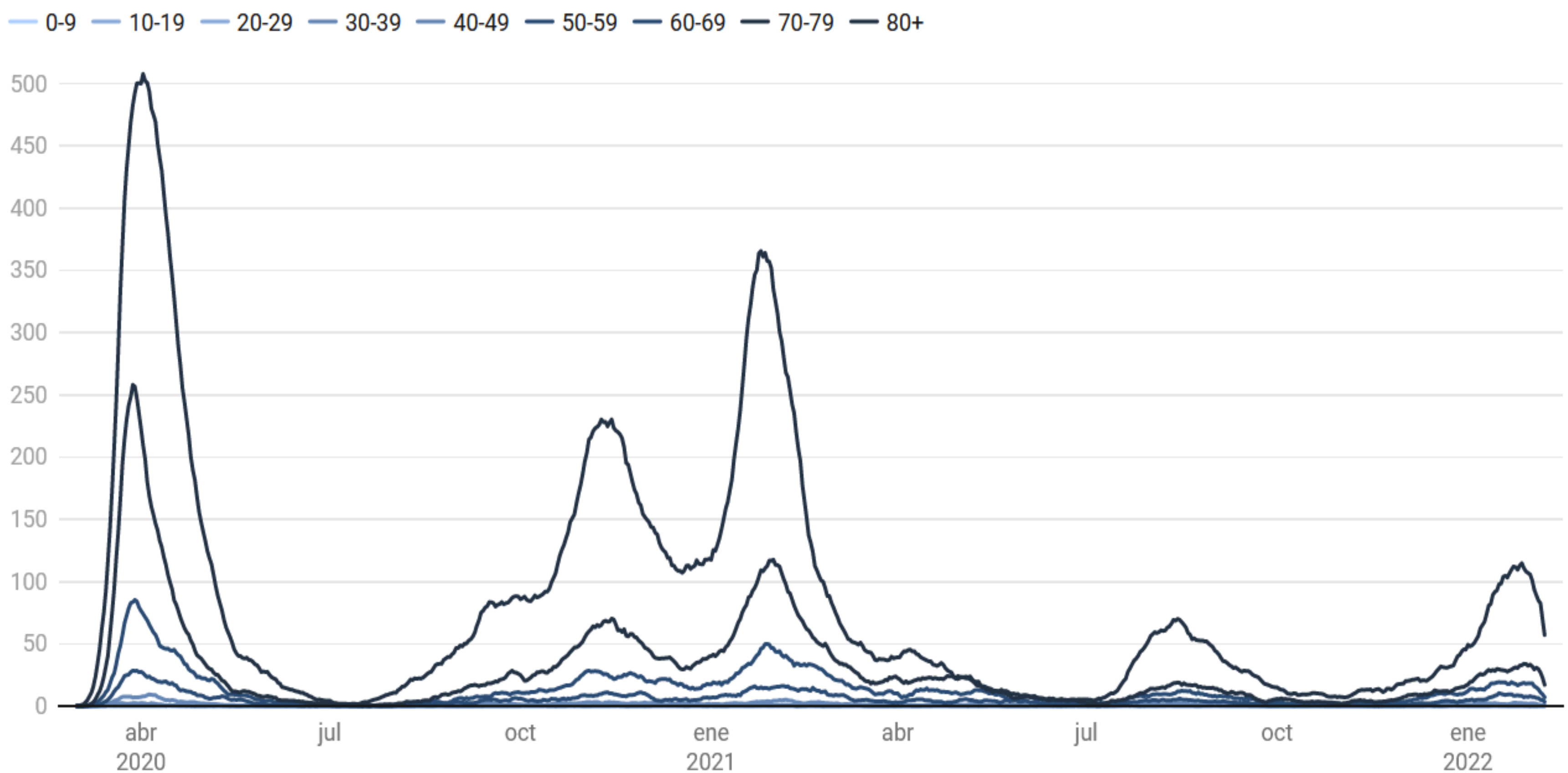
"Every behavior, intelligent or not, can be captured by a function that maps inputs to outputs."

