Lethality Factors for Positive COVID-19 Testing in Mexico

Final Project for Statistical Machine Learning (2019/2020)

Andrea Cicchini, Victor Plesco and Michele Rispoli

Presentation Outline

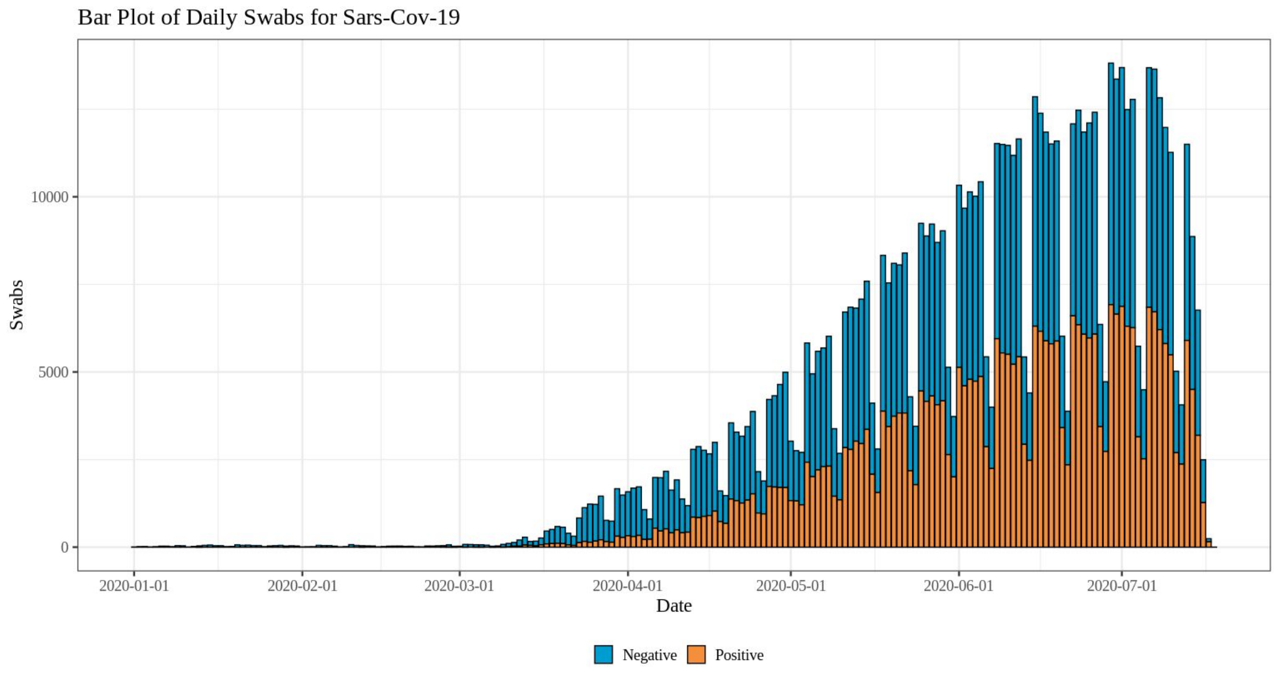
1. Problem Statement
2. Data
3. Models
   - Logistic Regression
   - Bayesian Logistic Regression
   - Neural Networks
4. Conclusions

1. Problem Statement

1 - Problem Statement
COVID-19 in Mexico - Situation

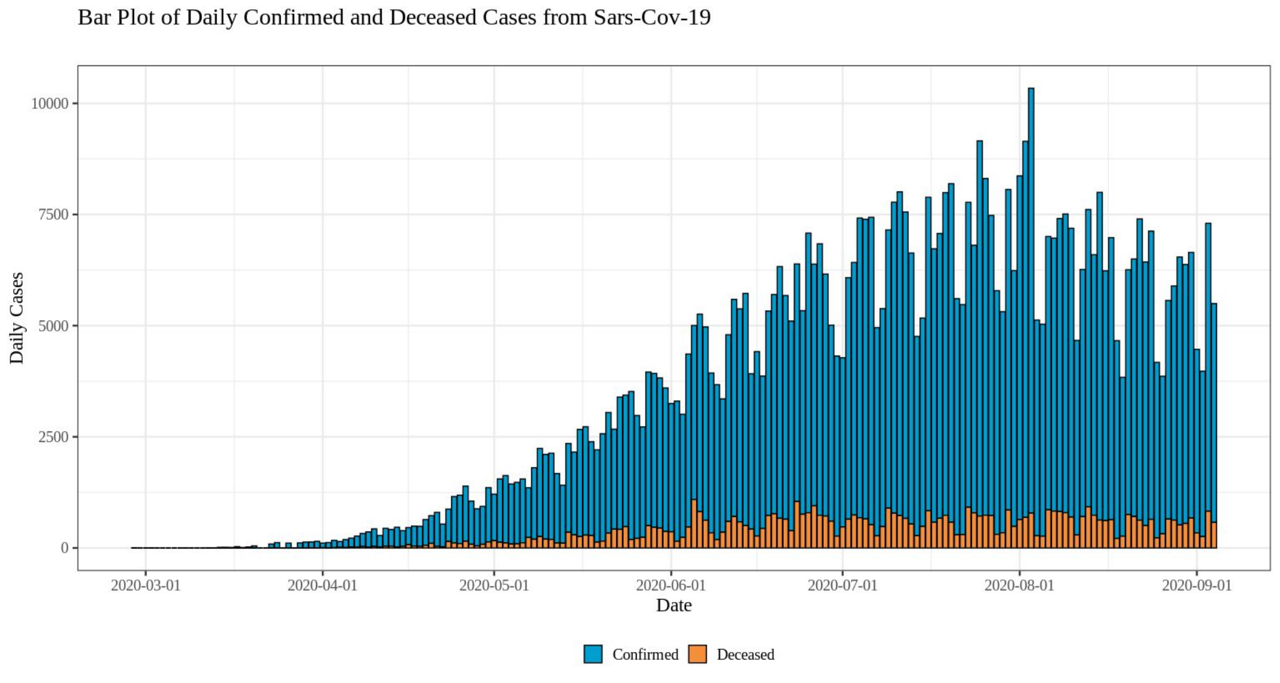
- First case reported on 13 January 2020
- Current situation (as of 05/09/2020, accordingly to [1.b])
  - 1'338'591 Total Swabs (~10’600 swabs/M. citizens, 153rd in the world [2]) *(Italy is 39th with ~151’300 swabs/M.citizens)*
  - 629’409 Confirmed Cases (47% of tested)
  - 67’326 Deceased (~10.7% mortality rate)

1 - Problem Statement
COVID-19 in Mexico - Daily Swabs



- Low number of tests per thousand
  (0.28, versus a median 6.28 per thousand for a group of 50 countries with updated information up to at April 18) [4]

