INMI Data Analysis

09/02/2020

O. Iorio[1] per :

Il team INFN

F. Baldo[2], D. Bonacorsi [2], F. Carnevali[1], L.Lista[1], A. Piovani[2], S. Rossi Tisbeni[3]

Il team INMI

C. Cimaglia, E. Girardi, S. Lanini, E. Nicastri, A.Navarra, P. Piselli, G. Rinonapoli

1) INFN Napoli e Università degli Studi Di Napoli Federico II
2) INFN Bologna e Università di Bologna Alma Mater Studiorum
3) INFN - CNAF

The DETECOVID project

Project of Ministero della Salute (COVID-2020-12371675):

https://www.inmi.it/servizio/progetto-ministero-della-salute

Goal:

Analysis of **clinical data** for a **"disease map"**

⇒ **A model for interpretation** and predictions on COVID-19 starting from clinical data and other input information

Members of the project

→ INFN in collaboration with Istituto Nazionale di Malattie Infettive Lazzaro Spallanzani, **INMI**

→ **INFN + 7 units**:

- INMI
- Policlinico di Milano
- Policlinico Umberto I di Roma
- Ospedale Papa Giovanni XXIII di Bergamo
- Ospedale del Sacro Cuore Don Calabria di Negrar (VR)
- Istituto Zooprofilattico Sperimentale di Puglia e Basilicata
- ASL Roma 1

ISTITUTO NAZIONALE PER LE MALATTIE INFETTIVE
"Lazzaro Spallanzani"
Istituto di Ricovero e Cura a Carattere Scientifico

INFN and DETECOVID

INFN
Istituto Nazionale di Fisica Nucleare

Epidemiological experience and data from Istituto Spallanzani

Data analysis and calculus from INFN

An analysis framework that can interpret epidemiological in the case of COVID, perform predictions, and could be generalized to other cases

What's done up to now

→ **Analysis framework** in python

→ **"Benchmark analyses"** that reproduce what done in classical epidemiology

→ '**Infrastructure** for data storage and processing: **Cloud INFN**

1 - Reading data

2 - Interpretazione statistica

3 - Comunicazione e riutilizzo dei dati

The data

⇒ Data from INMI between March and July 2020

➡ 785 patients (~ 495 M / 290F)

⇒ Age 61 years

⇒ ReCOVeRI protocol for info gathering from Istituto Spallanzani

⇒ Data in an homogeneous form

⇒ Python code using pandas/scipy

Available information

→ Anagrapics
→ Previous pathologies

Exams for COVID-19

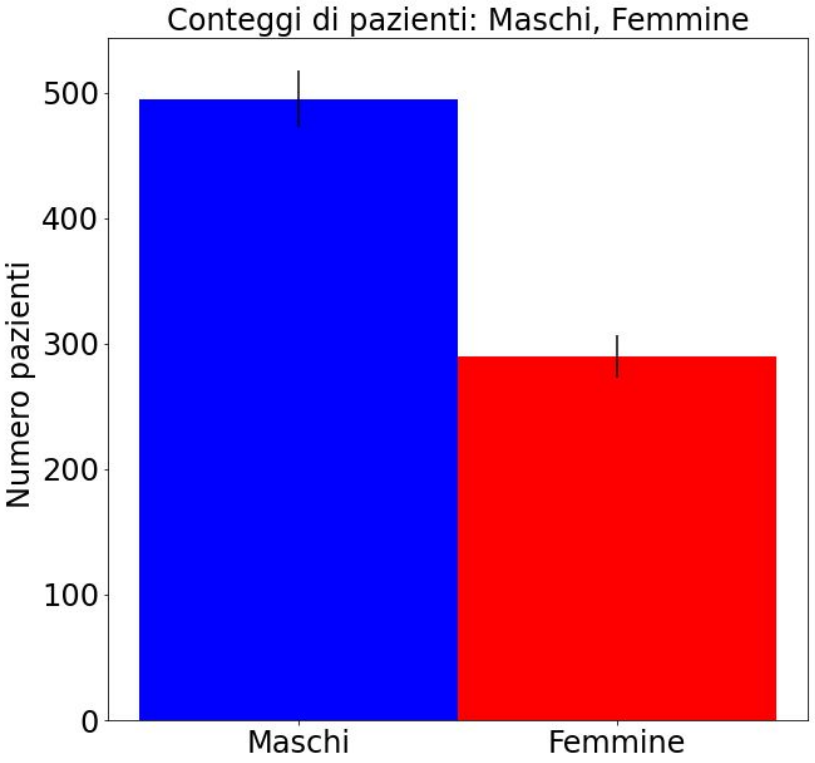
Treatment on the patient

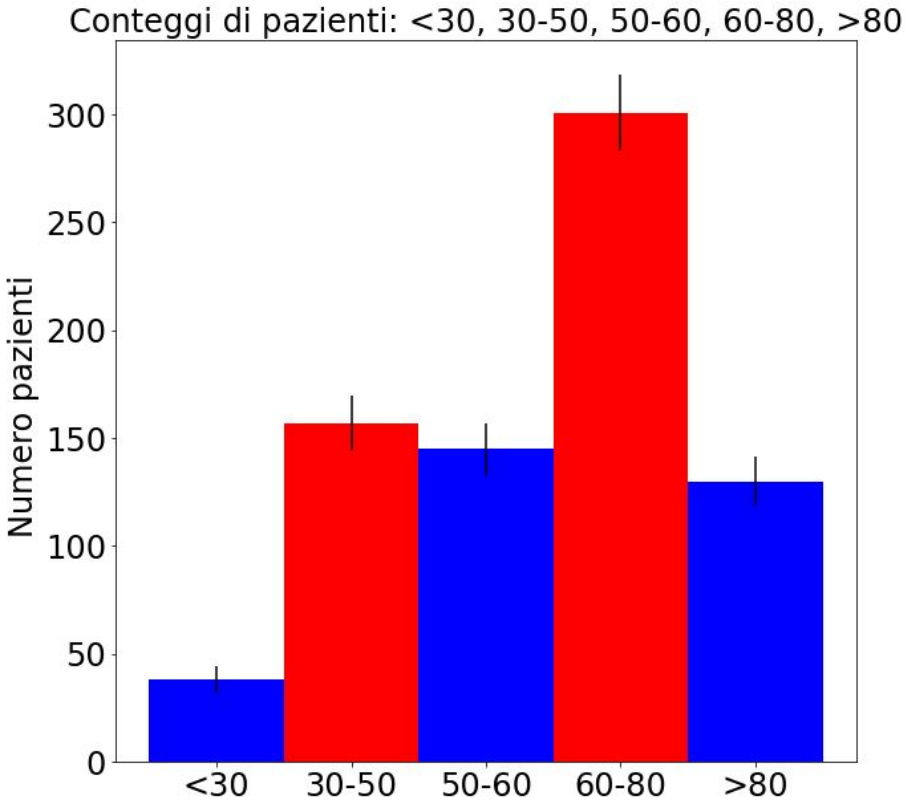
Blood and radiological exams

Daily monitoring data

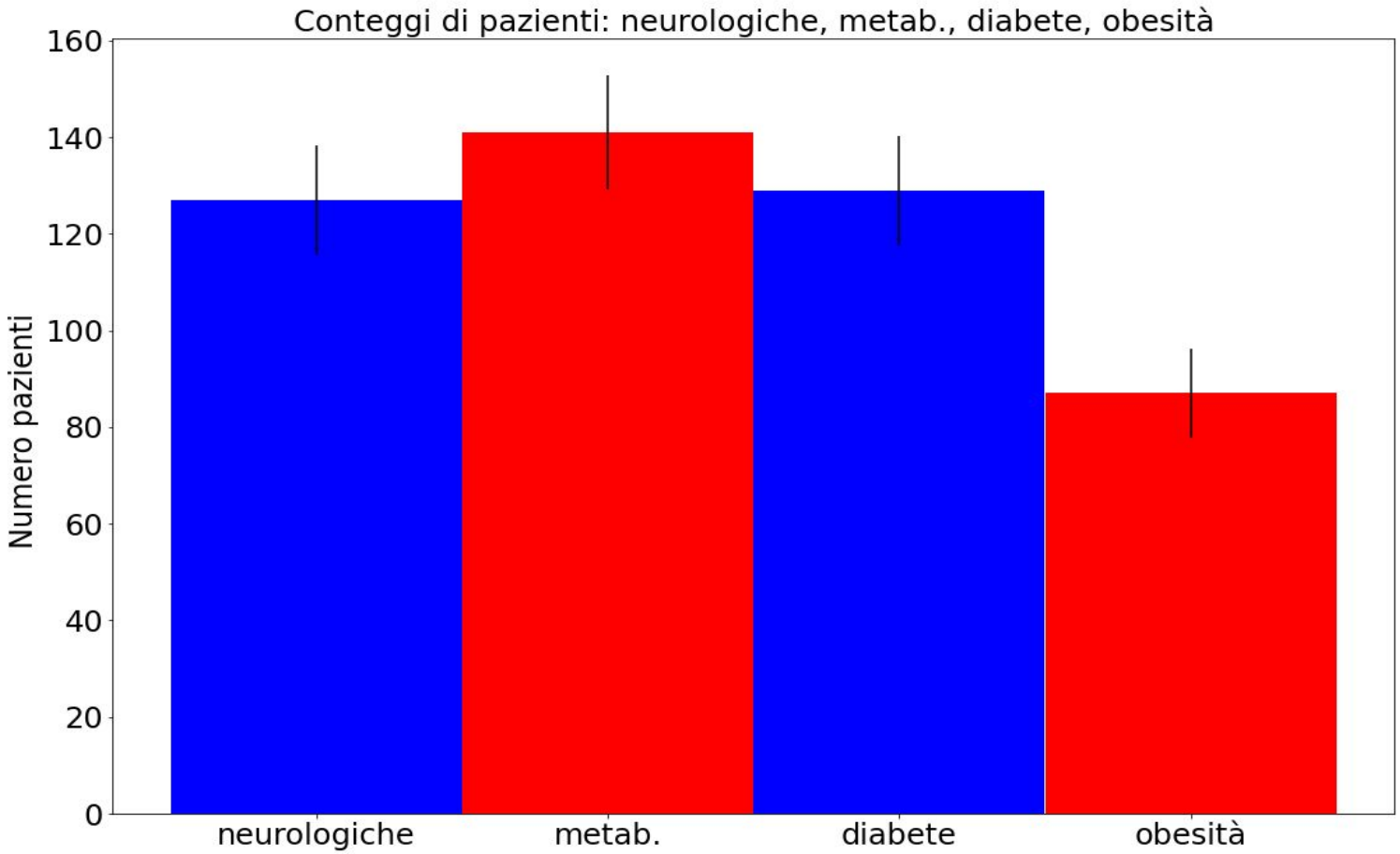
Outcome of the recovery

The data: gender/age

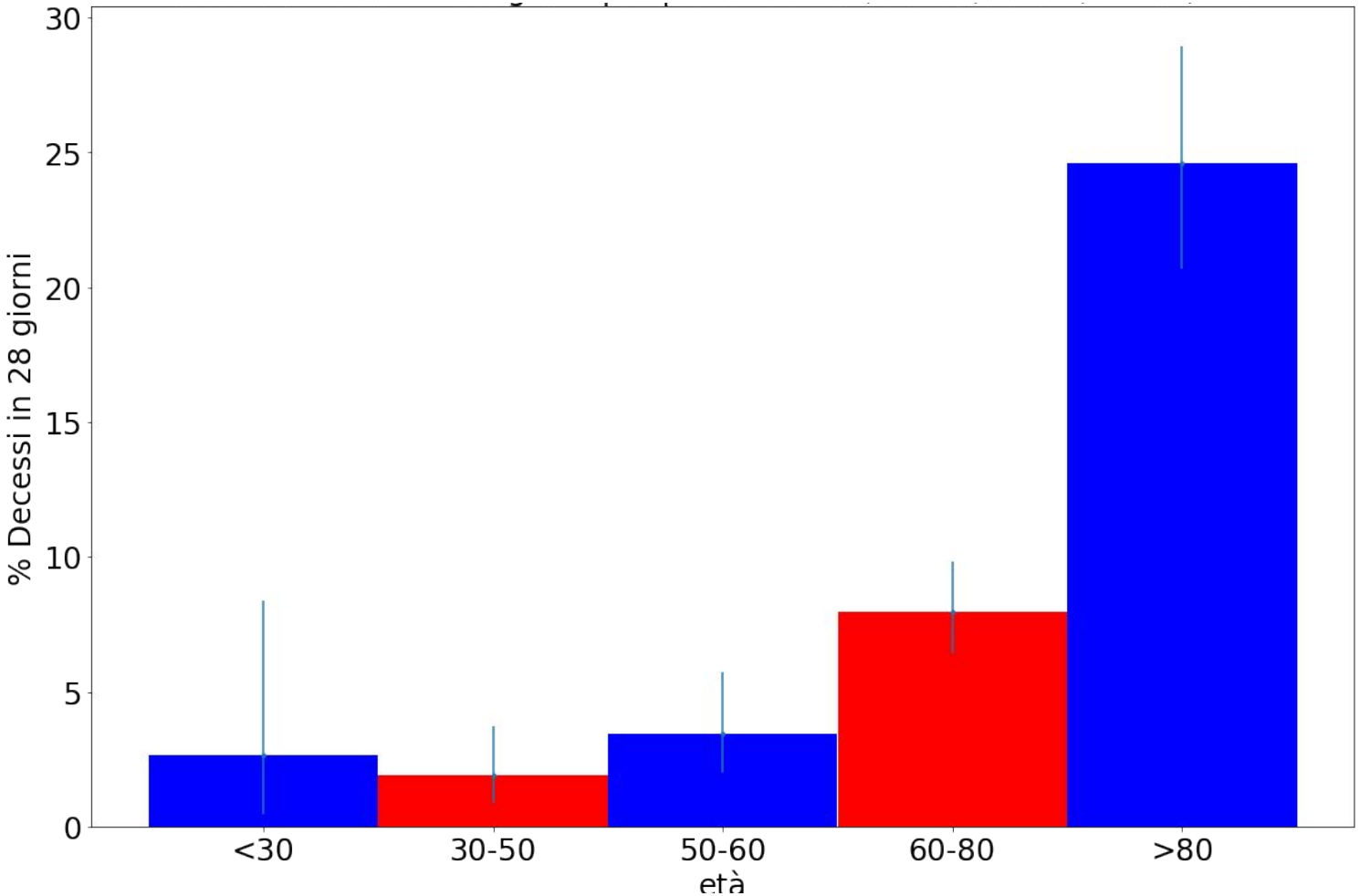




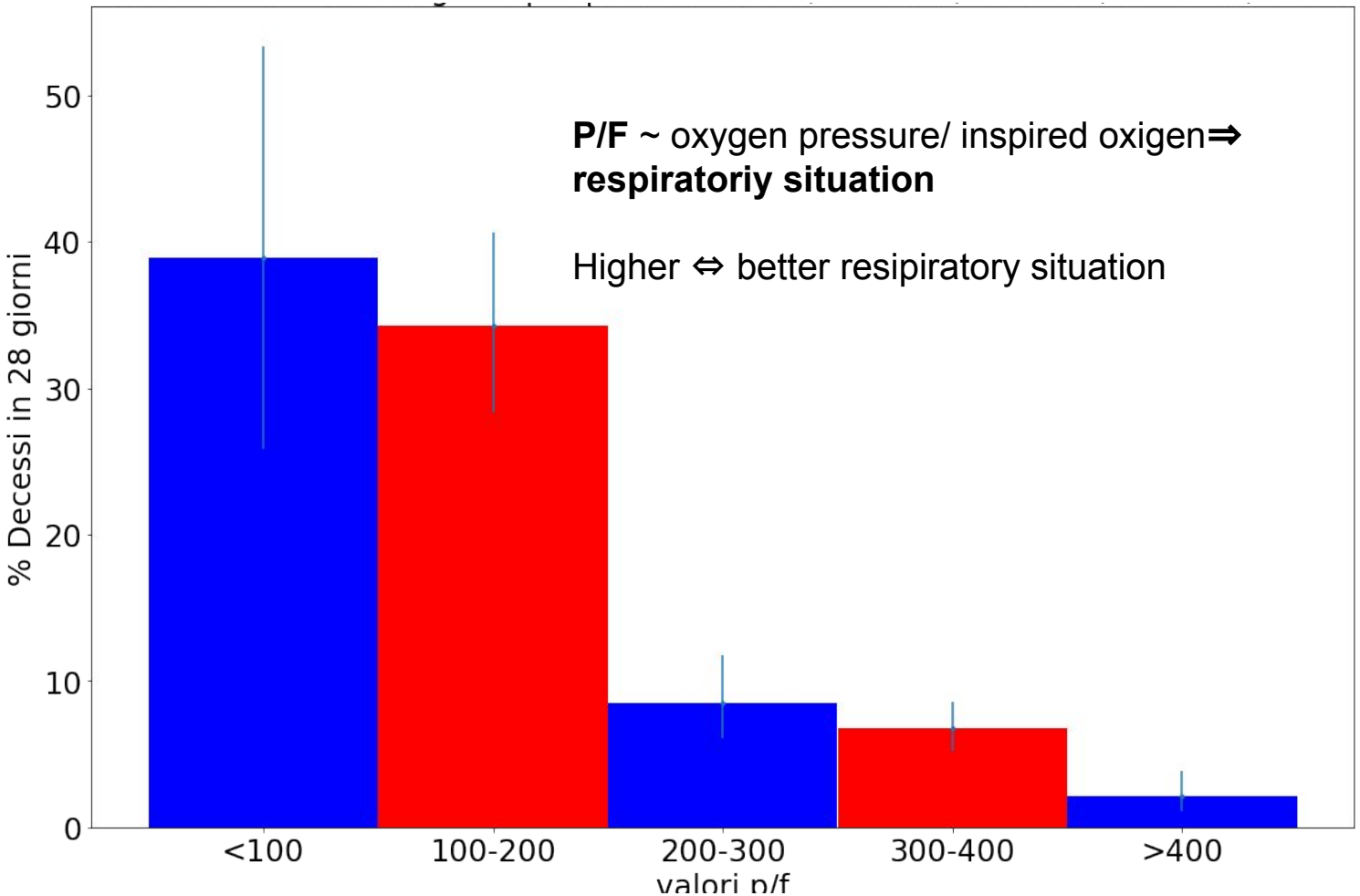
The data: some pathologies



Outcome percentage vs age



Outcome percentage: P/F



P/F ~ oxygen pressure/ inspired oxigen⇒ **respiratoriy situation**

Higher ⇔ better resipiratory situation

1 - Lettura e visualizzazione dei dati

2 - Statistical interpretation

3 - Comunicazione e riutilizzo dei dati

How to quantify those differences?

Different statistical models

→ Binomial

→ Functional dependence studies

→ Survival analysis

→ Logistic regression

→ Machine Learning (Work in progress!)

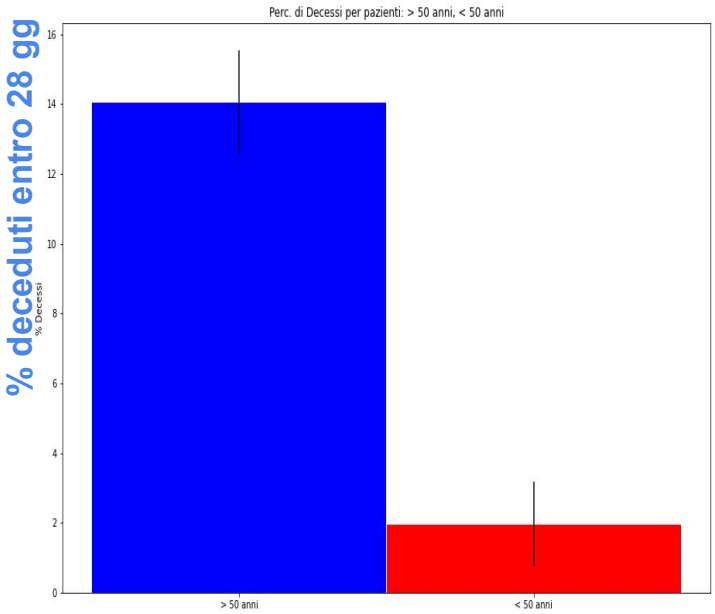
Binomial model: age differences

Two binomials, differing by the parameter ε.