- Admit a compact representation
- Allow accurate estimations from labeled data
- Be evaluated efficiently (no reasoning required)

Model-based approach

- Abstraction (Platonian realism)
- Inductive reasoning
- Causal inference
- Generalization

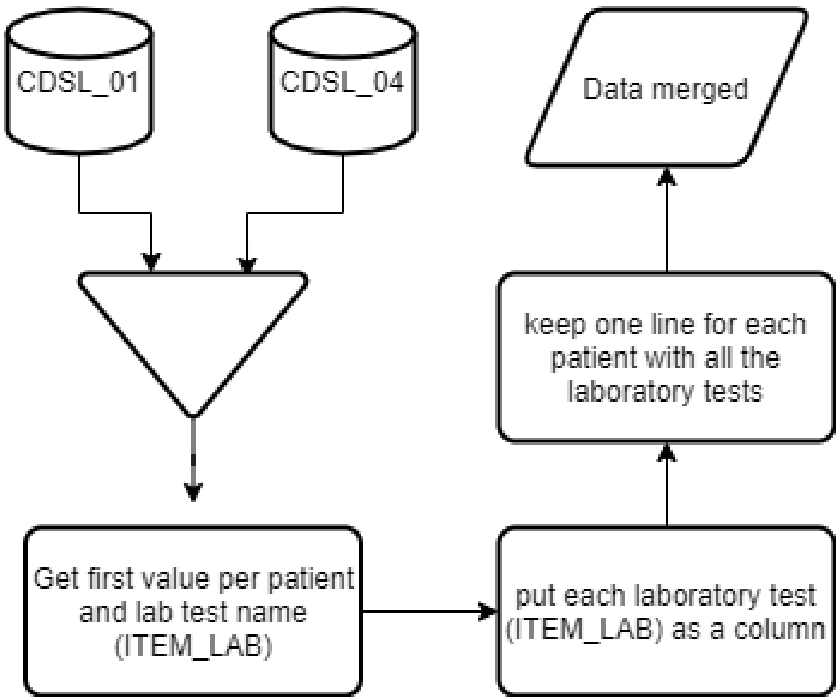
Healthcare collapse by COVID-19



Daily aggregate deaths from COVID-19 infection in Spain.

Source: Spanish Health Department

HM Hospitales CDSL - Data Filtering



Percentage of presence	# features
>= 90%	29
>= 80% & < 90%	4
>= 70% & < 80%	3
>= 60% & < 70%	0
>= 50% & < 60%	1
>= 40% & < 50%	4
>= 30% & < 40%	9
>= 20% & < 30%	17
>= 10% & < 20%	23
>= 0% & < 10%	348
Total	**438**

