Development and validation of prognostic models for hospitalization in the Basque Country: Analyzing the variability of non-deterministic algorithms

Alexander Olza[1] Eduardo Millán[23] María Xosé Rodríguez-Álvarez[45]

[1]Basque Center for Applied Mathematics (BCAM), Bilbao, Spain

[2]Osakidetza Basque Health Service, General Directorate for Healthcare

[3]Kronikgune, Institute for Health Services Research

[4]CINBIO, Universidade de Vigo, Dept. of Statistics and OR

[5]CITMAga, Center for Mathematical Research and Technology of Galicia, Spain

JCE-SEB September 2022

Contents

1. Background
2. Methods
3. Results
4. Discussion
5. Conclusions

Background & Objectives

- Basque Country: Ongoing stratification project since 2014
- Demographic ageing ⇒ Higher disease burden
- Predictive models ⇒ Proactive and efficient healthcare

Objectives

- To **develop** and externally **validate** population-based predictive models for **unplanned hospitalization** through
- **Comparing** a logistic regression model and three families of machine learning models.
- Evaluate the **variability** of performance → Discuss the best technique

- All Public Health System enrollees in 2016-2017 (2.2 million)
- Predictors: All year diagnoses & prescriptions, isomorbidity groups
- Outcome: Unplanned hospitalization the year following

Background & Objectives

- Basque Country: Ongoing stratification project since 2014
- Demographic ageing ⇒ Higher disease burden
- Predictive models ⇒ Proactive and efficient healthcare

Objectives

- To **develop** and externally **validate** population-based predictive models for **unplanned hospitalization** through
- **Comparing** a logistic regression model and three families of machine learning models.
- Evaluate the **variability** of performance → Discuss the best technique

- All Public Health System enrollees in 2016-2017 (2.2 million)
- Predictors: All year diagnoses & prescriptions, isomorbidity groups
- Outcome: Unplanned hospitalization the year following

Background & Objectives

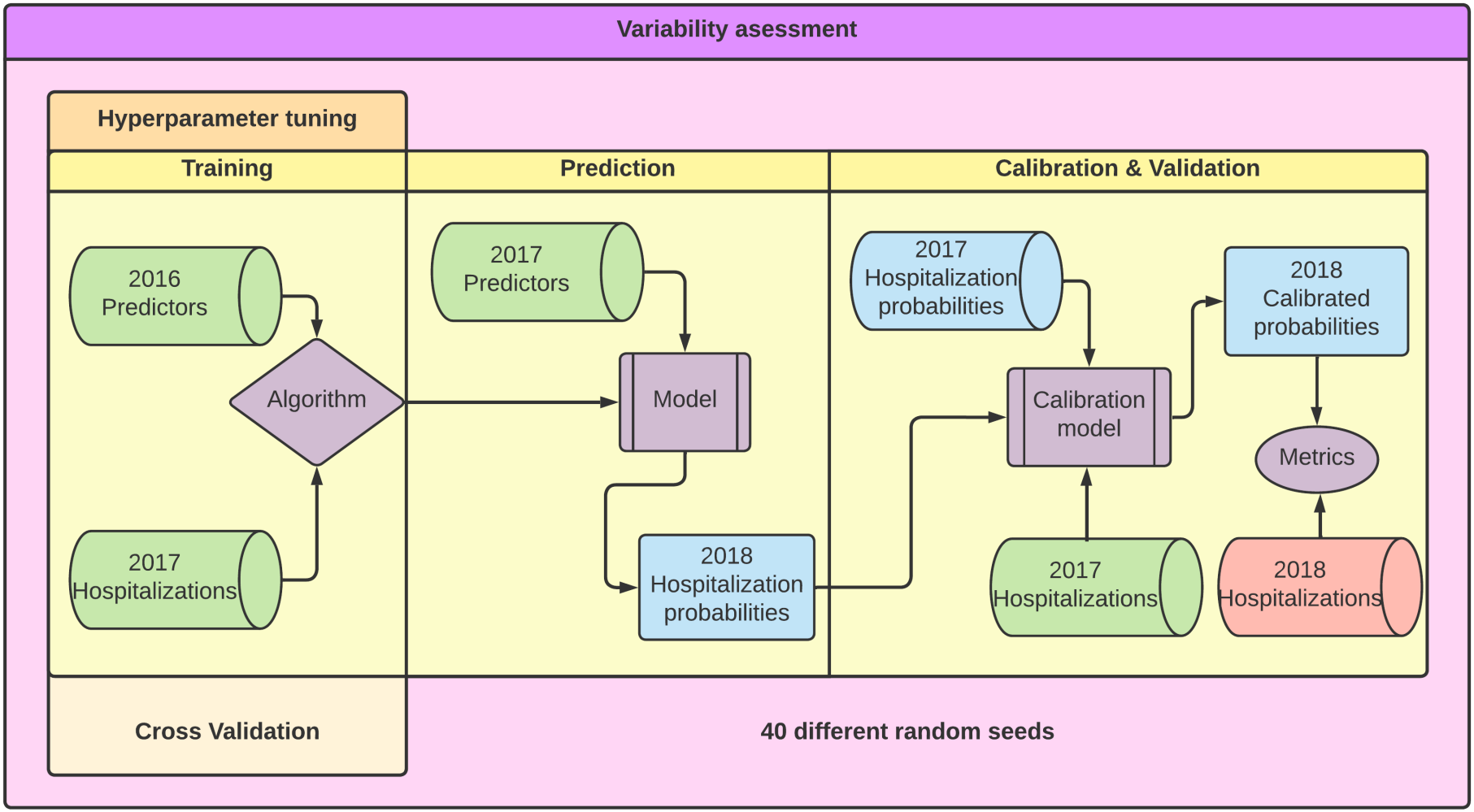
- Basque Country: Ongoing stratification project since 2014
- Demographic ageing $\Rightarrow$ Higher disease burden
- Predictive models $\Rightarrow$ Proactive and efficient healthcare