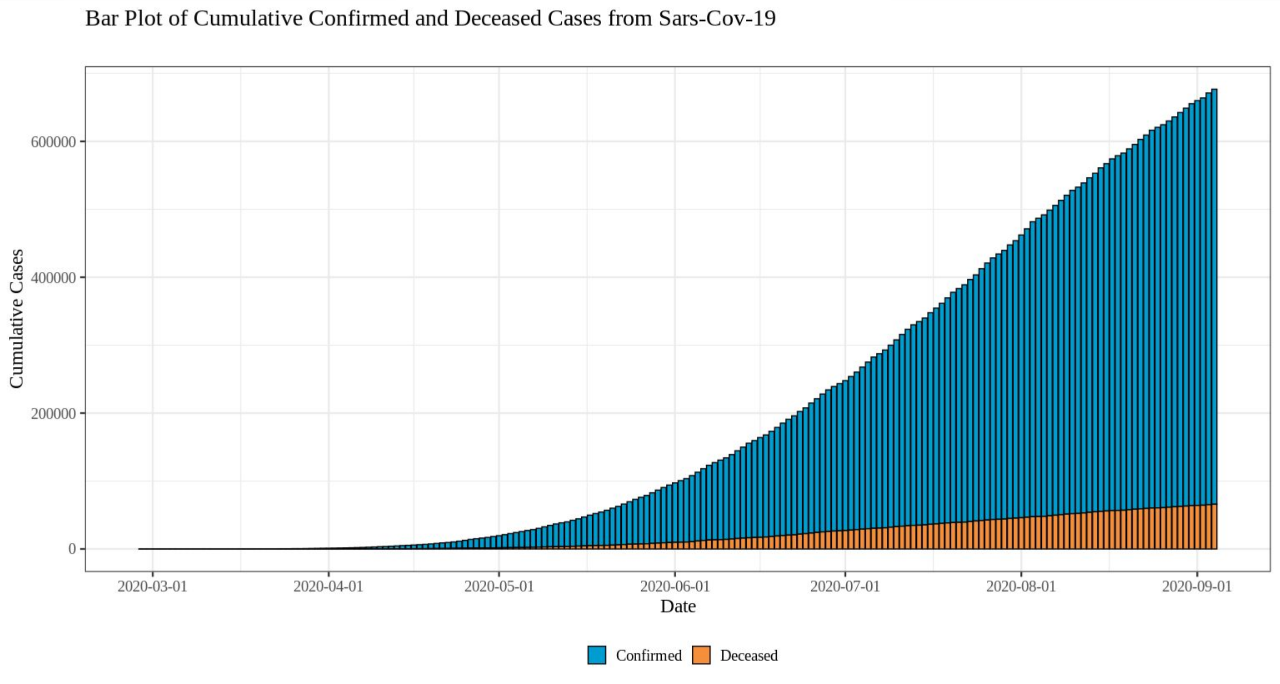
1 - Problem Statement
COVID-19 in Mexico - Daily Confirmed Cases



- Increasing of daily cases (high transmissibility rate)

source: wolrdmeters.infos[2]

1 - Problem Statement
COVID-19 in Mexico - Cumulative Confirmed Cases



- 38’310 deceased over 331’298 positive at 18/07/20

source: wolrdmeters.infos[2]

1 - Problem Statement
Previous studies on lethality

- [4] Highest hazzard factors (i.e. affected deceased/affected recovered in 3 weeks): CKD, COPD, Obesity (April 2020)
- [5] “The presence of pneumonia was also recorded but was considered part of the clinical picture of Covid-19 rather than comorbidity.”
- [6] Symptoms and risk factors as predictors of mortality

1 - Problem Statement
Objectives

- Predictive: Estimate lethality basing on previous pathologies
- Descriptive: Investigate influence of previous pathologies on lethality

Intent: aid in identifying more susceptible subjects (prevention strategy, treatment prioritization)

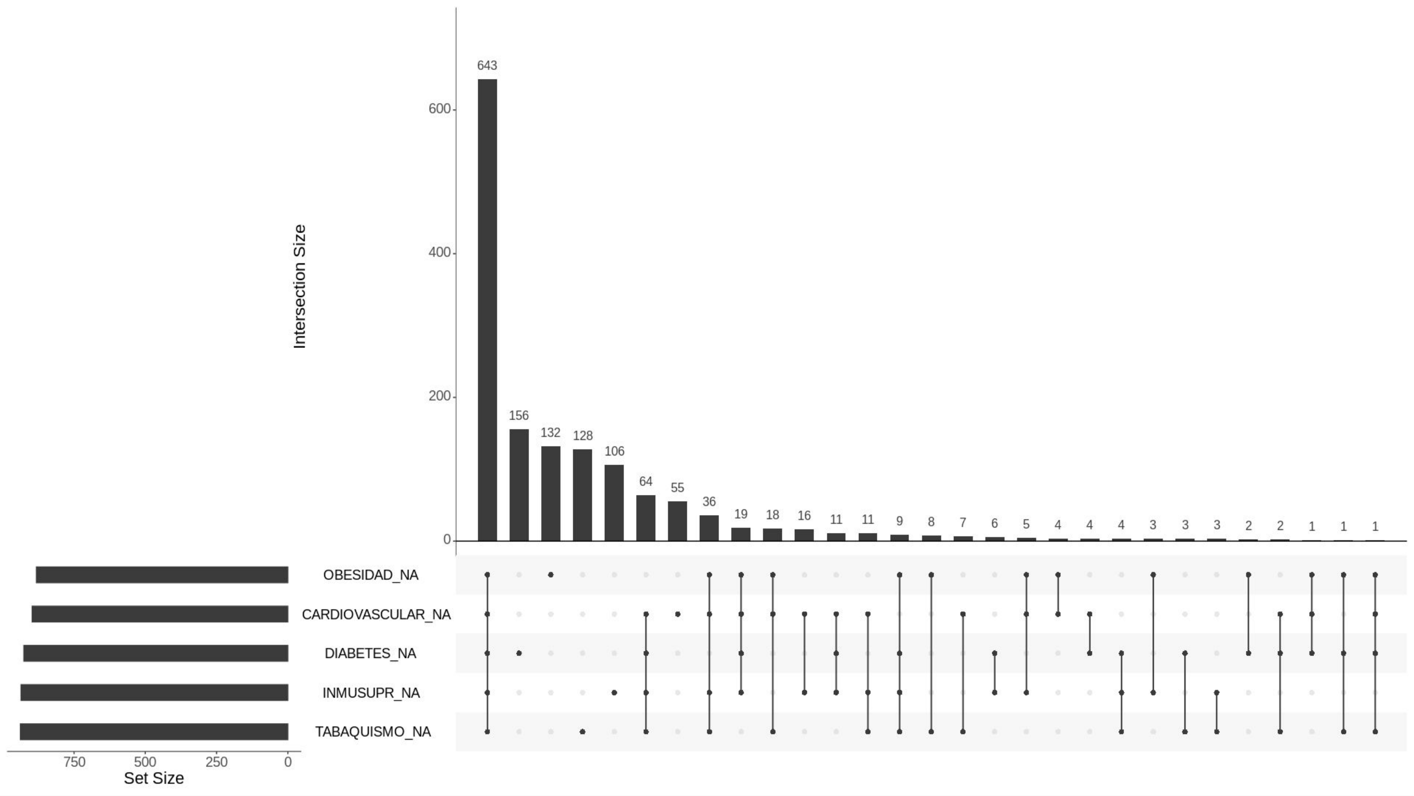
2. Data

2 - Data
Overview: Raw Data

- 727’549 patients
- Source: Secretaría de Salud, Gobierno de México [1.a]
- Last (downloadable) update 18/07/20

```
##       ENTIDAD_UM ID_REGISTRO   SEXO EDAD TIPO_PACIENTE  UCI INTUBADO NEUMONIA
## 1 AGUASCALIENTES      139c84 Female   38            No <NA>     <NA>       No
## 2 AGUASCALIENTES      1a9966 Female   92            No <NA>     <NA>       No
## 3 AGUASCALIENTES      0ef72d   Male   51            No <NA>     <NA>       No
## 4 AGUASCALIENTES      0e5c6f   Male   35            No <NA>     <NA>       No
## 5 AGUASCALIENTES      00748b   Male   14           Yes   No       No       No
## 6 AGUASCALIENTES      07c6ec Female   22            No <NA>     <NA>       No
##   EMBARAZO DIABETES EPOC ASMA INMUSUPR HIPERTENSION CARDIOVASCULAR OBESIDAD
## 1       No       No   No   No       No          Yes             No       No
## 2       No       No  Yes   No       No          Yes             No       No
## 3     <NA>       No   No   No       No           No             No       No
## 4     <NA>       No   No   No       No           No             No       No
## 5     <NA>      Yes   No   No       No           No             No       No
## 6       No       No   No   No       No           No             No       No
##   RENAL_CRONICA TABAQUISMO OTRO_CASO RESULTADO FECHA_INGRESO FECHA_SINTOMAS
## 1            No         No      <NA>  Negative    2020-05-27     2020-05-23
## 2            No         No        No  Negative    2020-05-12     2020-05-11
## 3            No        Yes        No  Negative    2020-05-25     2020-05-25
## 4            No        Yes        No  Negative    2020-05-31     2020-05-29
## 5            No         No      <NA>  Positive    2020-06-14     2020-06-14
## 6            No         No       Yes  Negative    2020-03-30     2020-03-25
##   FECHA_DEF FALLECIDO
## 1      <NA>        No
## 2      <NA>        No
## 3      <NA>        No
## 4      <NA>        No
## 5      <NA>        No
## 6      <NA>        No
```