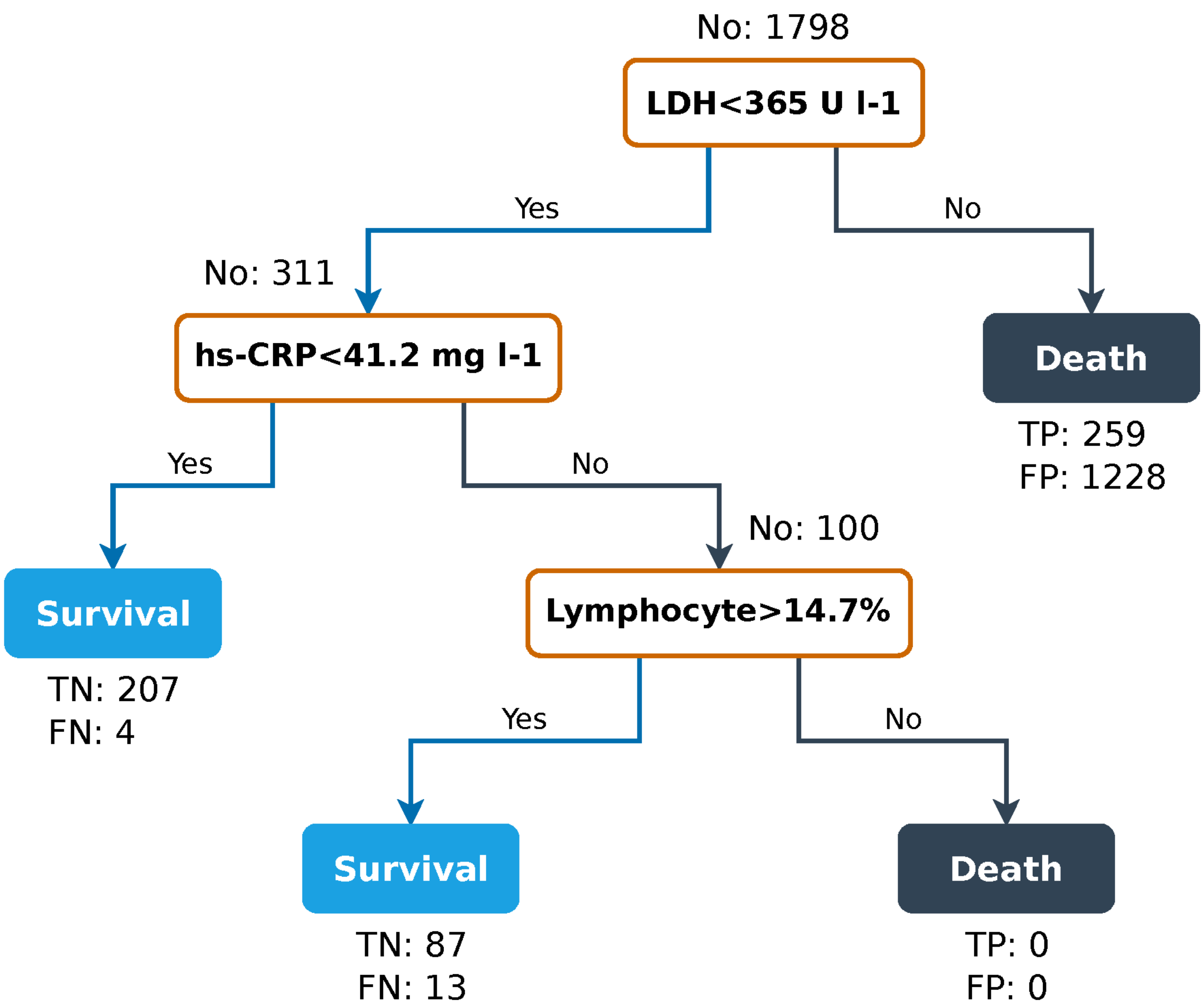
- Only numerical values
- Missing rate at 30% max.
- Missing values at 3 tests max.

Data Set

DEMOGRAPHIC CHARACTERISTICS FOR THE FINALIZED DATASET TO PREDICT IN HOSPITAL MORTALITY. TIME SPAN COVERS FROM DECEMBER 26^{th} (2019) TO JUNE 10^{th} (2020). VALUES FOR THE INTERQUARTILE RANGE (IQR) SHOW THE 25^{th} AND 75^{th} DATA PERCENTILES.