The likelihood defined L is defined as:

$$\binom{N_a}{n_a}((1-\varepsilon)p)^{n_a} \cdot (1-(1-\varepsilon)p)^{N_a-n_a} \cdot \binom{N_b}{n_b} p^{n_b} \cdot (1-p)^{N_b-n_b}$$

>50 anni **<=50 anni**



Na = **patients a**

na = **outcomes a**

Nb = **patients b**

nb = **outcome b**

p = probability of outcome b

ε = difference between two probabilities

$\Rightarrow p_a = (1-\varepsilon)\, p$

Binomial model: age differences

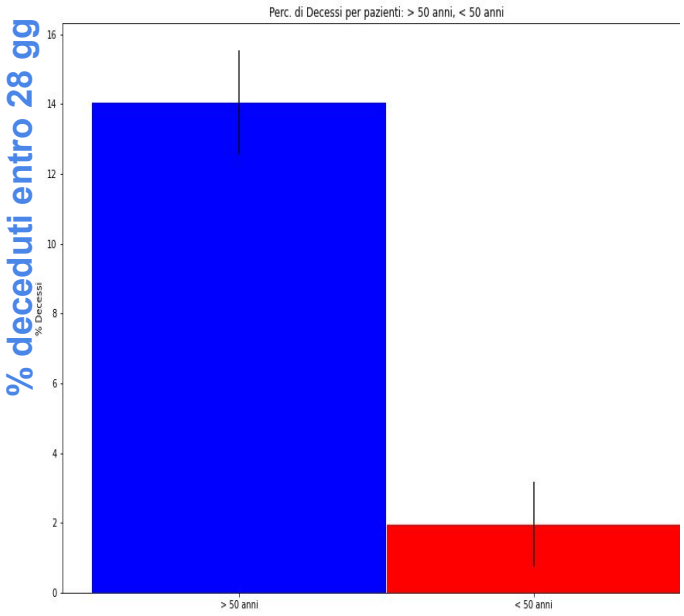
A variable λ is defined from the profile likelihood to establish if the probabilties are the same, i.e. $\varepsilon = 0$:

$$\lambda\varepsilon = -2\ln L_{\text{prof}} = -2\ln L(\varepsilon, \hat{\hat{p}})/L(\hat{\varepsilon}, \hat{p})$$

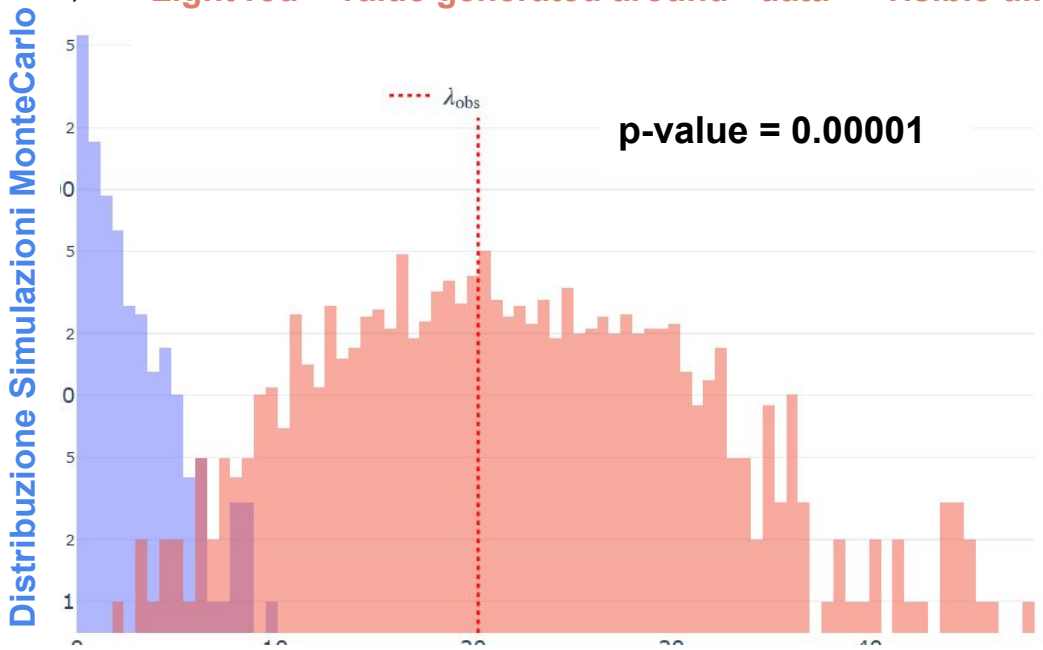
→ Greater difference - higher separation

Comparison of the two percentages

>50 anni **<=50 anni**



Blue = value when difference is not significant
Red line = value in data
Light red = value generated around ~data ⇒ visible difference



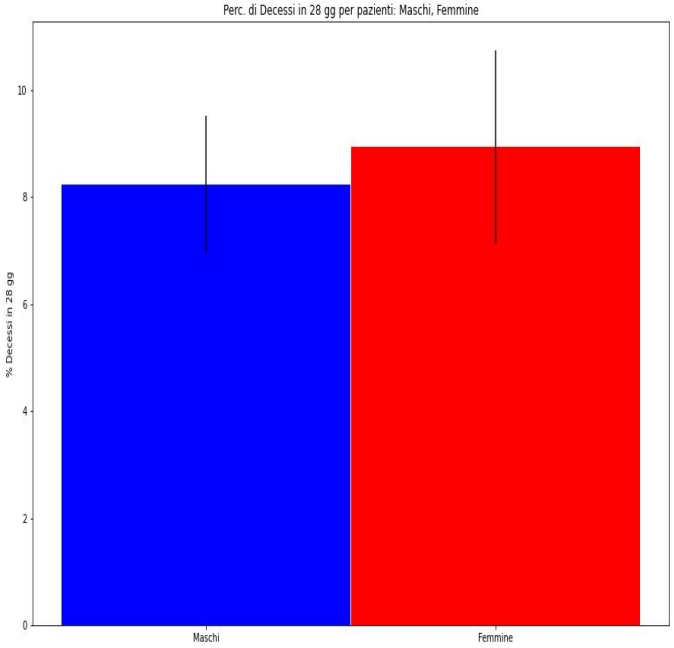
Valore statistica di test con (ε =0)

Binomial model: gender difference

In this case the two distributions are within the uncertainties → can't be distinguished

Comparison of the two percentages

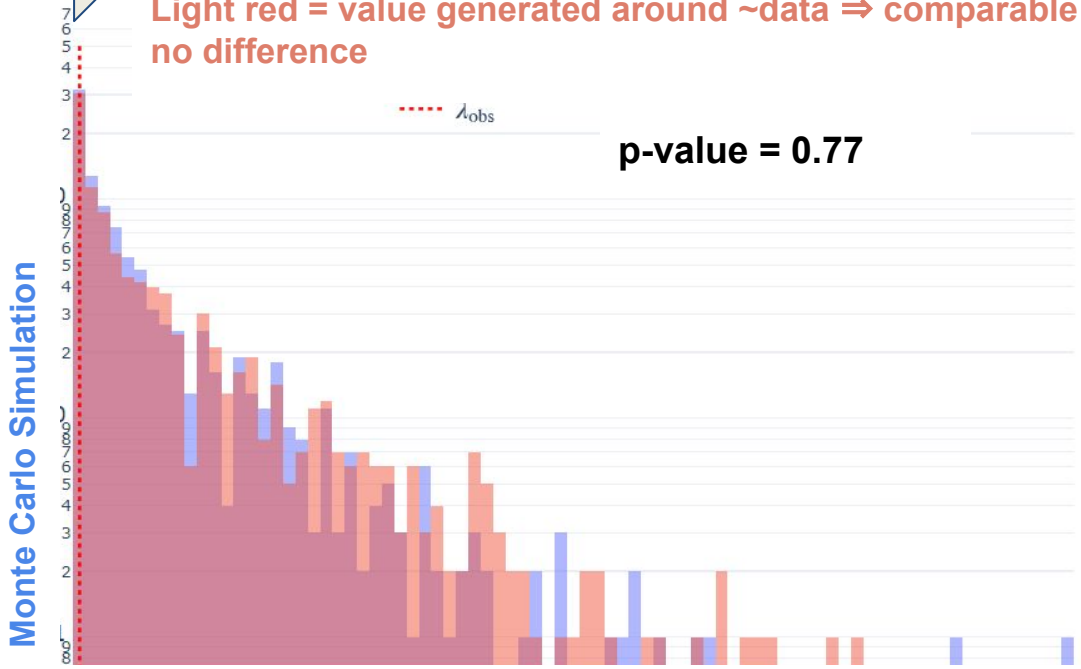
M **F**



Blue = value when difference is not significant

Red line = value in data

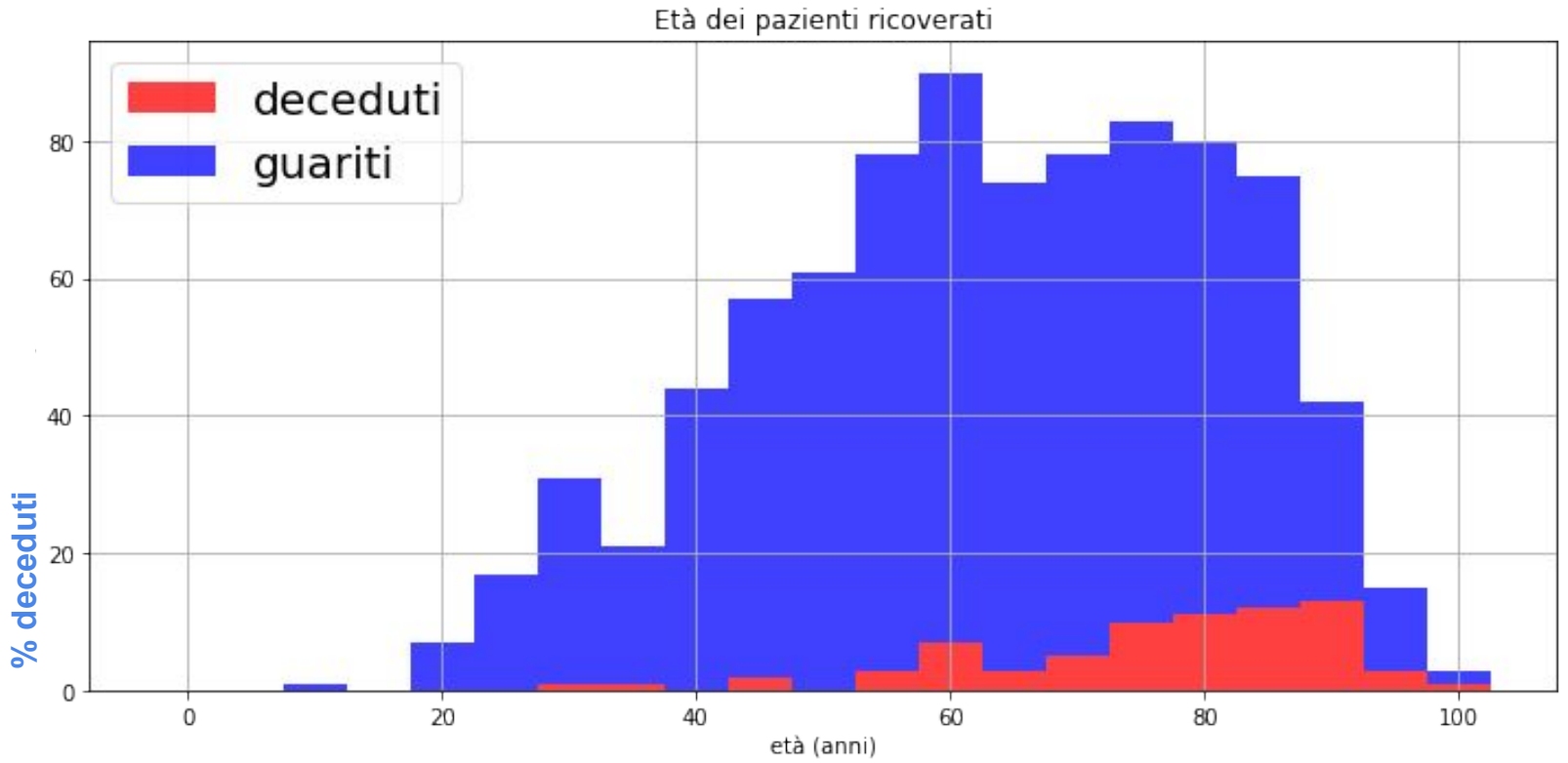
Light red = value generated around ~data ⇒ comparable with no difference



p-value = 0.77

Test statistic with (ε =0)

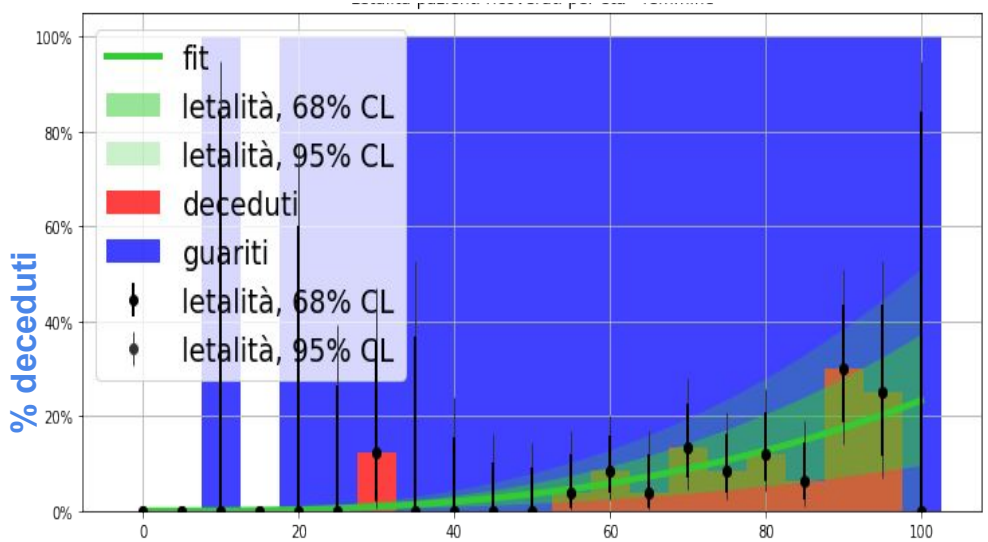
Functional dependencies: more detailed studies



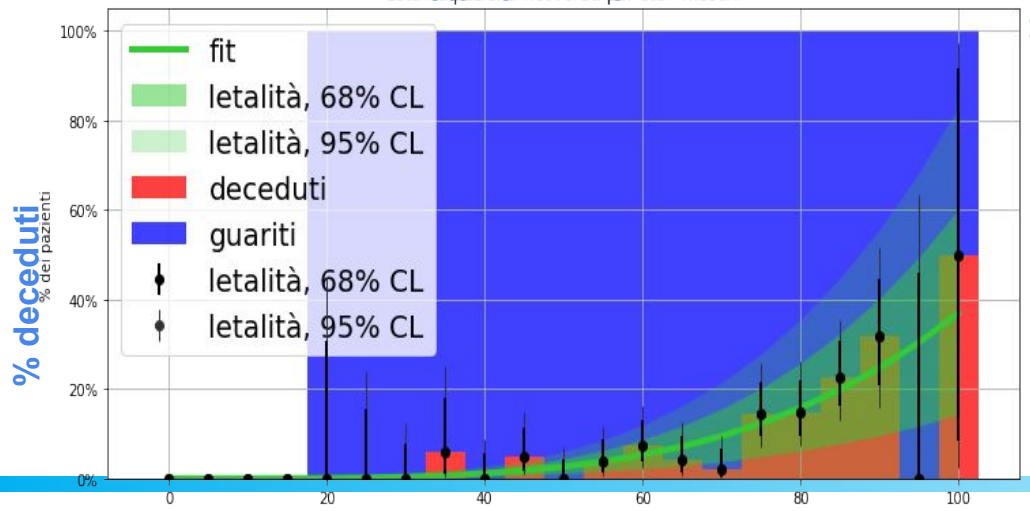
1) Study global behavior for continuous variables, like age, in two cases (M/F)