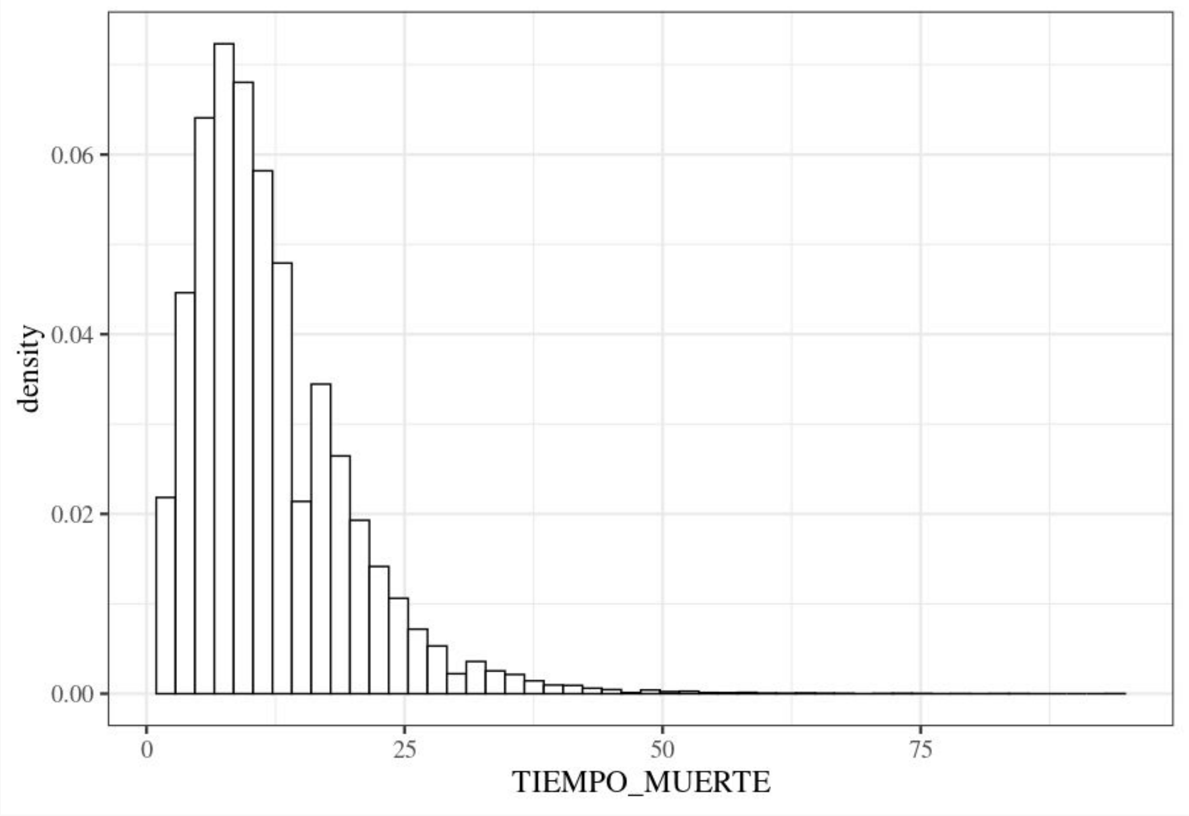
2 - Data
Overview: Missing Values



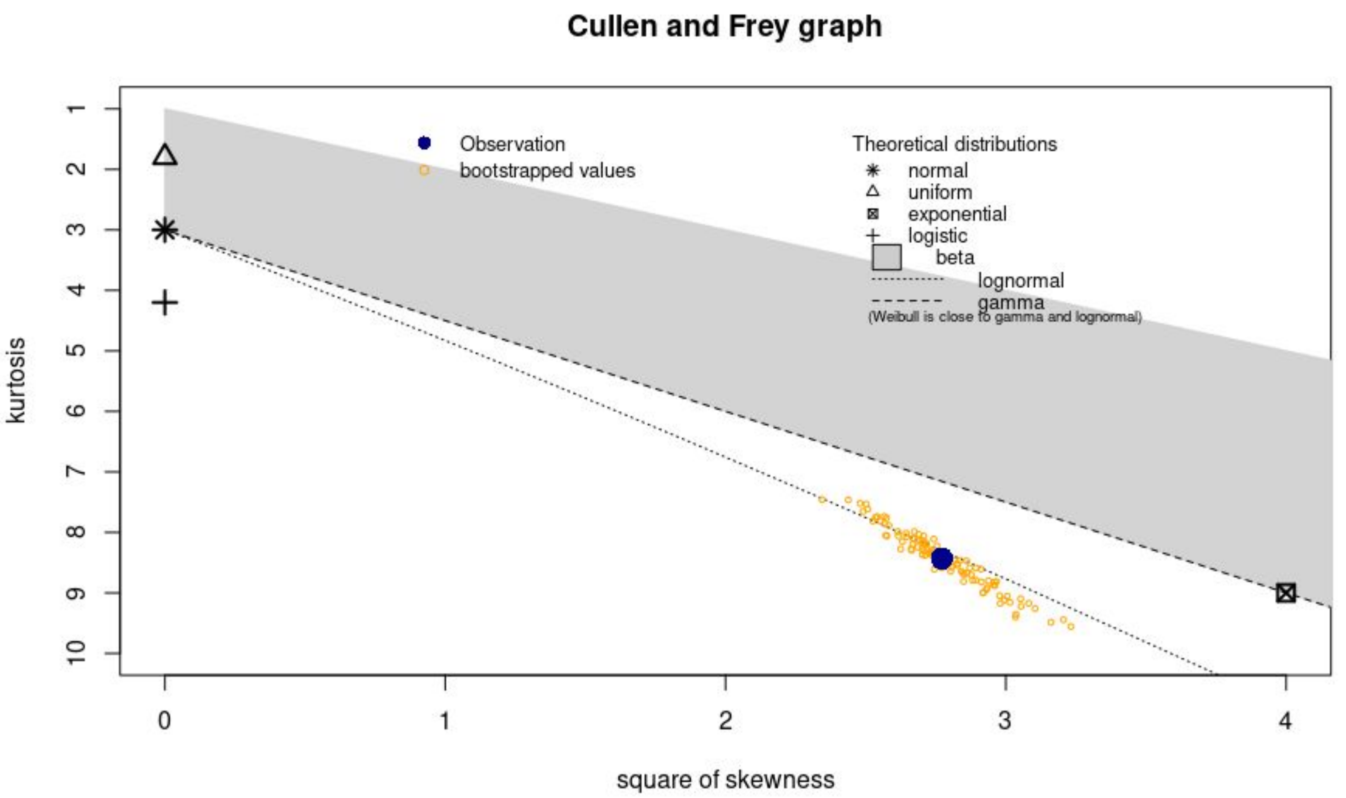
2 - Data Pre-processing

- Remove cases with NAs
- Considering only confirmed cases (i.e. exclude negatives)
- Deduce Recovered: estimate 99% decease time from symptoms detection