Characteristic	No. of patients (%)	Total No.
Inpatient mortality	276 (15.35%)	1798
Ventilator use	1035 (57.56%)	1798
Female patients	707 (39.32%)	1798
Male patients	1091 (60.68%)	1798
ICU admission	167 (9.29%)	1798
	Mean ± Std.	Median [IQR]
Age (years)	67.79 ± 15.67	69 [57,80]
No. of selected comorbidities	0.49 ± 0.77	0 [0,1]
ICU stay (days)	8.72 ± 10.50	5 [1,12]
Oxygen saturation	94.67 ± 4.81	95 [94,97]
Heart rate (bpm)	79.28 ± 14.75	78 [70,88]

Prediction Model from Yan et al., 2020



Prediction	Death	Survival
Positive	259	1228
Negative	17	294
Accuracy	0.307	
Precision	0.174	
Sensitivity	0.938	
Specificity	0.193	
F1 Score	0.294	

Yan et al. (2020) An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence,* 2(5):283-8

Prediction Model from ISARIC 4C, 2020

Variable	4C
Age (years)	
‹50	—
50-59	+2
60-69	+4
70-79	+6
≥80	+7
Sex at birth	
Female	—
Male	+1
No of comorbidities*	
0	—
1	+1
≥2	+2
Respiratory rate (breaths/min)	
‹20	—
20-29	+1
≥30	+2
Peripheral oxygen saturation on room air (%)	
≥92	—
‹92	+2
Glasgow coma scale score	
15	—
‹15	+2
Urea (mmol/L)	
‹7	—
7-14	+1
›14	+3
C reactive protein (mg/L)	
‹50	—
50-99	+1
≥100	+2

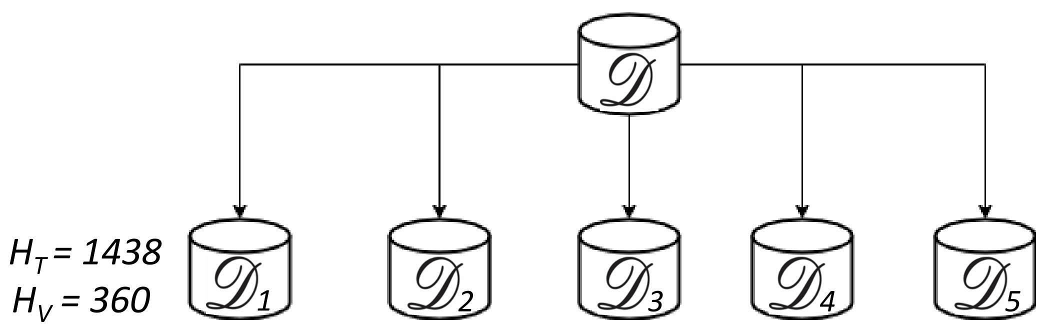
Prediction	Death	Survival
Positive (4C ≥ 9)	272	942
Negative (4C < 9)	4	580
Accuracy	0.474	
Precision	0.224	
Sensitivity	0.985	
Specificity	0.381	
F1 Score	0.365	

Knight et al. (2020) Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. *BMJ*, 370.

Data Selection – Additional Clinicals

- Compute comorbidities and symptoms indicators [0-3]
- Included Age and Gender
- Label
  - Positive class as death (1)
  - Negative class as survival (0)

- Multistart (5) configurations
  - 80% Training Data
  - 20% Hold-Out Data



Feature Selection

<u>Input:</u> 36 laboratory tests

1. Recursive Feature Elimination (5CV-RFE) -> <u>34</u> tests
   - Removed D-dimer and gamma-glutamyl transferase (GGT) test

<u>Input:</u> 34 laboratory tests & 4 epidemiological

2. Check univariate relevance using reg. coefficients from
   - LASSO (L1-norm penalization)
   - Logistic Regression (dichotomic, L1-norm penalization)