2) Compare the two behaviors

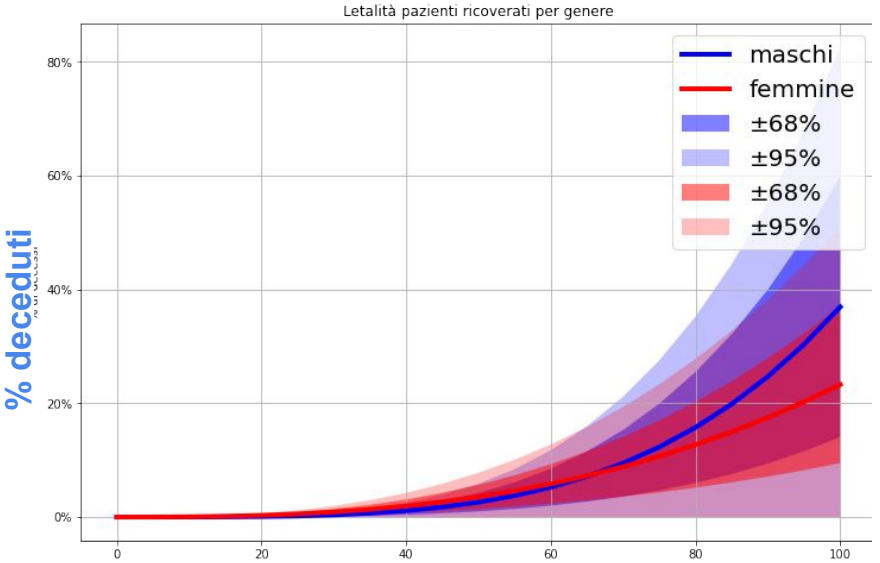
Functional dependence: sex and age



età femmine



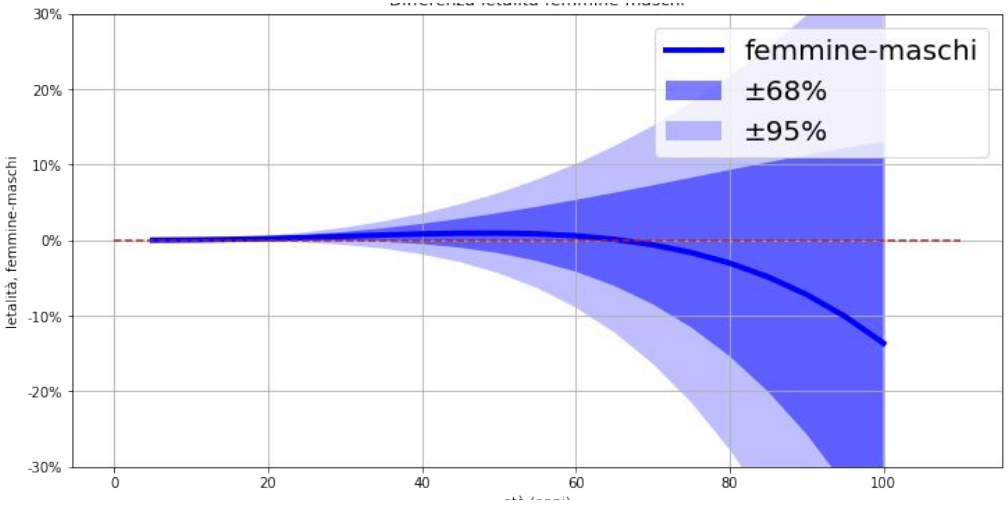
età maschi

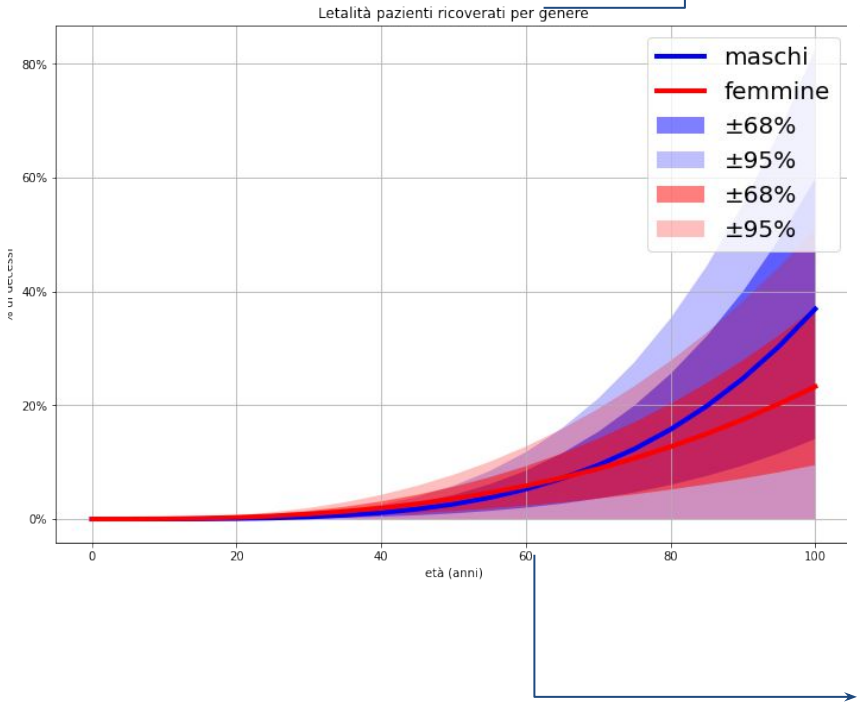


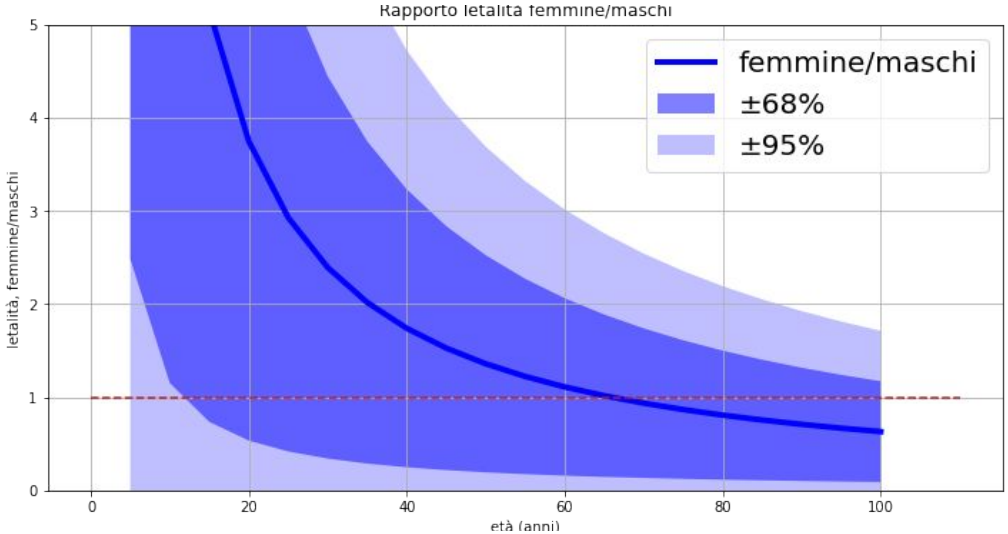
età

Functional dependence: sex and age

Difference blue/red

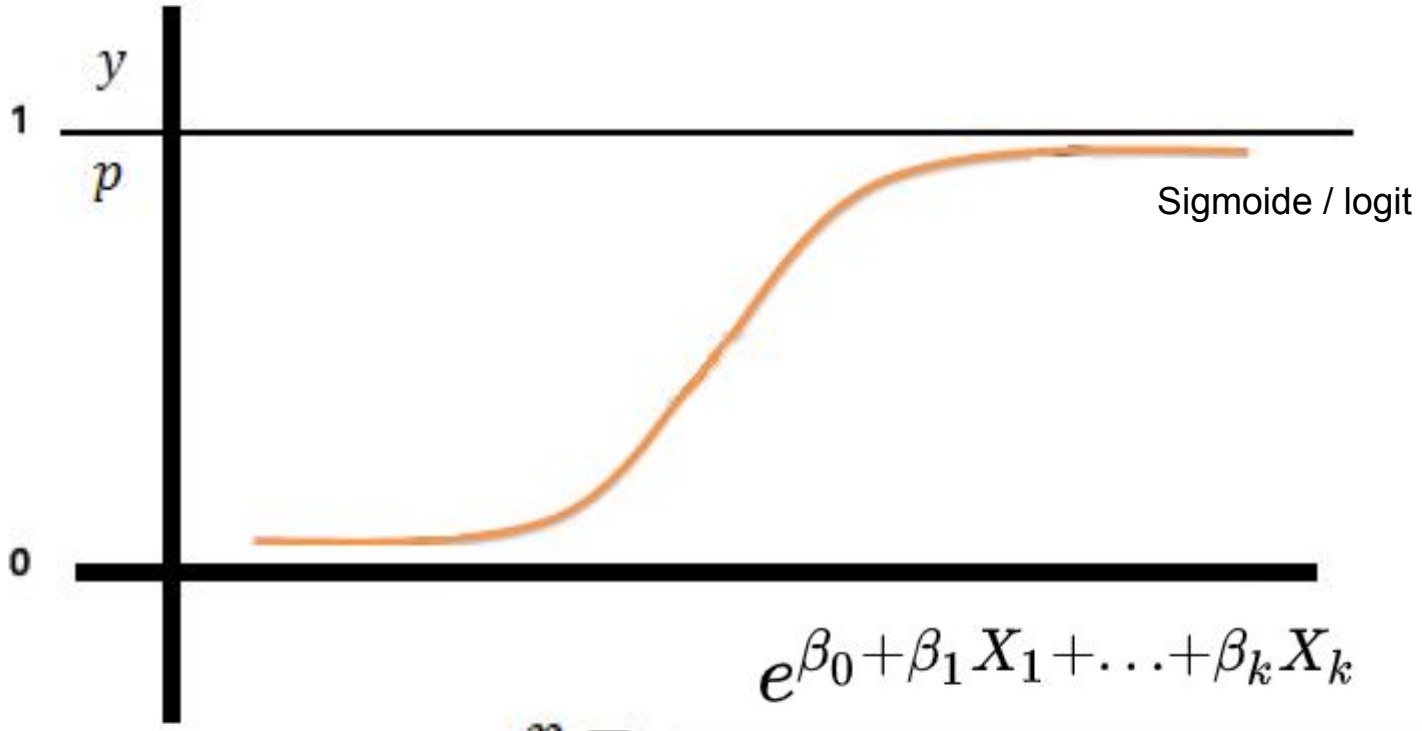






Ratio blue/red

Logistic regression models



$$p = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}}$$

→ Functional form for probabiltiy of outcome ***p***

→ X_1,X_2…X_k= age, gender, p/f, pathologies yes/no etc.

→ X is 1 or 0 for “binary” features, like the presence of a pathology

→ Parameters ***βx*** obtained with a maximum likelihood fit to data

Logistic regression: the odds ratios

Odds Ratio ==

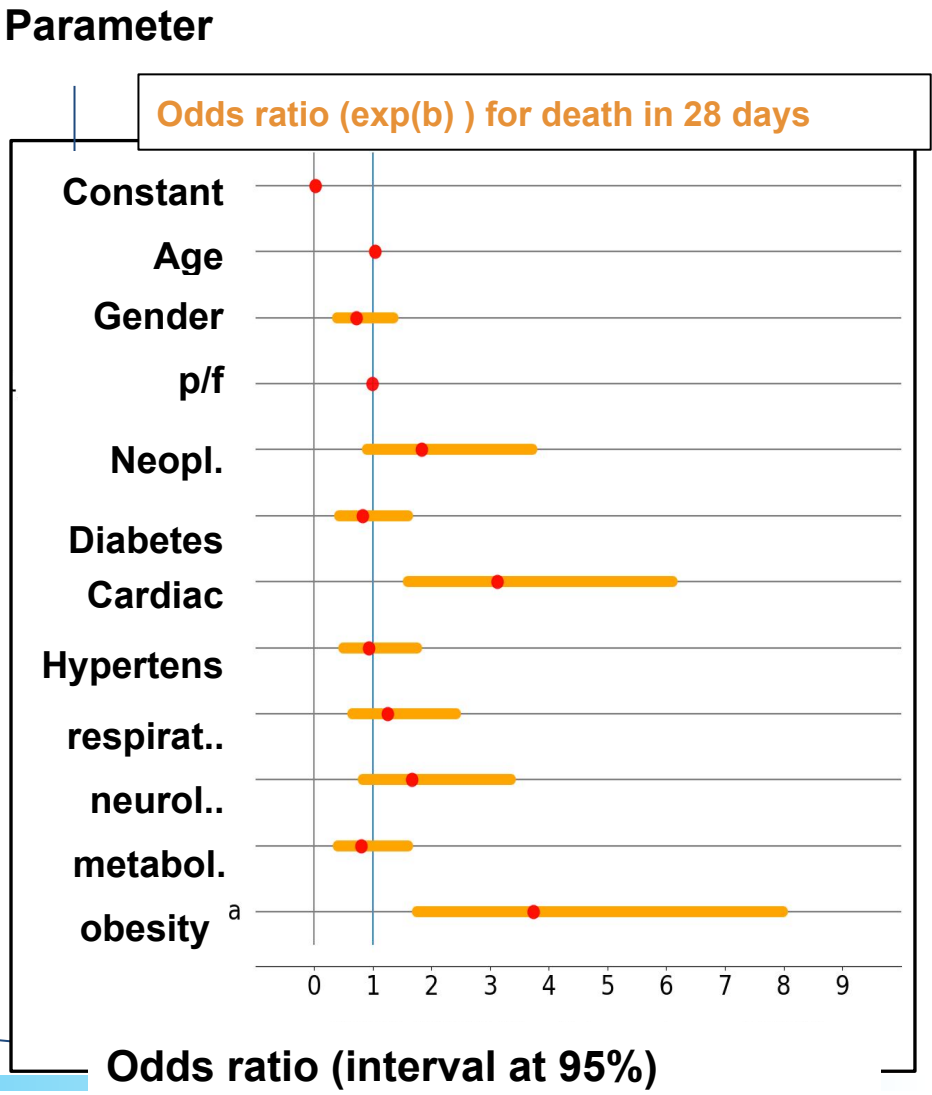
Ratio of the **conditional probabilities:**

$$\frac{P(Y=1 \mid X=1)/P(Y=0 \mid X=1)}{P(Y=1 \mid X=0)/P(Y=0 \mid X=0)}$$

By fixing the $\mathbf{Z_1,Z_p}$ **parameters:**

$$\frac{P(Y=1 \mid X=1, Z_1, \ldots, Z_p)/P(Y=0 \mid X=1, Z_1, \ldots, Z_p)}{P(Y=1 \mid X=0, Z_1, \ldots, Z_p)/P(Y=0 \mid X=0, Z_1, \ldots, Z_p)}$$

= e^(βx)



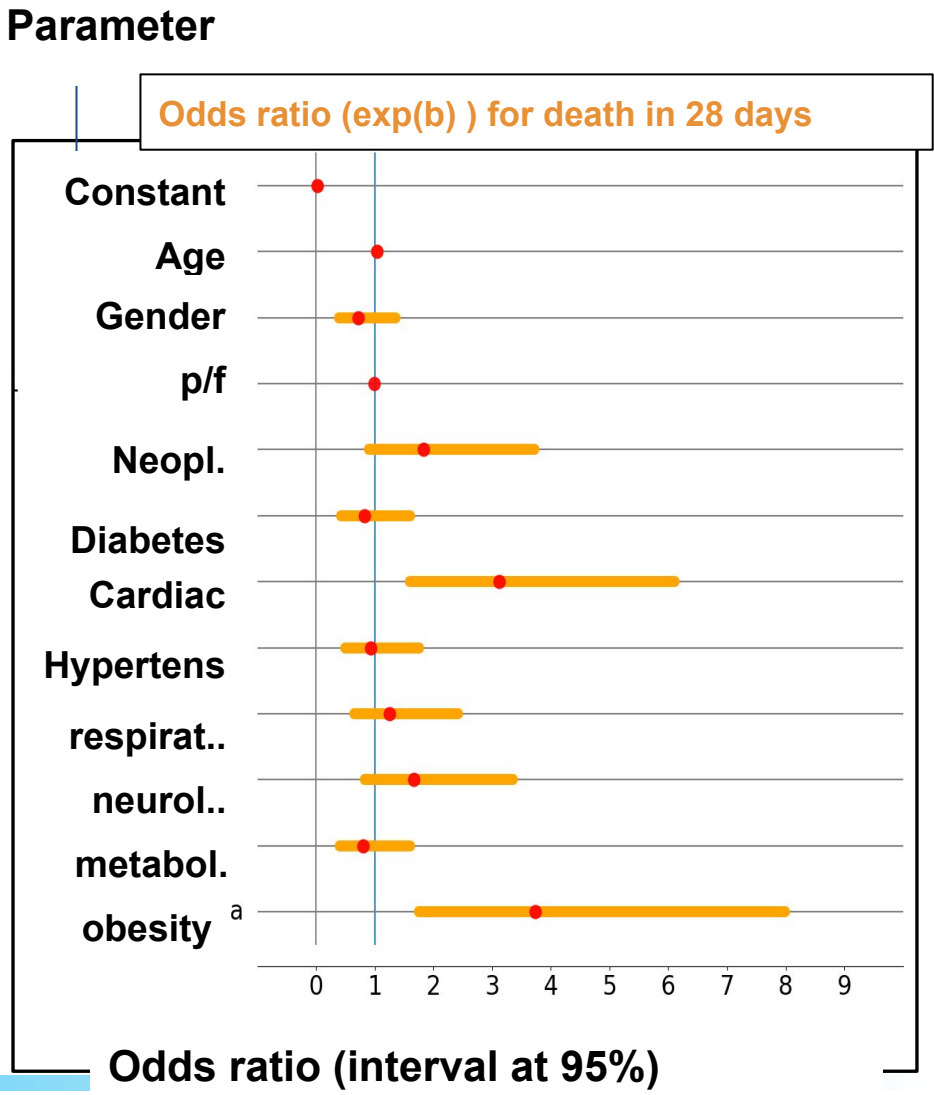
Logistic regression: other features

Quality criterion
Pseudo-R^2 = 0.28

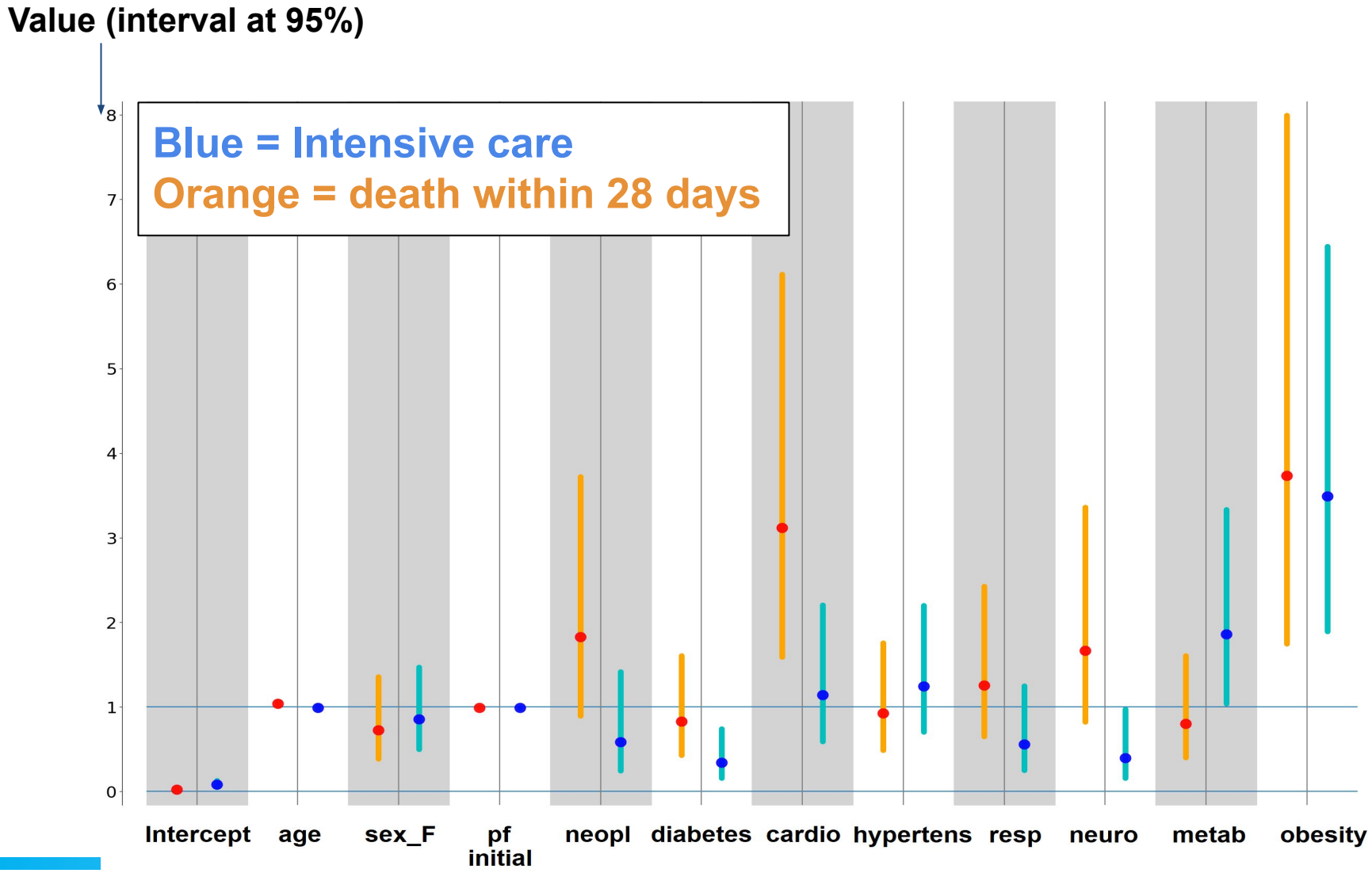
$$R^2 = 1 - \frac{LL_{full\ model}}{LL_{intercept}}$$

The constant is the **probability of an outcome** for the case with all X == 0

→ normalaized e p/f medi



Comparison of outcome: death vs intensive care



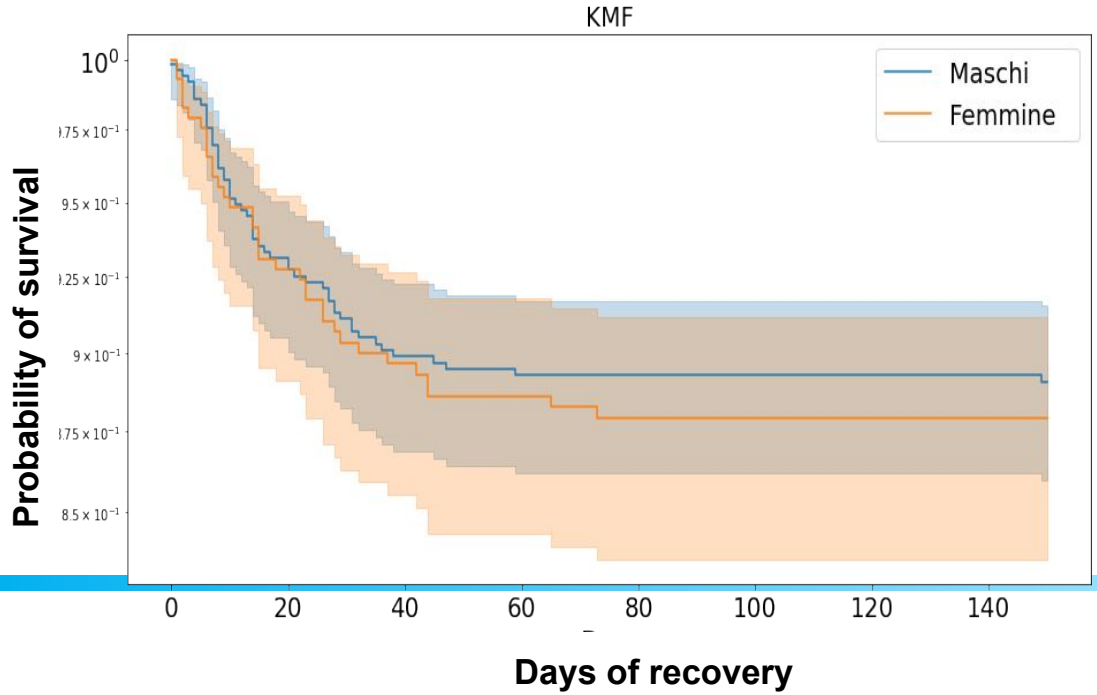
Survival analysis:

Performed to exploit the time information of the event

Function of "survival" over time - probability to survive up to time t:

$$S(t) = P[T > t] = 1 - P[T \leq t] = 1 - F(t) \longrightarrow \hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

Below: male vs female



We define **hazard :**

$$h(t) = \frac{f(t)}{S(t)}$$

f = Probability at time t

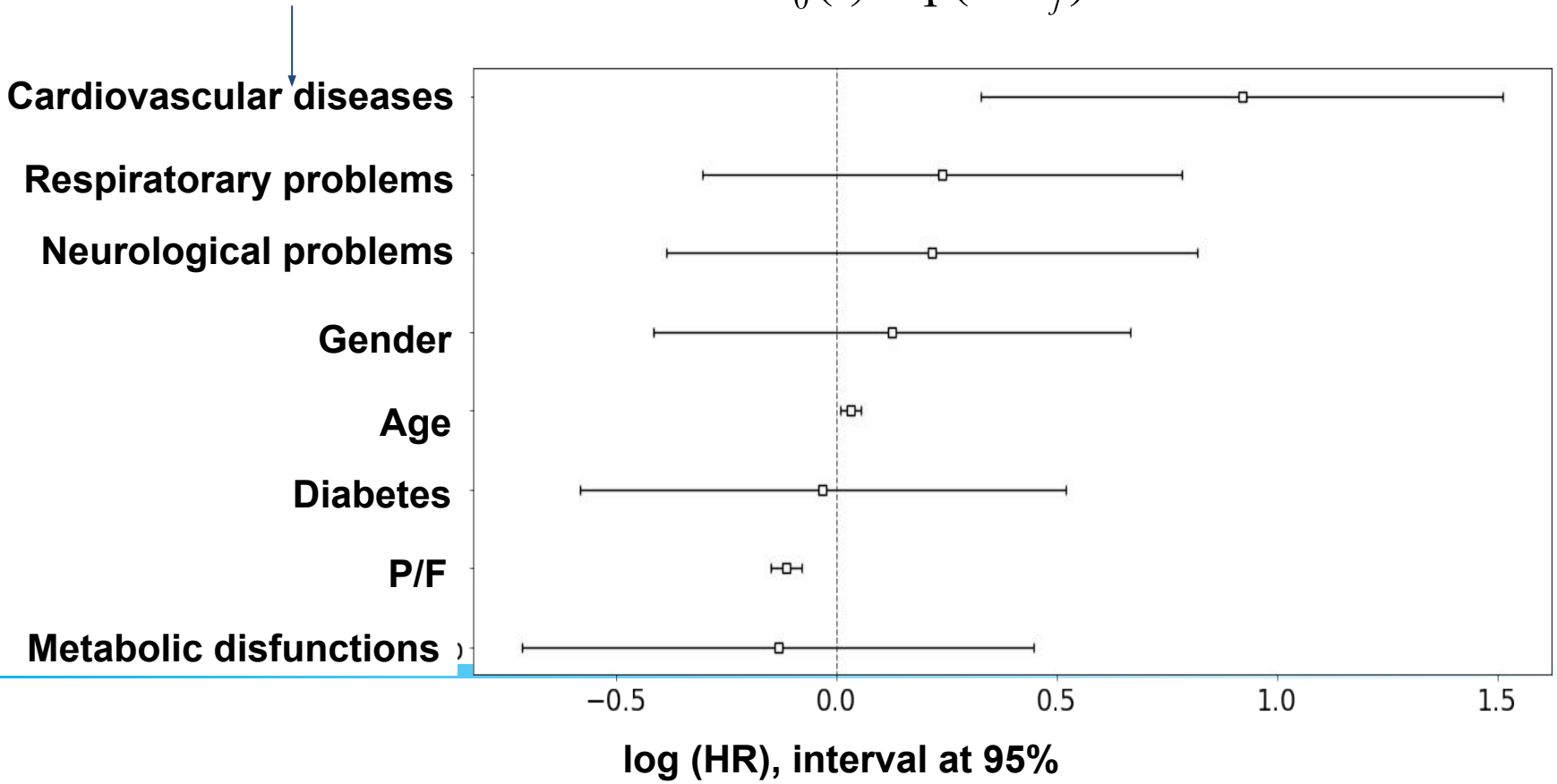
S = Probability up to time t

Survival analysis: Cox Model

*A **Hazard Ratio*** is defined by taking two hypotheses and assuming the same time dependency: **HR > 1 ⇒ greater risk**

Modeled as exponenitals

$$\mathrm{HR} = \frac{h_0(t)\exp(\vec{b}\cdot\vec{x_i})}{h_0(t)\exp(\vec{b}\cdot\vec{x_j})} = \exp(\vec{b}\cdot(\vec{x_i}-\vec{x_j}))$$

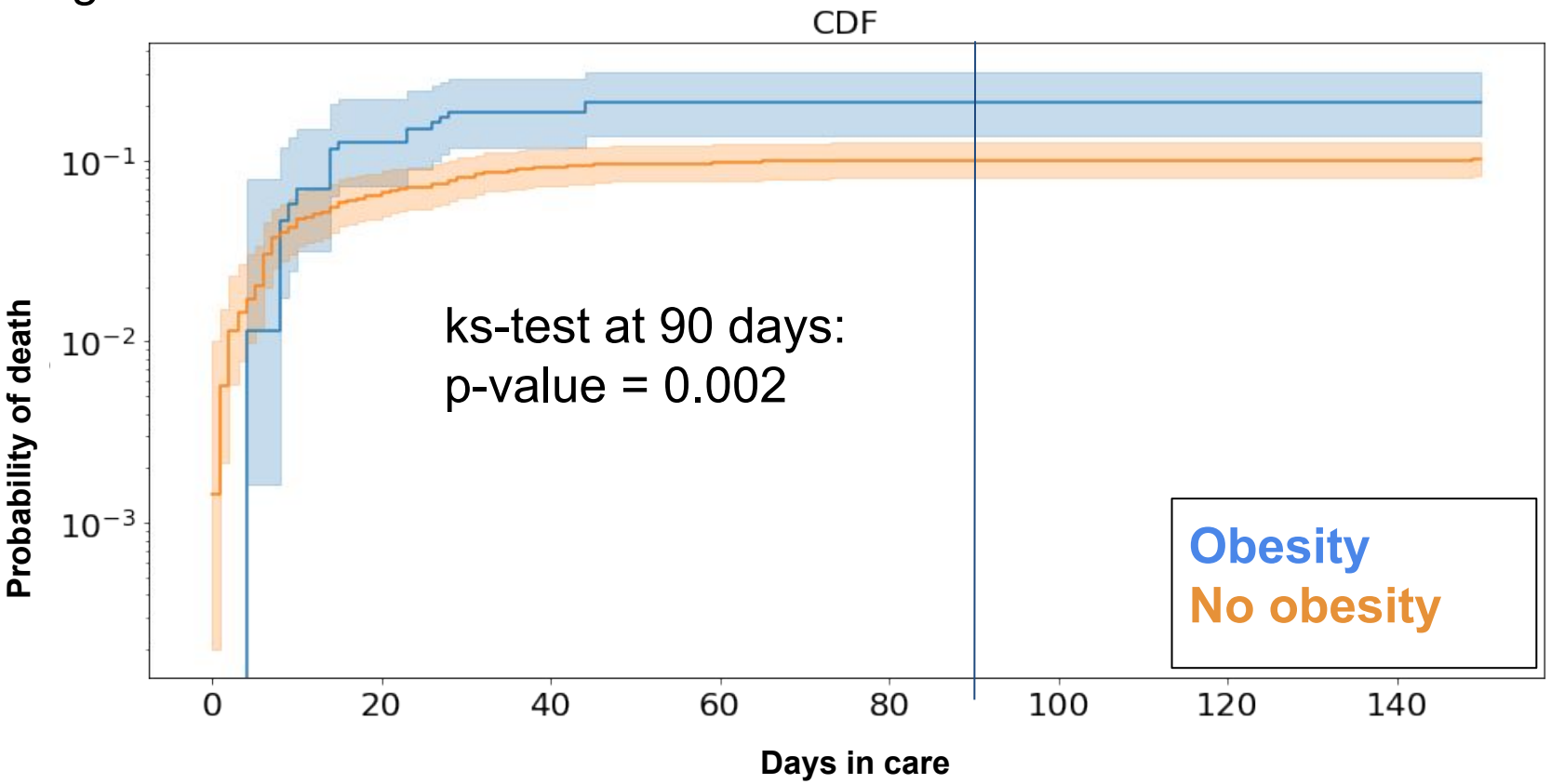


Cumulative functions for cases where standard Cox model is not applicable

Obesity does not respect the conditions of proportionality

→ Different time dependence

→ Cumulative distributions can be differentiated e.g. via a Kolmogorov-Smirnov test

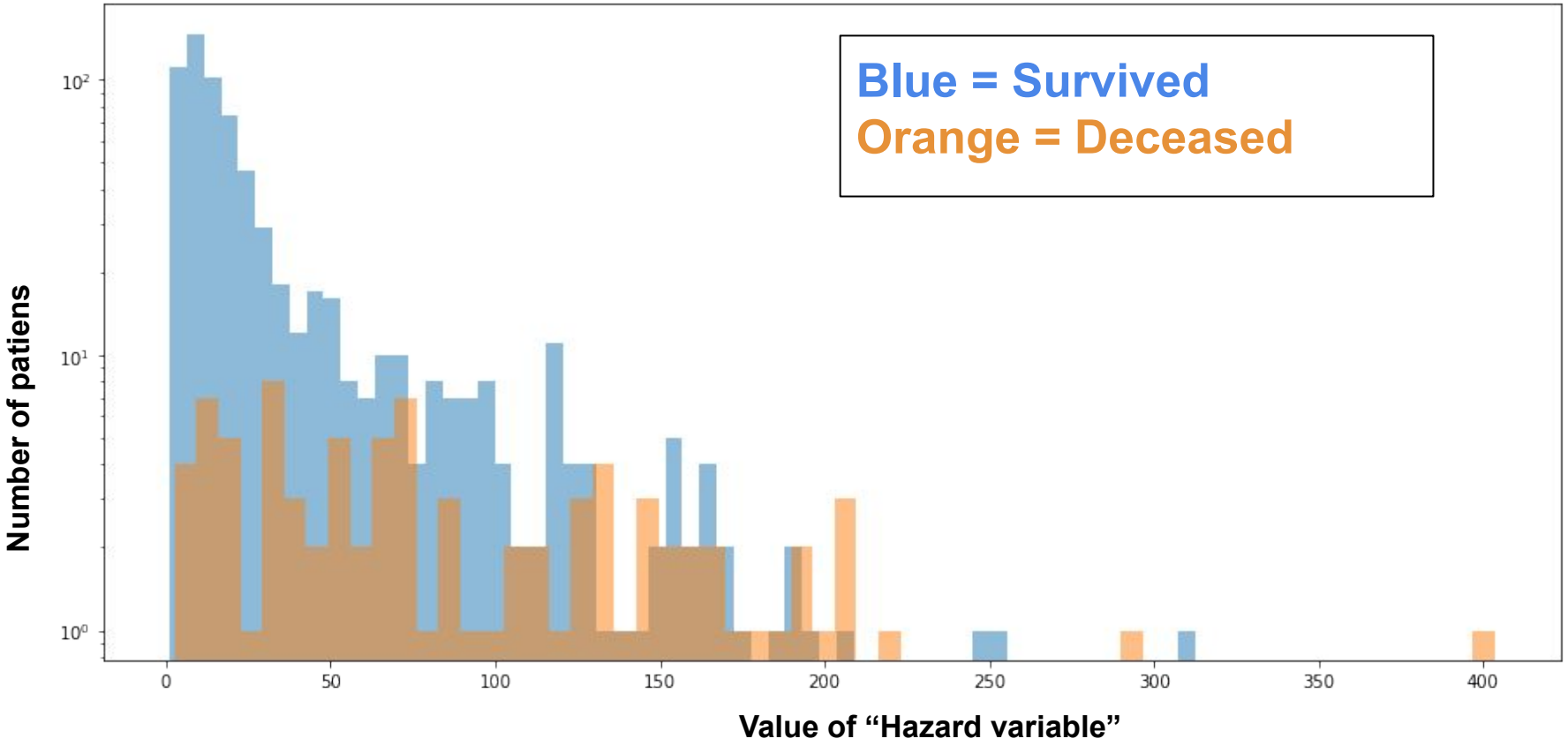


The “hazard” function

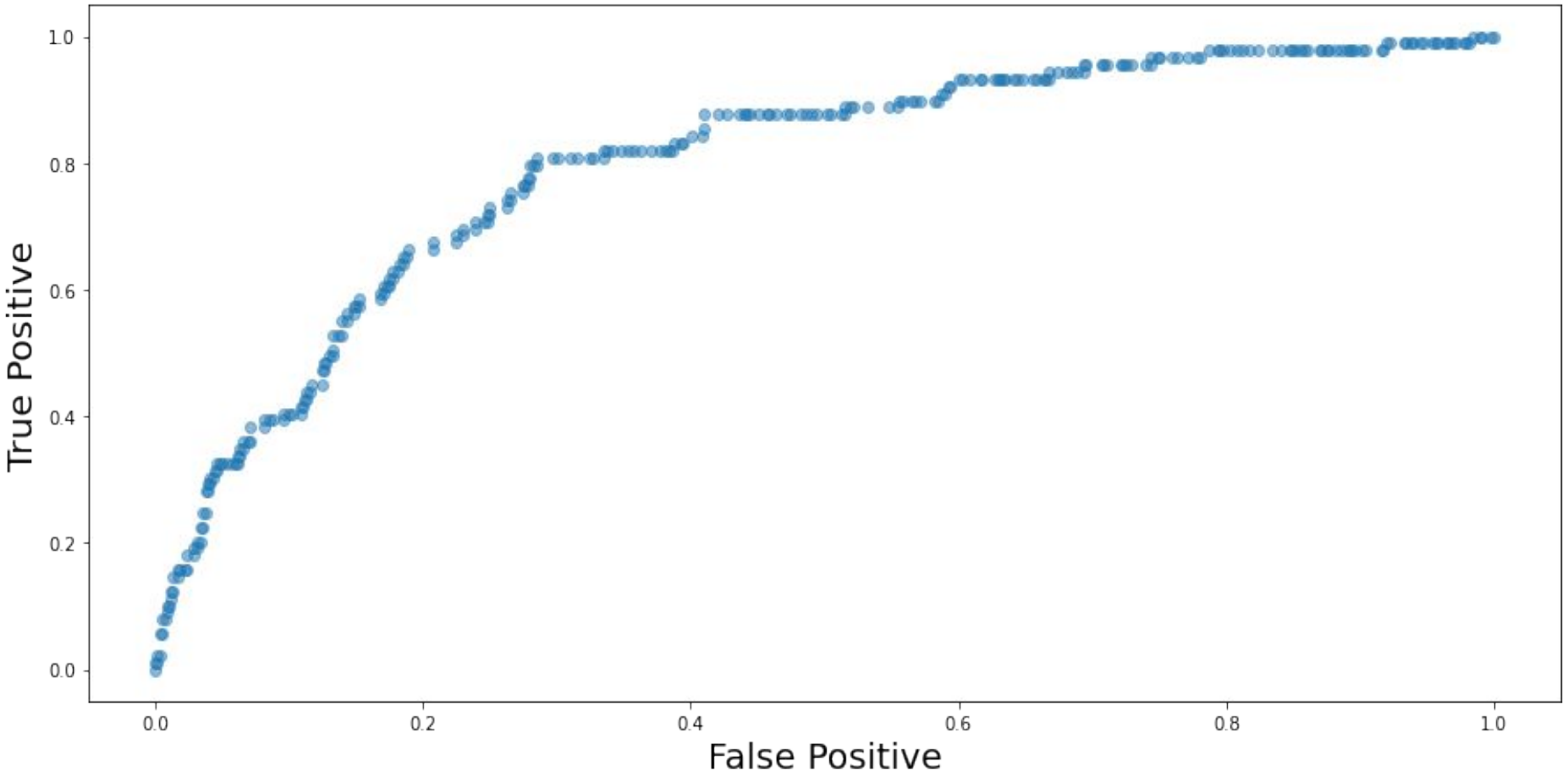
→ It is possible to obtain a “hazard” variable by fitting the parameters from data

→ **Not particularly discriminating** atm!