2 - Data
Pre-processing: Decease Time

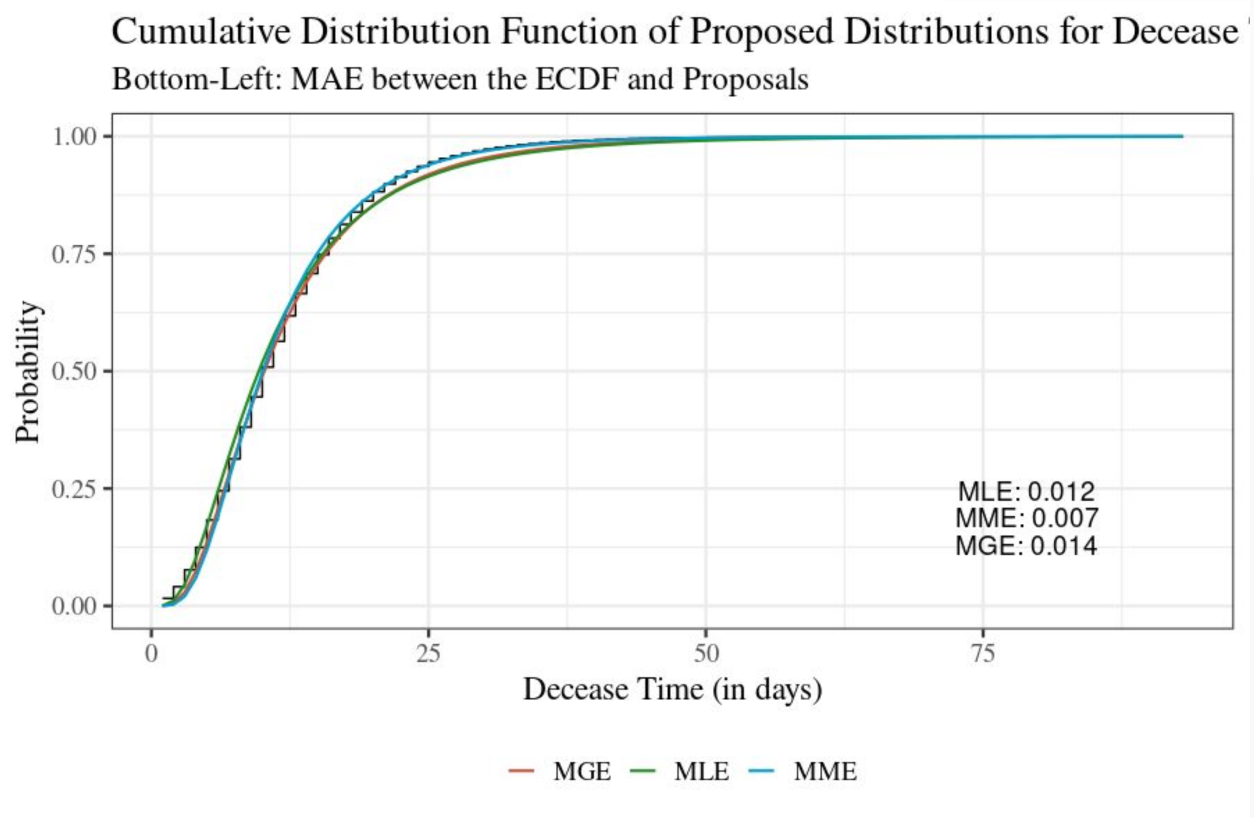


2 - Data
Pre-processing: Decease Time parametrization

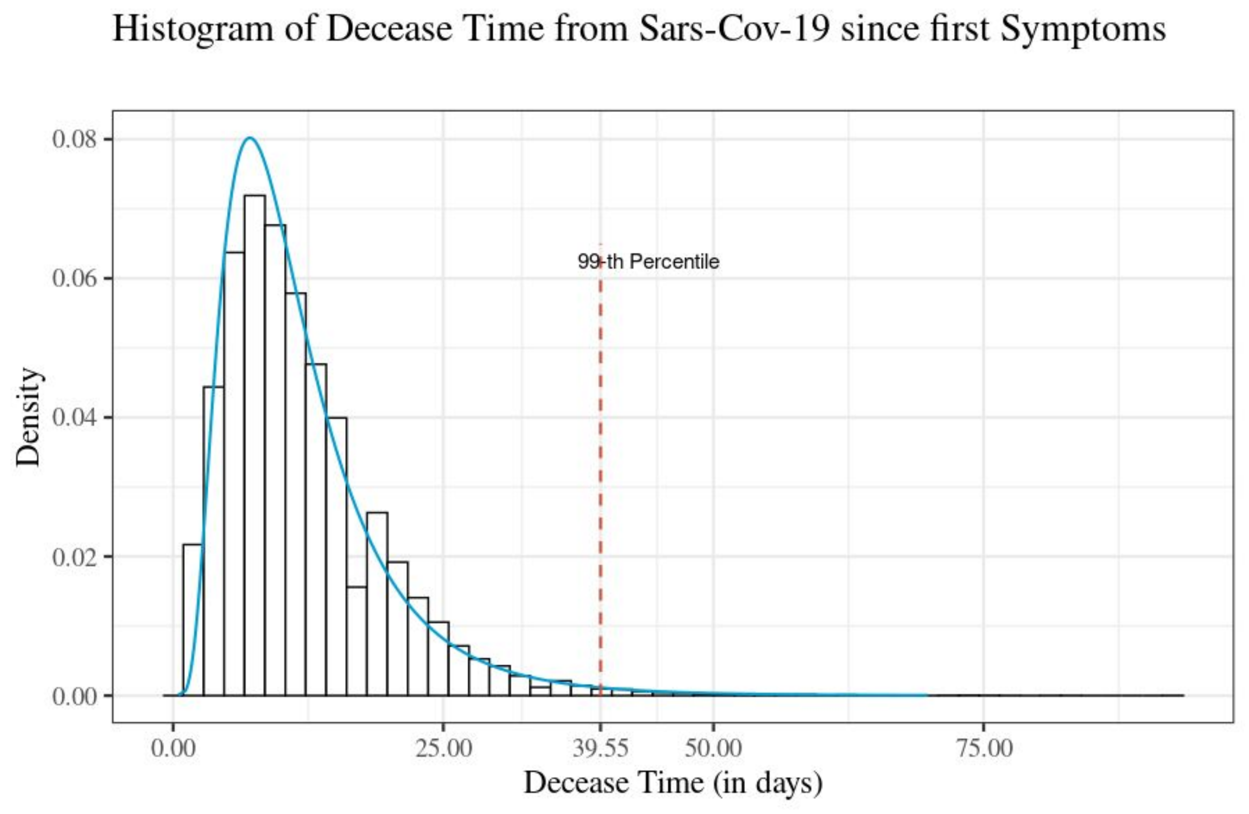


2 - Data
Pre-processing: Decease Time Parametrization

- Maximum likelihood estimation
- Method of moments
- Maximum goodness of fit

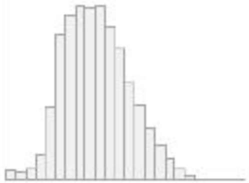


2 - Data
Pre-processing: Decease ?



2 - Data
Final Dataset: Overview

- 163'666 rows
- (EPOC, INMUSUPR) < 2%
- (ASMA, CARDIOVASCULAR, CKD) < 3%
- TABAQUISMO 7,9%

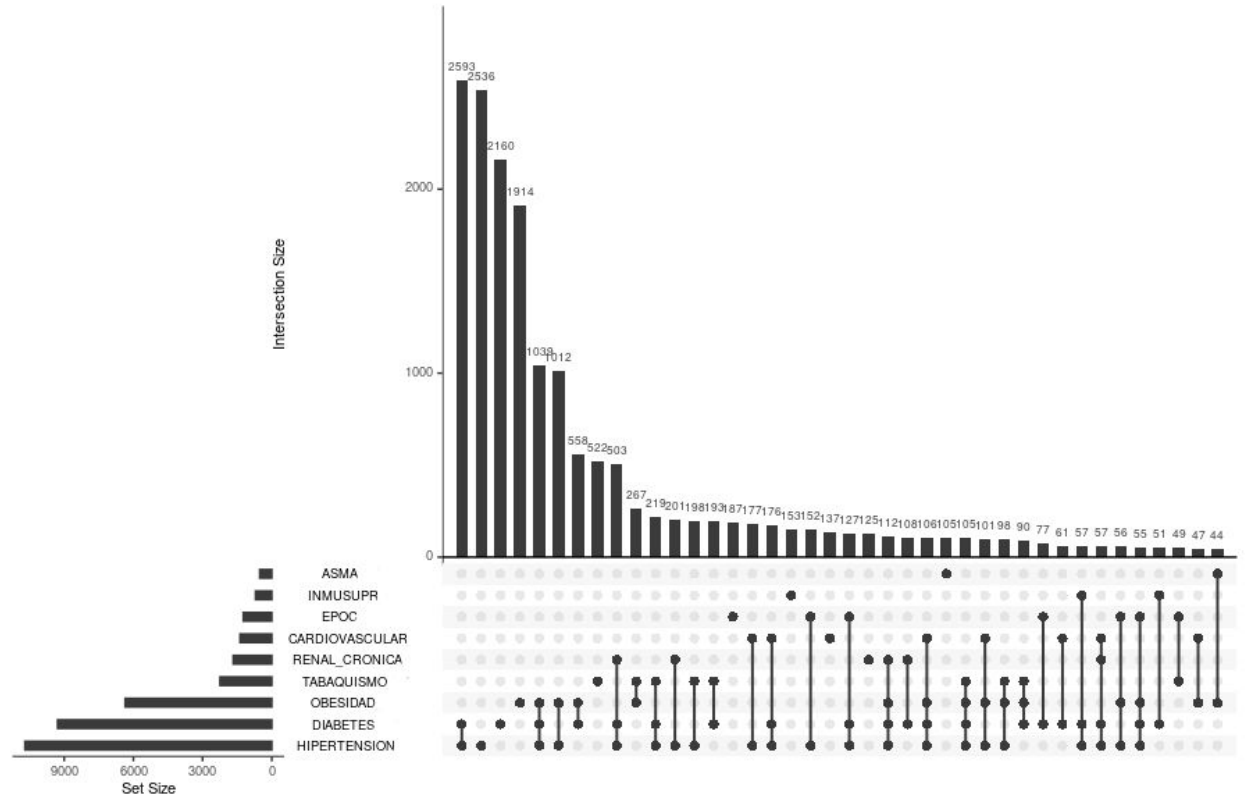
No	Variable	Stats / Values	Freqs (% of Valid)	Graph
1	SEXO [factor]	1. Female 2. Male	72990 (44.6%) 90676 (55.4%)	
2	EDAD [integer]	Mean (sd) : 46.1 (16.2) min < med < max: 0 < 45 < 120 IQR (CV) : 23 (0.4)	107 distinct values	
3	DIABETES [factor]	1. No 2. Yes	136082 (83.2%) 27584 (16.9%)	
4	EPOC [factor]	1. No 2. Yes	160655 (98.2%) 3011 (1.8%)	
5	ASMA [factor]	1. No 2. Yes	159120 (97.2%) 4546 (2.8%)	
6	INMUSUPR [factor]	1. No 2. Yes	161314 (98.6%) 2352 (1.4%)	

2 - Data
Final Dataset: Overview

- 163’666 rows
- (EPOC, INMUSUPR) < 2%
- (ASMA, CARDIOVASCULAR, CKD) < 3%
- TABAQUISMO 7,9%

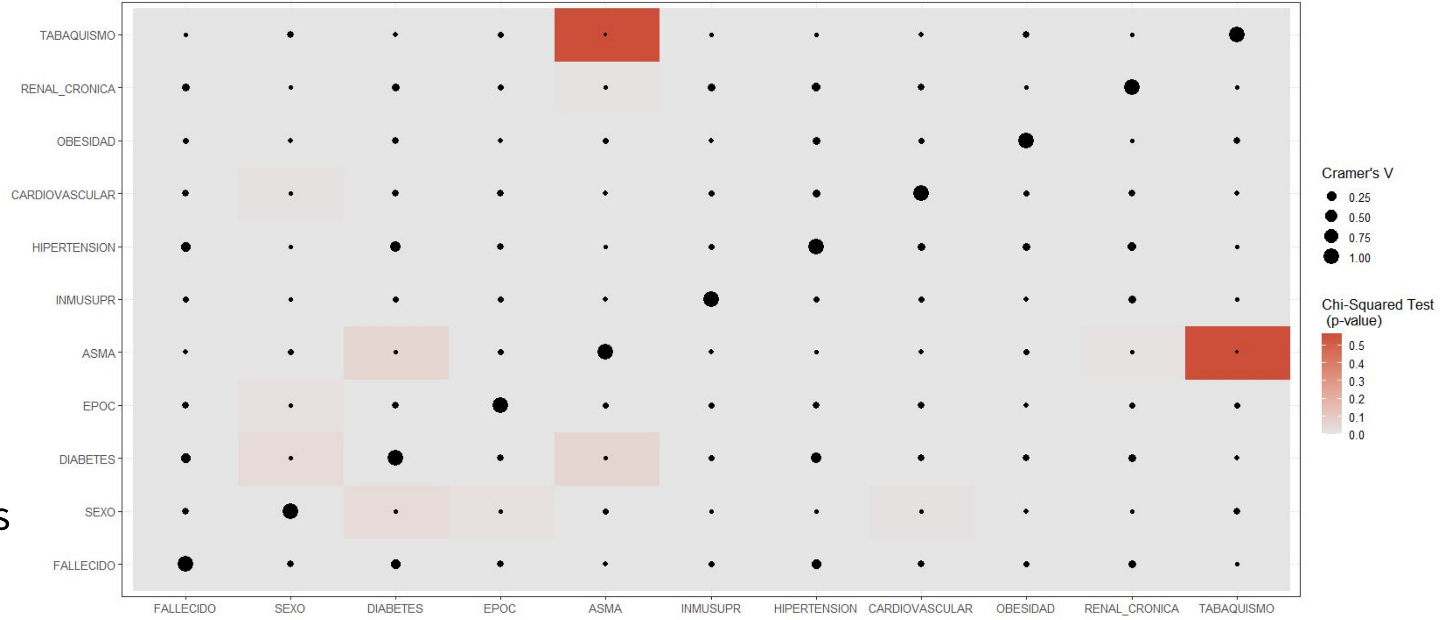
No	Variable	Stats / Values	Freqs (% of Valid)	Graph
7	HIPERTENSION [factor]	1. No 2. Yes	129938 (79.4%) 33728 (20.6%)	
8	CARDIOVASCULAR [factor]	1. No 2. Yes	159652 (97.5%) 4014 (2.5%)	
9	OBESIDAD [factor]	1. No 2. Yes	131046 (80.1%) 32620 (19.9%)	
10	RENAL_CRONICA [factor]	1. No 2. Yes	159980 (97.8%) 3686 (2.2%)	
11	TABAQUISMO [factor]	1. No 2. Yes	150699 (92.1%) 12967 (7.9%)	
12	FALLECIDO [factor]	1. No 2. Yes	138154 (84.4%) 25512 (15.6%)	

2 - Data
Final Dataset: Deceased risk factors



2 - Data
Contingency Analysis: Risk Factors

- Most pairs have Cr.V<0.25
- Hypertension and Diabetes slightly correlated
- Tabagism and Asthma correlation has very high p-value (asthmatics usually don't smoke)