Objectives

- To **develop** and externally **validate** population-based predictive models for **unplanned hospitalization** through
- **Comparing** a logistic regression model and three families of machine learning models.
- Evaluate the **variability** of performance $\rightarrow$ Discuss the best technique

- All Public Health System enrollees in 2016-2017 (2.2 million)
- Predictors: All year diagnoses & prescriptions, isomorbidity groups
- Outcome: Unplanned hospitalization the year following

Overall methodology



Variability asessment

Multilayer Perceptron, Random Forest and Gradient-Boosting Decision Trees are non-deterministic: Many possible models for the same data.

Challenge: Training is non-deterministic

Randomness sources $\left\{ \begin{array}{l} \text{Class balancing} \\ \text{Hyperparameter tuning} \\ \text{Intrinsic randomness} \end{array} \right.$ $\Rightarrow$ Different realizations $\rightarrow$ diverse selected models & rankings of techniques

(Hyperparameter tuning selects the "best" model... but we can not explore the whole space, we are sampling grid points)

Approach

Repeat model development process 40 times with different seeds $\rightarrow$ Obtain a sample of possible "best" models. Report performance with variability.

Variability asessment

Multilayer Perceptron, Random Forest and Gradient-Boosting Decision Trees are non-deterministic: Many possible models for the same data.

Challenge: Training is non-deterministic

$$\text{Randomness sources} \left\{ \begin{array}{l} \text{Class balancing} \\ \text{Hyperparameter tuning} \\ \text{Intrinsic randomness} \end{array} \right] \Rightarrow \text{Different realizations} \rightarrow \text{diverse selected models \& rankings of techniques}$$

(Hyperparameter tuning selects the "best" model... but we can not explore the whole space, we are sampling grid points)

Approach

Repeat model development process 40 times with different seeds → Obtain a sample of possible "best" models. Report performance with variability.