→ NB: same sample for train and test



Example of ROC on the hazard variable



Model interpretation: impact of the sample size

The power of the test depends on the

→ The precision of our test depends on the sample size:

→ In some case there are < 10 patients

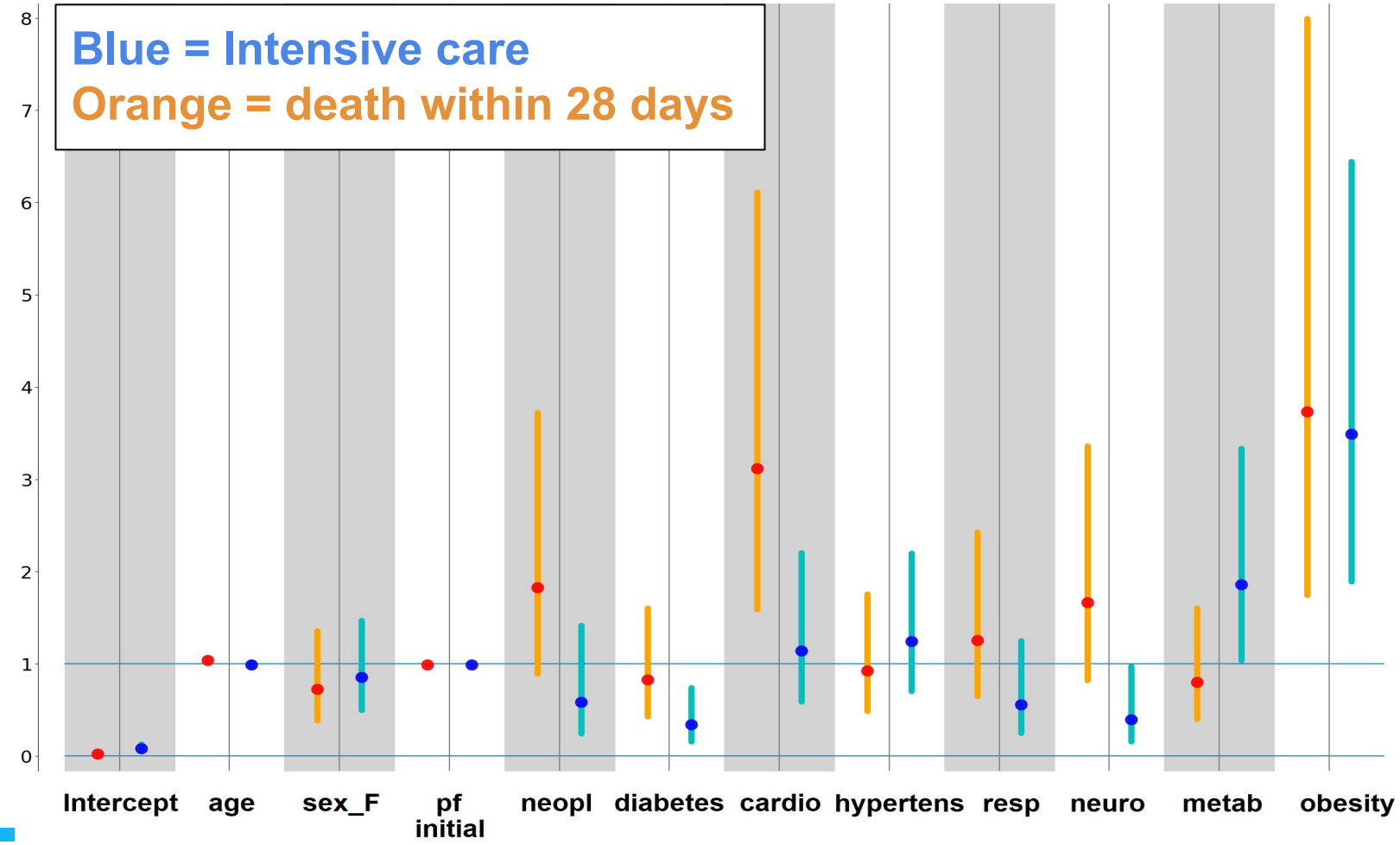
Effect:

→ “Fine” effects, like M/F are not visible

→ Some pathologies don’t have enough statistics to make difference between two effects

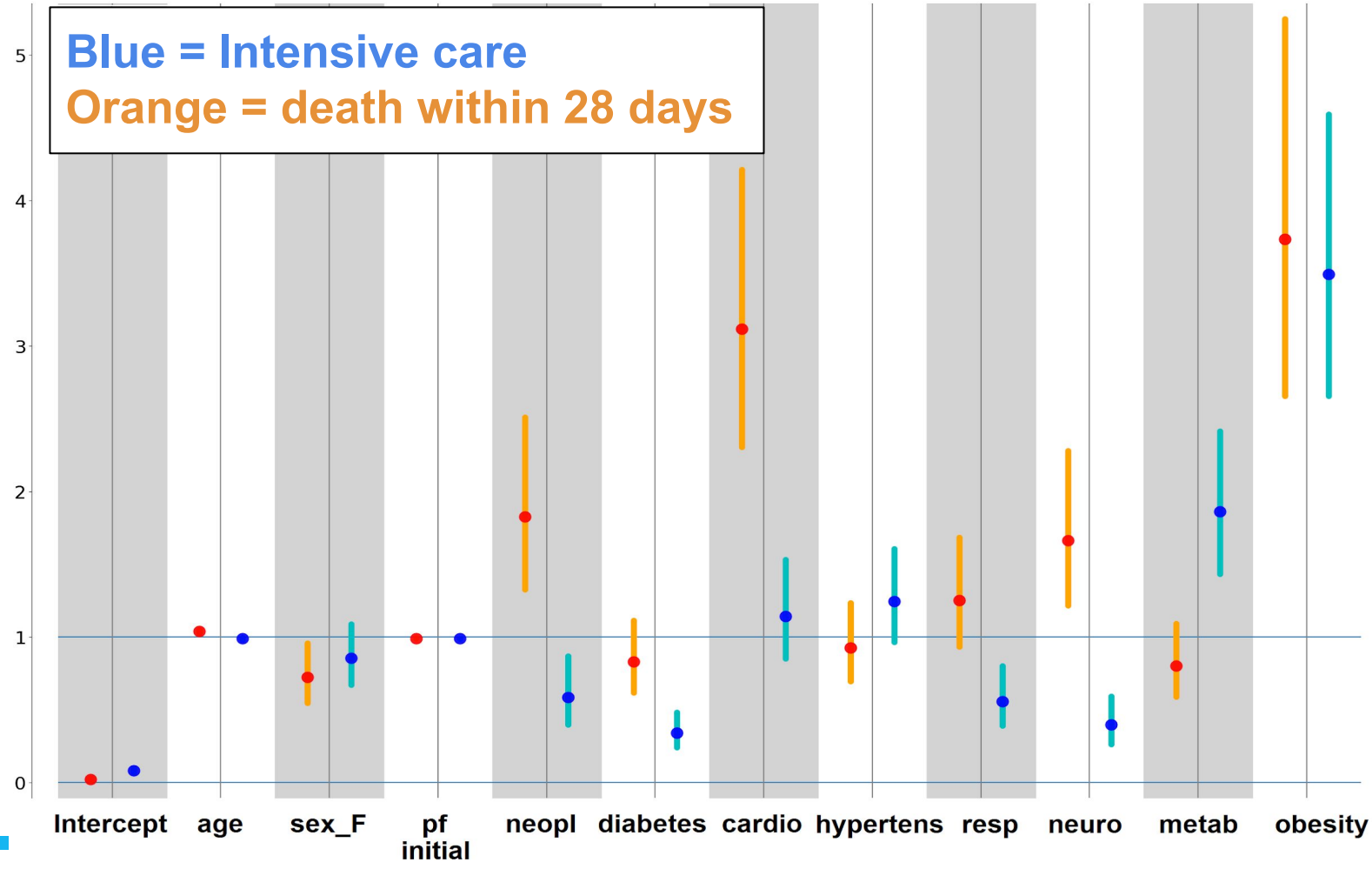
Sample size dependence

Current size

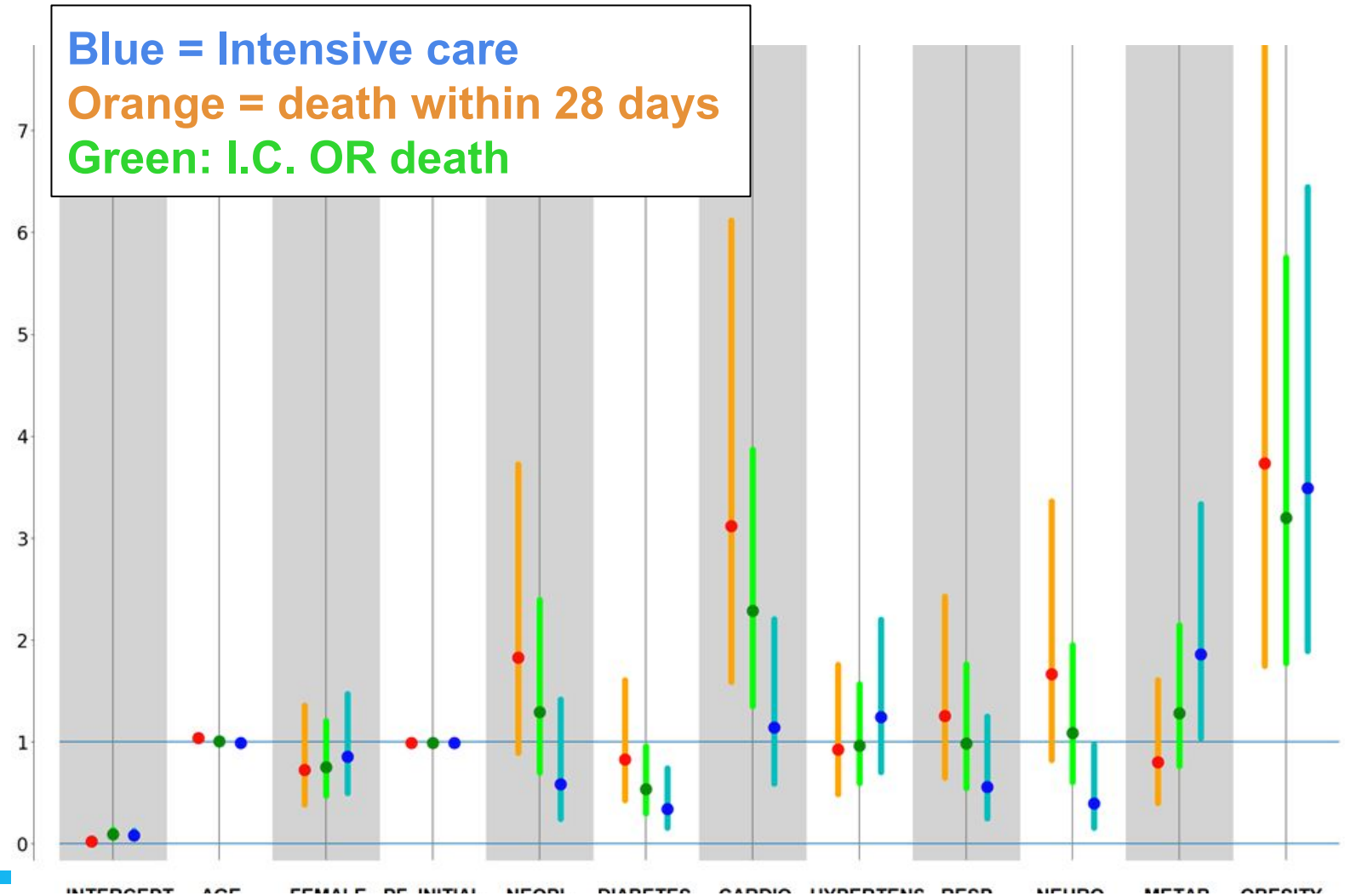


Sample size dependence

Prediction at 5 times current sample size



OR of two cases



Wok in progress: improving models

Logistic models

→ Limits on functional dependence and correlations between variables

Next goal: get “from data” the best functional dependence

Survival analysis

→ One currently needs very strong hypotheses

Next goal: make a fit to the time dependence to have a ***hazard vs time of recovery***

Wok in progress: improving predictions

Limited prediction: probabilities from “classical models” don’t have a strong discriminating power

Improvements we are working on:

→ Combine the information we have from previous knowledge and experience

→ Interpolate missing data

→ **Machine Learning:** low statistics makes this challenging ⇒ studies ongoing for feature selection and data augmentation

1 - Lettura e visualizzazione dei dati

2 - Interpretazione statistica

3 - Usability and communication of data

Data usability and communication

Cloud system INFN:

→ Data on authenticated server

→ Code that can run on local laptop and on notebook from site

Software tool for geneal use:

→ Allow members of the team to access it

→ Allow to use data and statistical tools

→ User can insert information and personalize models

→ Several steps to put this in place! Software infrastructure is consolidating

Future developments

→ Analysis framework in python

→ “Benchmark analyses” that reproduce what done in classical epidemiology

→ ’Infrastructure for data storage and processing: Cloud INFN

→ **Improving** statistical models we are using

→ Adding **Machine Learning techniques** to improve the prediction quality

→ **Allow an user** from the collaboration to use this framework (Input from a commonData Lake, personalization, etc.)

Thanks!

Workflow dell'analisi

Studi Prognosi:

1) Estendere ai valori degli altri esami iniziali / valori iniziali

2) Individuare maggiori fattori di rischio, e.g. età o p/f e separare in intervalli

3) ripetere l'analisi in maniera differenziale su questi intervalli

Studi Terapia

4) Estendere ai farmaci e valutare se qualcuno ha un impatto diverso

Ripetere 1-4 per i diversi modelli analizzati

Orizzonte temporale 2021

Gennaio:

→ Formalizzazione selezione features

→ Implementazione test d'ipotesi su features selezionate

Febbraio-Aprile :

→ Completamento studio feature selection

→ Studio dettagliato predittività modelli

→ Implementazione modelli alternativi

→ Implementazione framework generalizzato

Dopo Aprile:

→ Applicazione a potenziali campioni di dati più grandi

→ Studio software di uso generale

→ Documentazione risultati

Backup

Altri esempi lettura dei dati

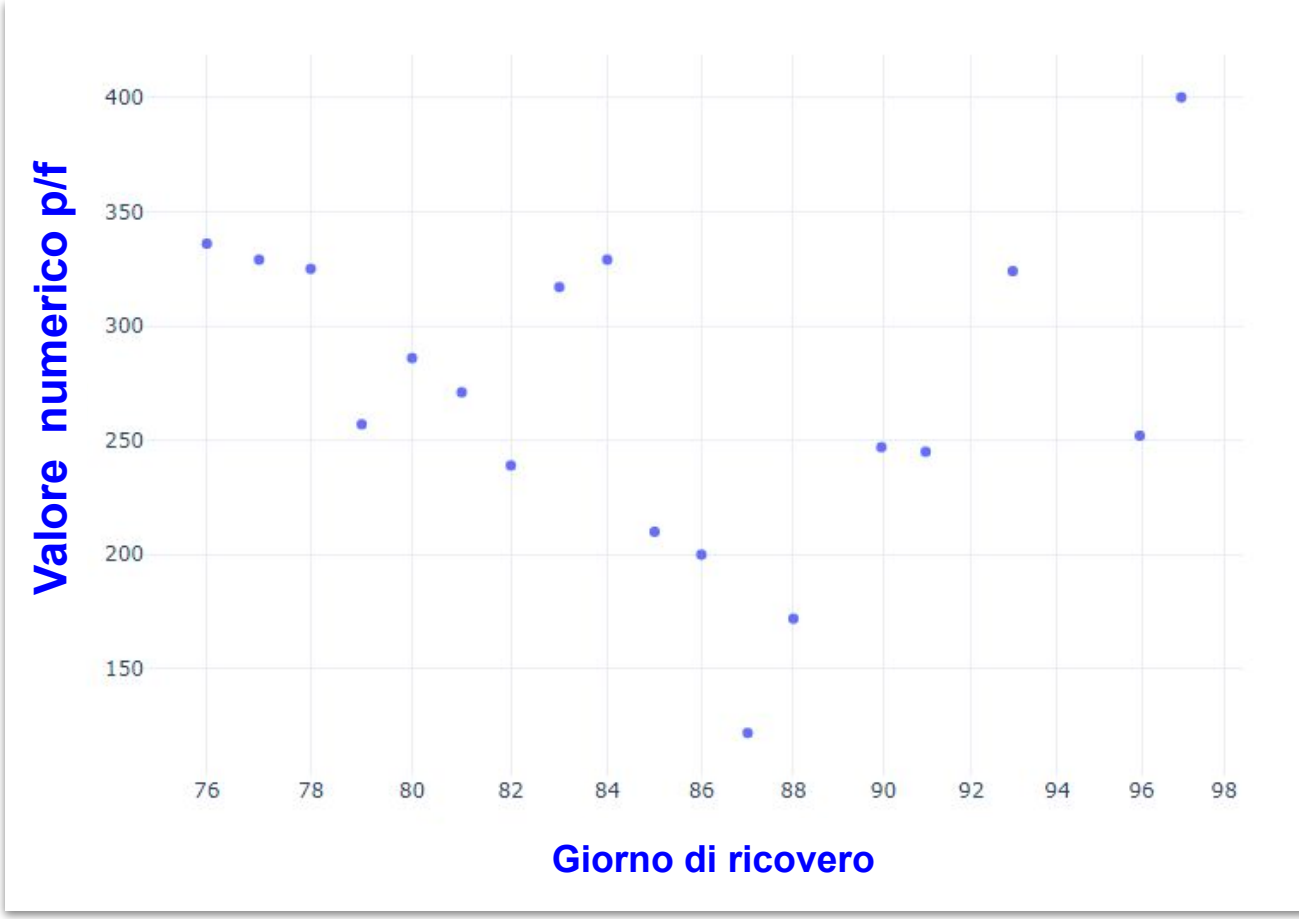
Esempio lettura dei dati: rapporto p/f per un singolo paziente