- Final feature set comprised by <u>19 variables</u>

Derivation Mortality Risk Indicator

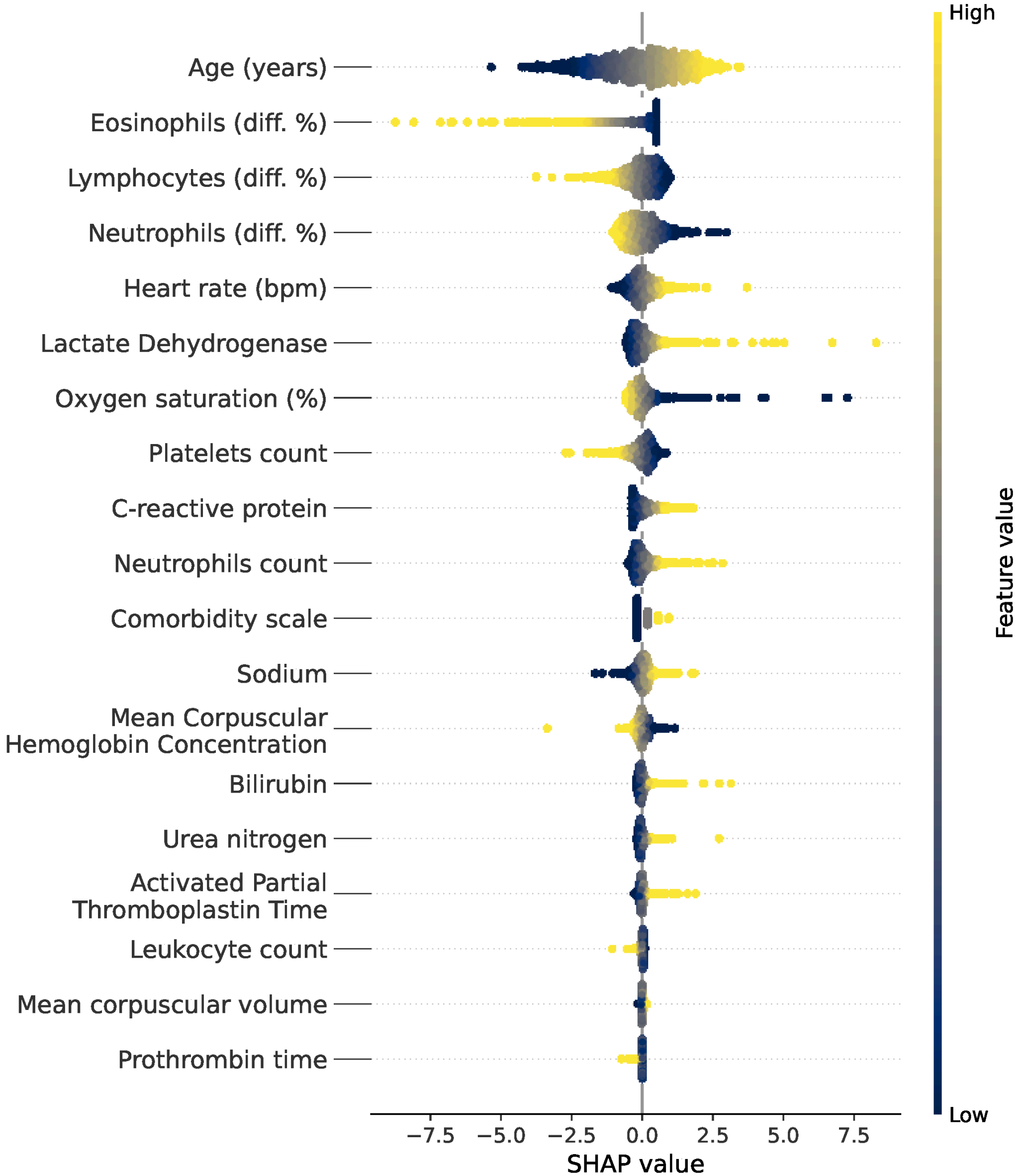
$$\phi : (x_1, \ldots, x_n) \to \{0, 1\}, \text{ where } \boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{R}^n$$

$$p_{C=1} = S(\underbrace{\cancel{\beta_0}}_{=0} + \sum_{i=1}^{n} \beta_i x_i)$$

$$\arg\max_{\boldsymbol{\beta}} \sum_{k=1}^{K} \theta \underbrace{\cancel{w_k}}_{=1} y_k \log_b(p(\boldsymbol{x_k})) + \sum_{k=1}^{K} \theta \underbrace{\cancel{w_k}}_{=1} (1 - y_k) \log_b(1 - p(\boldsymbol{x_k}))$$

$$\min_{\boldsymbol{\beta}} \sum_{k=1}^{K} |y_k - \beta_0 - X_k \beta| \text{ subject to } \sum_{i=1}^{n} |\beta_i| \le t$$