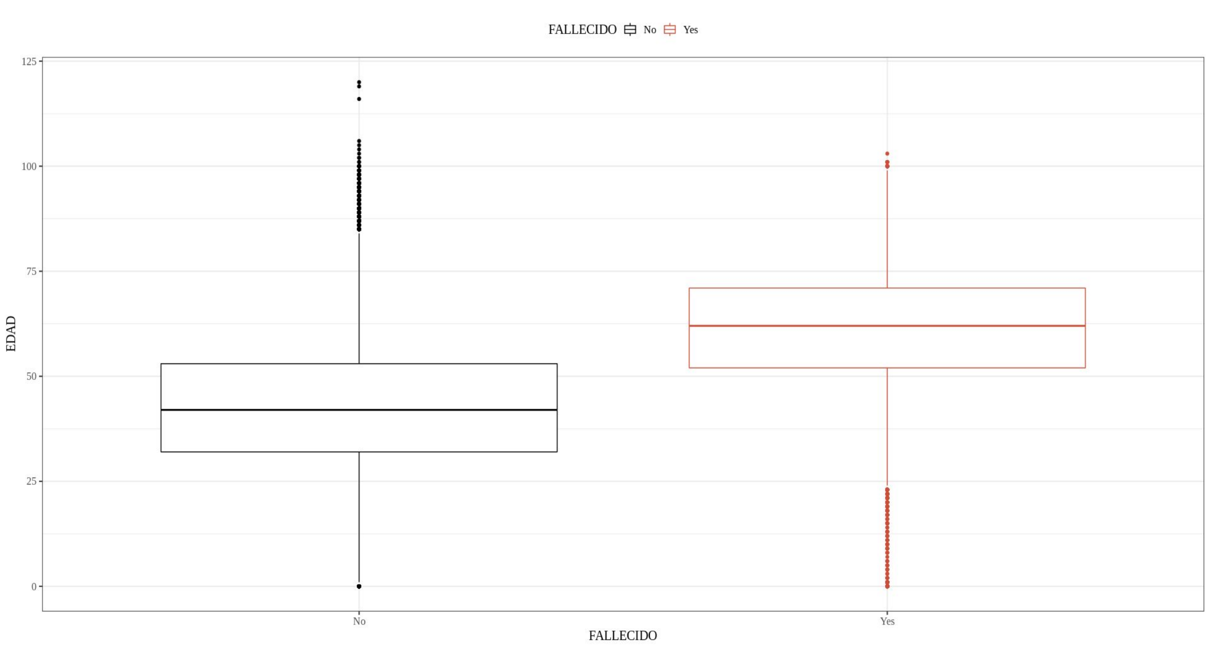


2 - Data
Contingency Analysis: Age

Statistically significant difference between mean ages of **deceased (~61)** and **recovered (~43)**.

T-TEST

t = -254.01, df = 74205,
95% CI = [-17.55558 , -17.28674]
p-value < 2.2 * e-16



3.1 Models: Logistic Regression

3 - Models
Logistic Regression: Introduction

- Structure
  - Input
    - <Factor> Gender = {“Female”, “Male”}
    - <Numeric> Age = [0, 120]
    - <Factor> Risk Factors = {“No”, “Yes”} (9/9)
  - Output
    - <Factor> Deceased = {“No”, “Yes”} | CUTOFF

3 - Models
Logistic Regression: Variable Selection

- Stepwise Forward Selection on a 5-fold Cross Validation

```
for i = 0 to K - 1 do                          with K = cross-validation runs
    for j = 0 to M - 1 do                      with M = model complexity
        for r = 0 to N - j do                  with N = regressors
            model [ j ] = min(DEVIANCE [ r_0 , ..., r_{N-j} ]);
            AIC, LRT;
        end for
    end for
end for
```

Note: NO STOPPING CRITERION

3 - Models
Logistic Regression: Variable Selection

Predictors	Deviance	Deviance (SE)	AIC	AIC (SE)	LRT	LRT (SE)	Wald Test	Wald Test (SE)
+ EDAD	92653.98	360.41	92657.98	360.41	0.0000	0.0000	0.0000	0.0000
+ DIABETES	91434.31	352.82	91440.31	352.82	0.0000	0.0000	0.0000	0.0000
+ SEXO	90407.51	358.89	90415.51	358.89	0.0000	0.0000	0.0000	0.0000
+ RENAL_CRONICA	89918.23	364.27	89928.23	364.27	0.0000	0.0000	0.0000	0.0000
+ OBESIDAD	89563.71	358.45	89575.71	358.45	0.0000	0.0000	0.0000	0.0000
+ HIPERTENSION	89432.67	352.01	89446.67	352.01	0.0000	0.0000	0.0000	0.0000
+ INMUSUPR	89378.93	341.04	89394.93	341.04	0.0000	0.0000	0.0000	0.0000
+ EPOC	89360.49	336.19	89379.69	336.88	0.0015	0.0011	0.0006	0.0004
+ ASMA	89362.62	344.73	89381.42	344.02	0.0017	0.0019	0.0018	0.0020
+ CARDIOVASCULAR	89353.34	339.73	89375.74	339.79	0.2160	0.2044	0.2233	0.2071
+ TABAQUISMO	89352.87	339.54	89376.47	339.49	0.2588	0.2118	0.2604	0.2103

+ (EPOC) + (EPOC) + (ASMA) + (ASMA)

3 - Models
Logistic Regression: Train and Test Errors

- Train & Test Errors over Complexity
- CUTOFF = max(Sensitivity, Accuracy)
- A model with complexity 9 (+ASMA) presents a strange behavior

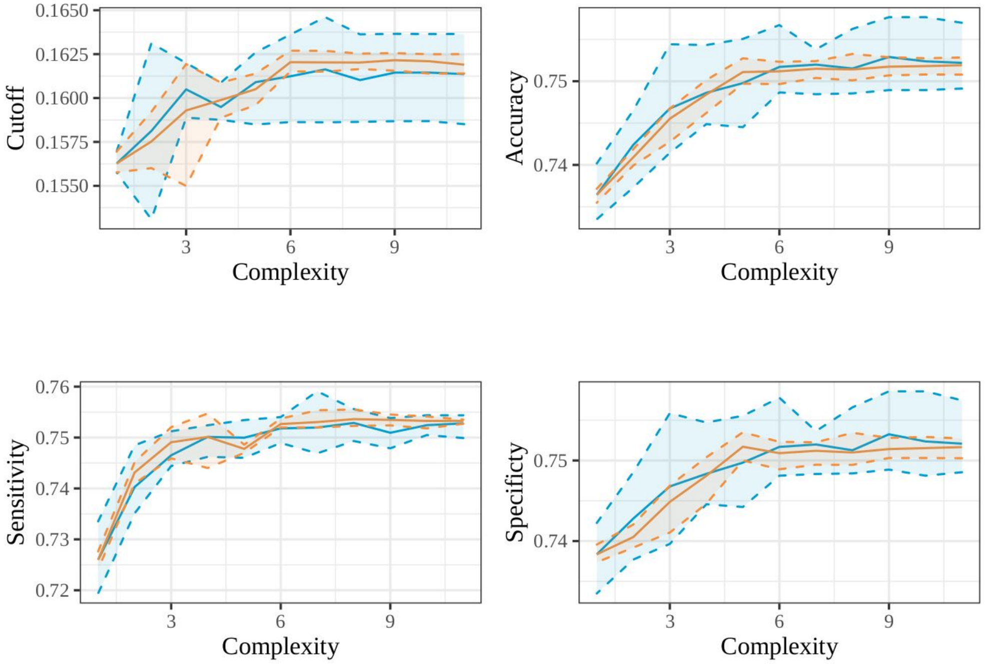


Figure 1: Train (orange) and test (blue) errors for model complexity as average of a 5-fold Cross Validation. The cutoff value is chosen to maximize both accuracy and sensitivity. Dashed lines represent min/max values of the cross validation runs. Top-Left: Cutoff values for model complexity. Top-Left: Accuracy values for model complexity. Bottom-Right: Sensitivity values for model complexity. Top-Right: Specificity values for model complexity.

3 - Models
Logistic Regression: Results

	Estimates	Standard Error	2.5 (%)	97.5 (%)
(Intercept)	-5.9283660	0.0348357	-5.9968496	-5.8602833
SEXOMale	$\boxed{0.5977156}$	0.0160538	0.5662831	0.6292166
EDAD	0.0683840	0.0005712	0.0672666	0.0695056
DIABETESYes	$\boxed{0.5262683}$	0.0178294	0.4912967	0.5611887
EPOCYes	0.1513203	0.0425994	0.0676822	0.2346836
INMUSUPRYes	0.4299422	0.0532283	0.3251884	0.5338675
HIPERTENSIONYes	0.2280410	0.0178166	0.1930922	0.2629341
OBESIDADYes	0.3490336	0.0181811	0.3133573	0.3846291
RENAL_CRONICAYes	$\boxed{0.8865093}$	0.0398067	0.8084539	0.9645048

	Accuracy	Accuracy (SE)	Sensitivity	Sensitivity (SE)	Specificity	Specificity (SE)
Train	0.7514053	0.0013720	0.7536154	0.0013112	0.7509969	0.0018114
Test	0.7515183	0.0030176	0.7528687	0.0025328	0.7512694	0.0033521

- All coefficients are positive (no protective factors)
- The most influential factors are RENAL_CRONICA, SEXOMale and DIABETES

3.2 Models: Bayesian Logistic Regression

3 - Models
Bayesian Logistic Regression: Introduction

- Structure
  - Input
    - <Factor> Gender = {“Female”, “Male”}
    - <Numeric> Age = [0, 120]
    - <Factor> Risk Factors = {“No”, “Yes”} (6/9)
      - Best Logistic Regression Model
  - Output
    - <Factor> Deceased = {“No”, “Yes”} | CUTOFF
  - Priors

3 - Models
Bayesian Logistic Regression: Priors

Model 1

weakly informative

- Student t prior with 7 degrees of freedom and a scale of 2.5 [7]
- Reasonable when coefficients:
  - are close to zero but have some chance of being large
  - are as likely to be positive as they are to be negative

Model 2

informative (?)

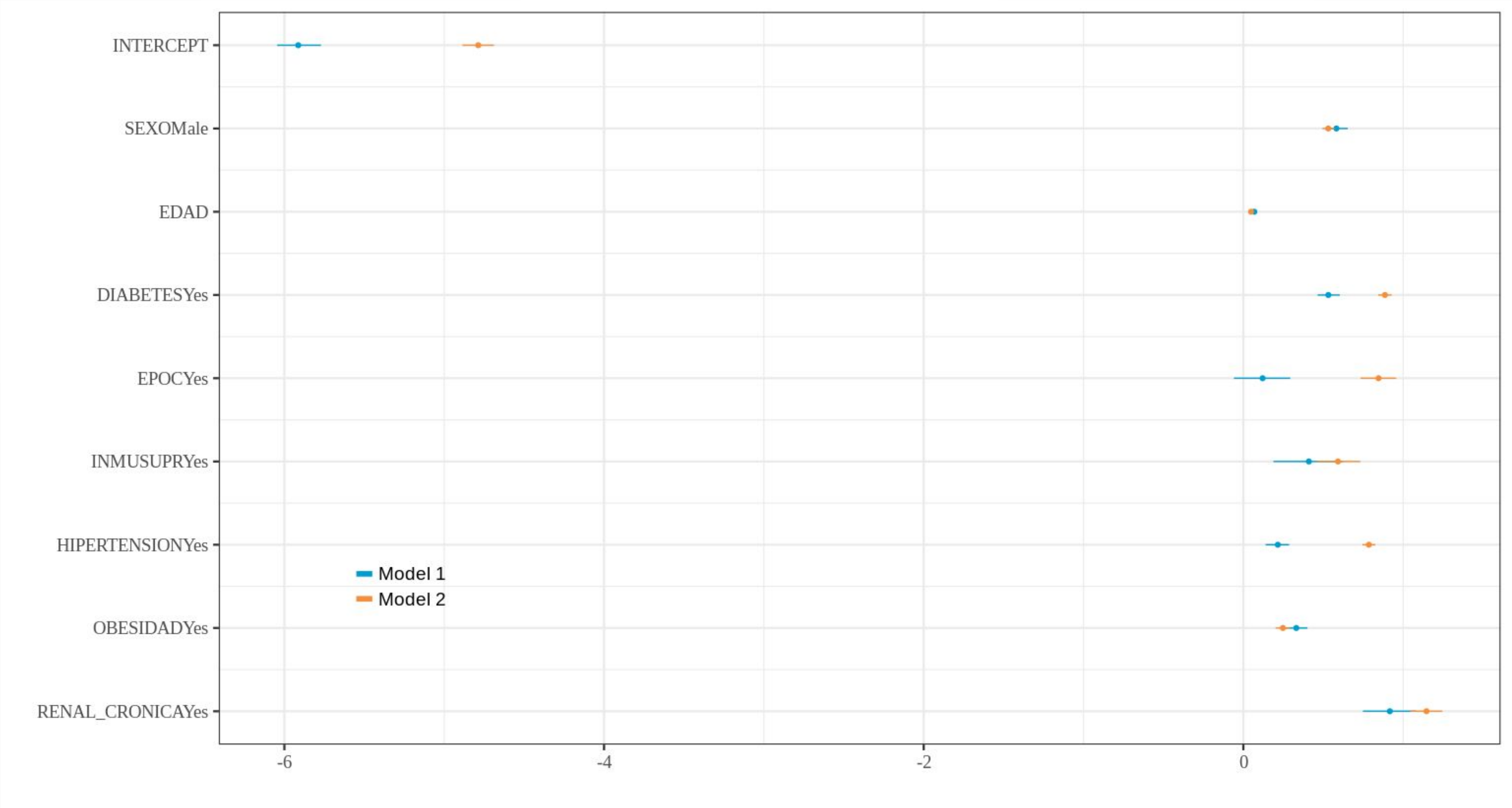
- Let’s test our data! Normal prior with mean and standard deviation defined by a bootstrap resampling on our dataset
- What to expect? If our data has to be an unbiased sample, we would expect faster convergence to the true values of the log(odds)

3 - Models
Bayesian Logistic Regression: Implementation

- stan_glm via MCMC
  - 8 chains
  - 2000 iterations
- 80/20 Train/Test Split
- CUTOFF = max(Sensitivity, Accuracy)

5.32 hours later…

3 - Models
Bayesian Logistic Regression: Posteriors



3 - Models
Bayesian Logistic Regression: Diagnostics

	Model 1		Model 2	
	n_eff	Rhat	n_eff	Rhat
(Intercept)	8435	1	8354	1
SEXOMale	10501	1	7772	1
EDAD	9218	1	8157	1
DIABETESYes	9178	1	7363	1
EPOCYes	11737	1	7295	1
INMUSUPRYes	10740	1	8648	1
HIPERTENSIONYes	9803	1	7962	1
OBESIDADYes	10460	1	8182	1
RENAL_CRONICAYes	10966	1	7589	1

	Model 1			Model 2		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Train	0.75252	0.751342	0.752739	0.738742	0.750659	0.736531
Test	0.752032	0.751694	0.752093	0.749496	0.737744	0.751624

- *n_eff* is smaller in Model 2, identifying a higher autocorrelation among samples (however high enough)
- In terms of accuracy and sensitivity Model 1 performs better than Model 2

3 - Models
Comparison - Classical vs Bayesian Logistic

- Standard errors are impressively high!
- Coefficients of the two models have almost the same values

Bayesian Logistic Regression

	Estimates	Standard Error	2.5 (%)	97.5 (%)
(Intercept)	-5.9136490	0.3633826	-7.0054048	-5.5775794
SEXOMale	0.5821162	0.1625301	0.1894201	0.8325200
EDAD	0.0685065	0.0058403	0.0641918	0.0872264
DIABETESYes	0.5318281	0.1859552	0.1793758	0.9088702
EPOCYes	0.1192632	0.4639468	-0.8589502	0.9572092
INMUSUPRYes	0.4092256	0.5691893	-1.1362457	1.0961667
HIPERTENSIONYes	0.2153859	0.1926825	-0.5560629	0.2046989
OBESIDADYes	0.3309566	0.1795412	0.0962327	0.7862833
RENAL_CRONICAYes	0.9166825	0.4114104	0.6259236	2.2382326

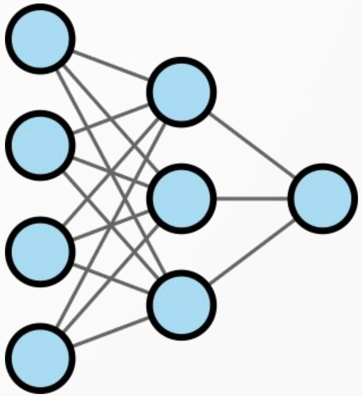
Logistic Regression

	Estimates	Standard Error	2.5 (%)	97.5 (%)
(Intercept)	-5.9283660	0.0348357	-5.9968496	-5.8602833
SEXOMale	0.5977156	0.0160538	0.5662831	0.6292166
EDAD	0.0683840	0.0005712	0.0672666	0.0695056
DIABETESYes	0.5262683	0.0178294	0.4912967	0.5611887
EPOCYes	0.1513203	0.0425994	0.0676822	0.2346836
INMUSUPRYes	0.4299422	0.0532283	0.3251884	0.5338675
HIPERTENSIONYes	0.2280410	0.0178166	0.1930922	0.2629341
OBESIDADYes	0.3490336	0.0181811	0.3133573	0.3846291
RENAL_CRONICAYes	0.8865093	0.0398067	0.8084539	0.9645048

3.3 Models: Neural Networks

3 - Models
Neural Networks: Architecture

- Architecture : FFNN
  - Input
    - Gender - One-hot
    - Age - rescaled in [0,1]
    - Risk factors - {0,1} (6/9 components)
  - Output = Sigmoid
  - Hidden
    - Activation: Hyperbolic Tangent/ReLU
    - Trial and error refinement of no. layers, units



3 - Models
Neural Networks: Training

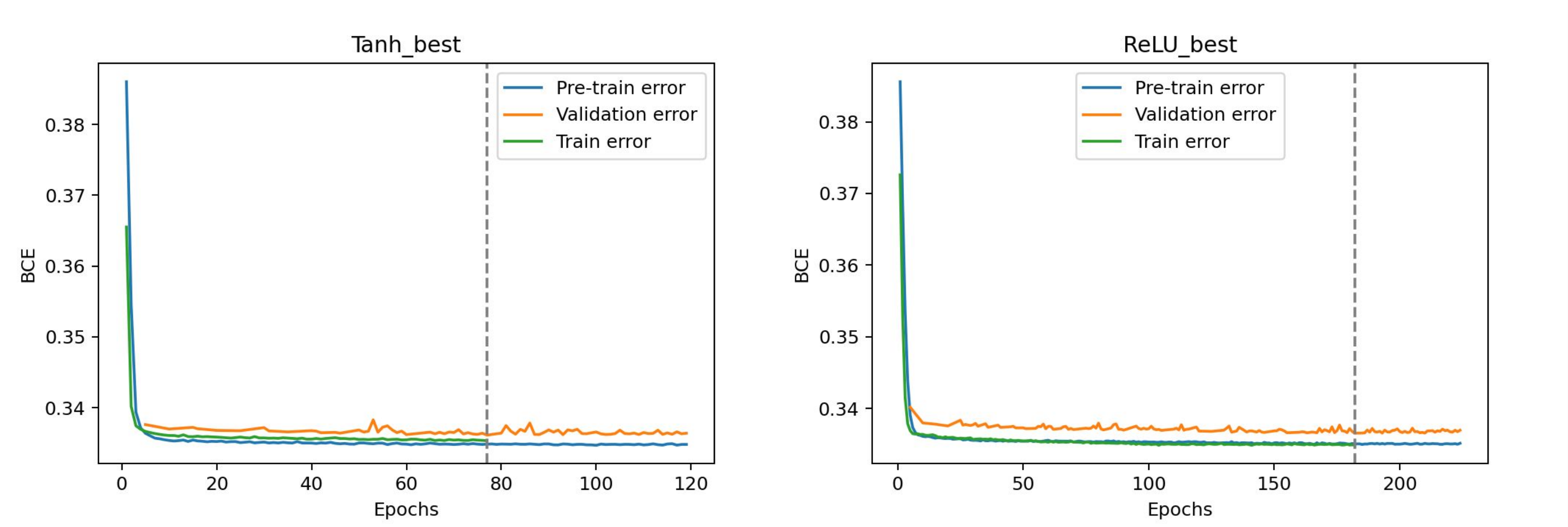
- Loss: Binary Cross Entropy

$$\mathrm{BCE}(Y, \hat{Y}) = -\frac{1}{N}\sum_{n=1}^{N} y_n \log(\hat{y}_n) + (1 - y_n)\log(1 - \hat{y}_n)$$

- Optimization:
  - Adam optimizer, default parameters [3]
  - Training mini-batches (size=100)

3 - Models
Neural Networks: Regularization

- Regularization: Early Stopping
  - 70/30 split of the training set
  - Patience = 40



3 - Models
Neural Networks: Evaluation Metrics

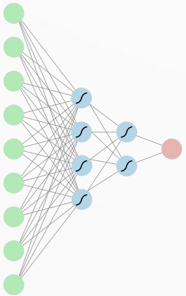
- Evaluation:
  - 80/20 split of the original dataset (Test set)
  - Test Loss (BCE)
  - Binary predictions: threshold selection
    - Best accuracy and sensitivity
    - Confusion Matrix

3 - Models
Neural Networks: Results

- ## Final Models

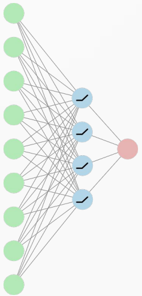
```
Architecture:  9, 4, 2, 1
Activation:    Tanh
Train ep:      77
Train loss:    0.3354018032550812
Test loss:     0.3394022285938263
Best thresh:   0.1849365234375
Accuracy @ t:  0.750198570293884
Sens @ t:      0.7455968688845401

Confusion Matrix
        0      1
P0 [20747   1300]
P1 [ 6877   3810]
```



```
Architecture:  9, 4, 1
Activation:    ReLU
Train ep:      182
Train loss:    0.3349879682064056 4
Test loss:     0.338871359825134 3
Best thresh:   0.200073242187 5
Accuracy @ t:  0.749770880430133 8
Sens @ t:      0.747553816046966 7

Confusion Matrix
        0      1
P0 [20723   1290]
P1 [ 6901   3820]
```

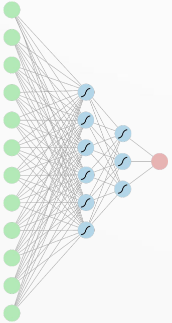


3 - Models
Neural Networks: Results

- Final Models, including ALL covariates

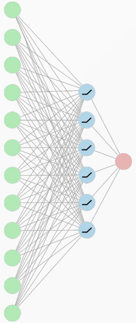
```
Architecture: 12, 6, 3, 1
Activation:   Tanh
Train ep:     107
Train loss:   0.3344671130180359
Test loss:    0.3391265571117401
Best thresh:  0.1878662109375
Accuracy @ t: 0.7504429645017413
Sens @ t:     0.7497064579256361

Confusion Matrix
         0        1
P0 [20734    1279]
P1 [ 6890    3831]
```



```
Architecture: 12, 6, 1
Activation:   ReLU
Train ep:     90
Train loss:   0.3346908092498779З
Test loss:    0.3392436504364013
Best thresh:  0.193237304687
Accuracy @ t: 0.752153723956742
Sens @ t:     0.744031311154598

Confusion Matrix
         0        1
P0 [20819    1308]
P1 [ 6805    3802]
```



3 - Models
Model Comparison

Model	Logistic	Bayesian Logistic (1)	NN - ReLU	NN - Tanh
Test Accuracy	0.7515*	**0.7520**	0.7498	0.7502
Test Sensitivity	**0.7529***	0.7517	0.7476	0.7456

*: 5-fold CV estimate

Conclusions

Prediction: Best model achieves ~75% accuracy

Description: CKD and Diabetes appear to be the most influential Risk Factors ([4] used ~50’000 cases)

Next Steps

- Include latest data from [1.a] (whenever available again!)
- Include data from other countries
- Extend by accounting for days from symptoms manifestation
- Extend for “on hospital” usage: account for symptoms and treatments

References

1. Official Mexican government COVID-19 Data
    a. https://www.gob.mx/salud/documentos/datos-abiertos-152127
    b. https://coronavirus.gob.mx/datos/
2. https://www.worldometers.info/coronavirus/
3. D. P. Kingma, J. Ba - *Adam: A Method for Stochastic Optimization (2015)*
4. P. Solís, H. Carreño - *COVID-19 Fatality and Comorbidity Risk Factors among Diagnosed Patients in Mexico,Patricio Solís, El Colegio de México (2020)*
5. E. Hernández-Garduño - *Obesity is the comorbidity more strongly associated for Covid-19 in Mexico. A case-control study (2020)*
6. R.Du et al - *Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study* (2020)
7. https://mc-stan.org/rstanarm/articles/binomial.html

Thanks for your attention!