Rappresentazione p/f

→ Asse y: valore

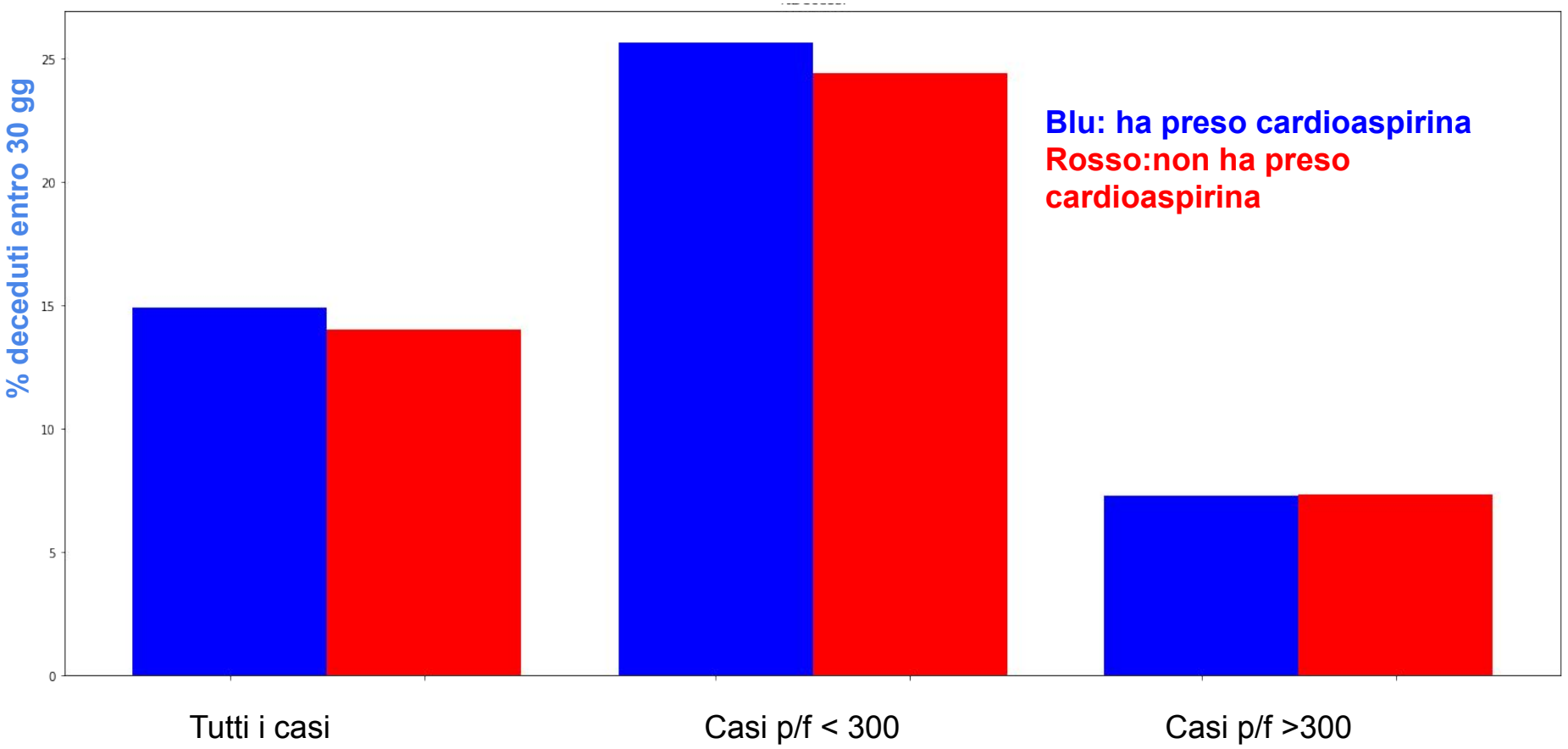
→ Asse x: giorno dall'inizio del 2020

→ Nelle slide future ci riferiamo al primo valore di p/f misurato



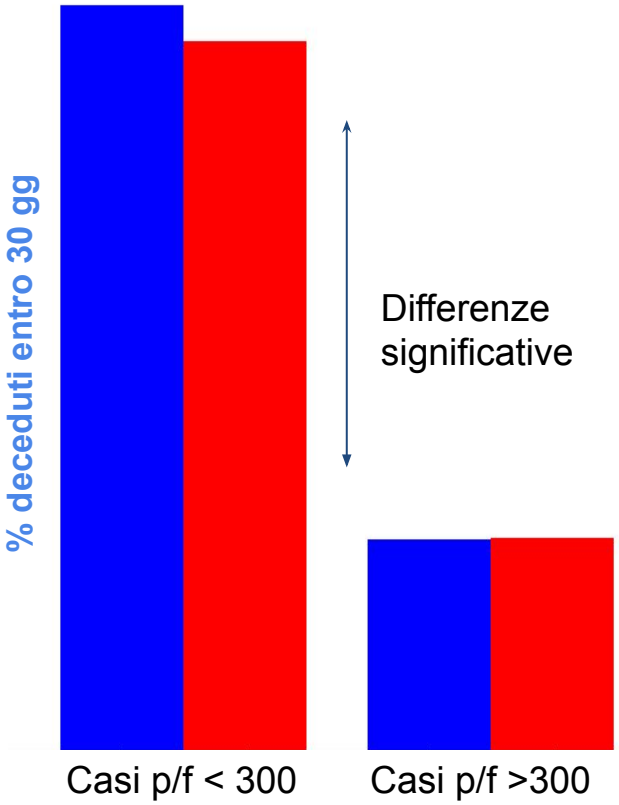
Altri esempi modelli 1-1

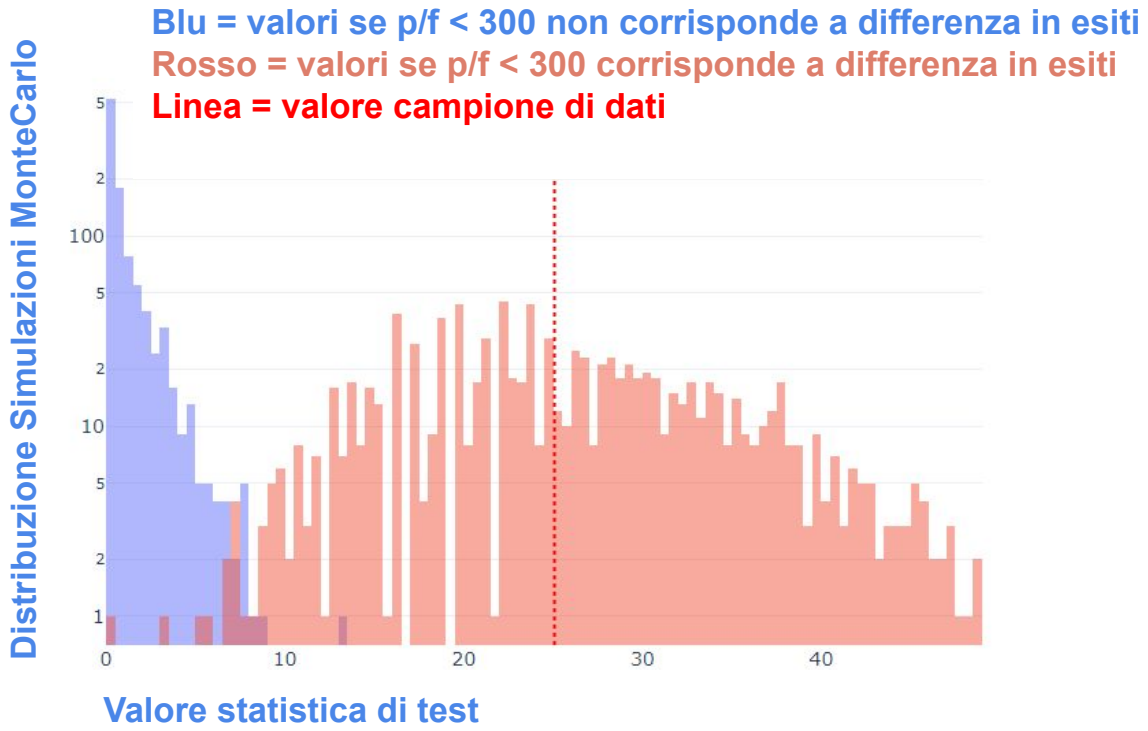
Modello “binomiale”: prototipo test d’ipotesi



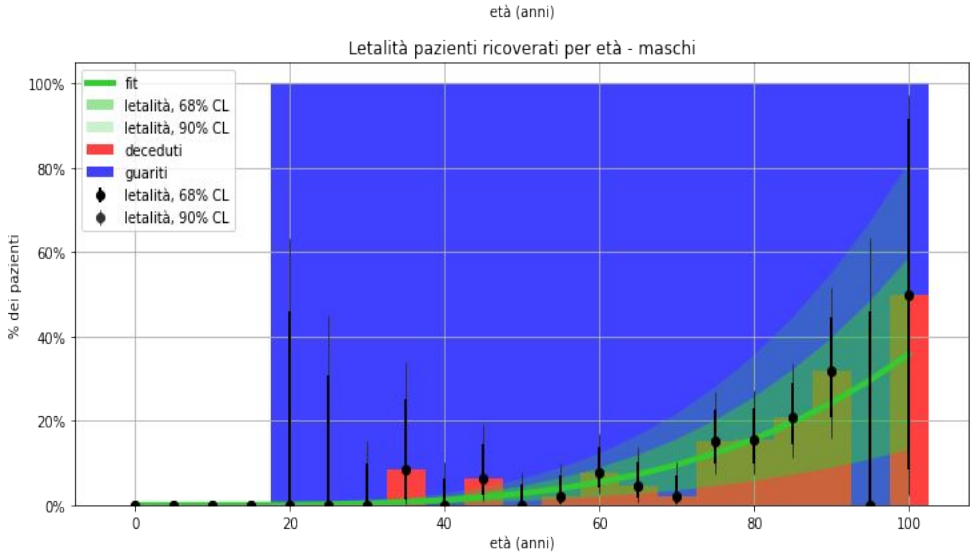
Modello “binomiale”: ipotesi ben separate

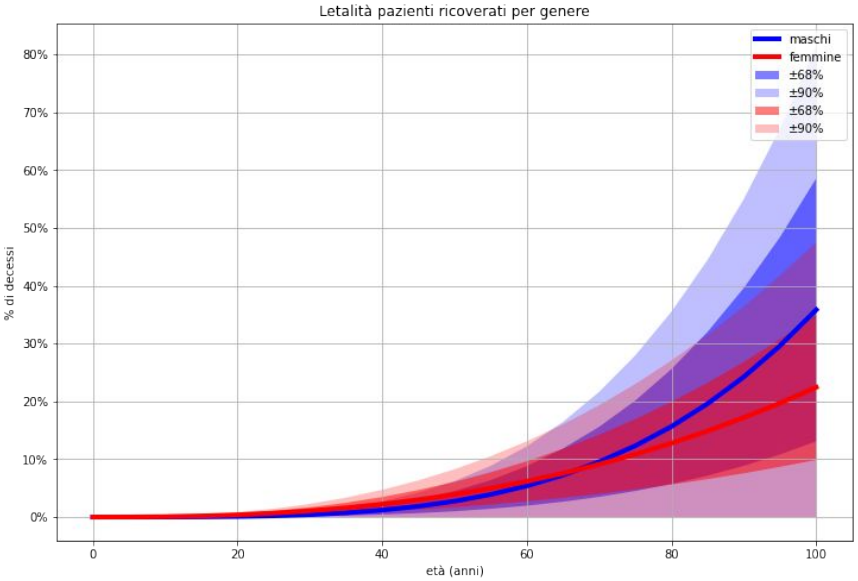
Valori p/f maggiori o minori di 300: differenza significativa negli esiti

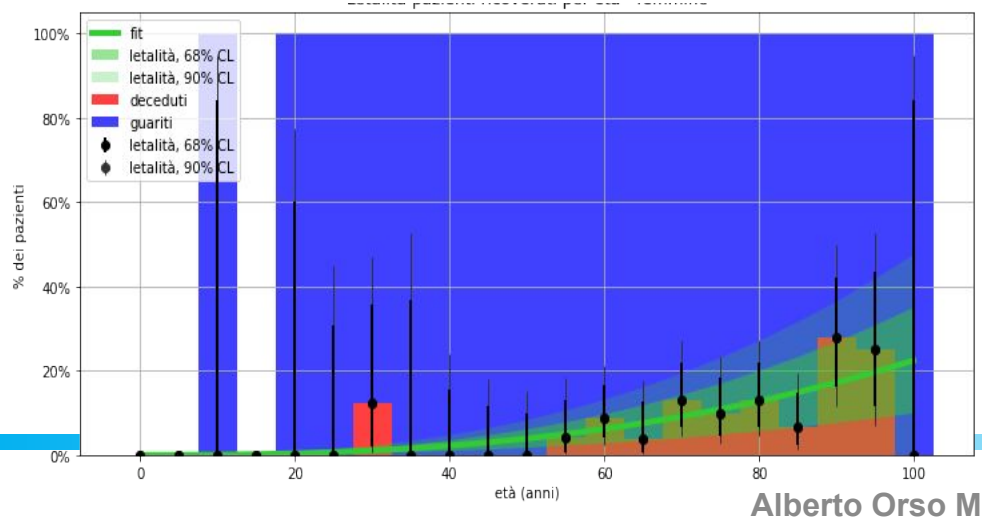




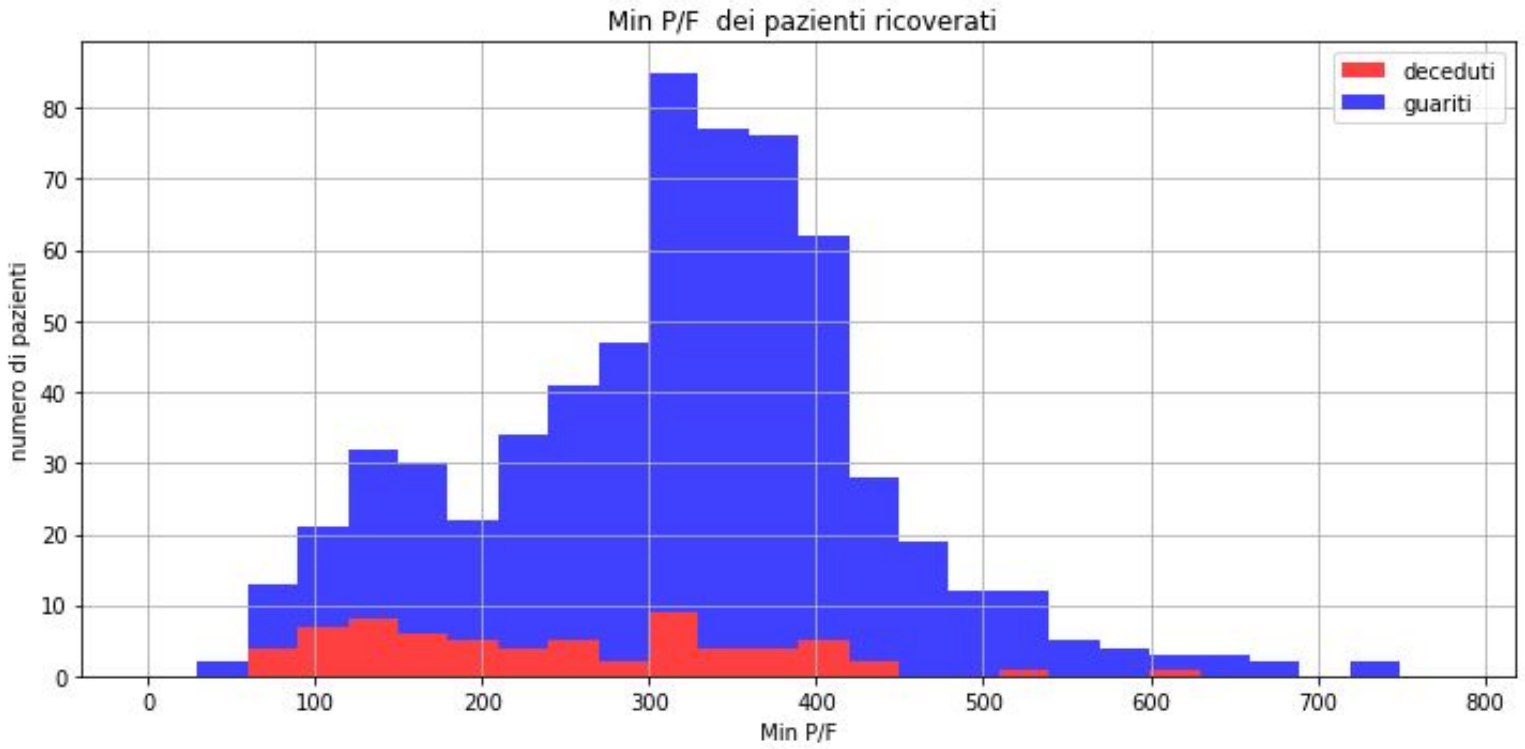
Dipendenze funzionali: caso età







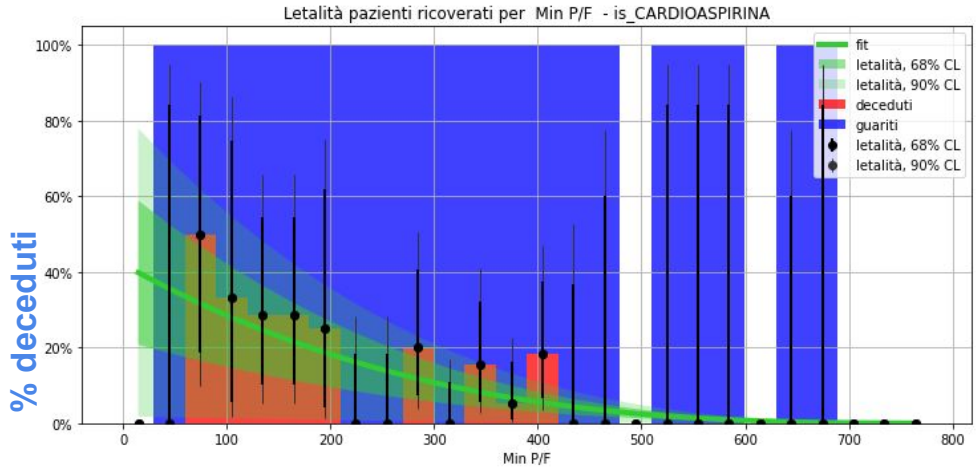
Modello di fit: facendo una separazione più fine



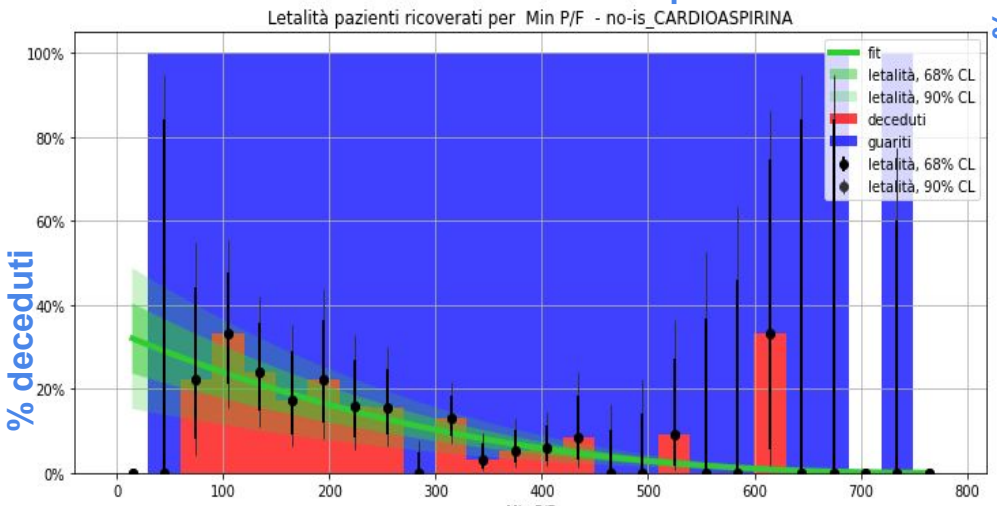
1) Studiare andamento globale dell'esito in funzione di variabile confondente

2) Confrontare andamento in due casi

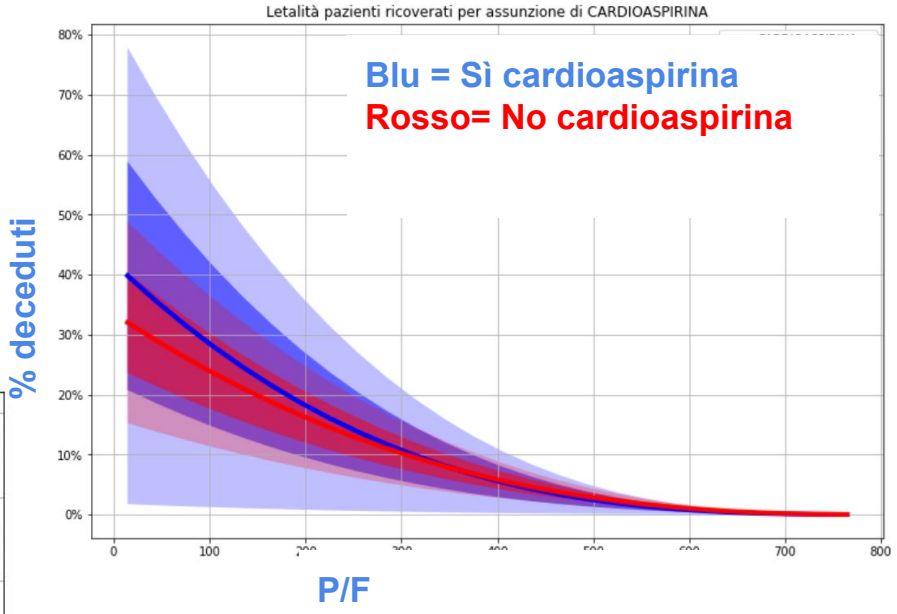
Dipendenze funzionali: caso cardioaspirina p/f