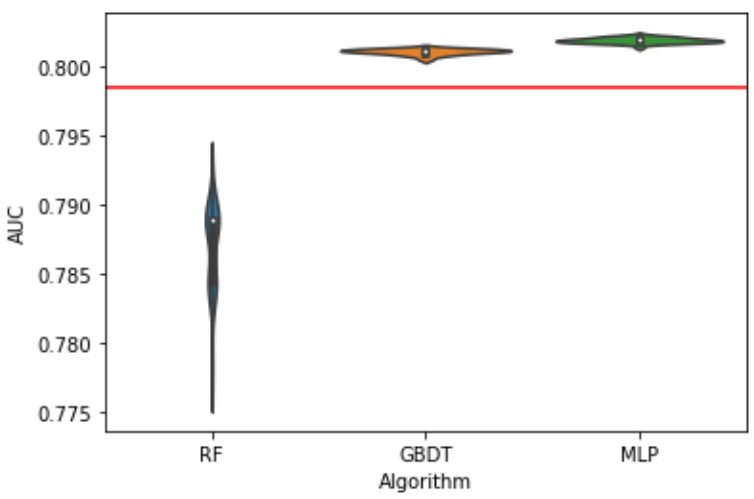
Performance metrics

- Population-wide $\rightarrow$ Threshold-independent metrics: AUC, AP
  - AP is more informative under class imbalance
  - $p \leq AP \leq 1$; p: Prevalence
- Risk groups (top 20000 probabilities) $\rightarrow$ Threshold-dependent metrics: R@20K (sensitivity), PPV@20K (positive predictive value)

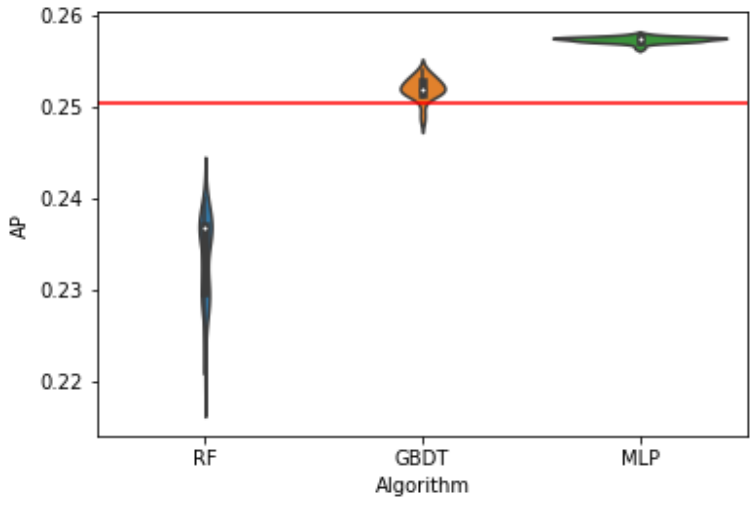
$$R@K = \frac{|L_K \cap H|}{|H|} \times 100 \qquad PPV@K = \frac{|L_K \cap H|}{K} \times 100$$

L_K : Patients with top K predicted probabilities
H: Hospitalized patients
$K \in \mathbb{N}$

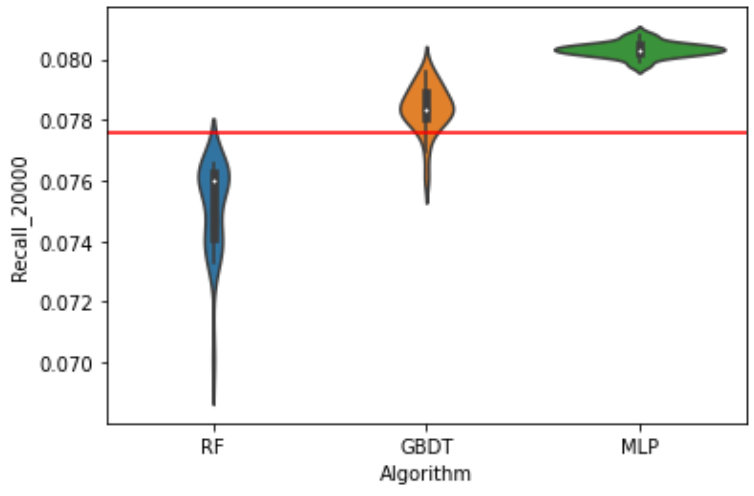
Model performance



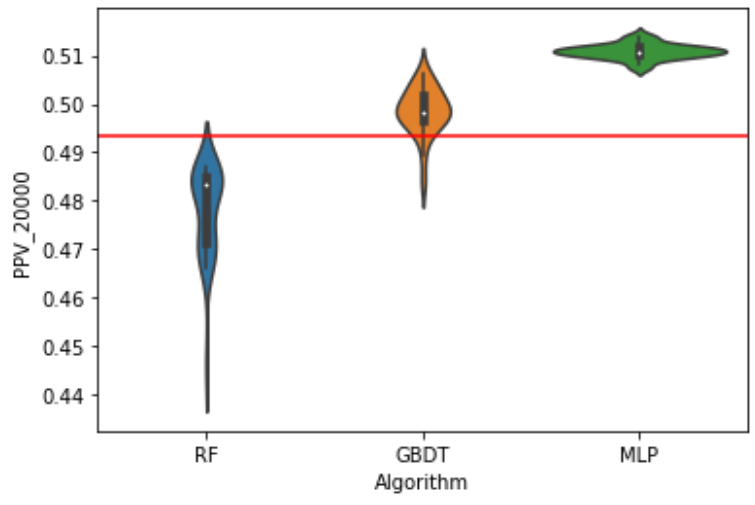
(a) AUC



(b) AP



(c) R@20k

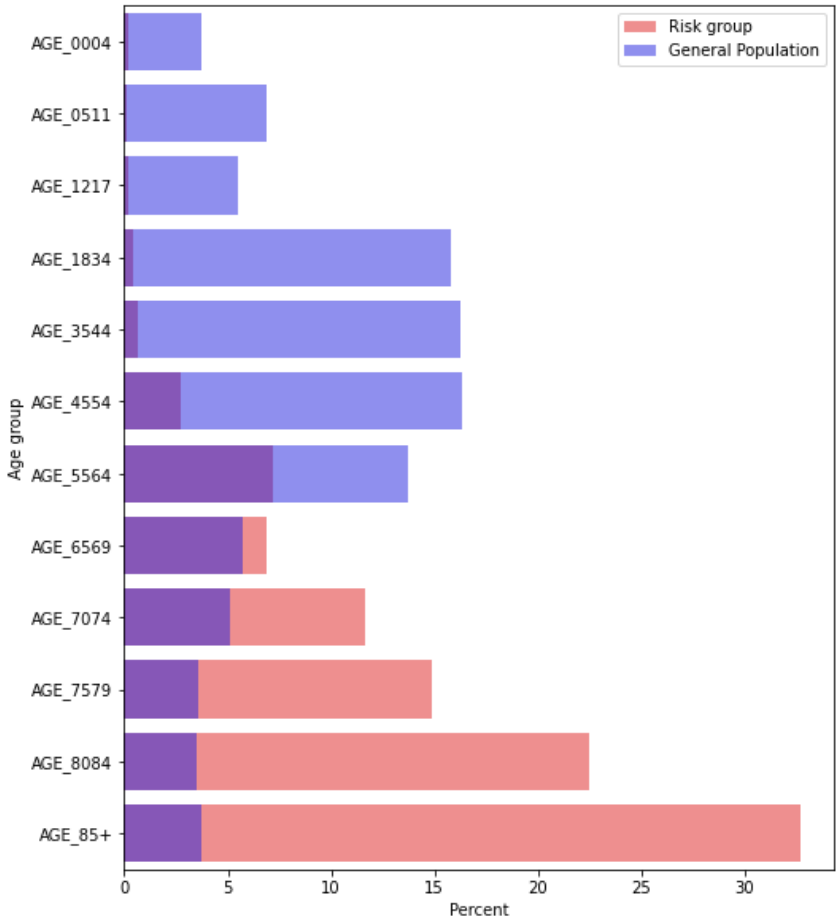


(d) PPV@20k

Model performance

	25 %	50 %	75 %	std
AUC				
MLP	0.8017	0.8019	0.8020	1.89e-04
LR		0.7985		
AP				
MLP	0.2571	0.2574	0.2575	3.04e-04
LR		0.2504		
R@20k				
MLP	0.0802	0.0803	0.0805	2.38e-04
LR		0.0776		
PPV@20k				
MLP	0.5101	0.5108	0.5117	1.51e-03
LR		0.4932		