SHAP values



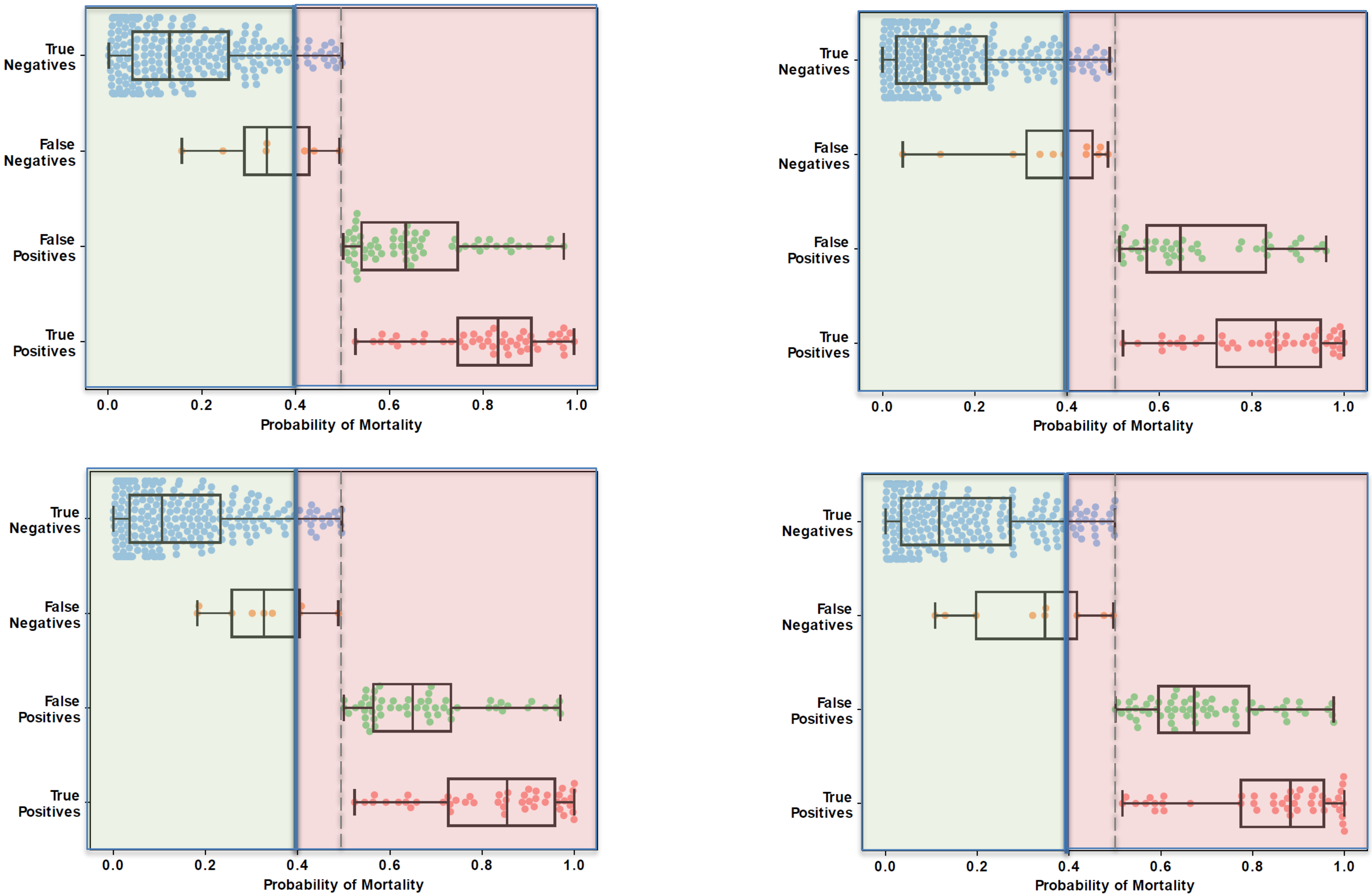
__Training__

- 5 Cross-Validation search
- Coordinate descent
- Class weight proportional to relative frequencies