P/F sì Cardioaspirina

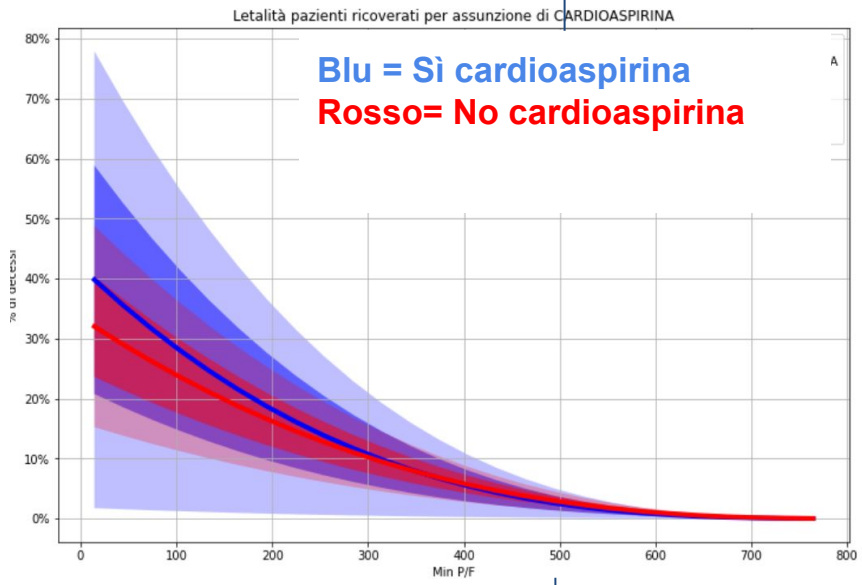


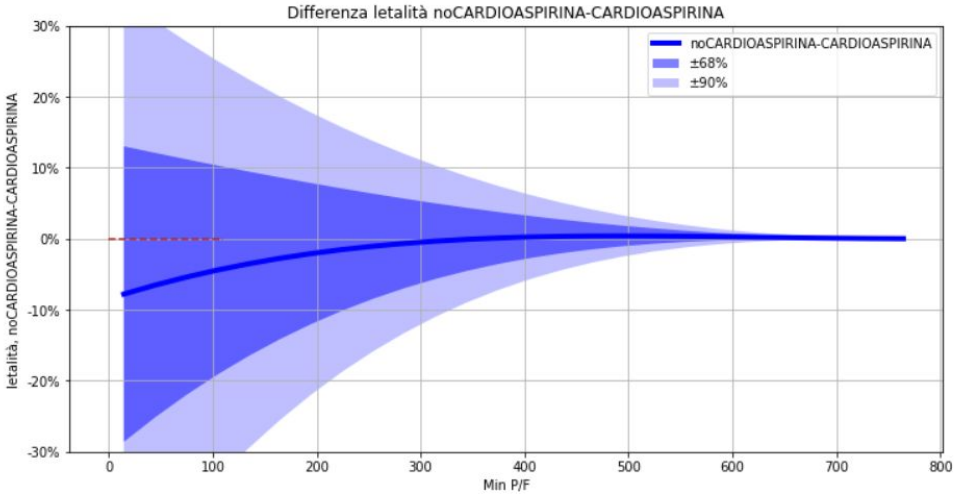
P/F no Cardioaspirina

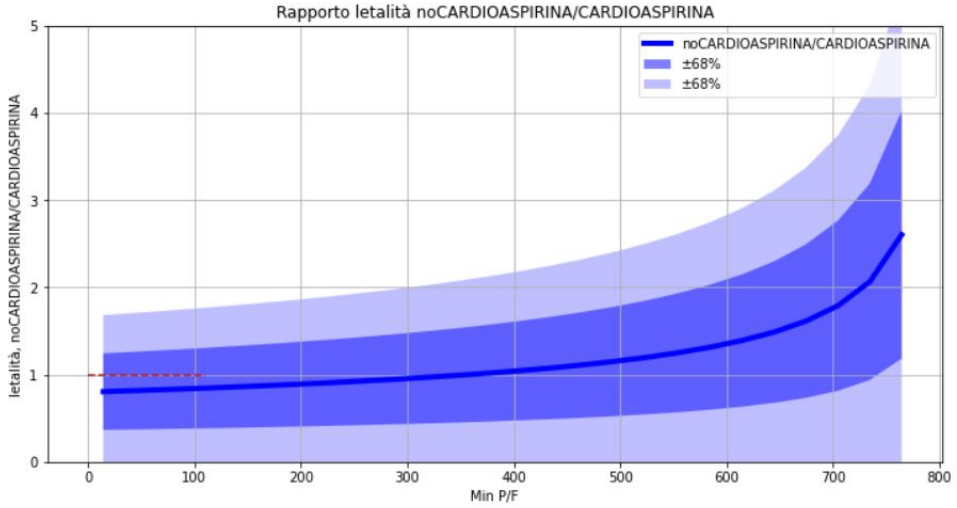


Dipendenze funzionali: caso cardioaspirina p/f

Differenza blu/rosso







Rapporto blu/rosso

Modello di andamento p/f per un singolo paziente

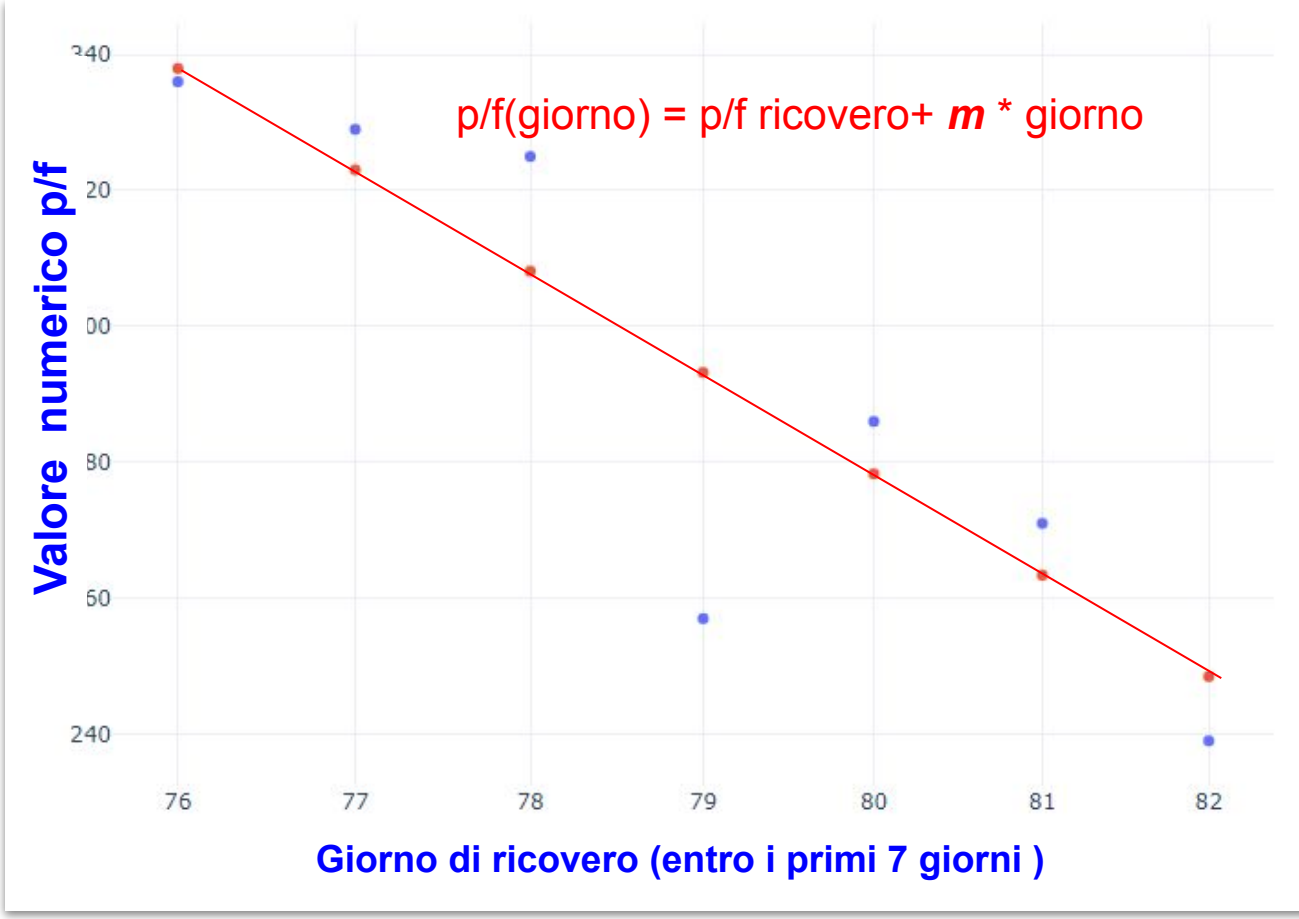
Modello lineare

→ Valore iniziale:

p/f al ricovero

→ pendenza ***m***:

Velocità di aggravamento



Dettagli logistica

Regressione logistica: metriche per il modello

→ Pseudo- R^2

$$R^2 = 1 - \frac{LL_{full\ model}}{LL_{intercept}}$$

→ AIC o BIC

Per modelli “nested”:
minore AIC/BIC = modello migliore

$$\mathrm{AIC}_i = -2\log L_i + 2p_i$$

$$\mathrm{BIC}_i = -2\log L_i + p_i \log n$$

LOGISTIC REGRESSION (DECEASE) DETAILS

Model:	GLM	AIC:	385.8984
Link Function:	logit	BIC:	-5094.7223
Dependent Variable:	DECEASE_28_DAYS	Log-Likelihood:	-181.95
Date:	2020-12-26 12:28	LL-Null:	-239.54
No. Observations:	824	Deviance:	363.90
Df Model:	10	Pearson chi2:	974.
Df Residuals:	813	Scale:	1.0000
Method:	IRLS	Pseudo R^2:	0.24

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-3.5815	0.3010	-11.8993	0.0000	-4.1714	-2.9916
eta_scaled	0.0405	0.0134	3.0271	0.0025	0.0143	0.0668
genere	-0.2871	0.3111	-0.9230	0.3560	-0.8969	0.3226
p_f_val_iniziale_scaled	-0.0073	0.0014	-5.1049	0.0000	-0.0102	-0.0045
diabete	-0.0948	0.3304	-0.2870	0.7741	-0.7423	0.5527
cardio	1.2089	0.3337	3.6230	0.0003	0.5549	1.8629
ipertens	-0.1038	0.3184	-0.3261	0.7443	-0.7279	0.5202
resp	0.2232	0.3261	0.6844	0.4937	-0.4160	0.8625
neuro	0.3631	0.3453	1.0516	0.2930	-0.3137	1.0400
metab	-0.2084	0.3456	-0.6032	0.5464	-0.8857	0.4688
obesita	1.1408	0.3723	3.0640	0.0022	0.4111	1.8706

Constant interpretation

$$\log\left(\frac{P\left(Y=1\middle|\vec{X}\right)}{P\left(Y=0\middle|\vec{X}\right)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Setting $\vec{X} = \vec{0}$ we have

$$\log\left(\frac{P\left(Y=1\middle|\vec{X}=\vec{0}\right)}{P\left(Y=0\middle|\vec{X}=\vec{0}\right)}\right) = \beta_0$$

$$\frac{P\left(Y=1\middle|\vec{X}=\vec{0}\right)}{P\left(Y=0\middle|\vec{X}=\vec{0}\right)} = e^{\beta_0}$$

e^{β_0} (reported in the plot of the previous page as «constant») is the ODDS when all the covariates are 0. That is, the odds to die if a patient is male (sex = 0) do not have any chronic disease, **Age = 0, P/F_initial = 0**.

Constant interpretation

$$\frac{P\left(Y=1\middle|\vec{X}=\vec{0}\right)}{P\left(Y=0\middle|\vec{X}=\vec{0}\right)} = e^{\beta_0}$$

e^{β_0} (reported in the plot of the previous page as «constant») is the ODDS when all the covariates are 0. That is, the odds to die if a patient is male (sex = 0) do not have any chronic disease, **Age = 0, P/F_initial = 0**.

Age = 0, P/F_initial = 0 do not have any sense

To make e^{β_0} explainable I use the Age and P/F_initial as the difference from their mean.

$$\mathbf{Age_i = Age_i - Mean(Age)}$$

$$\mathbf{P/F\_initial_i = P/F\_initial_i - Mean(P/F\_initial)}$$

Constant interpretation

$$\mathbf{Age_i = Age_i - Mean(Age)}$$

$$\mathbf{P/F\_initial_i = P/F\_initial_i - Mean(P/F\_initial)}$$

Thanks to these transformations, e^{β_0} (reported in the barplotof the previous slide as «constant») is the ODDS when all the covariates are 0.

That is, e^{β_0} is the odds to die if a patient is male (sex = 0), do not have any chronic disease, have the **average Age (≈ 62 y.o.)** and **average P/F_initial (≈ 347)**.

$$\frac{P\left(Y = 1 \middle| \vec{X} = \vec{0}\right)}{P\left(Y = 0 \middle| \vec{X} = \vec{0}\right)} = e^{\beta_0} = 3\%$$

Idee per la feature selection

Possibilità: CFS (Correlation based Feature Selection) trovare variabili di interazione (proposta). CFS dà un punteggio a un subset di features in base alla correlazione fra loro e a quella con la variabile dipendente (nel nostro caso DECESSO o TI)

→ Interazione fra farmaci (es metformina*cardioaspirina)

→ Comorbilità (es ipertensione*diabete)

→ Farmaco*Malattia

Alternativa: trovare le variabili di interazione guardando la matrice di correlazione.

Altre idee e risultati

Modelli a molte variabili

Modelli che tengono conto di più fattori simultaneamente:

1) **Identificare** il sottoinsieme delle **variabili interessanti**

2) **Comprendere** i nessi tra più variabili e le correlazioni

3) **Predire** un esito in funzione delle caratteristiche di un paziente

Idee per la feature selection

Possibilità: CFS (Correlation based Feature Selection) trovare variabili di interazione (proposta). CFS dà un punteggio a un subset di features in base alla correlazione fra loro e a quella con la variabile dipendente (nel nostro caso DECESSO o TI)

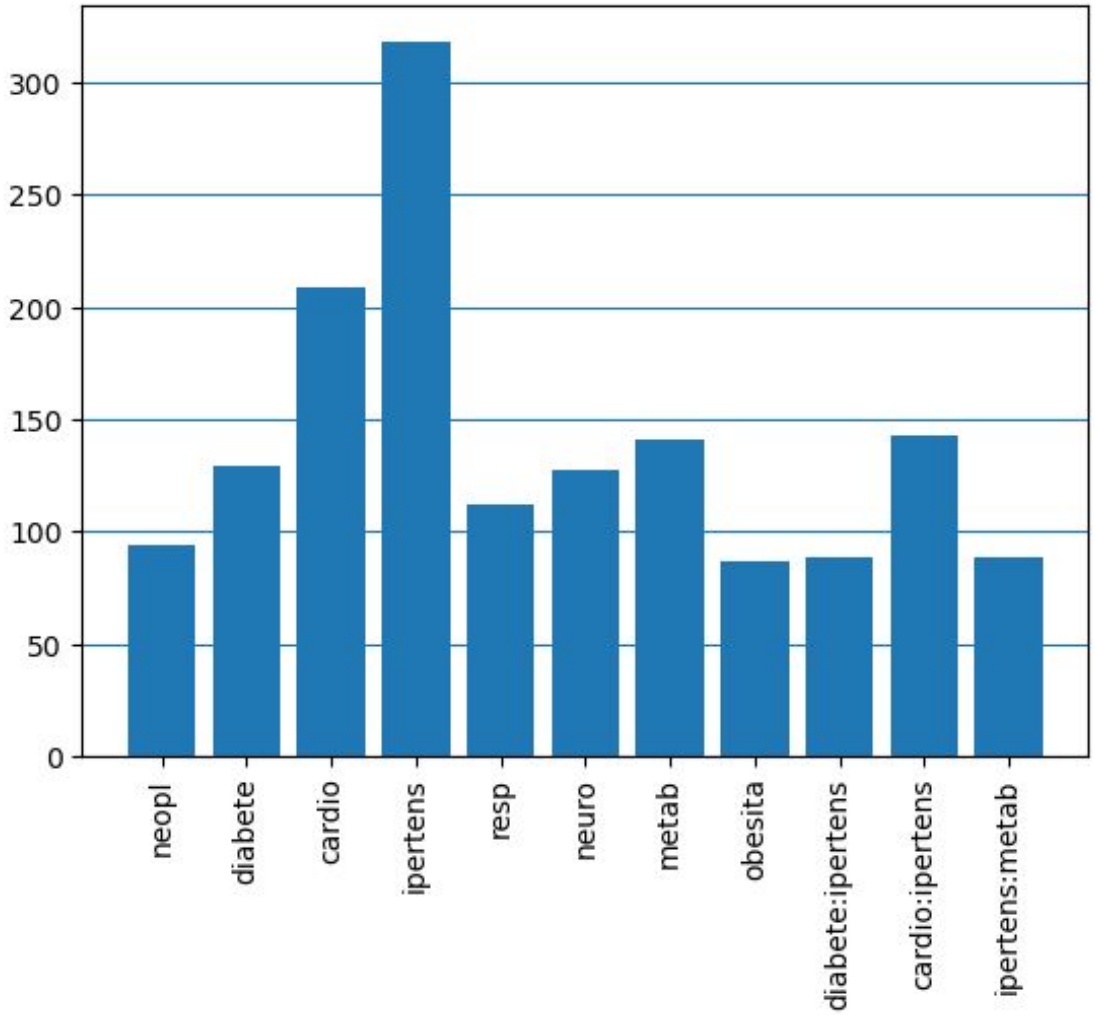
→ Interazione fra farmaci (es metformina*cardioaspirina)

→ Comorbilità (es ipertensione*diabete)

→ Farmaco*Malattia

Alternativa: trovare le variabili di interazione guardando la matrice di correlazione.