Characteristics of the risk groups

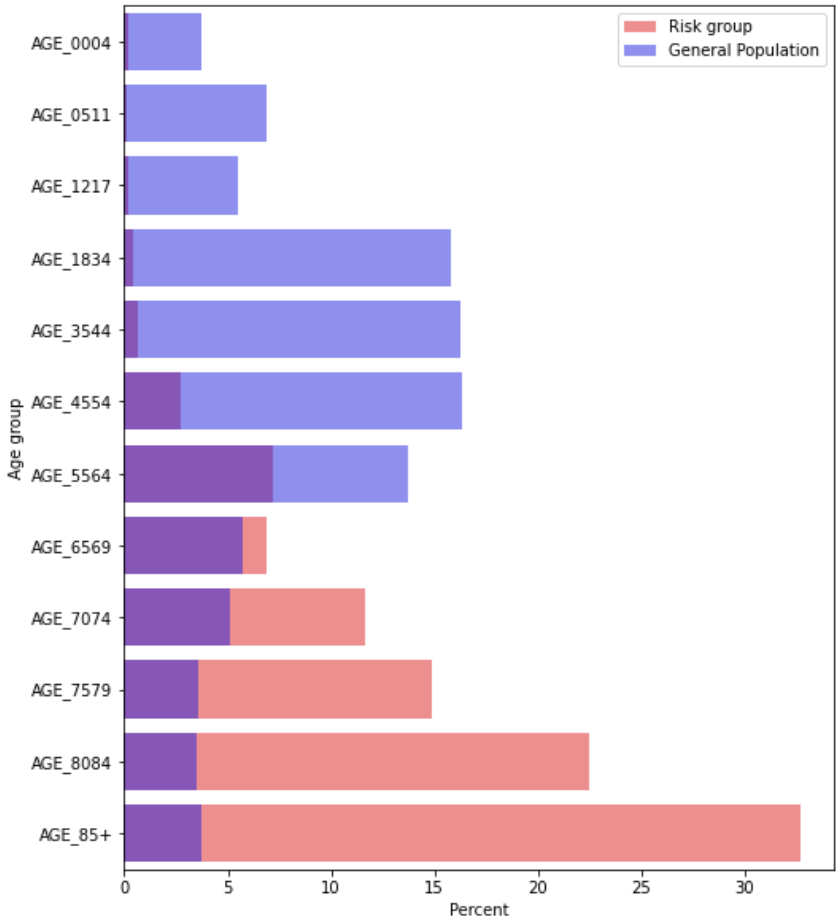


Comorbidities among the top 20000 patients (%)

	MLP	LR	GP
Hospit. in 2018	51.08	49.32	5.68
% of Women	33.95	35.20	51.06
Hypertension	80.66	74.53	9.11
Diabetes Mellitus	44.04	39.02	4.73
COPD	32.62	30.59	1.39
Chronic Renal Fail.	34.27	29.90	0.90
Heart Failure	38.63	34.60	0.73
Ischemic Heart Dis.	18.87	17.54	0.96
Low back pain	17.48	17.86	7.63

GP: General Population

Characteristics of the risk groups

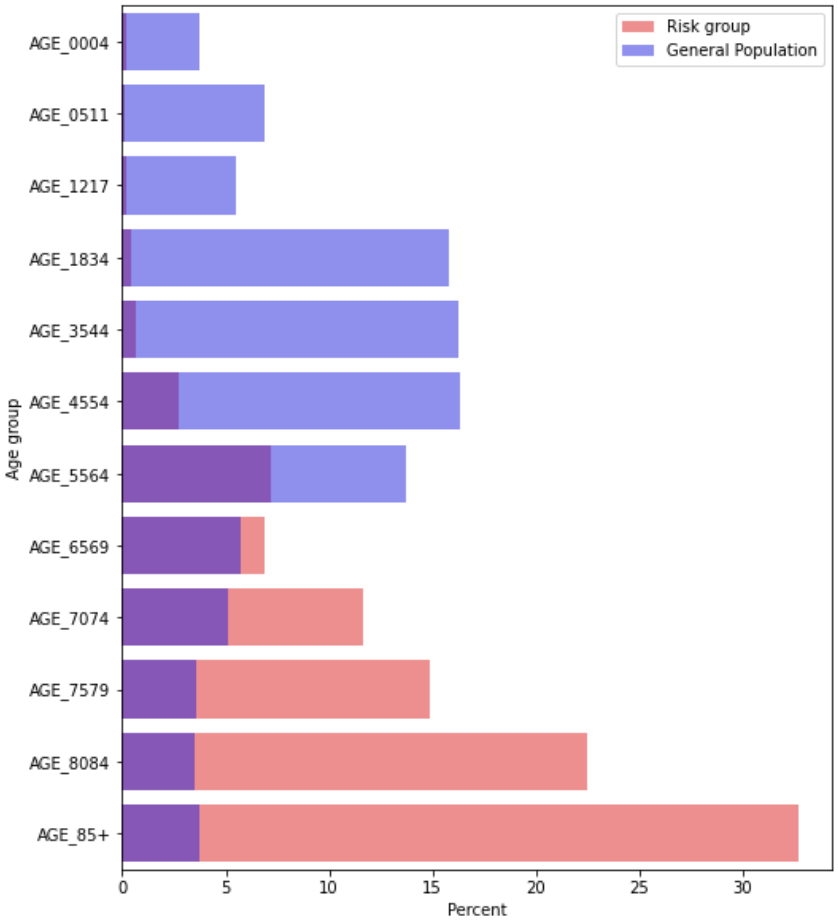


Comorbidities among the top 20000 patients (%)

	MLP	LR	GP
Hospit. in 2018	51.08	49.32	5.68
% of Women	33.95	35.20	51.06
Hypertension	80.66	74.53	9.11
Diabetes Mellitus	44.04	39.02	4.73
COPD	32.62	30.59	1.39
Chronic Renal Fail.	34.27	29.90	0.90
Heart Failure	38.63	34.60	0.73
Ischemic Heart Dis.	18.87	17.54	0.96
Low back pain	17.48	17.86	7.63

GP: General Population

Characteristics of the risk groups

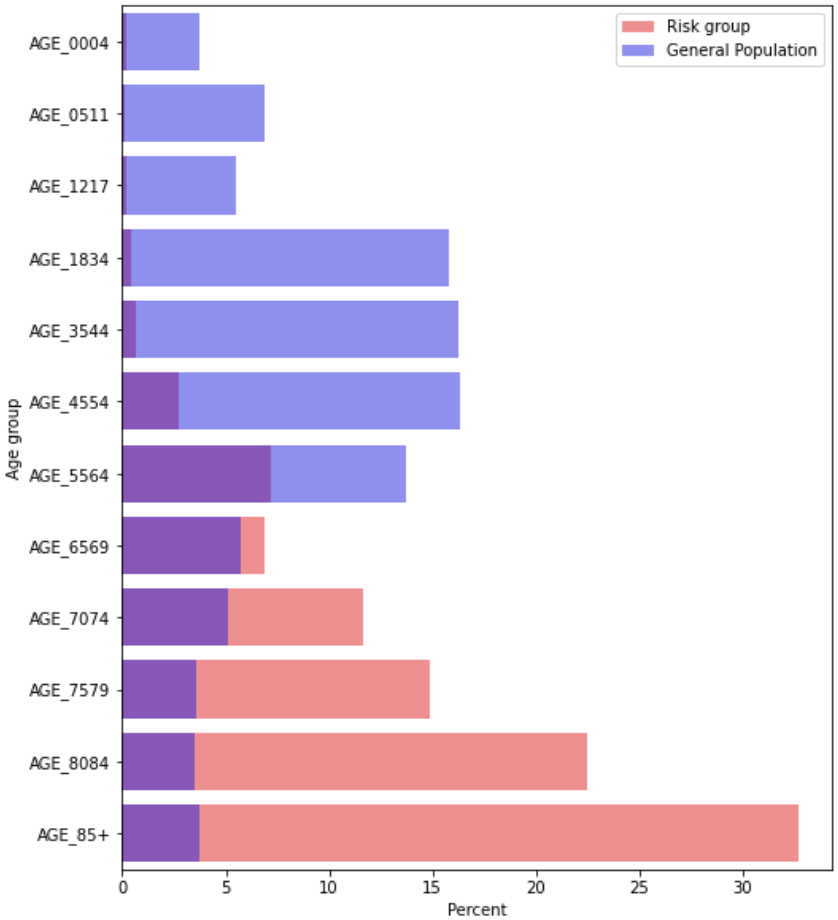


Comorbidities among the top 20000 patients (%)

	MLP	LR	GP
Hospit. in 2018	51.08	49.32	5.68
% of Women	33.95	35.20	51.06
Hypertension	80.66	74.53	9.11
Diabetes Mellitus	44.04	39.02	4.73
COPD	32.62	30.59	1.39
Chronic Renal Fail.	34.27	29.90	0.90
Heart Failure	38.63	34.60	0.73
Ischemic Heart Dis.	18.87	17.54	0.96
Low back pain	17.48	17.86	7.63

GP: General Population

Characteristics of the risk groups



Comorbidities among the top 20000 patients (%)

	MLP	LR	GP
Hospit. in 2018	51.08	49.32	5.68
% of Women	33.95	35.20	51.06
Hypertension	80.66	74.53	9.11
Diabetes Mellitus	44.04	39.02	4.73
COPD	32.62	30.59	1.39
Chronic Renal Fail.	34.27	29.90	0.90
Heart Failure	38.63	34.60	0.73
Ischemic Heart Dis.	18.87	17.54	0.96
Low back pain	17.48	17.86	7.63

GP: General Population

Discussion

Possibilities

- Choose model → Build ready-to-use application
- Display predictions in electronic health record
- Risk group identification: Integrate into alert systems already in place
- Easily change "high-risk" labelling criterion

Limitations

- Highly processed binary predictors → Little room to excel
- No way to account for the different healthcare use patterns of subpopulations → Fairness concerns

Further research

- Why is there such a **low % of women** in risk groups despite similar prevalence? Are there any other **under-represented groups** we can think of? (Work in progress)
- Trends in **model structure** and **variable importance**
- **New data**: Additional variables, different preprocessing
- Introducing the **time** component: Is the weight of recent diagnoses bigger? Does this reduce noise?

Further research

- Why is there such a **low % of women** in risk groups despite similar prevalence? Are there any other **under-represented groups** we can think of? (Work in progress)
- Trends in **model structure** and **variable importance**
- **New data**: Additional variables, different preprocessing
- Introducing the **time** component: Is the weight of recent diagnoses bigger? Does this reduce noise?

Further research

- Why is there such a **low % of women** in risk groups despite similar prevalence? Are there any other **under-represented groups** we can think of? (Work in progress)
- Trends in **model structure** and **variable importance**
- **New data**: Additional variables, different preprocessing
- Introducing the **time** component: Is the weight of recent diagnoses bigger? Does this reduce noise?

Further research

- Why is there such a **low % of women** in risk groups despite similar prevalence? Are there any other **under-represented groups** we can think of? (Work in progress)
- Trends in **model structure** and **variable importance**
- **New data**: Additional variables, different preprocessing
- Introducing the **time** component: Is the weight of recent diagnoses bigger? Does this reduce noise?

Conclusions

We have compared several techniques to predict the probability of unplanned hospitalization in the Basque Country, and we have been able to rank them based on performance by taking variability into account.

- **Best performance** and lowest variability: **Multilayer perceptrons** (median **AUC = 0.802**).
- The **observed overlap** between some of the techniques highlights the importance of a variability analysis

Thank You

Code: AlexOlza/estratificacion

Slides: AlexOlza/ConferenceSlides

bcam basque center for applied mathematics

Osakidetza

EUSKO JAURLARITZA GOBIERNO VASCO
OSASUN SAILA
DEPARTAMENTO DE SALUD

Appendix: Post-calibration

Definición

Calibrated classifier: Its predictions correspond to the true probability distribution for each class.

$$\left.\begin{array}{l}\text{True class } y \in \{0,1\} \\ \text{Predicted probability } \hat{p}\end{array}\right\} \xrightarrow{f} p_i = f(\hat{p}_i)$$

Requirements:

- **Monotone** to preserve discrimination
- **Smooth** to allow stratification

Isotonic Regression
+
PCHIP

Piecewise Cubic Hermite Interpolating Polynomial

Appendix: Preprocessing (ACG)

ACG Johns Hopkings: Private software that groups the thousands of possible diagnostic codes into a few hundred variables, and also clusters patients into morbidity groups.

- ACG: Isomorbidity groups (mutually exclusive)
- EDC: Diagnoses
- Rx-Mg: Pharmacy prescriptions

Examples

ACG 4910: 6-9 Diagnoses, Age>34, 0-1 Major diagnoses
EDC FRE03: Endometriosis
Rx-Mg PSYx030: Psychosocial/Anxiety