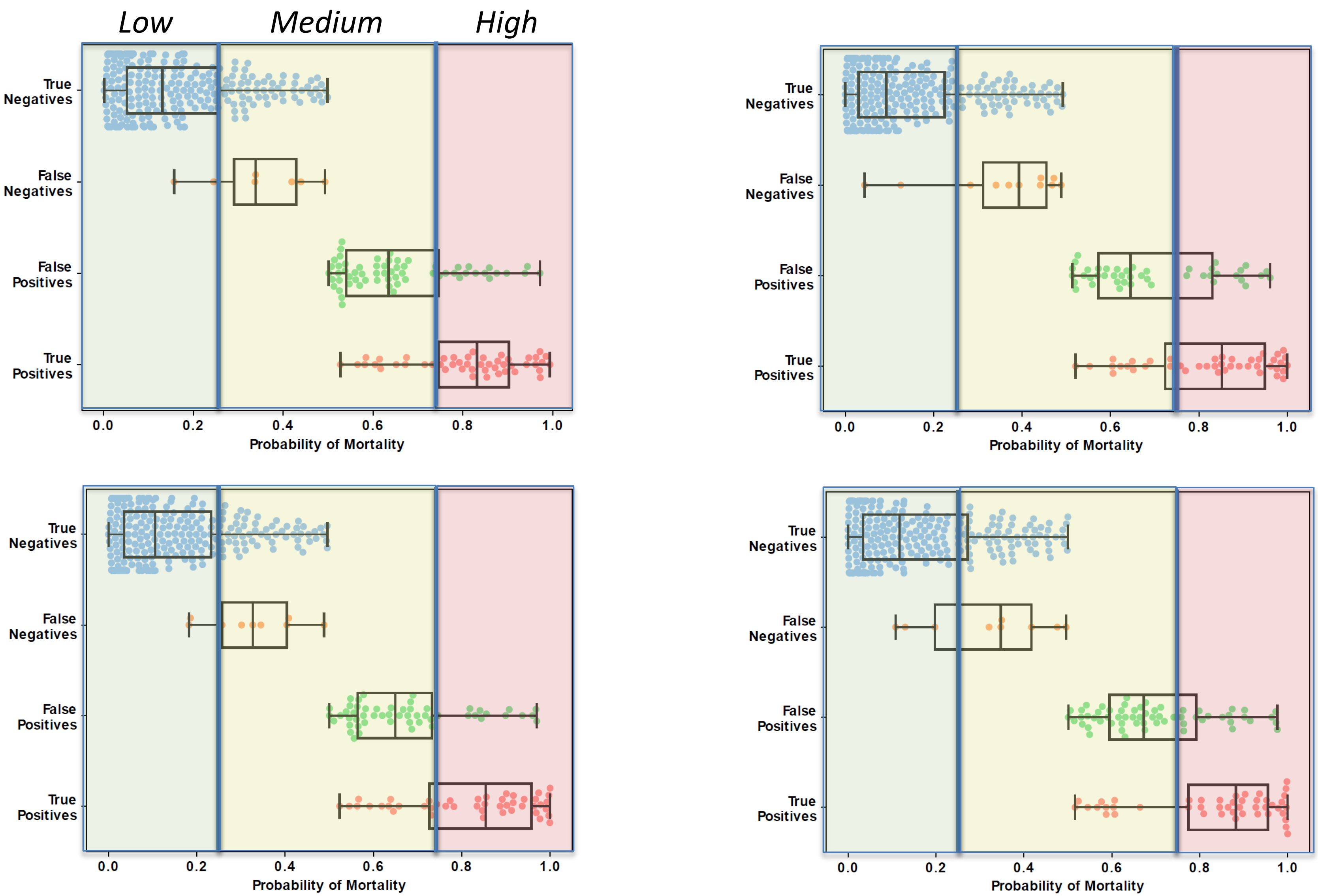
__Outcome__

- SHAP positive -> Influence towards death (class 1)
- SHAP negative -> Influence towards survival (class 0)

Cost-Sensitive Calibration



Risk Prediction



Prediction Performance

Multistart Hold-Out Subsets (360)

Prediction	Death	Surv.	Death	Surv.	Death	Surv.	Death	Surv.	Death	Surv.
Positive	51	81	49	73	49	63	49	86	49	85
Negative	4	224	6	232	6	242	6	219	6	220
AUC	0.922		0.916		0.915		0.895		0.886	
Accuracy	0.764		0.780		0.808		0.744		0.747	
Sensitivity	0.927		0.891		0.891		0.891		0.891	
Specificity	0.734		0.761		0.793		0.718		0.721	
F_1 Score	0.545		0.553		0.587		0.516		0.519	

Independent Hold-Out group (121)

Prediction	Death	Survival
Positive	13	26
Negative	0	82
AUC	0.880	
Accuracy	0.785	
Sensitivity	1.000	
Specificity	0.759	
F_1 Score	0.5	

Time Horizon – Risk Assessment

[0,7) days	Death	Survival
Low Risk	$0.40 \pm 0.37\%$	$64.38 \pm 5.28\%$
Medium Risk	$5.00 \pm 1.29\%$	$16.19 \pm 3.49\%$
High Risk	$11.80 \pm 2.02\%$	$2.22 \pm 1.34\%$
[7,14) days	Death	Survival
Low Risk	$0.27 \pm 0.61\%$	$51.13 \pm 3.33\%$
Medium Risk	$3.38 \pm 0.92\%$	$33.37 \pm 5.10\%$
High Risk	$6.49 \pm 1.09\%$	$5.33 \pm 2.32\%$
[14, 21) days	Death	Survival
Low Risk	$0.40 \pm 0.91\%$	$35.10 \pm 5.10\%$
Medium Risk	$5.78 \pm 3.68\%$	$38.42 \pm 2.99\%$
High Risk	$9.67 \pm 3.32\%$	$10.60 \pm 4.77\%$
$\geq$20 days	Death	Survival
Low Risk	$4.34 \pm 2.52\%$	$23.05 \pm 4.20\%$
Medium Risk	$13.22 \pm 9.47\%$	$44.35 \pm 8.60\%$
High Risk	$9.60 \pm 3.37\%$	$5.41 \pm 3.48\%$

Closing

- The model was trained from routinely blood panel data.
- The model is easily adaptable to healthcare systems to aid decision making.
- Successful validation of the model with patients several months after the onset of the pandemic.

- Multivariate time-series classification.
- Range of predictors based on population dynamics.
  - Inclusion of vaccination status.
- Healthcare system cost estimation.

Acknowledgements

- Armañanzas et al. (2021). Derivation of a Cost-Sensitive COVID-19 Mortality Risk Indicator Using a Multistart Framework. *In IEEE International Conference on Bioinformatics and Biomedicine*, pages 2179-2186.

EUSKO JAURLARITZA
GOBIERNO VASCO
HEZKUNTZA, HIZKUNTZA POLITIKA ETA KULTURA SAILA
DEPARTAMENTO DE EDUCACIÓN, POLÍTICA LINGÜÍSTICA Y CULTURA

AXA Research Fund

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HM HOSPITALES

basque center for applied mathematics

Osakidetza