Frequenza patologie

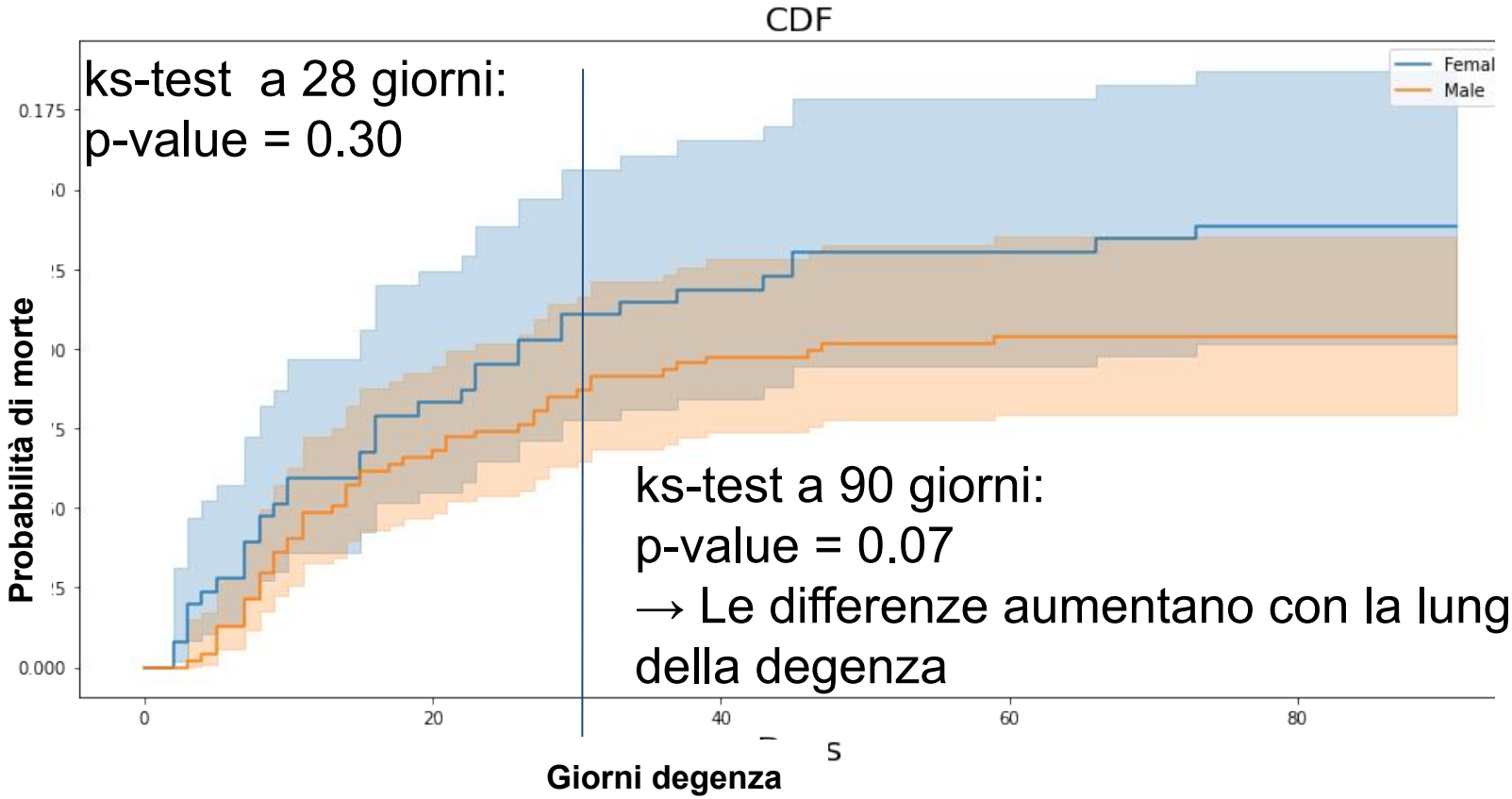


Analisi di sopravvivenza: funzioni cumulative

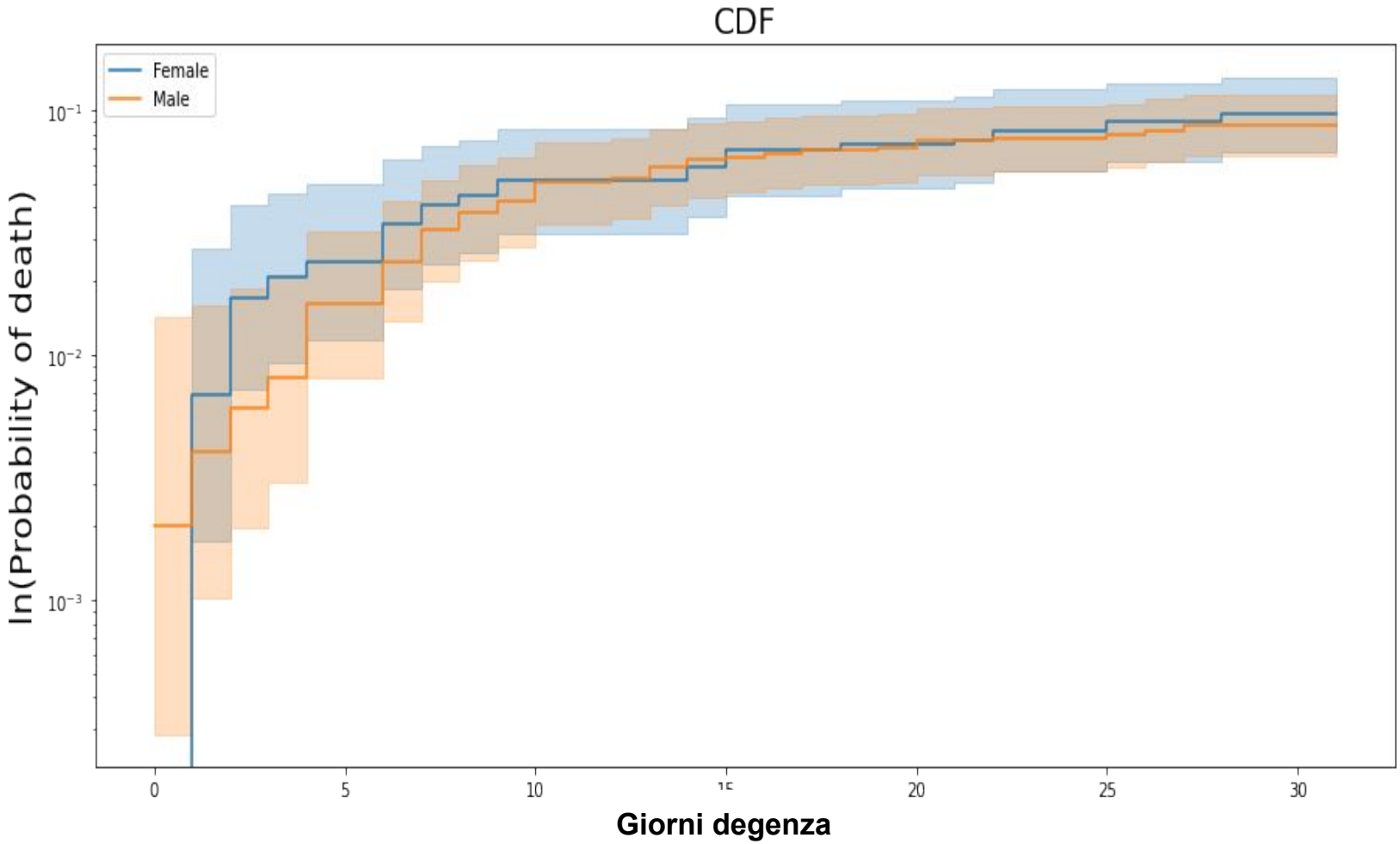
Funzione di probabità cumulativa = numero totale di decessi in funzione del tempo

→ Si può calcolare anche se il modello non è lineare

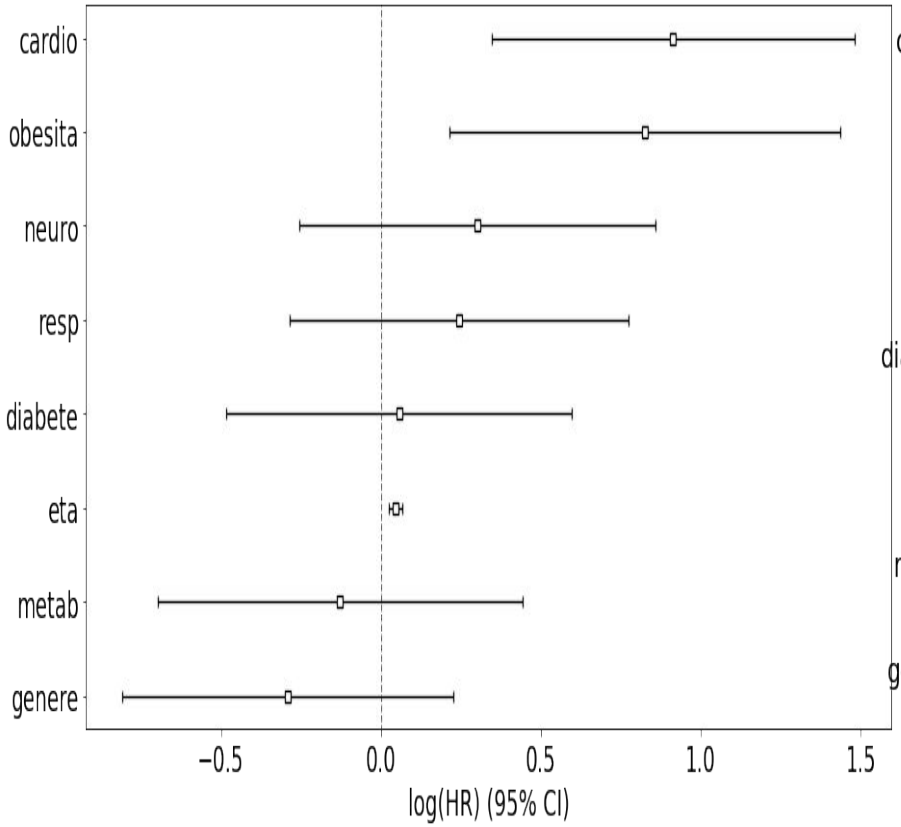
→ Differenze nelle cumulative ⇔ possibile test sulla forma della distribuzione

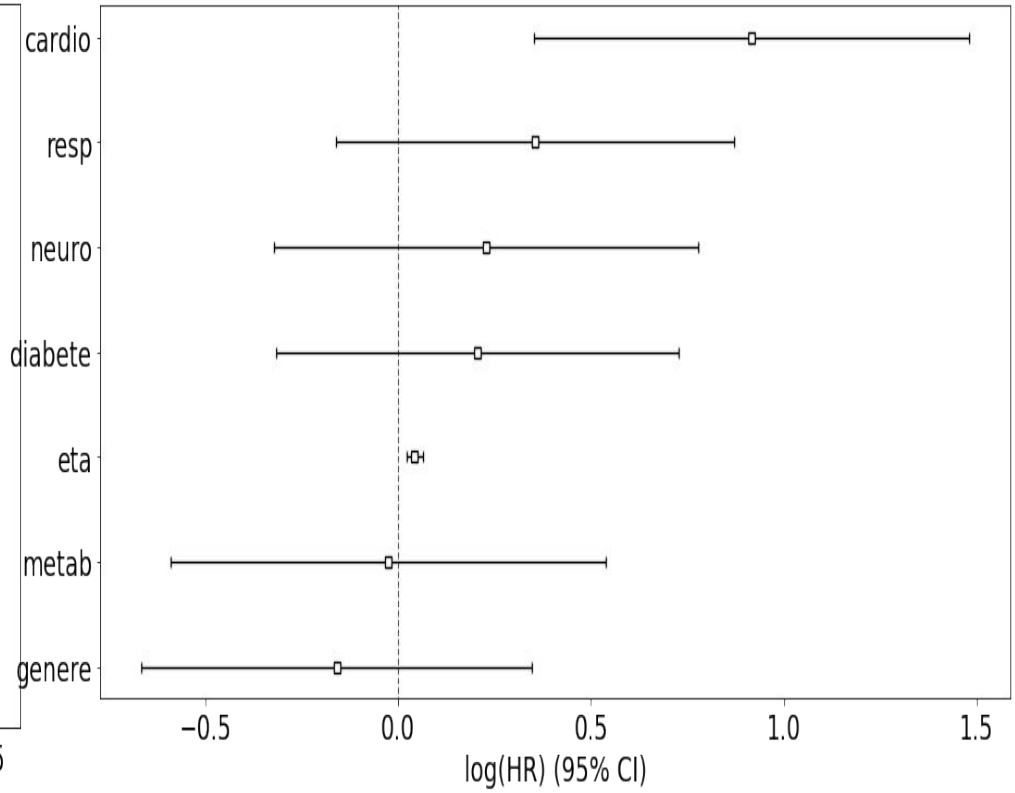


Analisi di sopravvivenza CDF m/f:

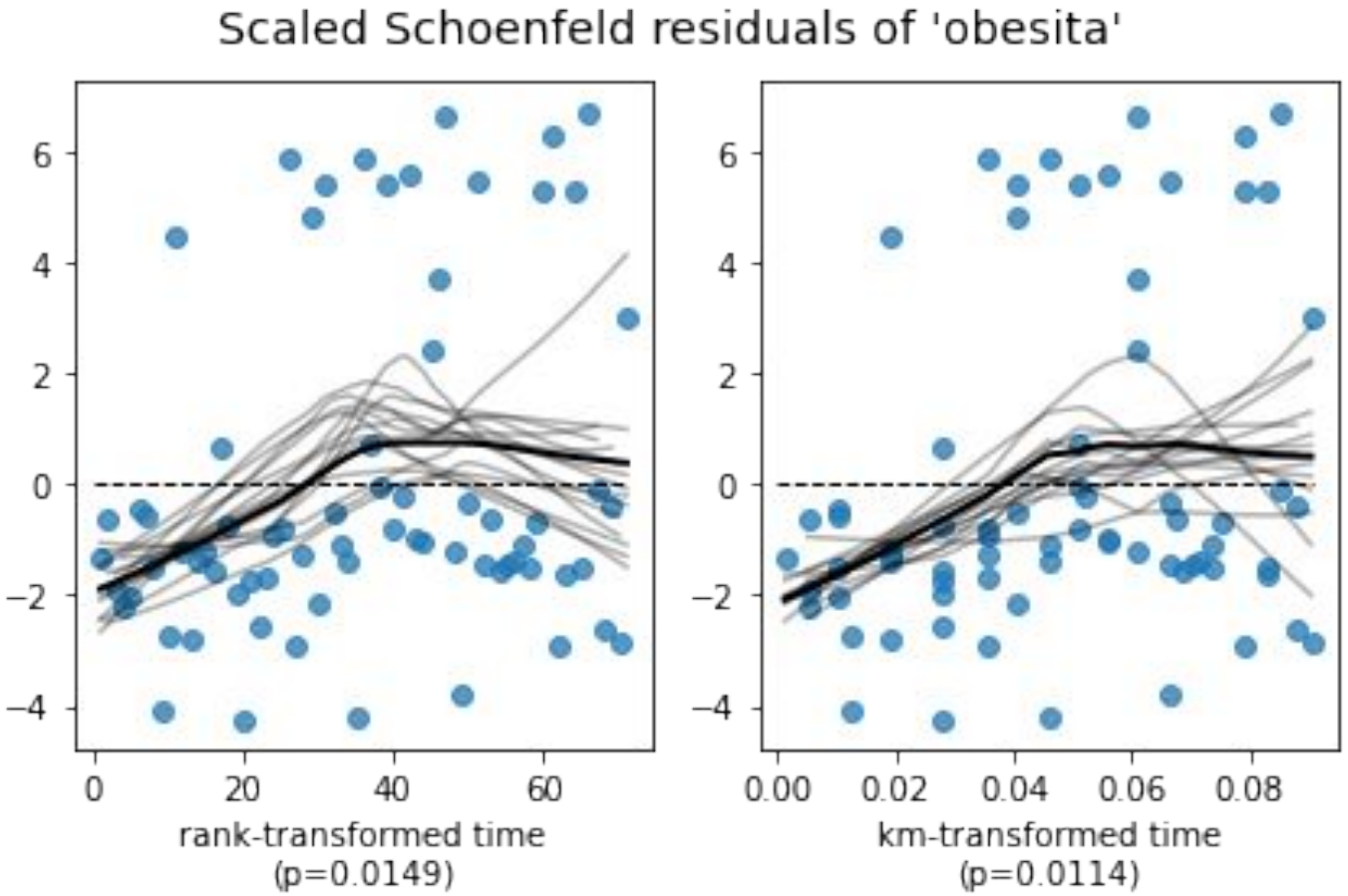


Analisi di sopravvivenza obesi non obesi:

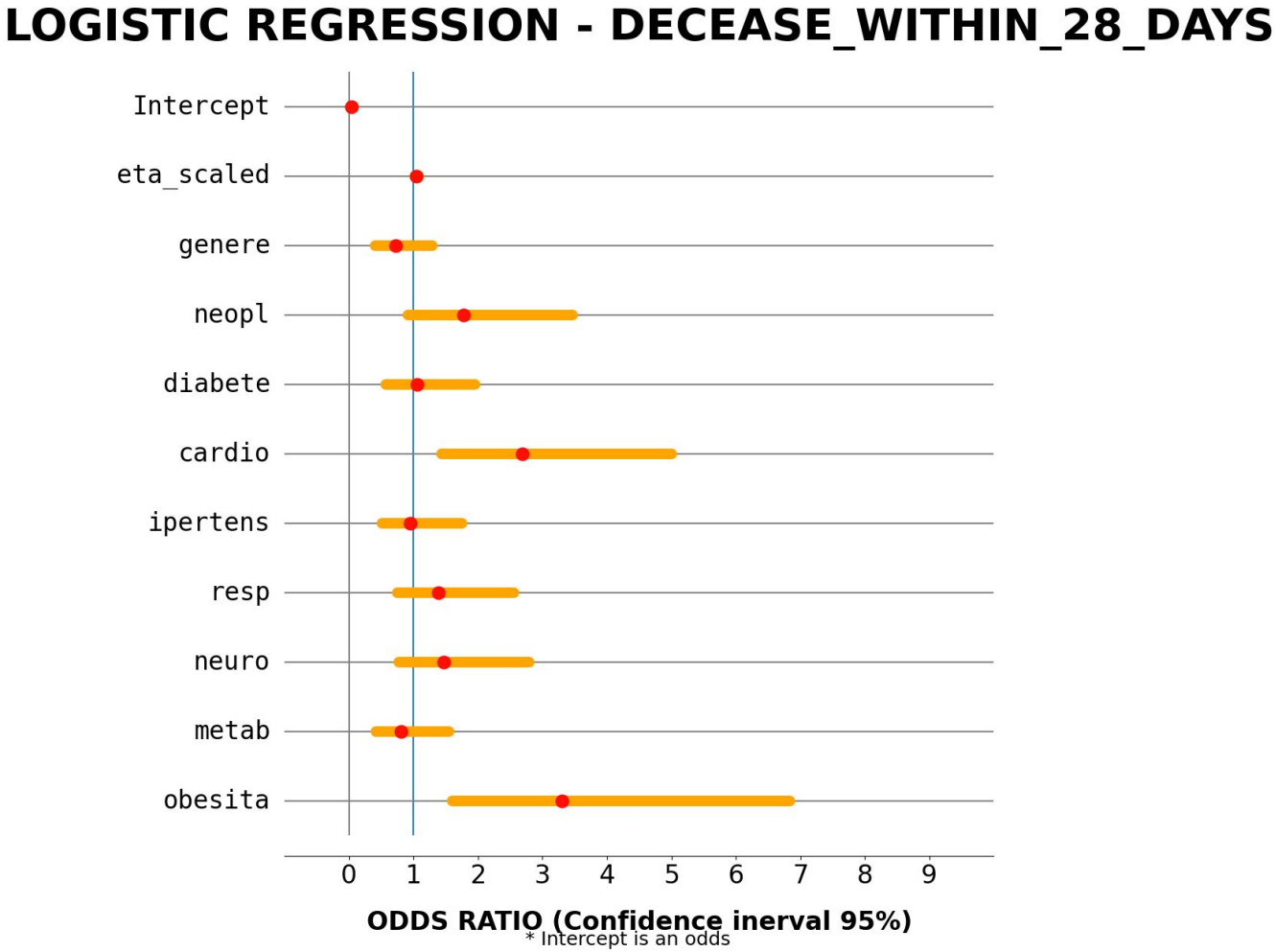




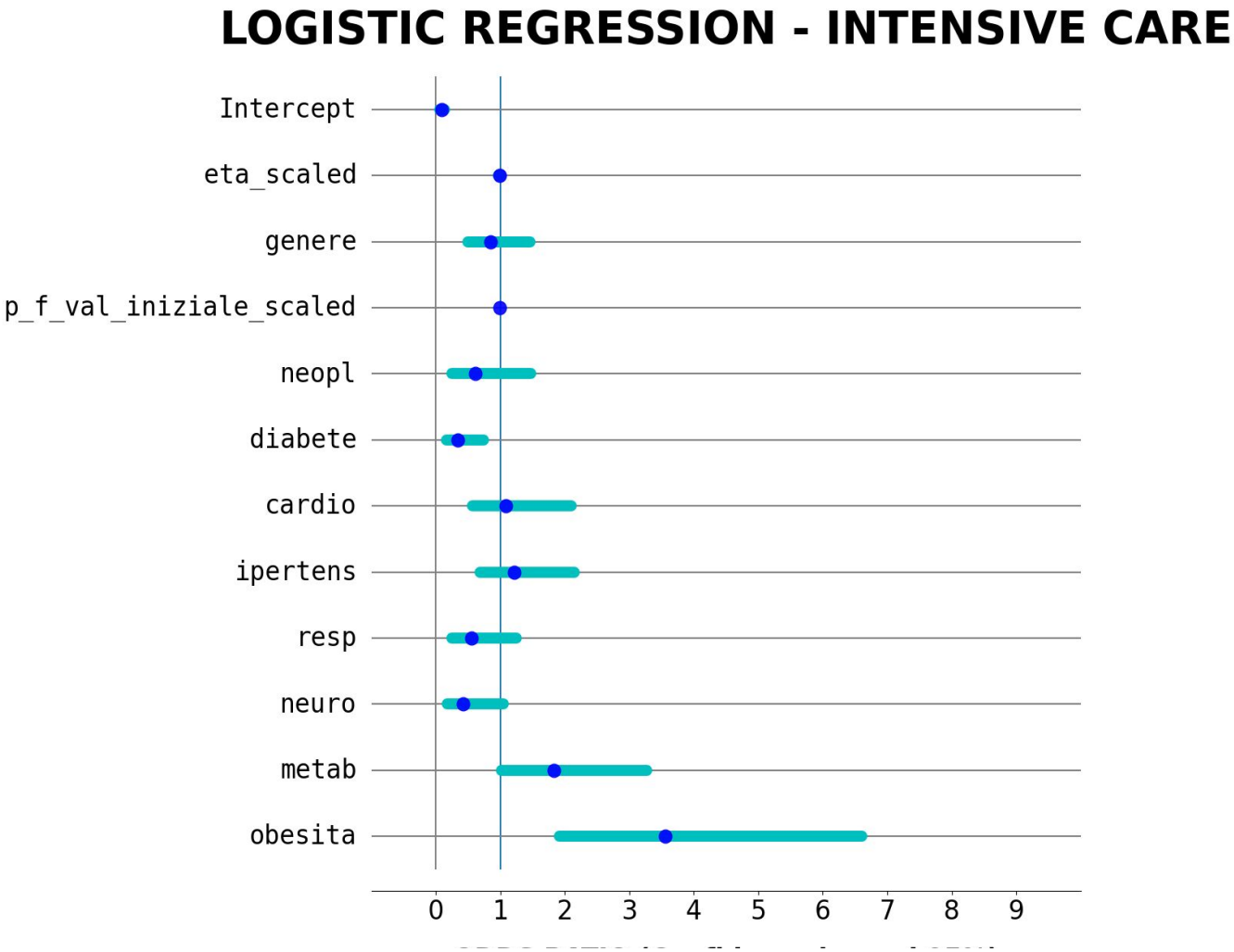
Analisi di sopravvivenza residui obesità:



Logistica senza p/f Imputation:



Logistica senza p/f Imputation:



Logistica senza p/f Imputation: