UAlg
UNIVERSIDADE DO ALGARVE

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Analysis of high throughput antibody data for better understanding of immunogenetics and epidemiology of malaria

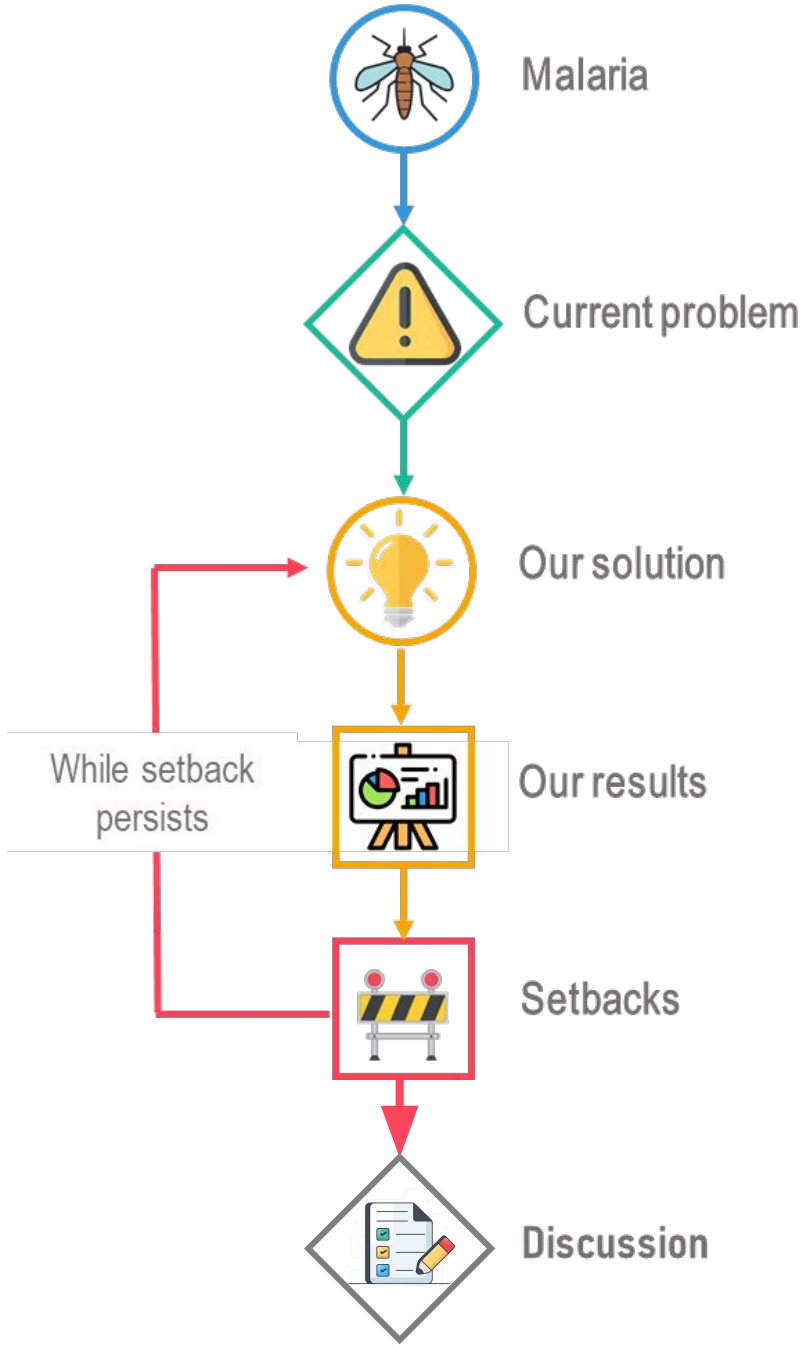
Ref FCT : SFRH/BD/147629/2019

Supervisor: Dr Nuno Sepúlveda
Co-supervisor: Dr Clara Cordeiro

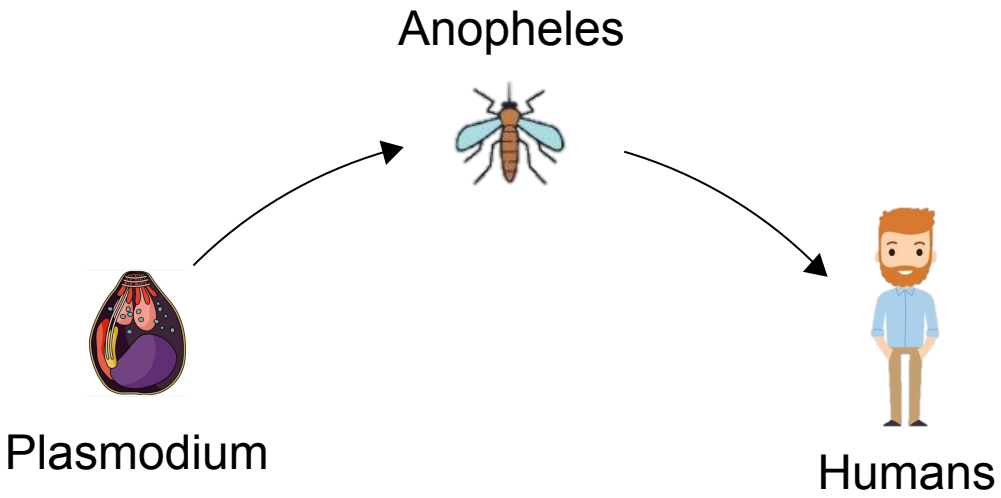
André Fonseca

06/12/2021

Agenda


Malaria
Current problem
Our solution
While setback persists
Our results
Setbacks
Discussion

Malaria



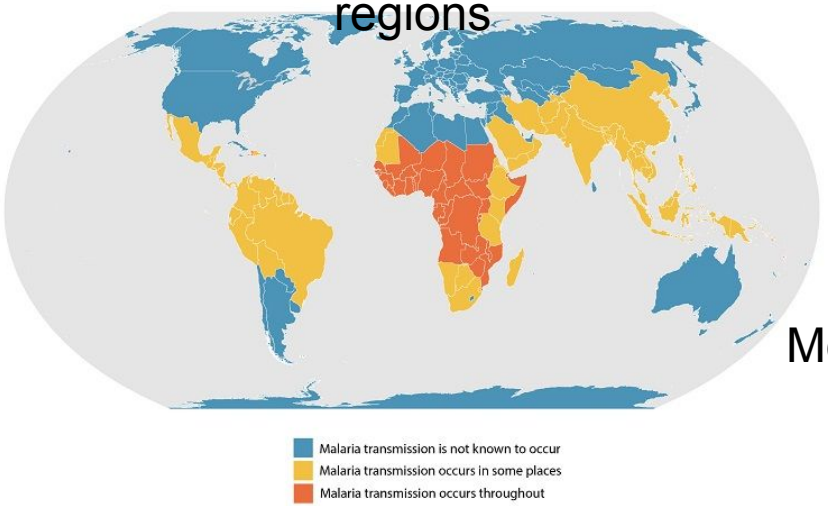
Malaria

Anopheles

Plasmodium

Humans

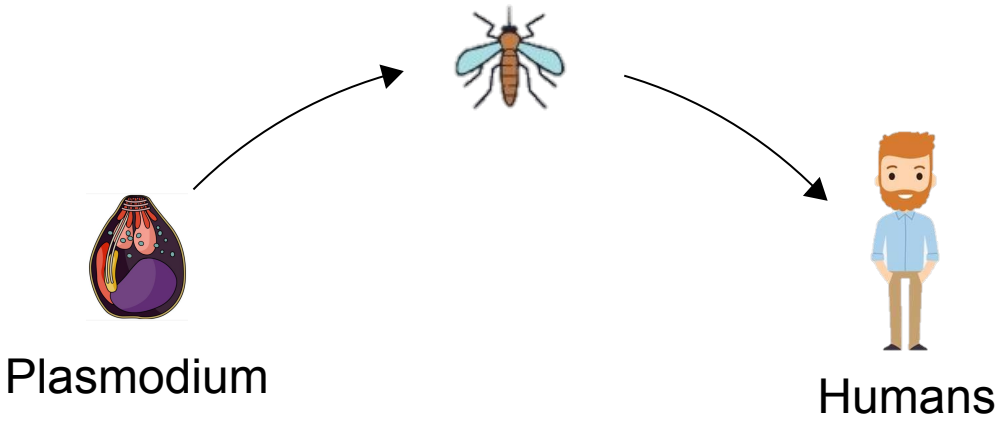
Prevalent in tropical regions



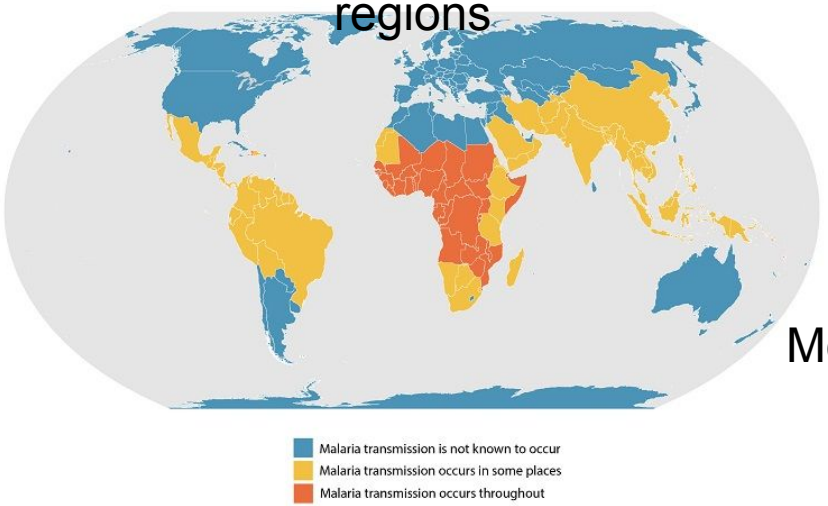
Age < 5

Most affected

Malaria



Prevalent in tropical regions



Age < 5

Most affected

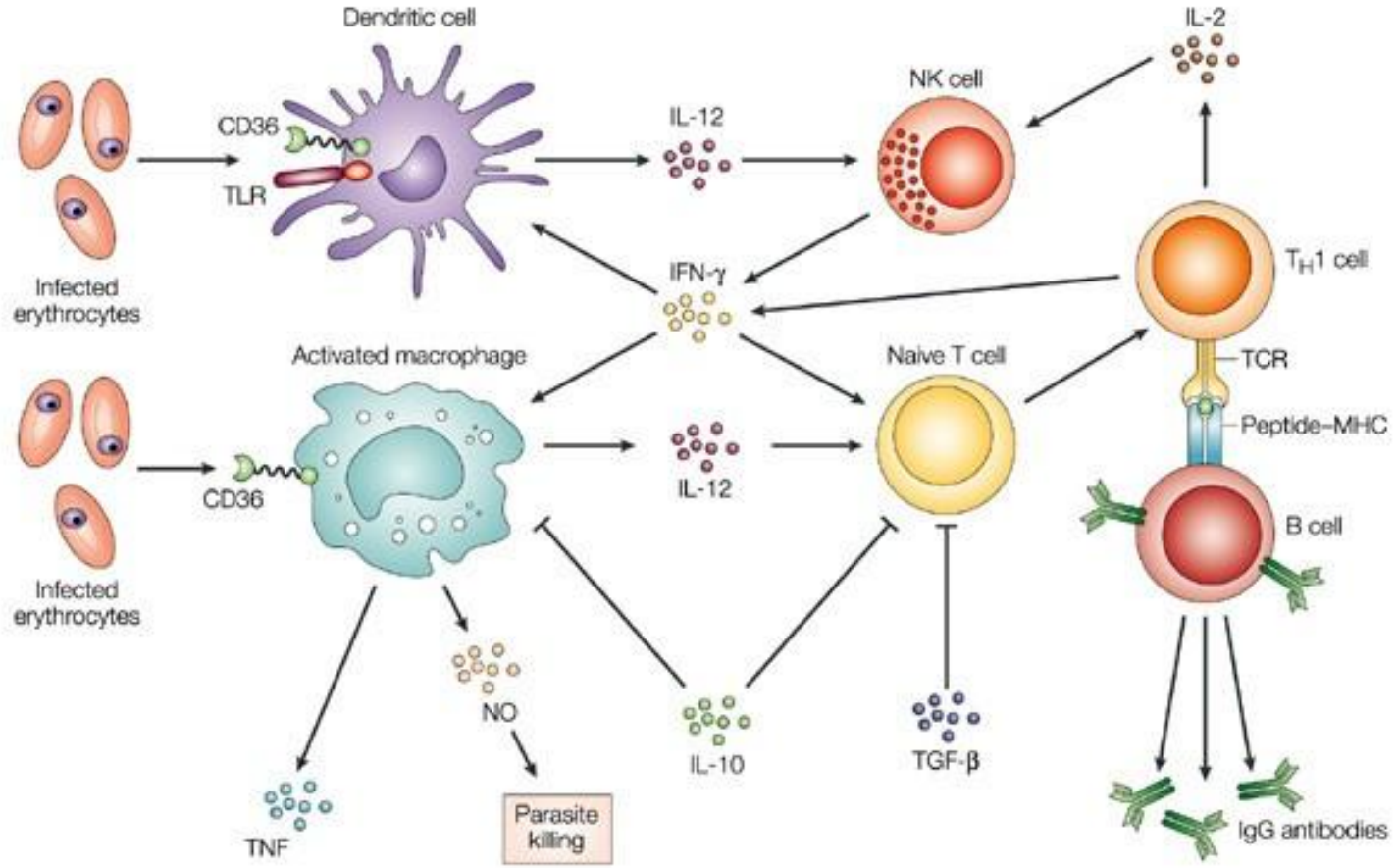
Childre n
Lack of immunity

Infectio n
Repeated infections with the malaria parasite

Antibodie s
Increase in antibodies against malaria

Protectio n
Dicrease in clinical episodes

Antibody protection



Problem

Unable to identify antibodies for **vaccine** development against malaria

Problem

Unable to identify antibodies for **vaccine** development against malaria

RTS,S/AS01

Efficacy on children (36%)

Problem

Unable to identify antibodies for **vaccine** development against malaria

RTS,S/AS01

Efficacy on children (36%)

Difficulties in indetifying vaccine candidate

- **Biological factors**
- **Technical factors**

Problem

Unable to identify antibodies for **vaccine** development against malaria

RTS,S/AS01

Efficacy on children (36%)

Difficulties in indetifying vaccine candidate

Biological factors

- Evolution of the parasite
- Complex life cycle
- Individual imune response

Technical factors

Problem

Unable to identify antibodies for **vaccine** development against malaria

RTS,S/AS01

Efficacy on children (36%)

Difficulties in indetifying vaccine candidate

Biological factors

- Evolution of the parasite
- Complex life cycle
- Individual imune response

Technical factors

- Experimental design

Problem

Unable to identify antibodies for **vaccine** development against malaria

RTS,S/AS01

Efficacy on children (36%)

Difficulties in indetifying vaccine candidate

Biological factors

- Evolution of the parasite
- Complex life cycle
- Individual imune response

Technical factors

- Experimental design
- Sub-optimal analysis

Incosistent results

failure of the statistical assumptions **(normality)**

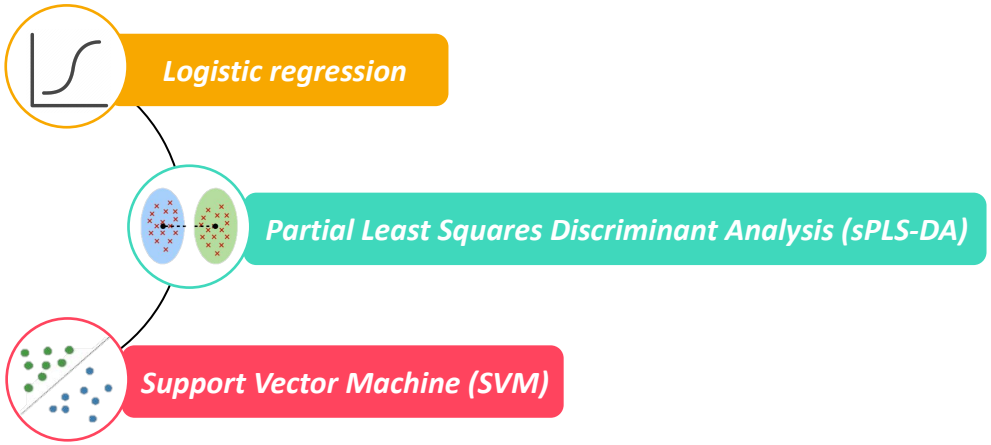
Problem

Research

Immune Signature Against *Plasmodium falciparum* Antigens Predicts Clinical Immunity in Distinct Malaria Endemic Communities

Authors

Carla Proietti, Lutz Krause, Angela Trieu, Daniel Dodoo, Ben Gyan, Kwadwo A. Koram, William O. Rogers, Thomas L. Richie, Peter D. Crompton, Philip L. Felgner, and Denise L. Doolan

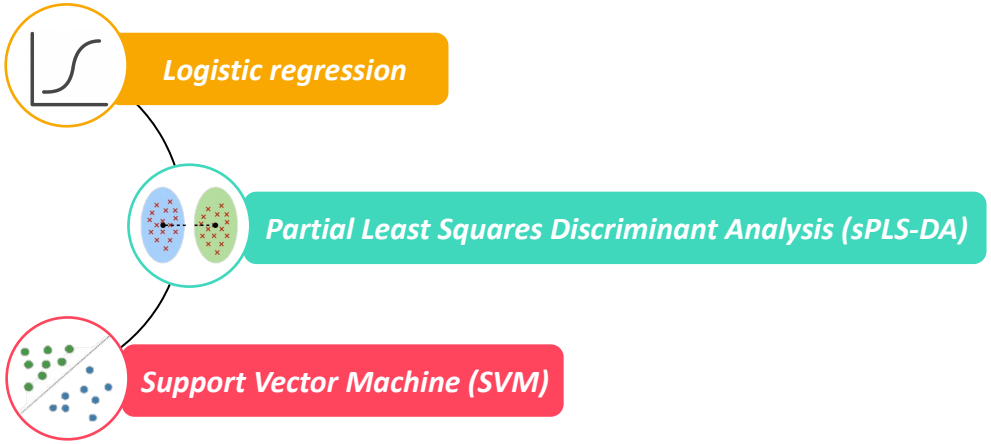


Problem

Research

Immune Signature Against *Plasmodium falciparum* Antigens Predicts Clinical Immunity in Distinct Malaria Endemic Communities

Authors
Carla Proietti, Lutz Krause, Angela Trieu, Daniel Dodoo, Ben Gyan, Kwadwo A. Koram, William O. Rogers, Thomas L. Richie, Peter D. Crompton, Philip L. Felgner, and Denise L. Doolan



PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Identification of immune signatures predictive of clinical protection from malaria

John Joseph Valletta, Mario Recker*

Centre for Mathematics and the Environment, University of Exeter, Penryn Campus, Penryn, United Kingdom

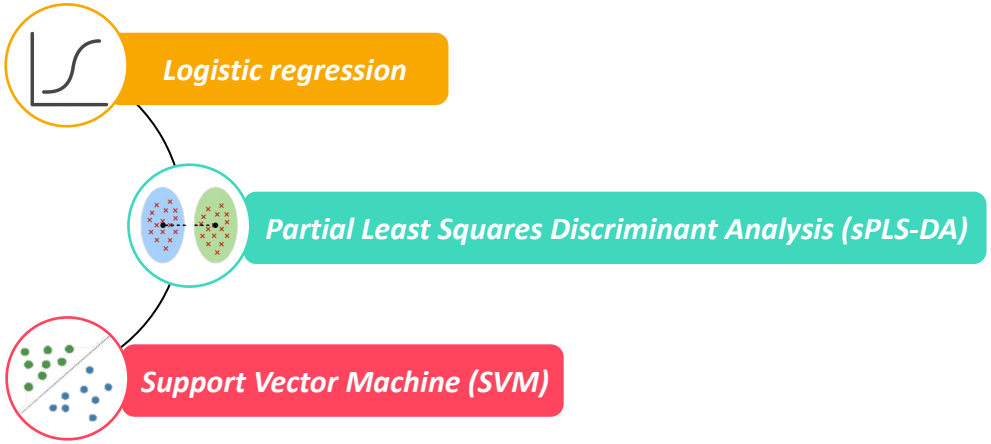
* m.recker@exeter.ac.uk

Problem

Research

Immune Signature Against *Plasmodium falciparum* Antigens Predicts Clinical Immunity in Distinct Malaria Endemic Communities

Authors
Carla Proietti, Lutz Krause, Angela Trieu, Daniel Dodoo, Ben Gyan, Kwadwo A. Koram, William O. Rogers, Thomas L. Richie, Peter D. Crompton, Philip L. Felgner, and Denise L. Doolan

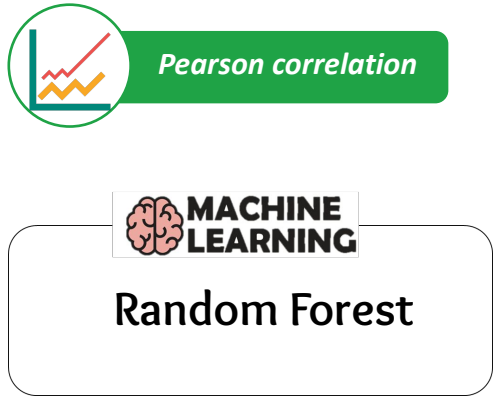


PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Identification of immune signatures predictive of clinical protection from malaria

John Joseph Valletta, Mario Recker*
Centre for Mathematics and the Environment, University of Exeter, Penryn Campus, Penryn, United Kingdom
* m.recker@exeter.ac.uk



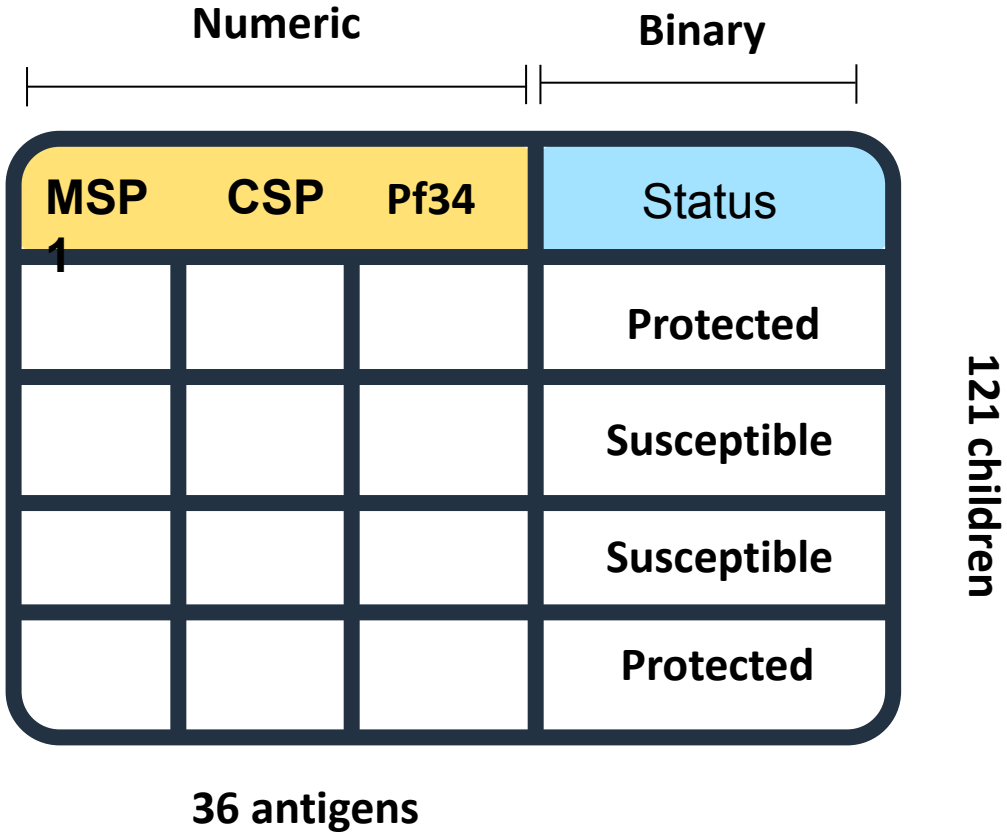
Statistical pipelines to analyze immunological data consistently and reliably

Our dataset

IgG antibodies against **36 Plasmodium Falciparum antigens**

Antibody quantity information for **121 Kenian children** (aged 1-10 years)

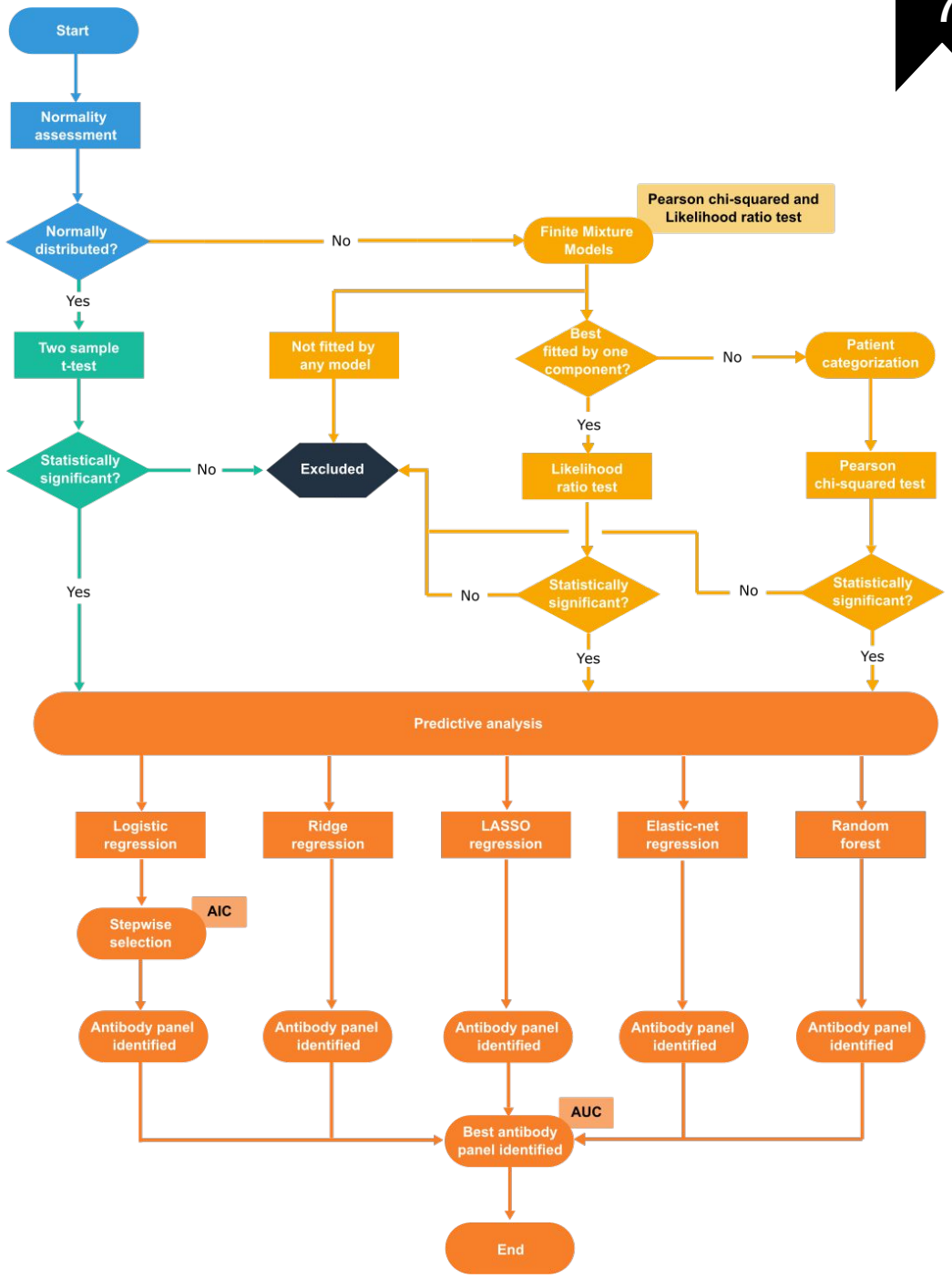
The **response variable** “Status” was binary (Protected vs. Susceptible)



Solution

Parametric pipeline

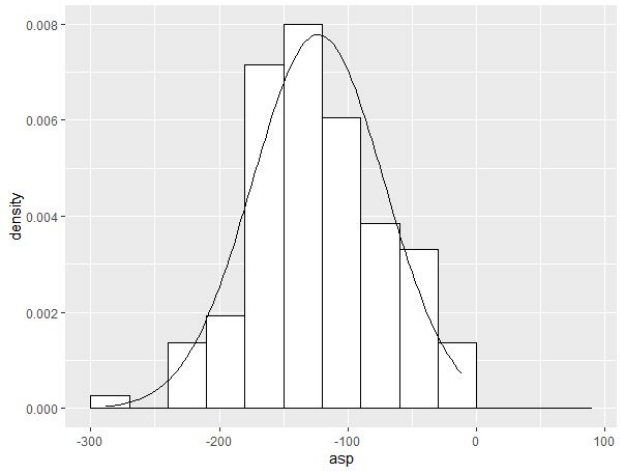
Normal distribution and flexible **mixture models** for seropositivity determination

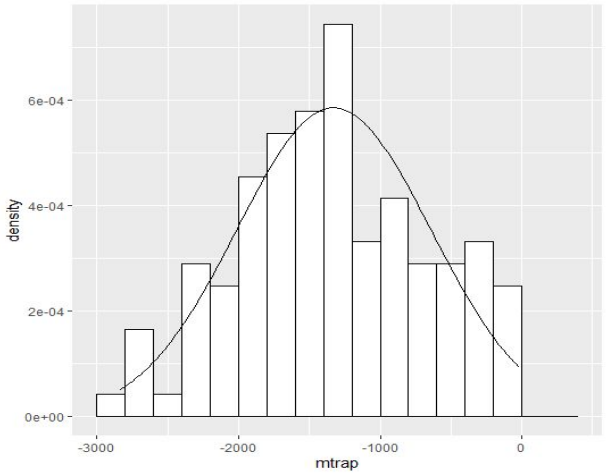


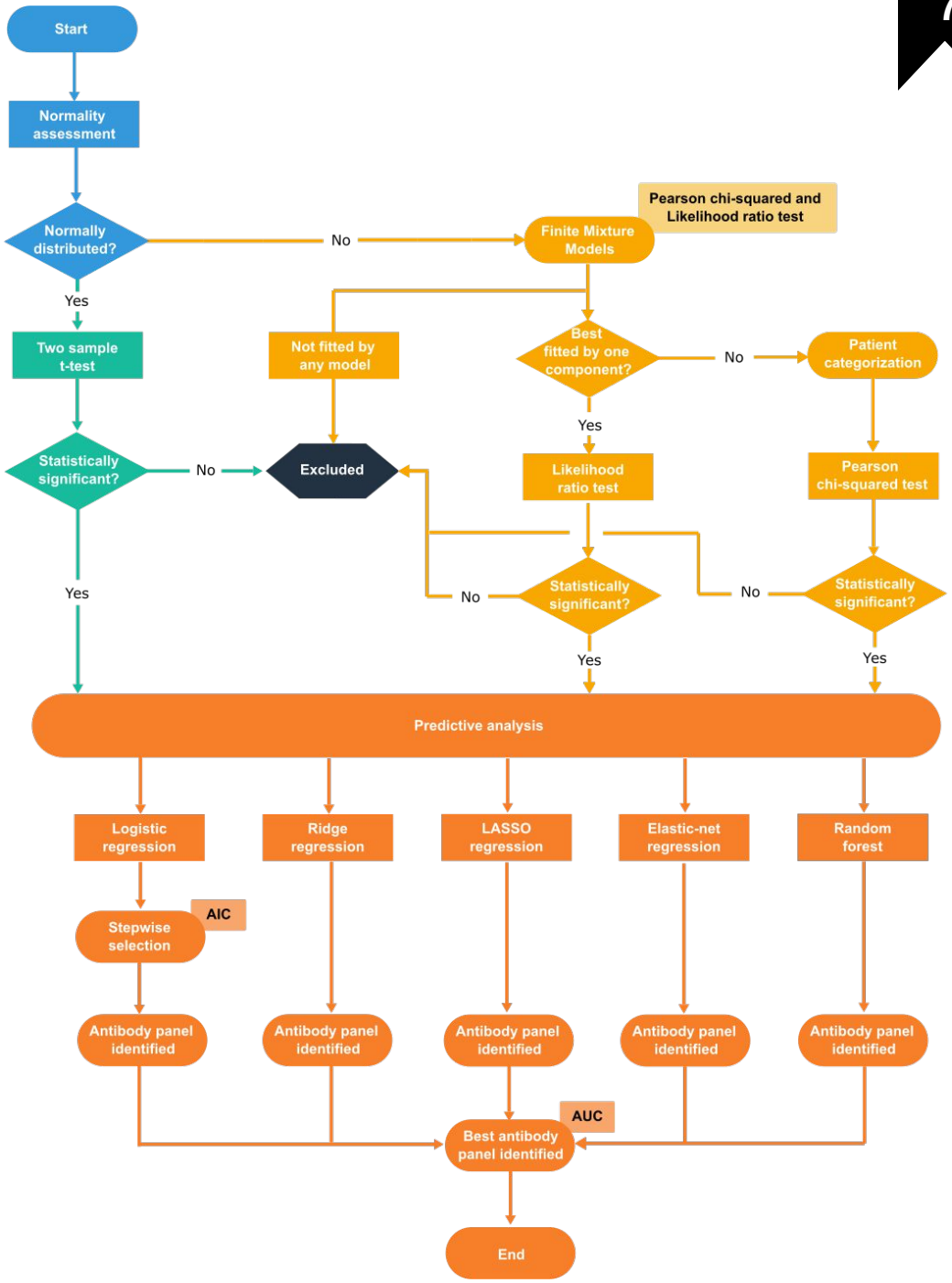
Solution

Parametric pipeline

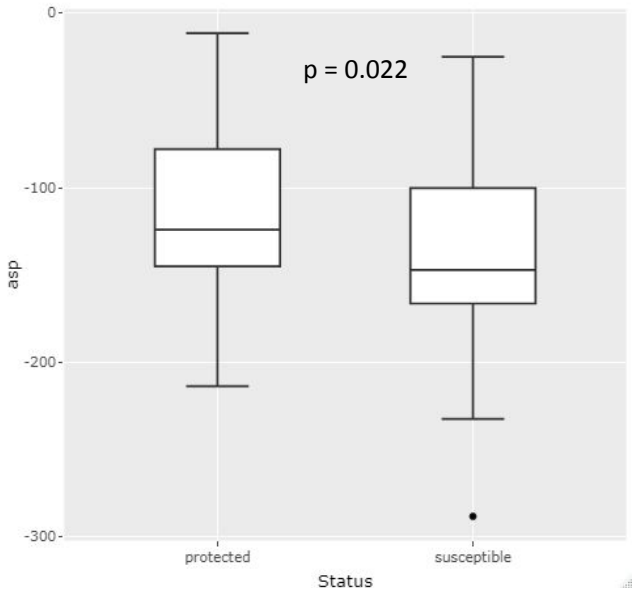
Normal distribution and flexible **mixture models** for seropositivity determination





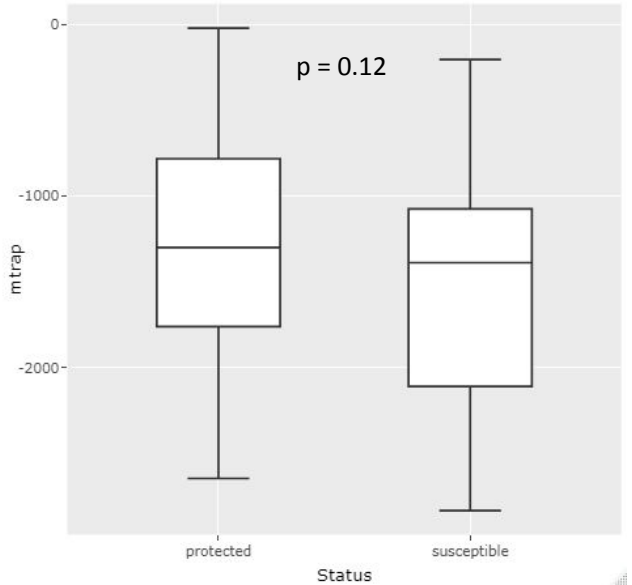


Solution



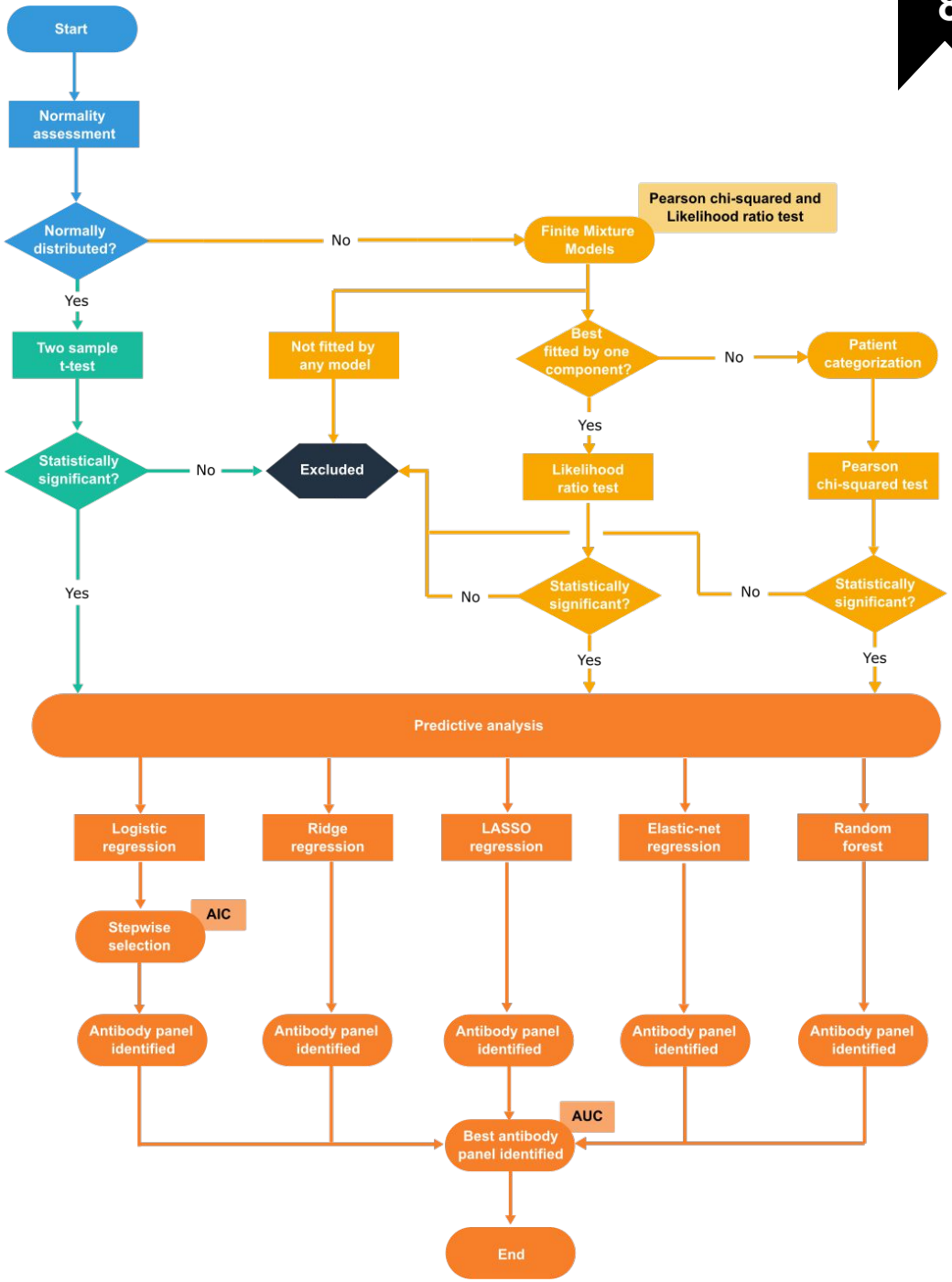
Predictive analysis

P-value < 0.05

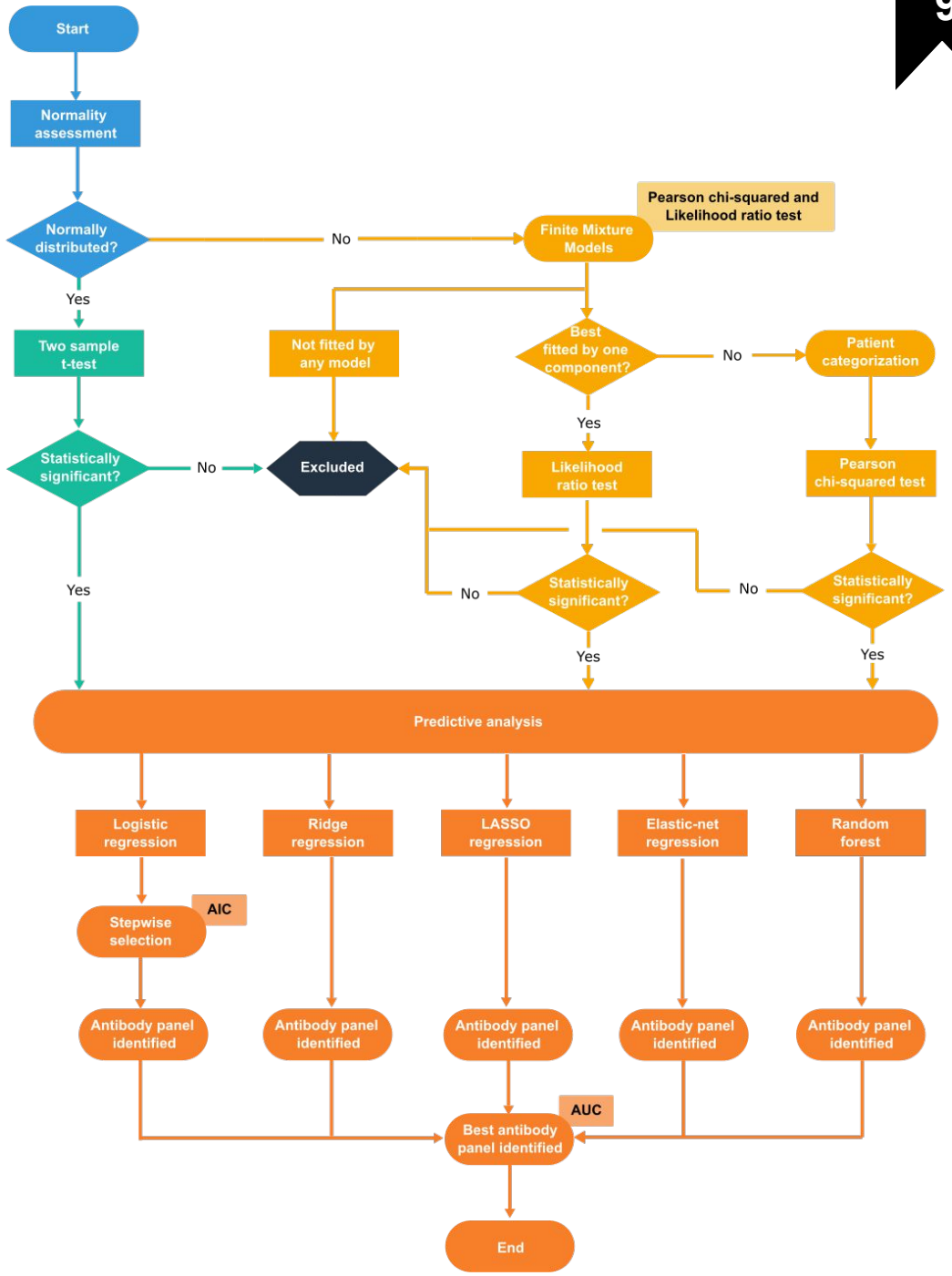


Excluded

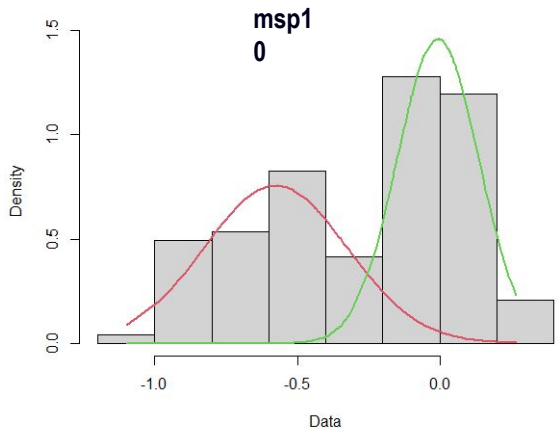
P-value > 0.05

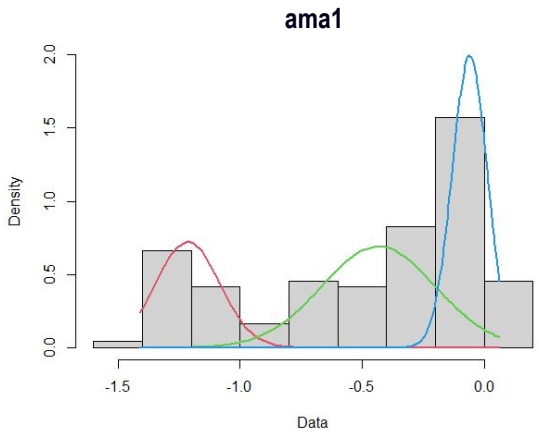


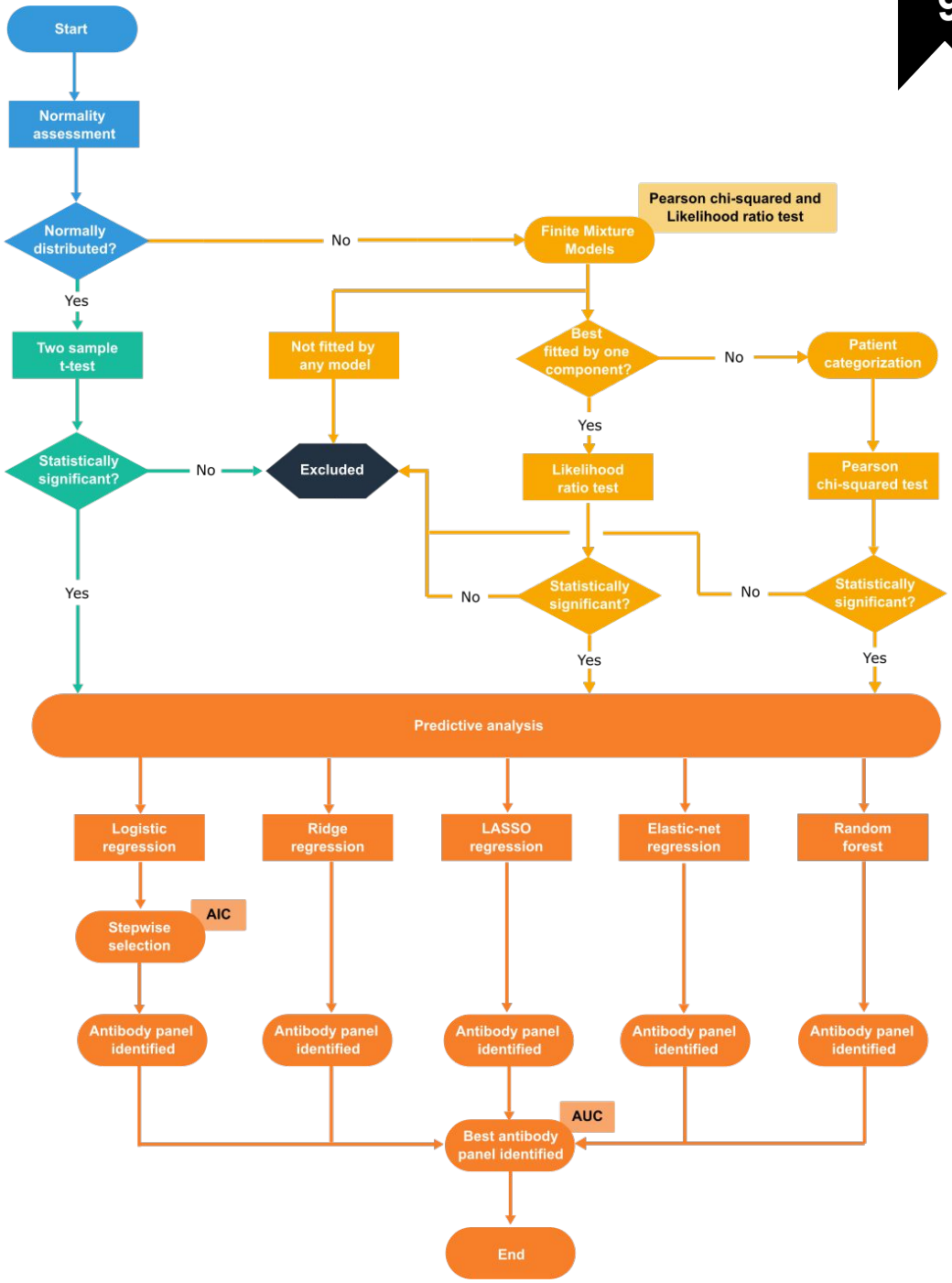
Solution


Start
Normality assessment
Normally distributed?
No
Yes
Pearson chi-squared and Likelihood ratio test
Finite Mixture Models
Two sample t-test
Not fitted by any model
Best fitted by one component?
No
Patient categorization
Yes
Statistically significant?
No
Excluded
Likelihood ratio test
Pearson chi-squared test
Yes
No
Statistically significant?
No
Statistically significant?
Yes
Yes
Predictive analysis
Logistic regression
Ridge regression
LASSO regression
Elastic-net regression
Random forest
Stepwise selection
AIC
Antibody panel identified
Antibody panel identified
Antibody panel identified
Antibody panel identified
Antibody panel identified
AUC
Best antibody panel identified
End

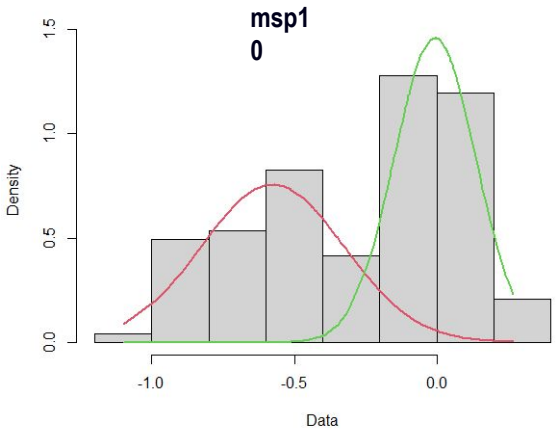
Solution

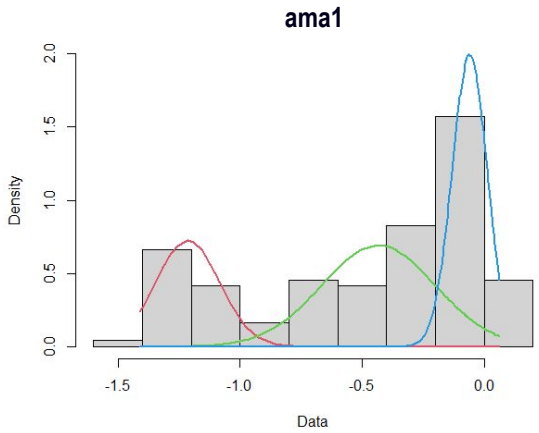

msp1
0
Density
Data
-1.0
-0.5
0.0
0.0
0.5
1.0
1.5


ama1
Density
Data
-1.5
-1.0
-0.5
0.0
0.0
0.5
1.0
1.5
2.0

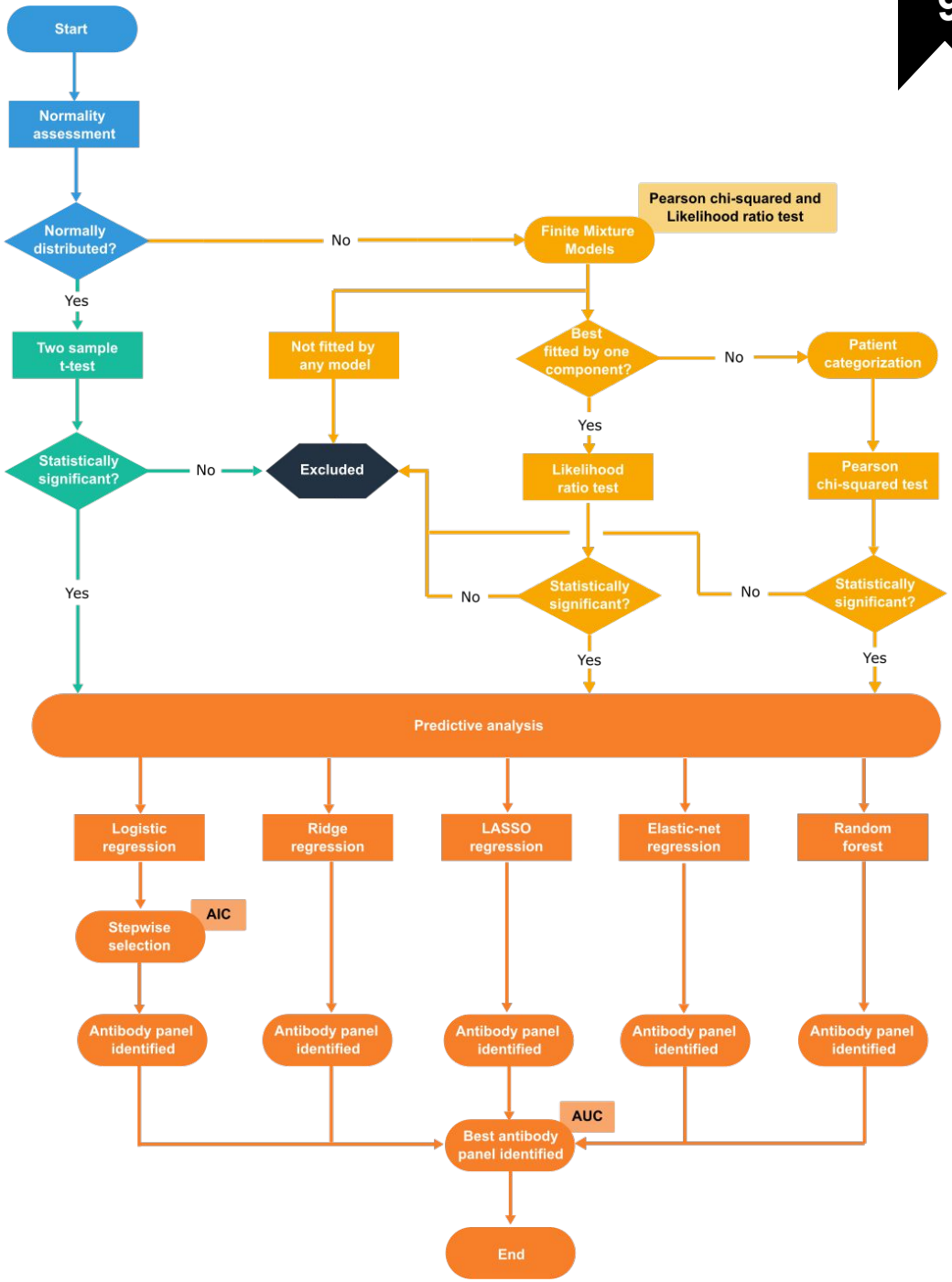

Start
Normality assessment
Normally distributed?
No
Pearson chi-squared and Likelihood ratio test
Finite Mixture Models
Yes
Two sample t-test
Not fitted by any model
Best fitted by one component?
No
Patient categorization
Yes
Statistically significant?
No
Excluded
Likelihood ratio test
Pearson chi-squared test
Yes
No
Statistically significant?
No
Statistically significant?
Yes
Yes
Predictive analysis
Logistic regression
Ridge regression
LASSO regression
Elastic-net regression
Random forest
Stepwise selection
AIC
Antibody panel identified
Antibody panel identified
Antibody panel identified
Antibody panel identified
Antibody panel identified
AUC
Best antibody panel identified
End

Solution

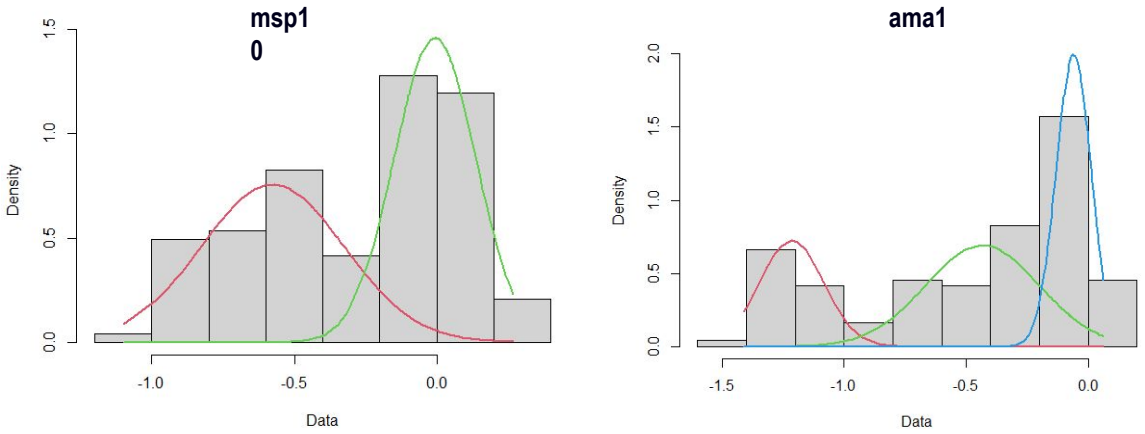




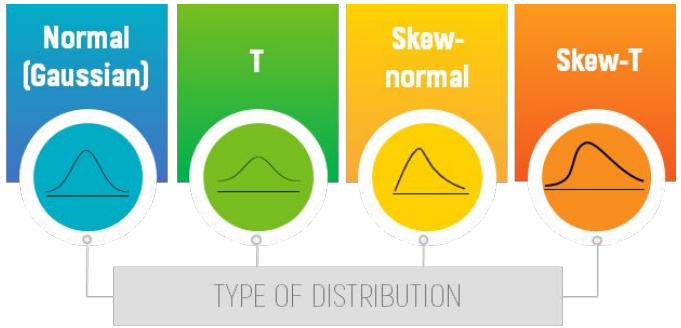
Mixture models are often used in seroepidemiologic studies

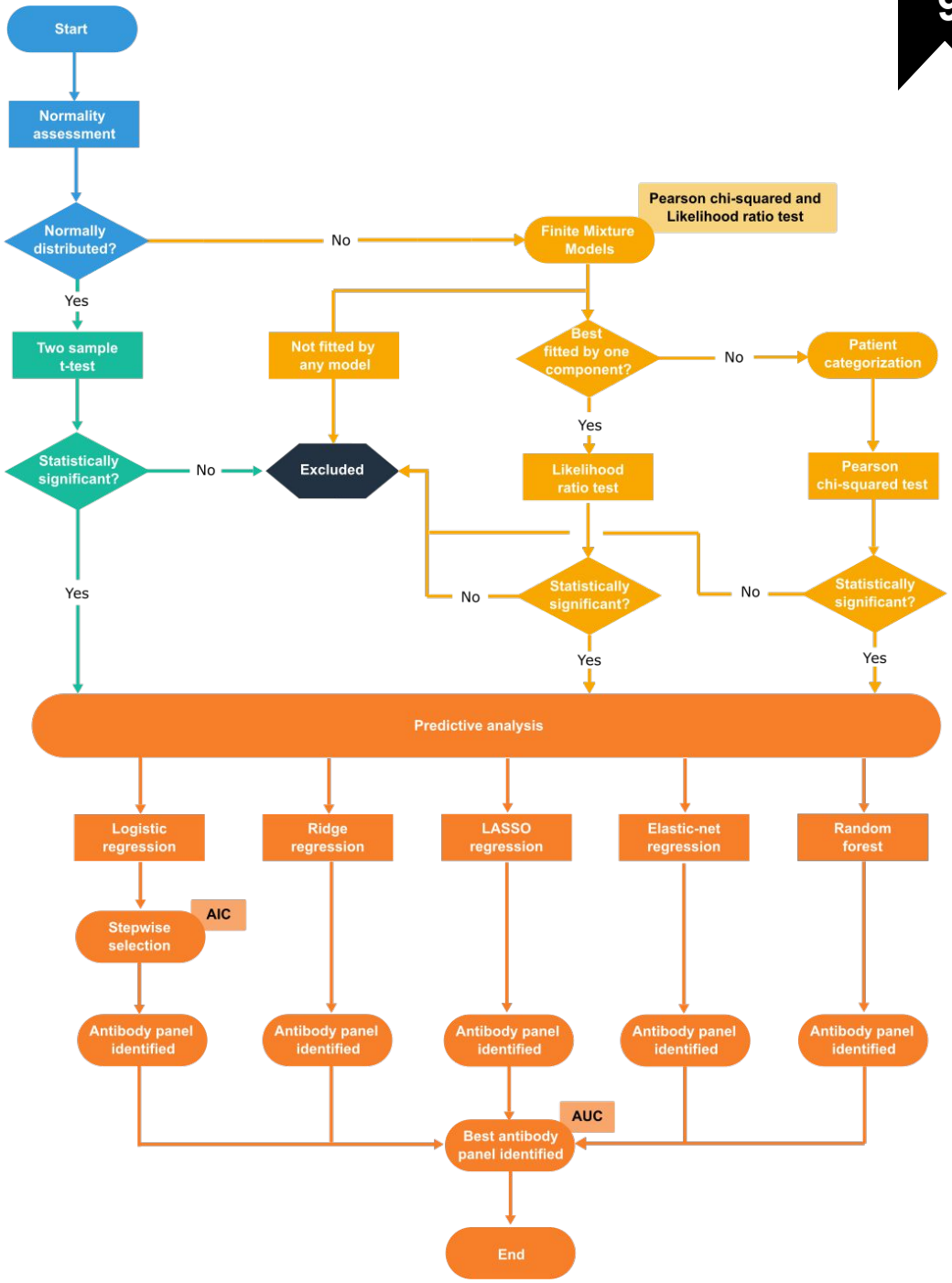


Solution

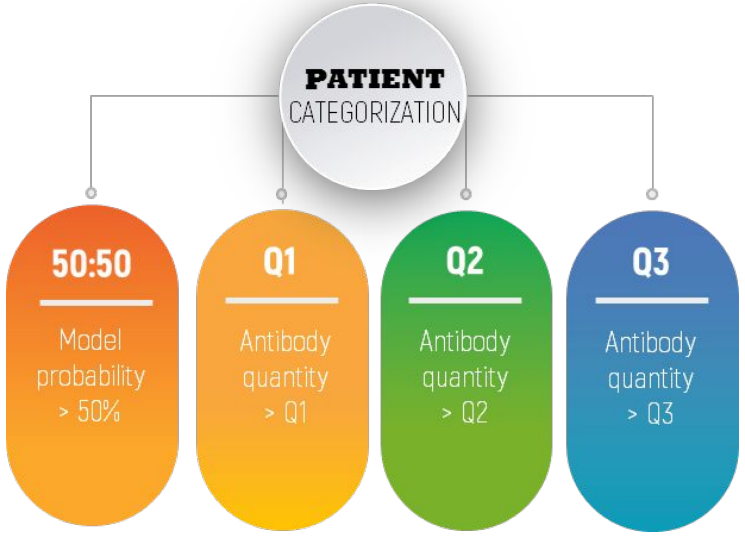


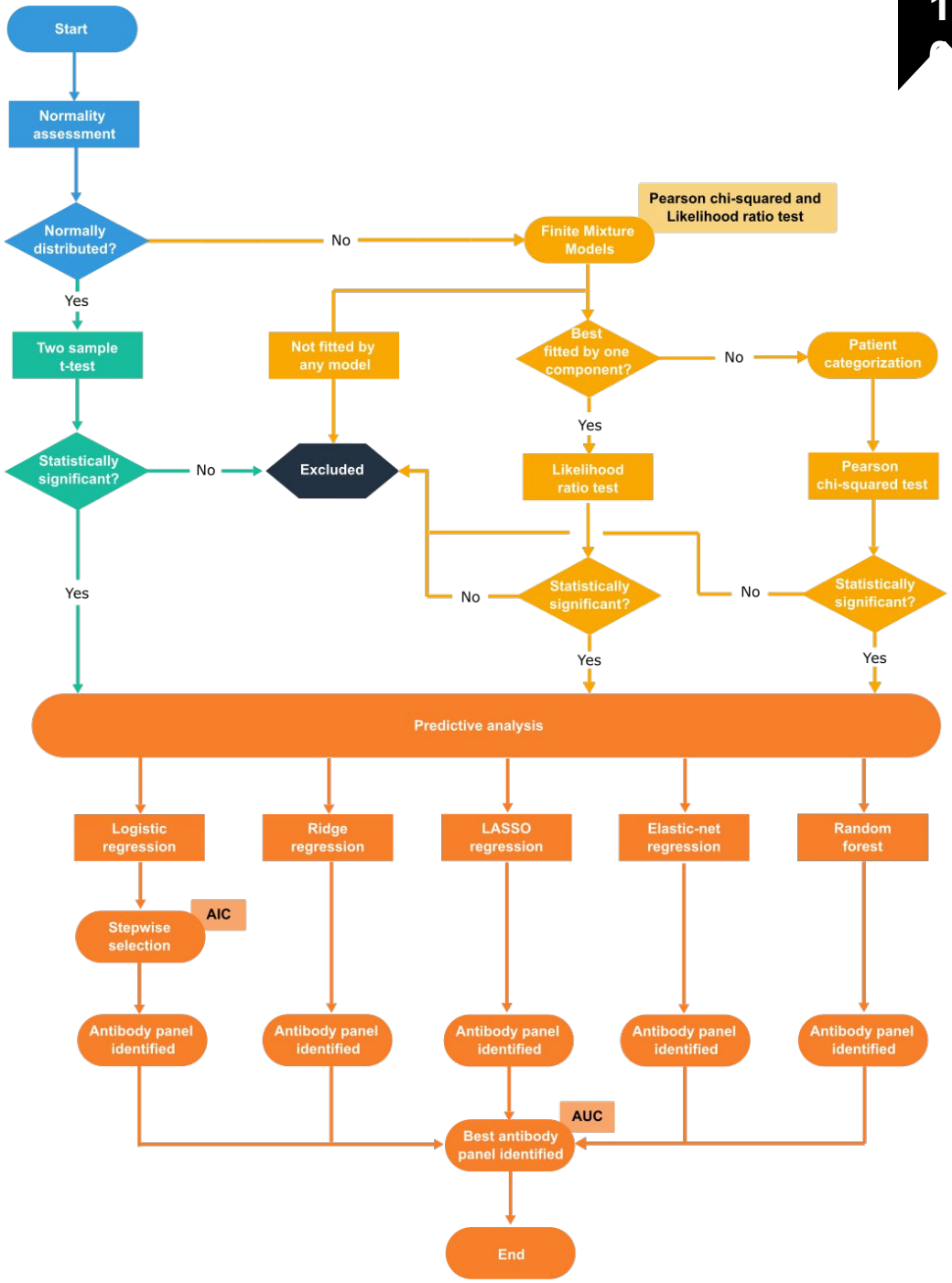
Mixture models are often used in seroepidemiologic studies



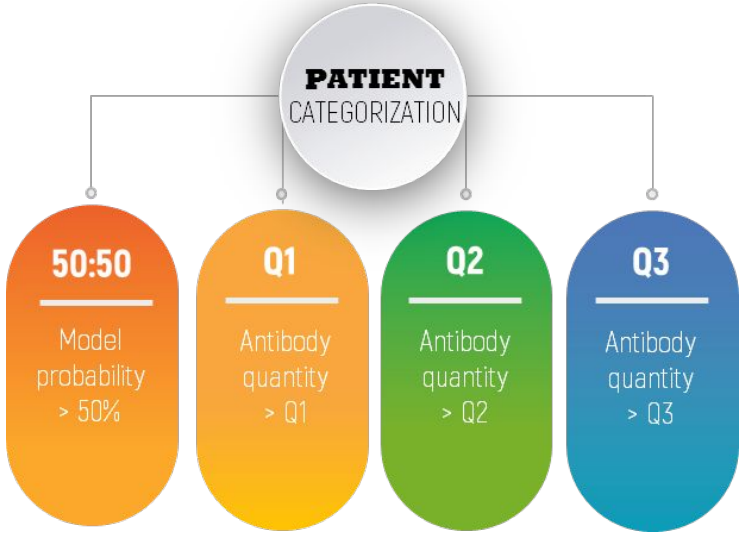


Solution


PATIENT
CATEGORIZATION
50:50
Model probability > 50%
Q1
Antibody quantity > Q1
Q2
Antibody quantity > Q2
Q3
Antibody quantity > Q3

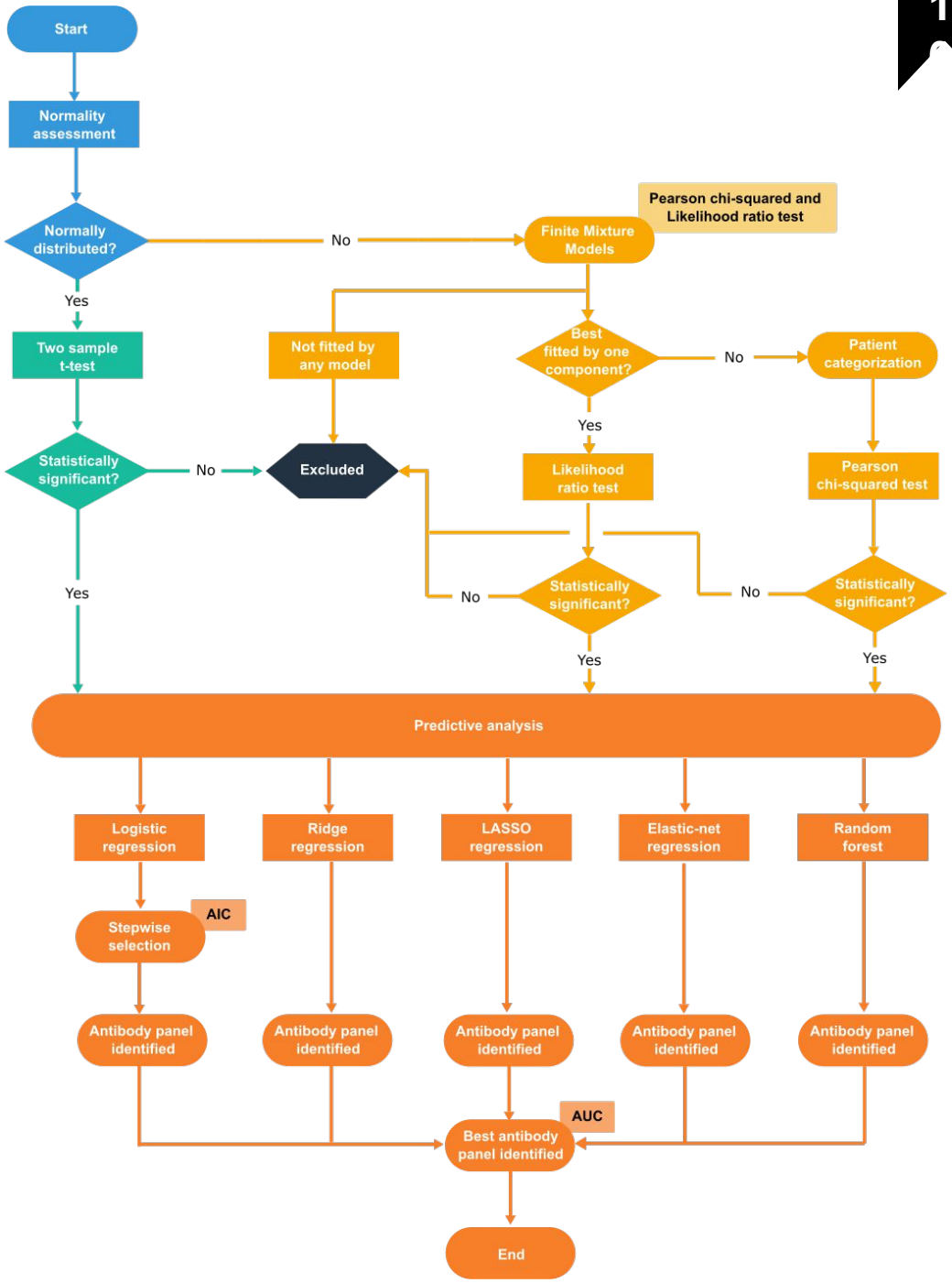

Start
Normality assessment
Normally distributed?
No
Pearson chi-squared and Likelihood ratio test
Finite Mixture Models
Yes
Two sample t-test
Not fitted by any model
Best fitted by one component?
No
Patient categorization
Yes
Statistically significant?
No
Excluded
Likelihood ratio test
Pearson chi-squared test
Yes
No
Statistically significant?
No
Statistically significant?
Yes
Yes
Predictive analysis
Logistic regression
Ridge regression
LASSO regression
Elastic-net regression
Random forest
Stepwise selection
AIC
Antibody panel identified
Antibody panel identified
Antibody panel identified
Antibody panel identified
Antibody panel identified
AUC
Best antibody panel identified
End

Solution

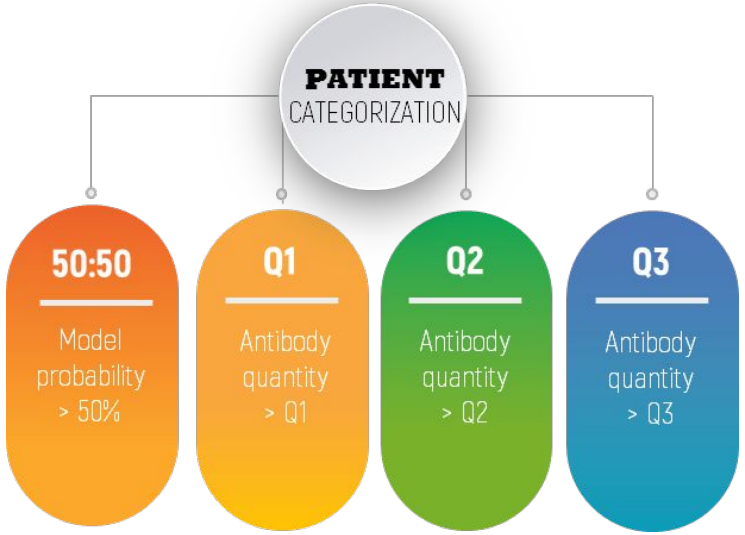


DATA DICHOTOMIZATION

SERONEGATIVE SEROPOSITIVE



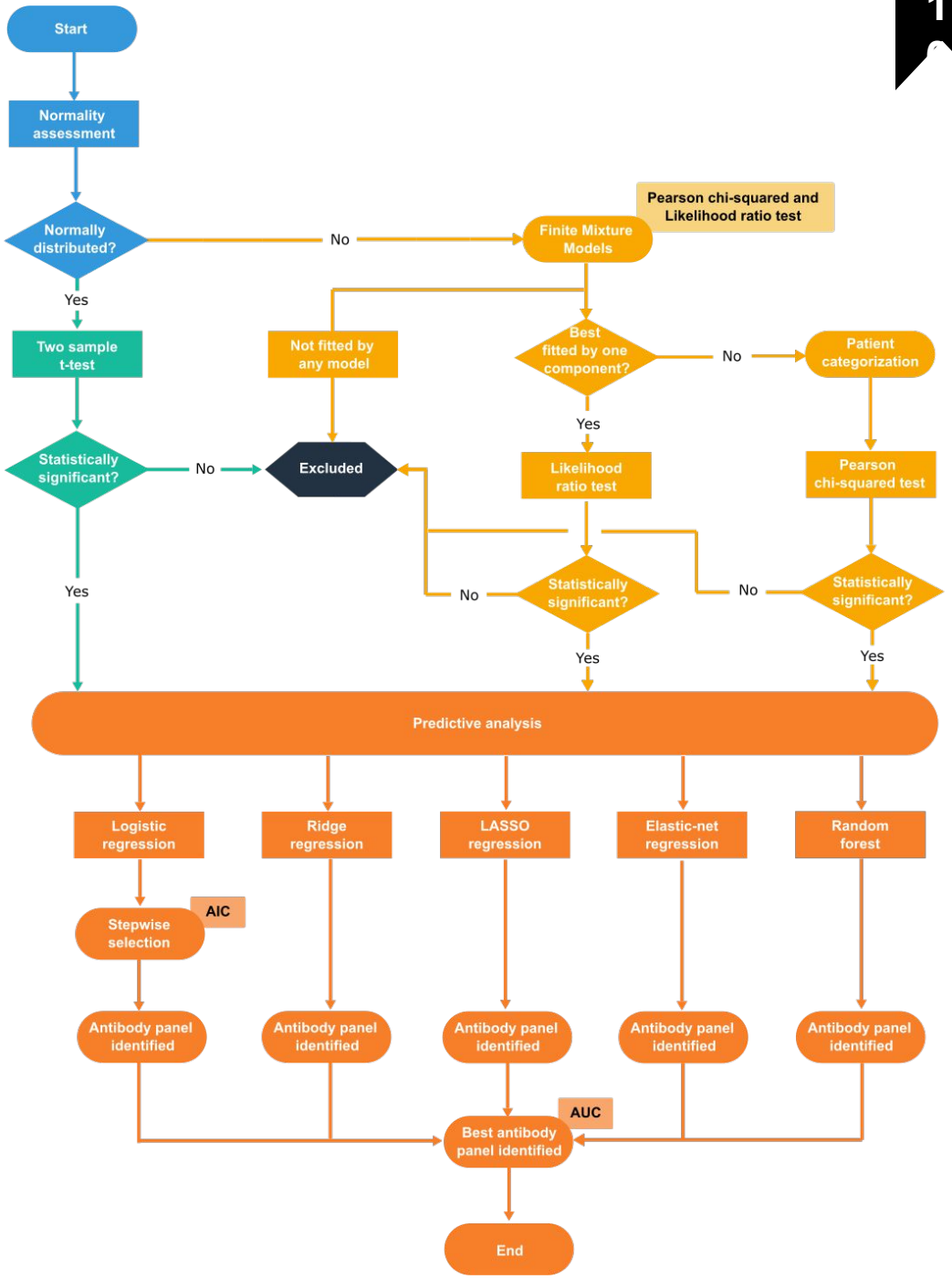
Solution



DATA DICHOTOMIZATION

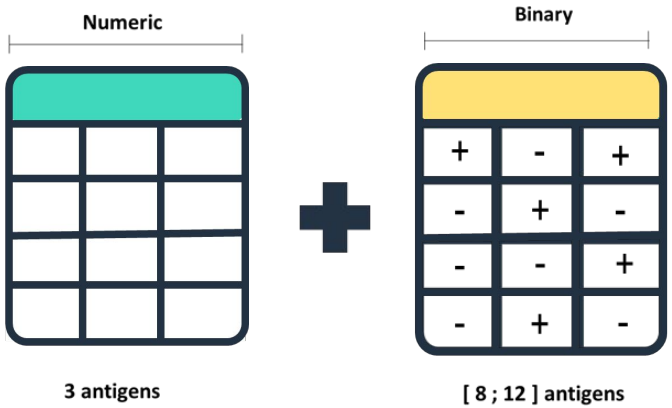
SERONEGATIVE SEROPOSITIVE

Chi-squared test of Seropositivity against Status

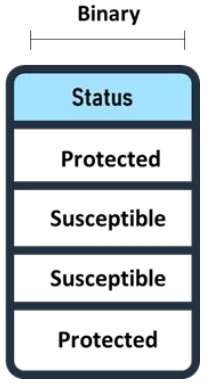


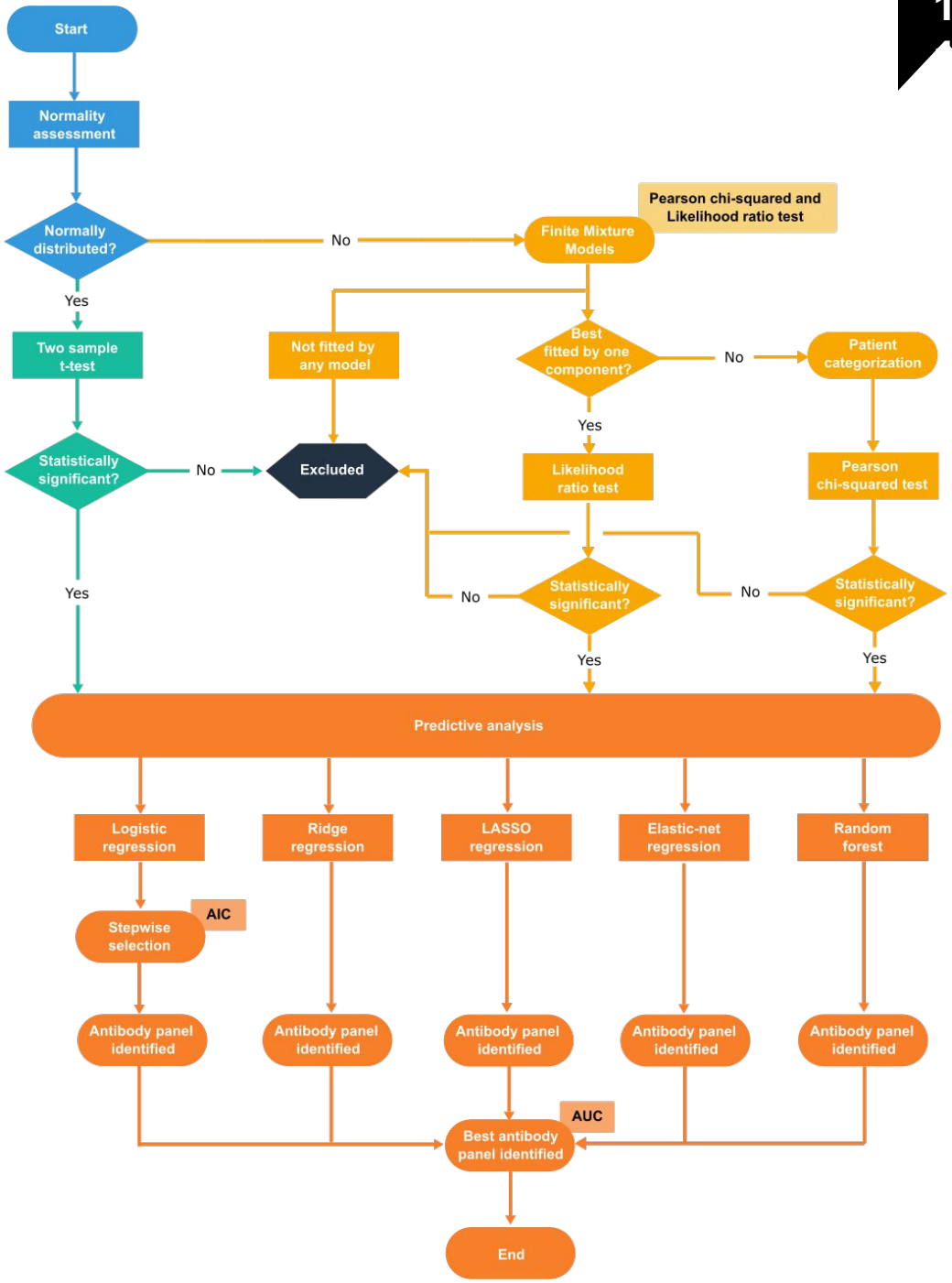
Solution

Predictor variables



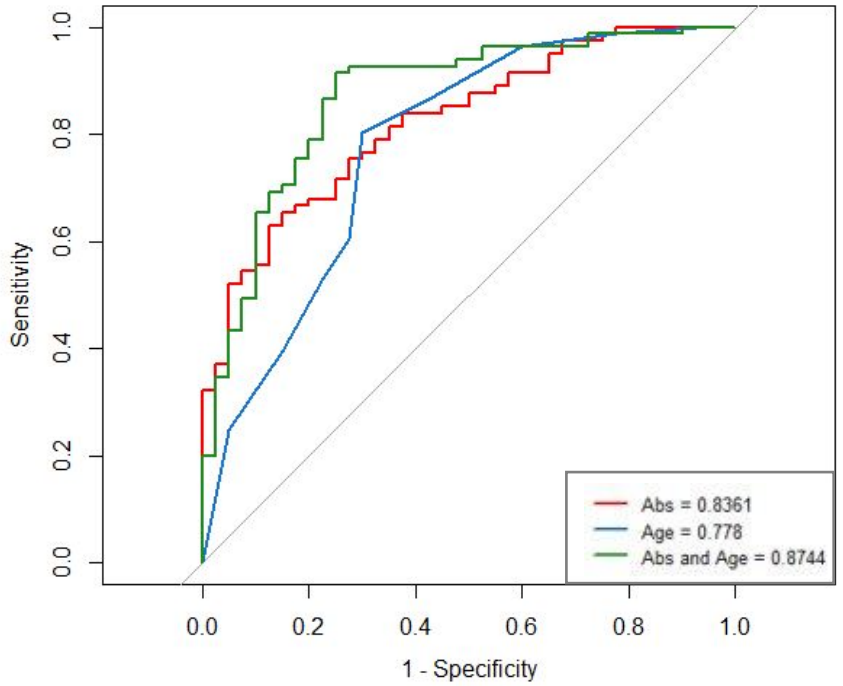
Response variable

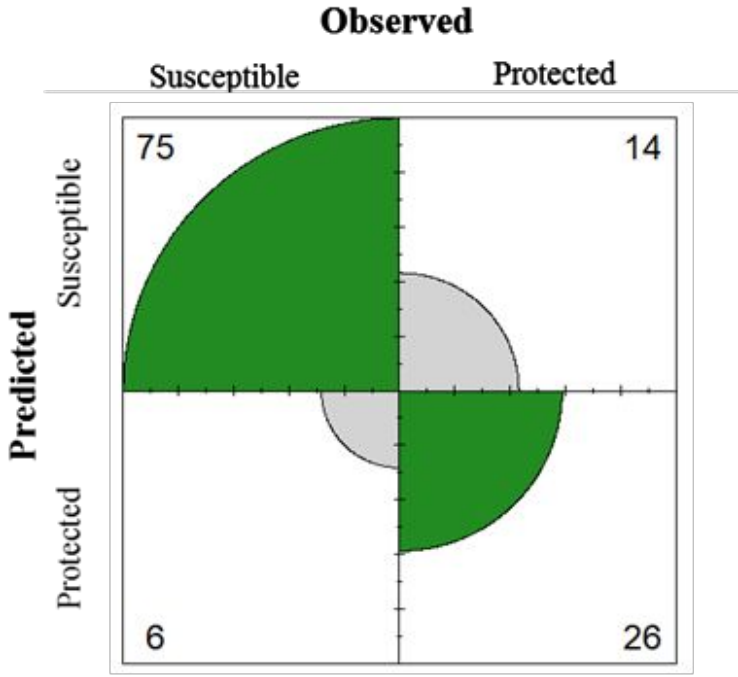




Results

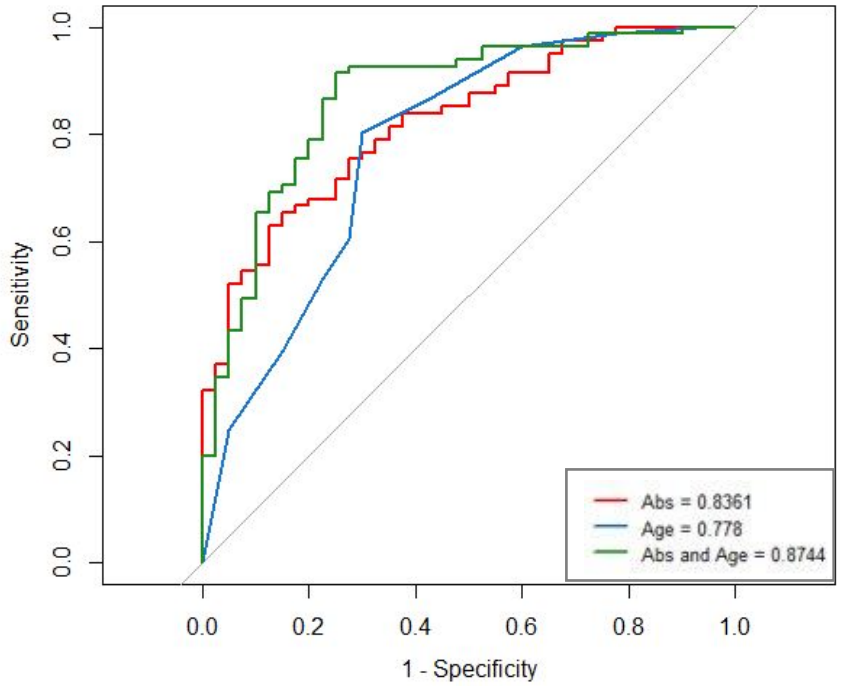
Stepwise Logistic regression

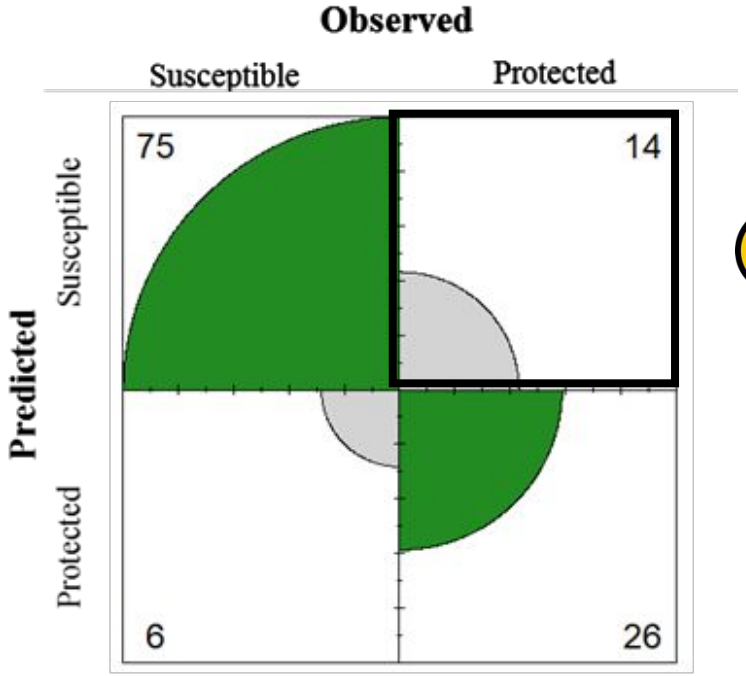




Results

Stepwise Logistic regression





Unable to correctly classify the protected individuals

Results

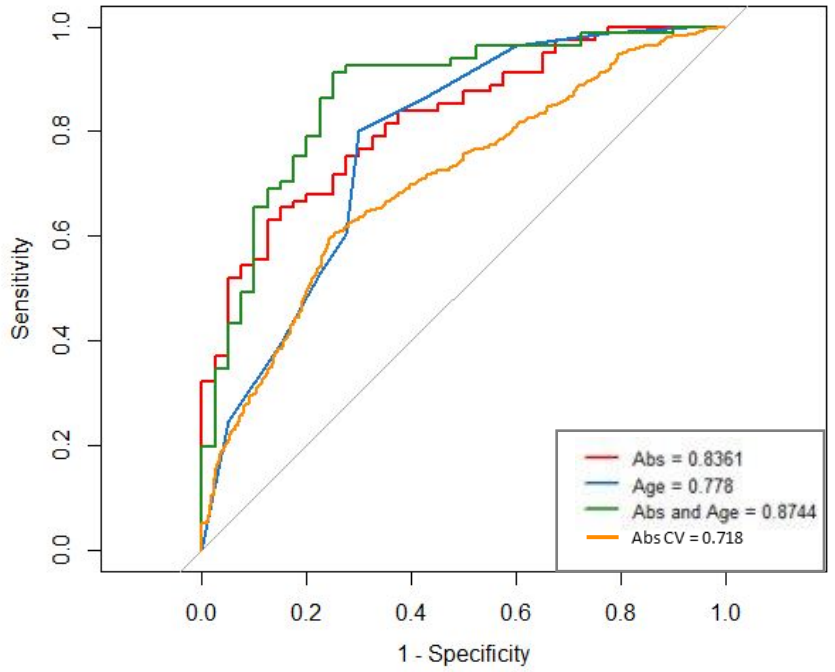
10 fold CV

Stepwise Logistic regression

Results

10 fold CV

Stepwise Logistic regression

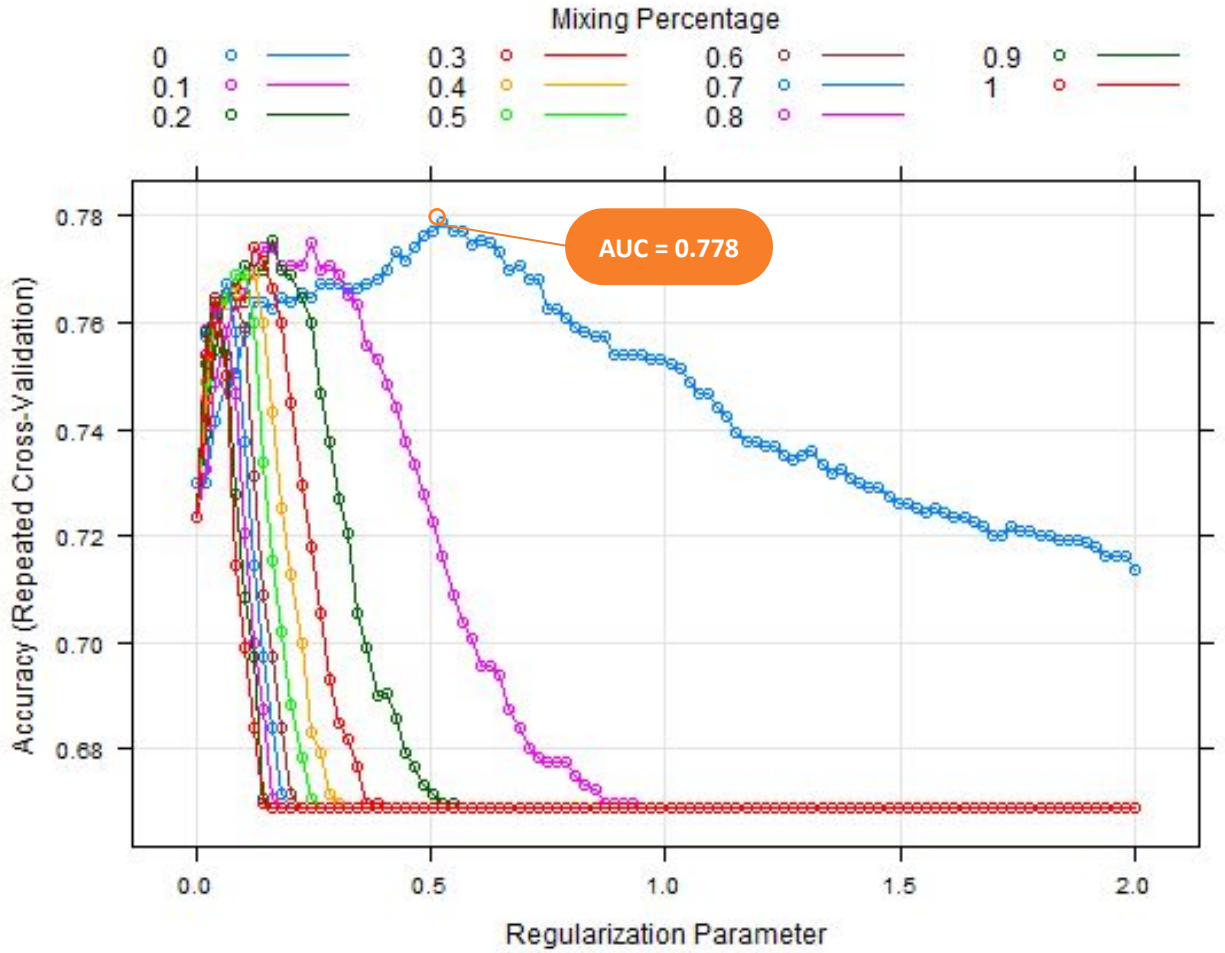


ow sample size (n=121)

Results

10 fold CV

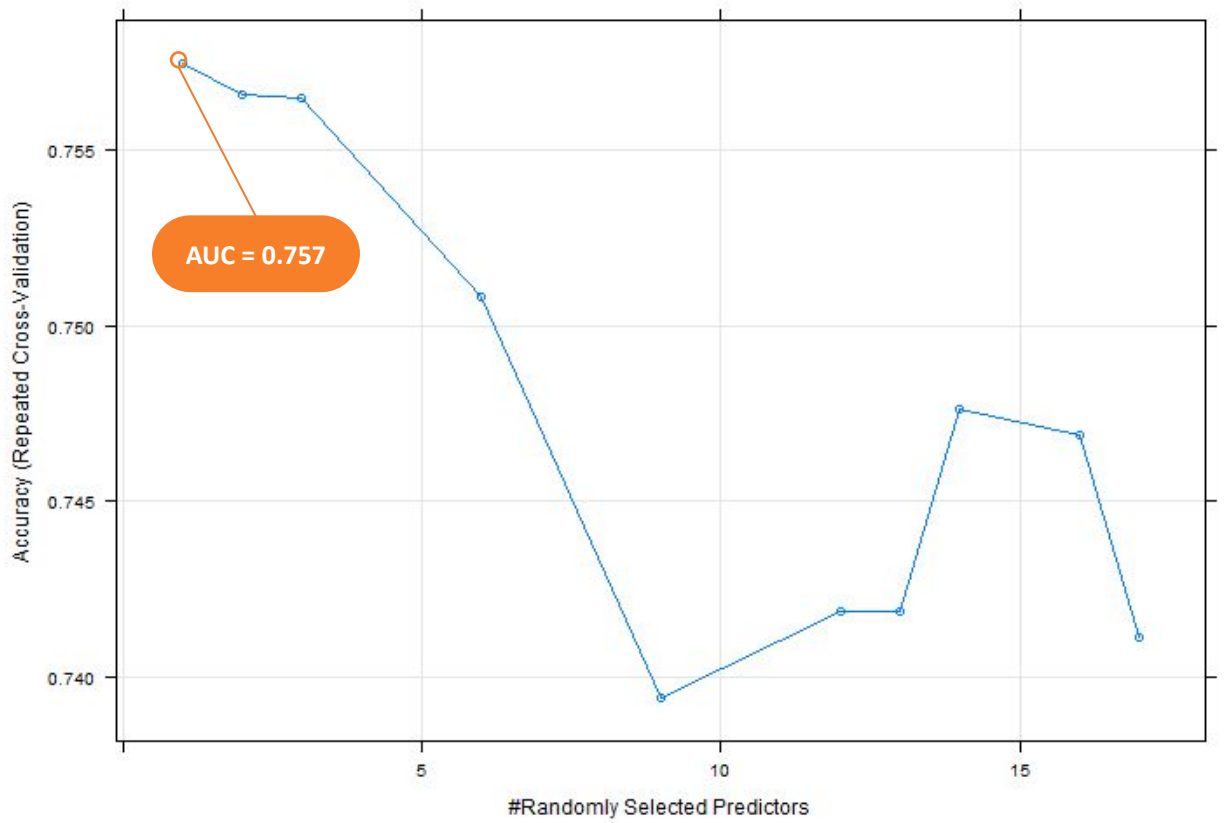
Elastic-net regression



Results

10 fold CV

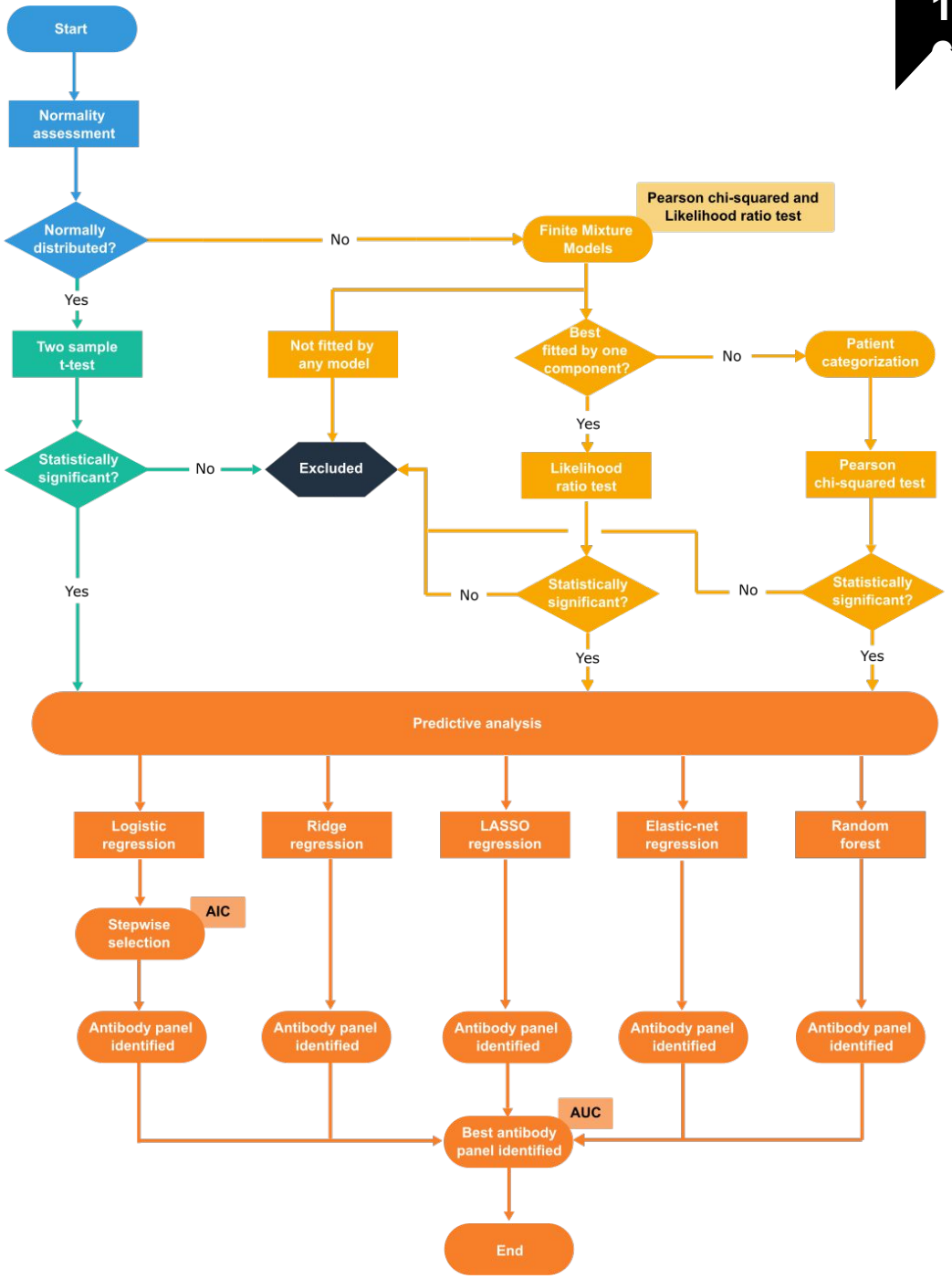
Random Forest



Solution

Parametric pipeline

Normal distribution and flexible **mixture models** for seropositivity determination



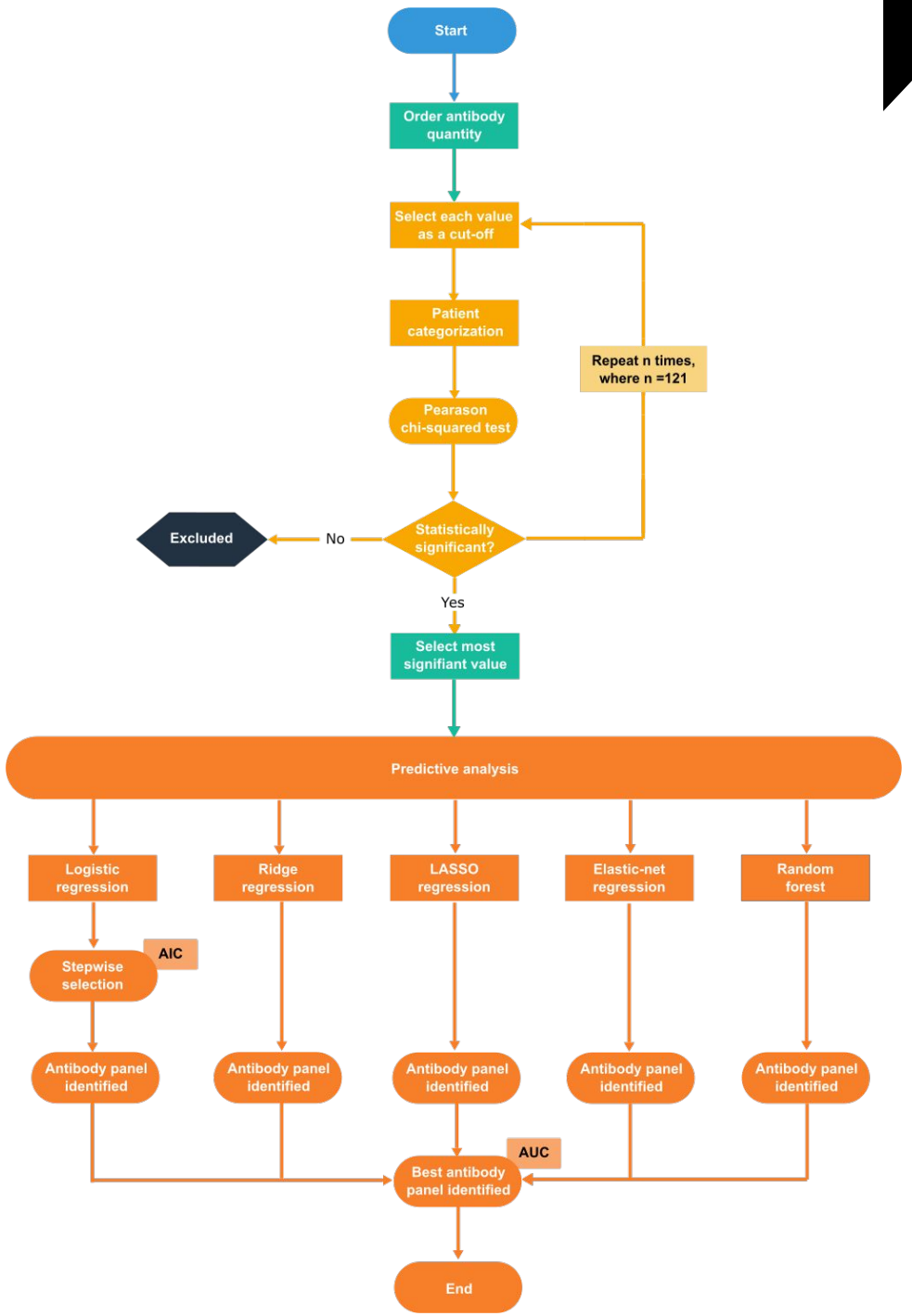
Solution

Parametric pipeline

Normal distribution and flexible **mixture models** for seropositivity determination

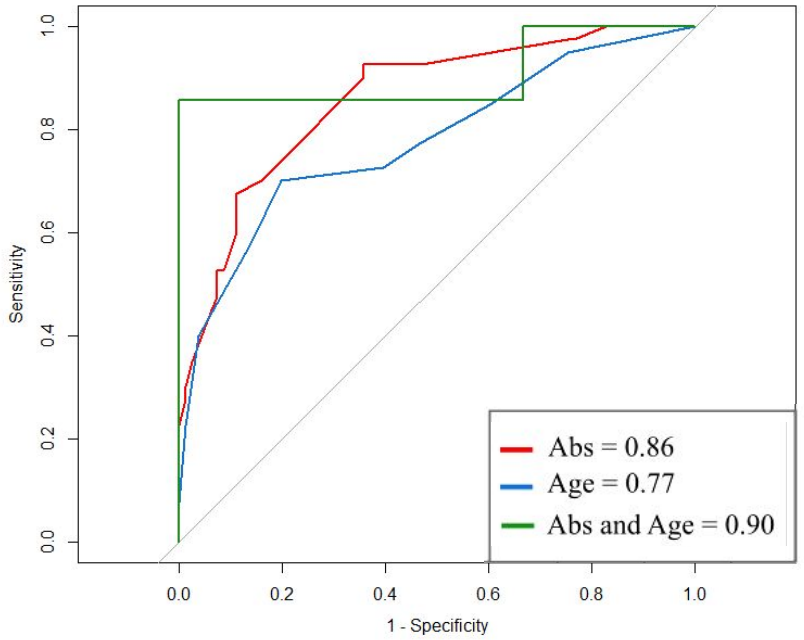
Pragmatic pipeline

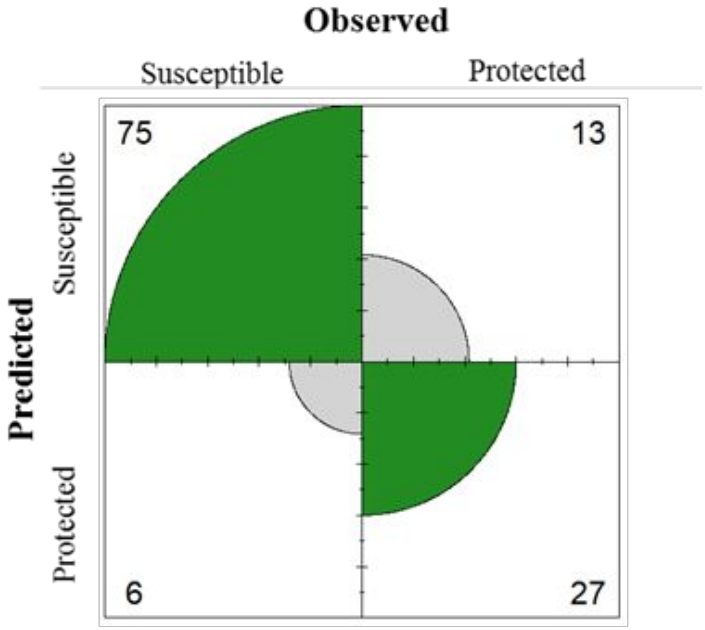
Cut-off that maximized the distinction between susceptible and protected individuals



Results

Stepwise Logistic regression

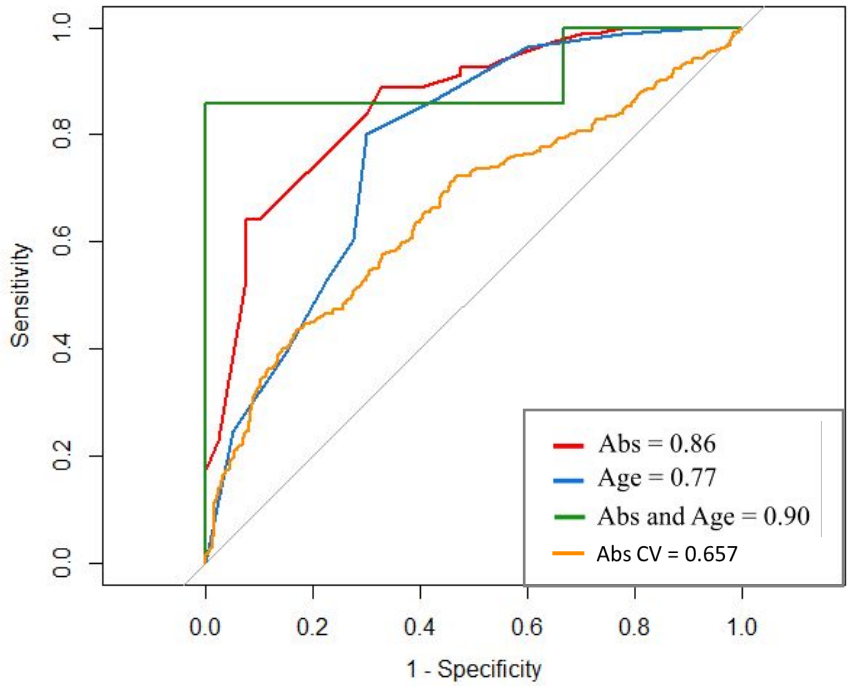




Results

10 fold CV

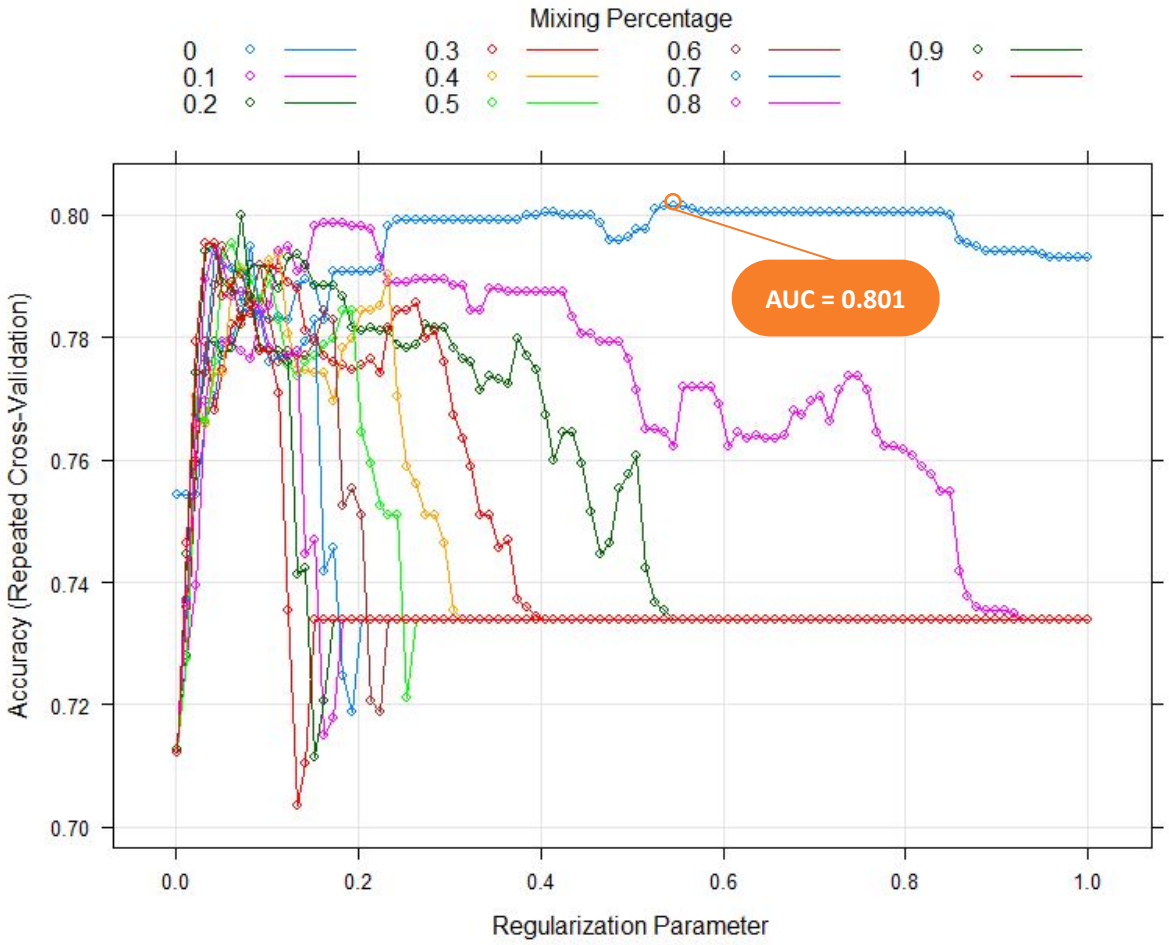
Stepwise Logistic regression



Results

10 fold CV

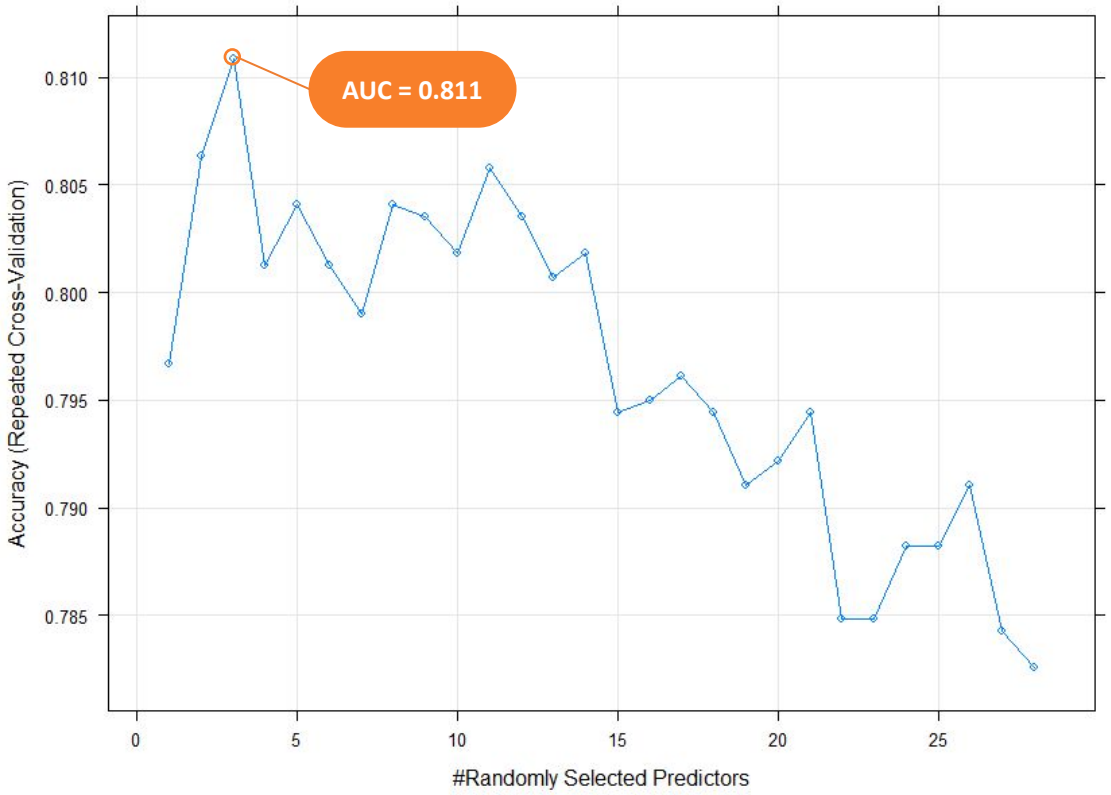
Elastic-net regression



Results

10 fold CV

Random Forest



Setbacks

- Scale to higher dimension dataset

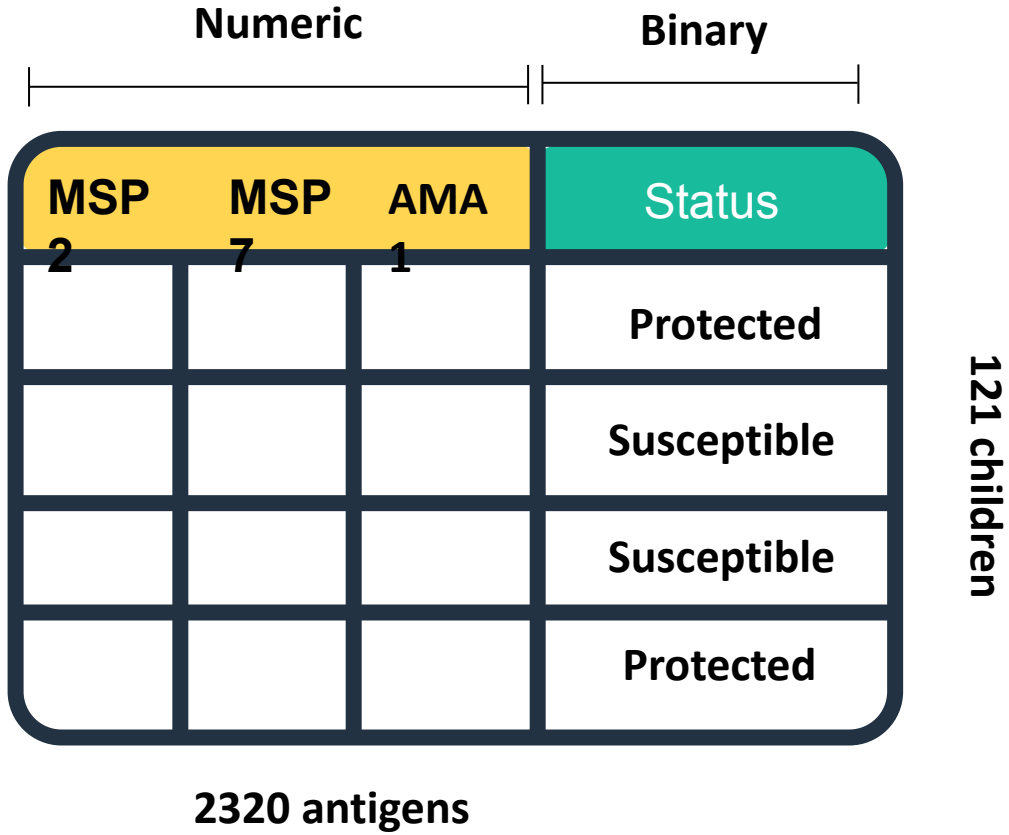
Setbacks

Scale to higher dimension dataset

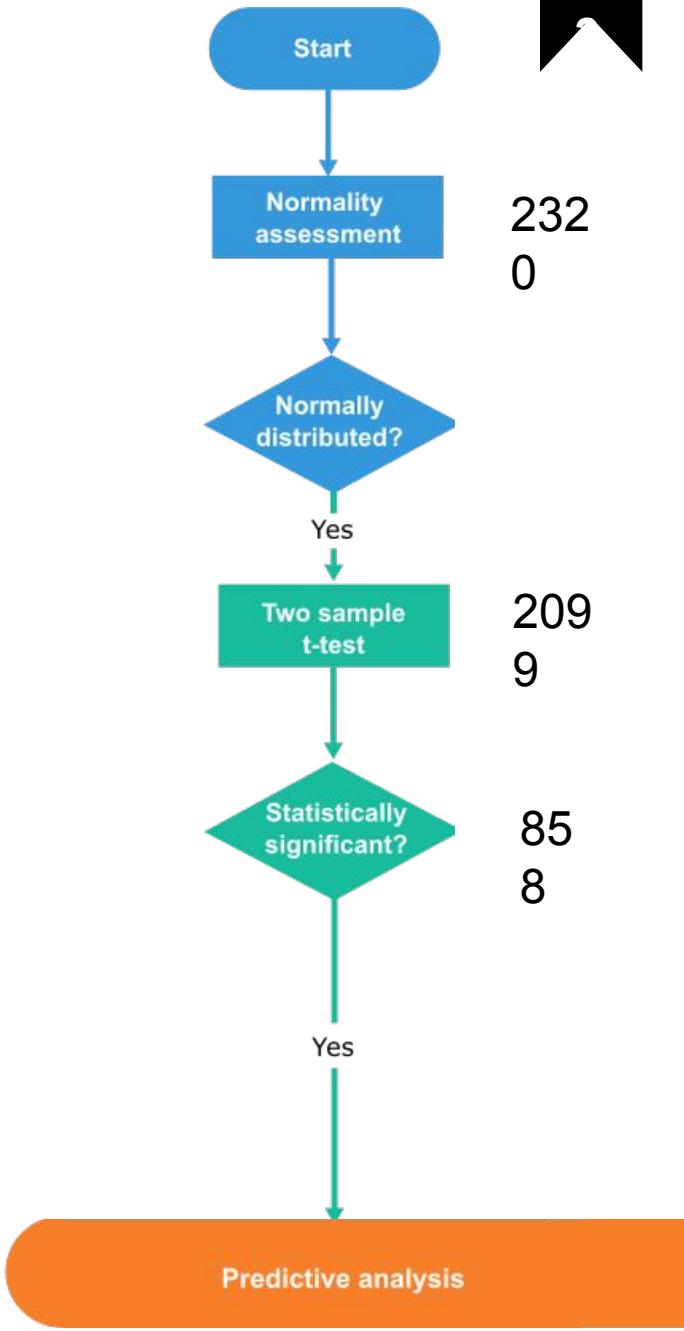
IgG antibodies against **2320 Plasmodium Falciparum antigens**

Antibody quantity information for **186 Malian individuals** (aged 2-25 years)

The **response variable** "Status" was binary (Protected vs. Susceptible)



Setbacks


Start
Normality assessment
2320
Normally distributed?
Yes
Two sample t-test
2099
Statistically significant?
858
Yes
Predictive analysis

Solution

209

Two sample independent t-test

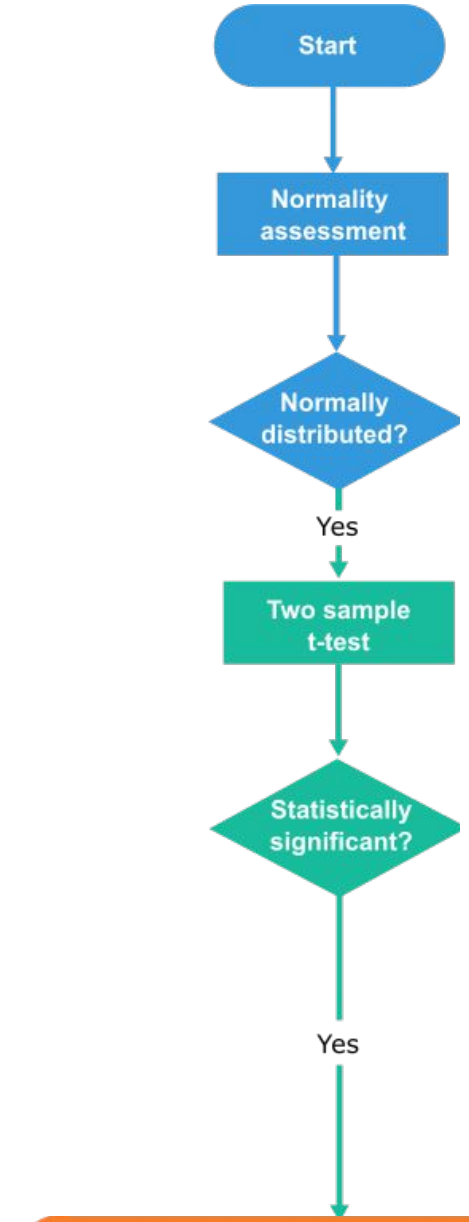
Multiple testing corrections

Bonferroni

$$\text{p-value} < \frac{\alpha}{m}$$

Dunn-Sidak

$$\text{p-value} < 1 - (1 - x)^{\frac{1}{m}}$$



232
0

209
9

85
8

Solution

209

Two sample independent t-test

Multiple testing corrections

Bonferroni

$$\text{p-value} < \frac{\alpha}{m}$$

Dunn-Sidak

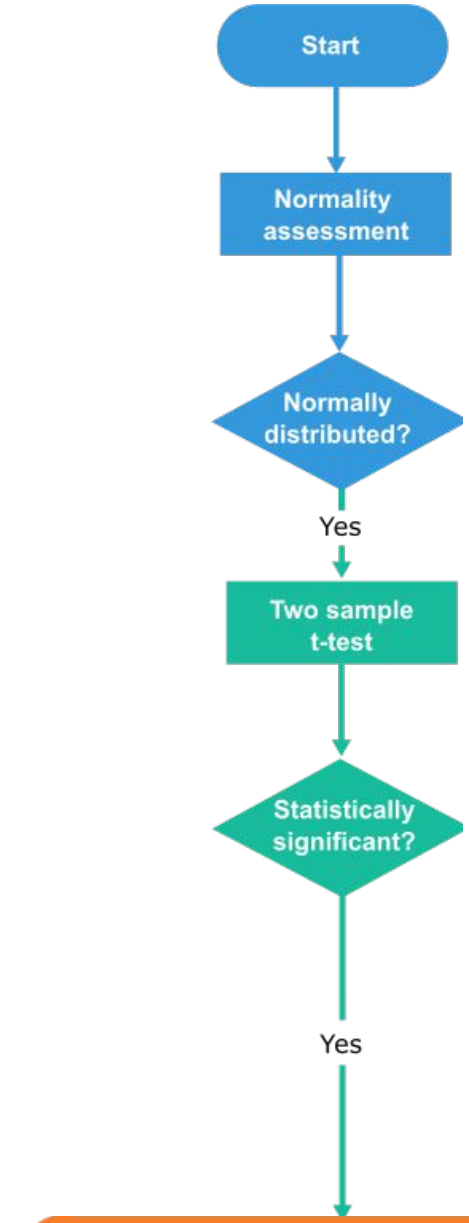
$$\text{p-value} < 1 - (1 - x)^{\frac{1}{m}}$$

m = number of

α = significance level

m = 2099

$\alpha = 0.05$



232
0

209
9

85
8

Solution

209

Two sample independent t-test

Multiple testing corrections

Bonferroni

$$\text{p-value} < \frac{\alpha}{m}$$

Dunn-Sidak

$$\text{p-value} < 1 - (1 - x)^{\frac{1}{m}}$$

m = number of

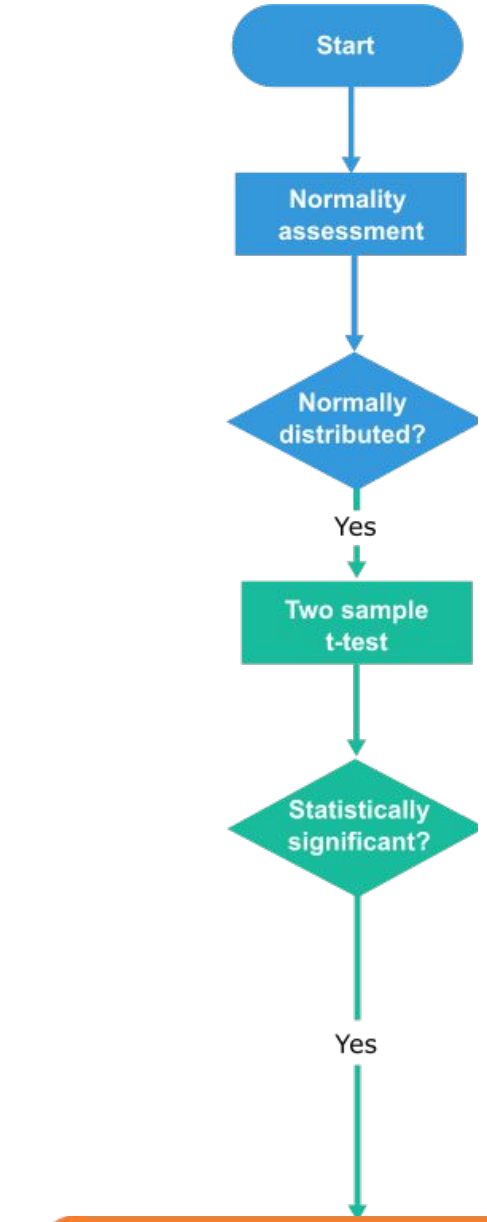
α = significance level

m = 2099

$\alpha = 0.05$

$$\text{p-value} < \frac{0.05}{2099} \quad (=)$$

$$\text{p-value} < 2.38 * 10^{-5}$$



232 0

209 9

85 8

Solution

209

Two sample independent t-test

Multiple testing corrections

Bonferroni

$$\text{p-value} < \frac{\alpha}{m}$$

Dunn-Sidak

$$\text{p-value} < 1 - (1 - x)^{\frac{1}{m}}$$

m = number of

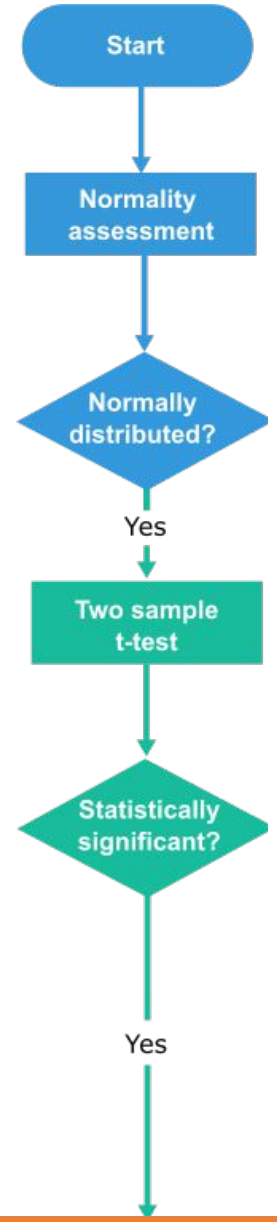
α = significance level

m = 2099

$\alpha = 0.05$

$$\text{p-value} < \frac{0.05}{2099} \left(= \right) \text{p-value} < 2.38 * 10^{-5}$$

$$\text{p-value} < 2.44 * 10^{-5} \left(= \right) \text{p-value} < 1 - (1 - 0.05)^{\frac{1}{2099}}$$



232
0

209
9

85
8

Solution

209

Two sample independent t-test

Multiple testing corrections

Bonferroni

$$p\text{-value} < \frac{\alpha}{m}$$

Dunn-Sidak

$$p\text{-value} < 1 - (1 - x)^{\frac{1}{m}}$$

m = number of α = significance level

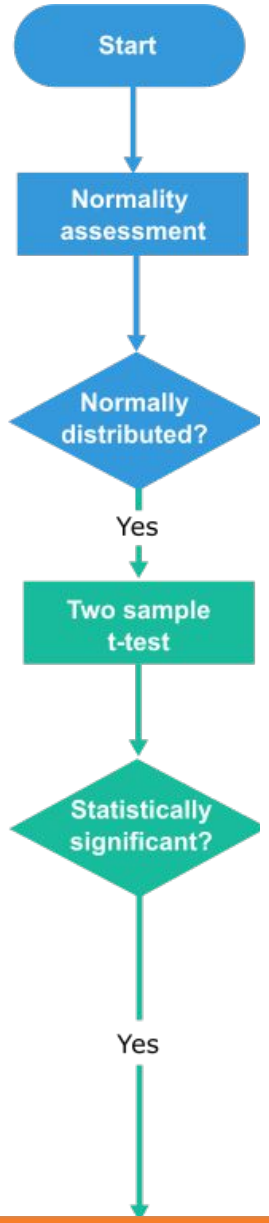
m = 2099

$\alpha = 0.05$

22 antibodies

85 8

22 antibodies



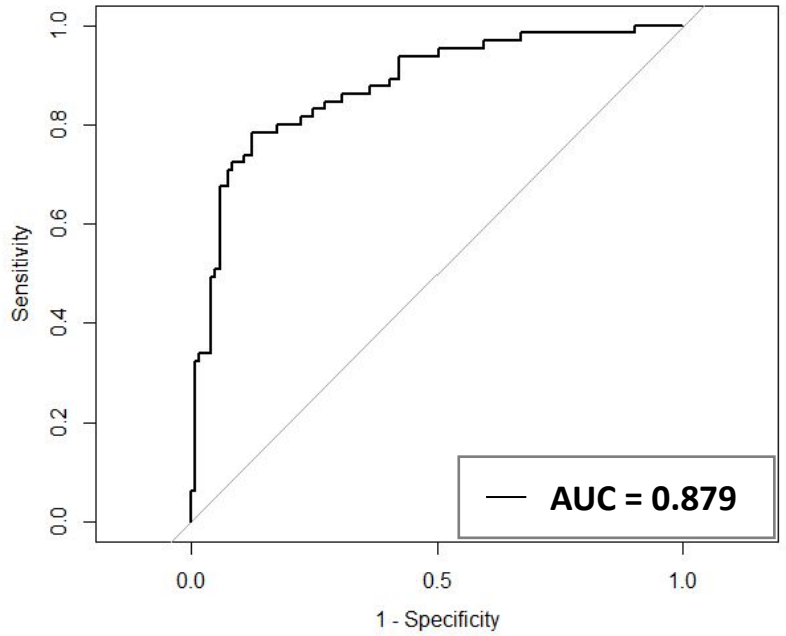
232 0

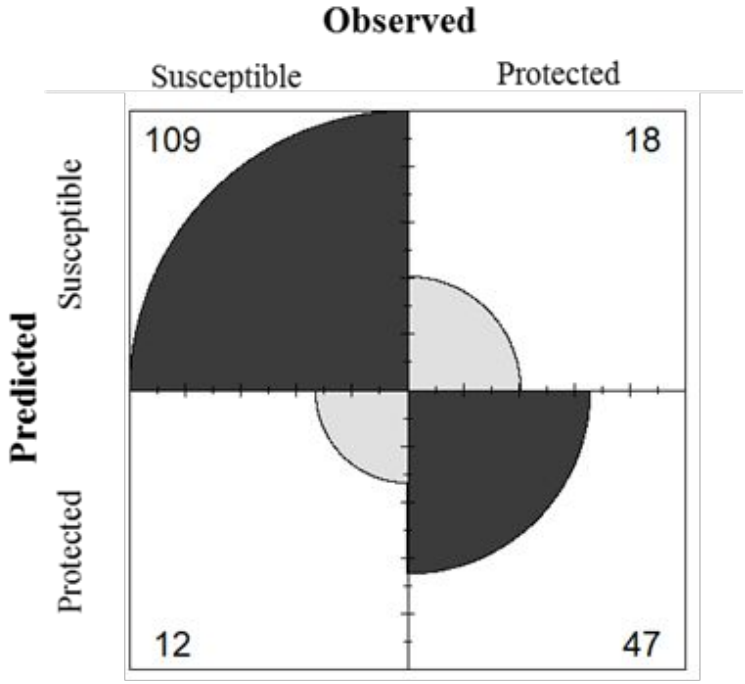
209 9

85 8

Results

Stepwise Logistic regression

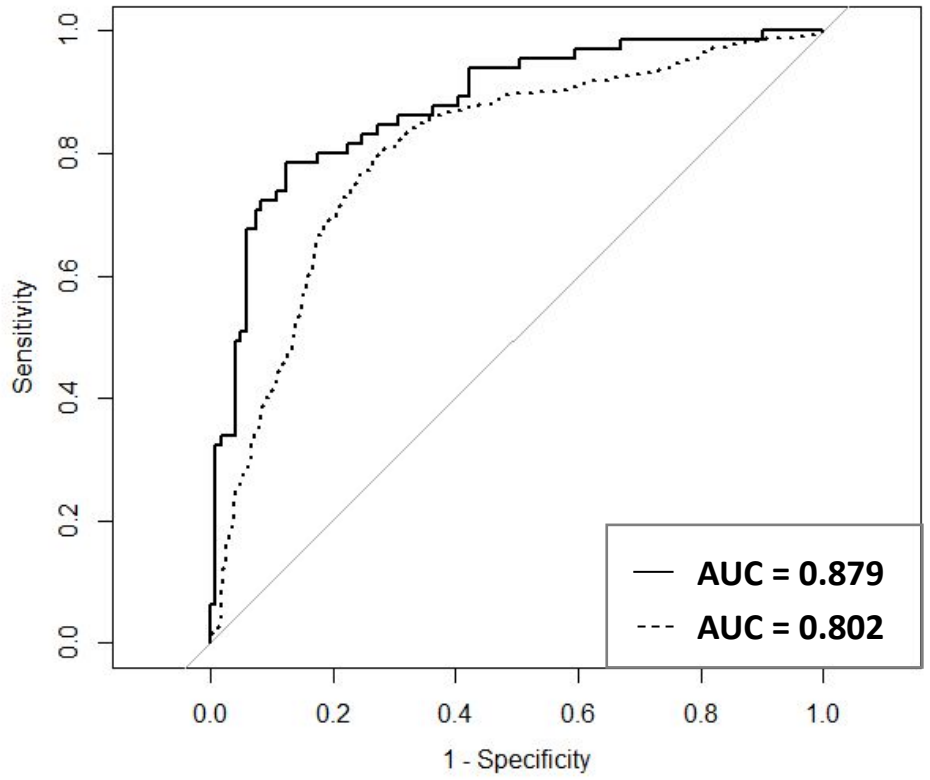




Results

10 fold CV

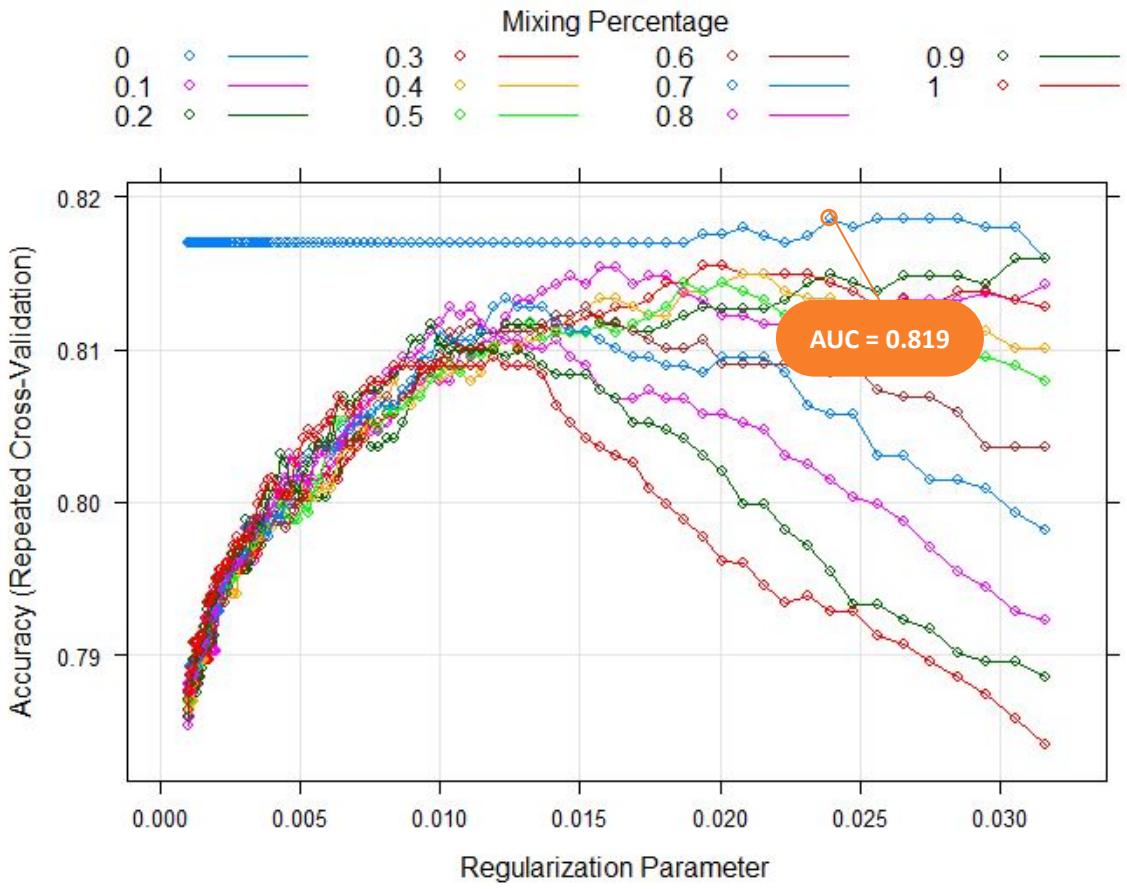
Stepwise Logistic regression



Results

10 fold CV

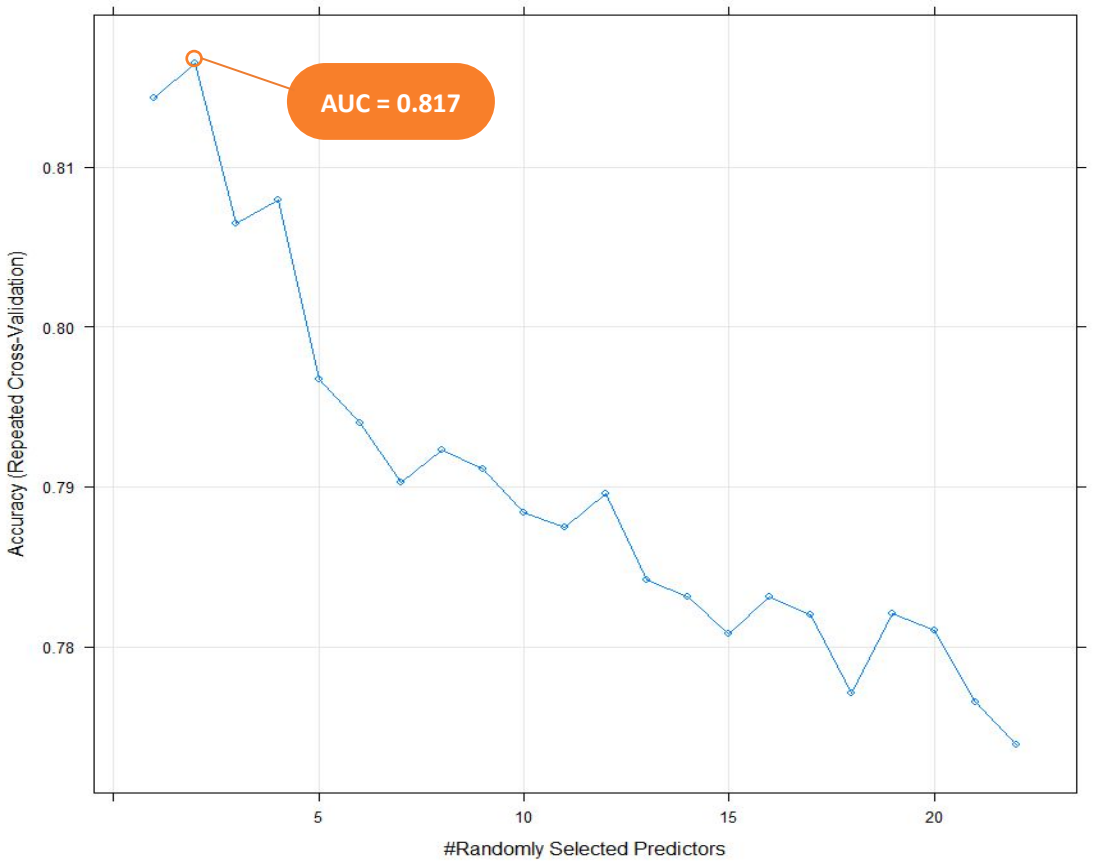
Elastic-net regression



Results

10 fold CV

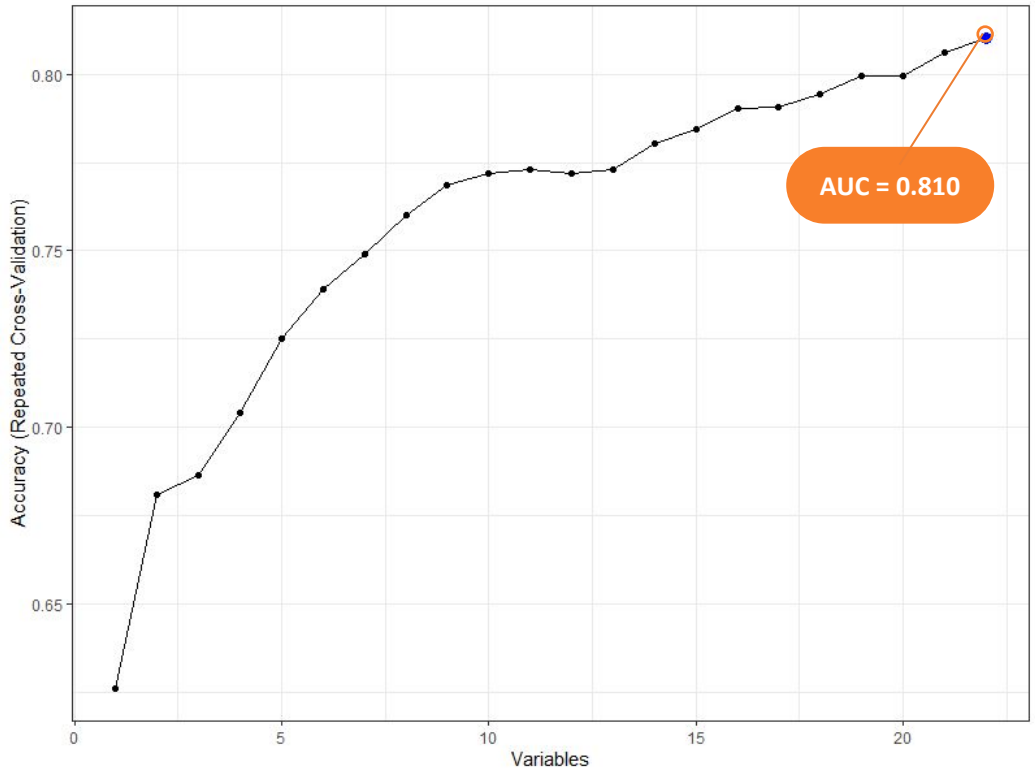
Random Forest



Results

10 fold CV

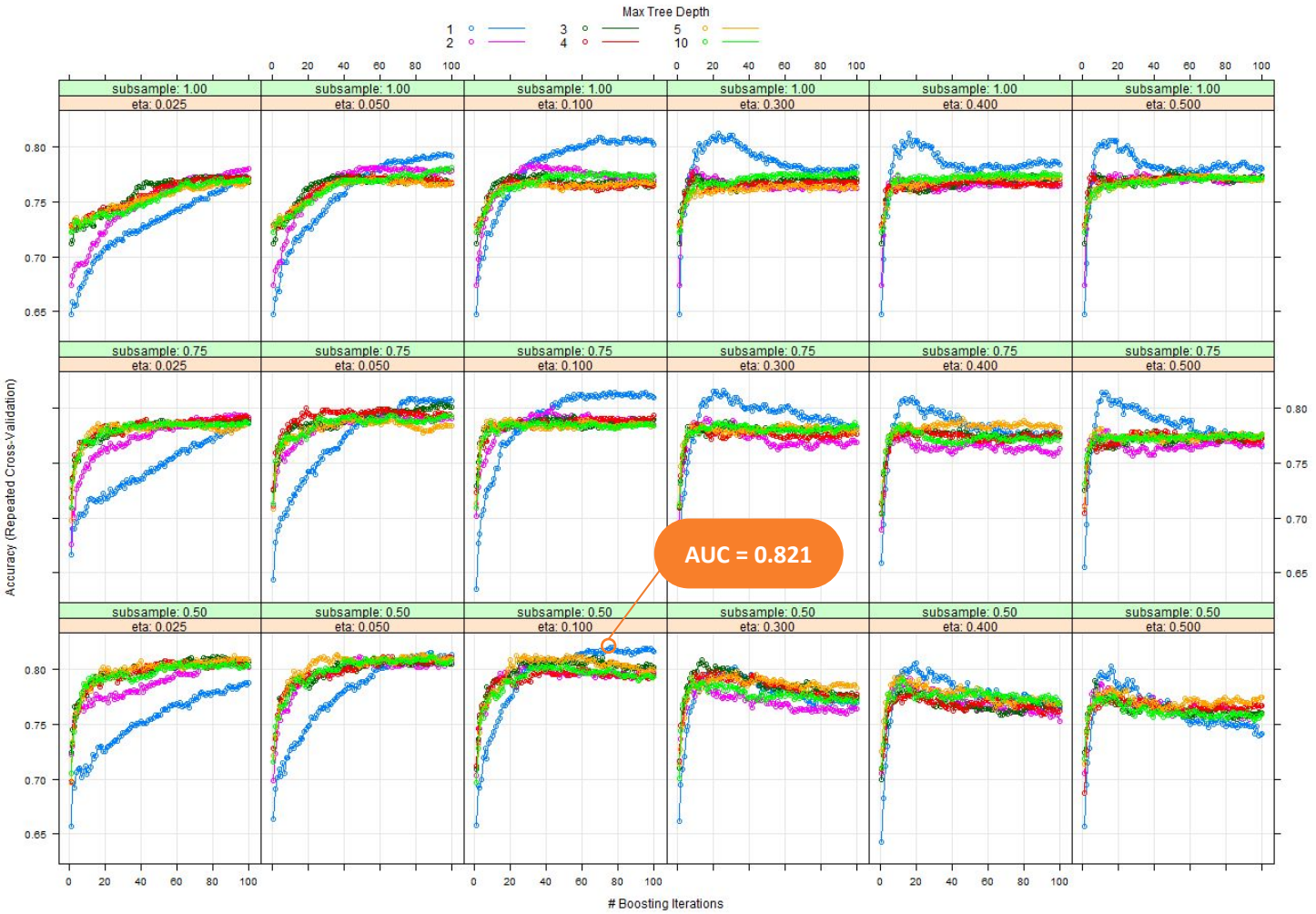
Recursive Feature Elimination



Results

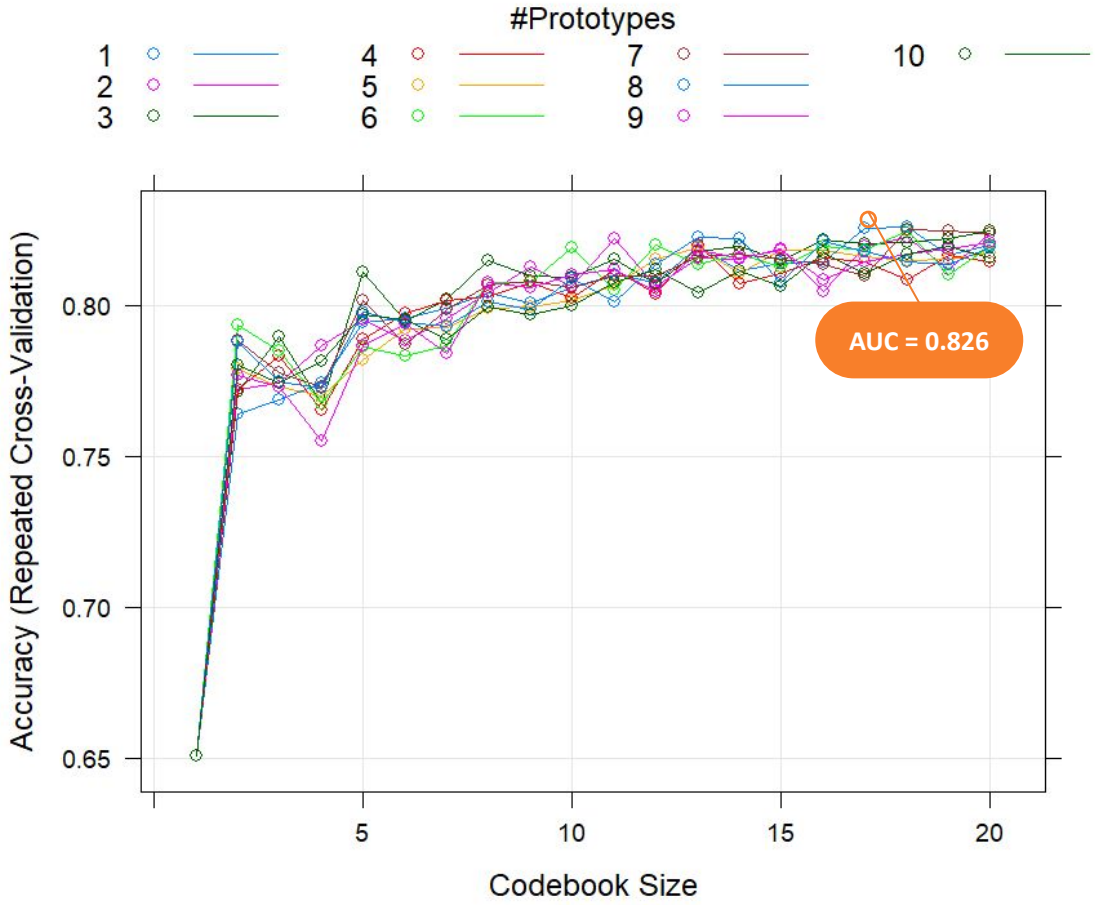
10 fold CV

Stochastic gradient boosting



Results

Learning Vector Quantization



Solution

What if instead of relying solely on the corrected p-value of the T-test we also rely on the power of the result?

Results

library(pwr)

Power calculations for two samples (different sizes) t-tests of means

Description

Compute power of tests or determine parameters to obtain target power (similar to as power.t.test).

Usage

```
pwr.t2n.test(n1 = NULL, n2= NULL, d = NULL, sig.level = 0.05, power = NULL,
 alternative = c("two.sided",
        "less","greater"))
```

Arguments

n1	Number of observations in the first sample
n2	Number of observations in the second sample
d	Effect size
sig.level	Significance level (Type I error probability)
power	Power of test (1 minus Type II error probability)
alternative	a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less"

Arguments

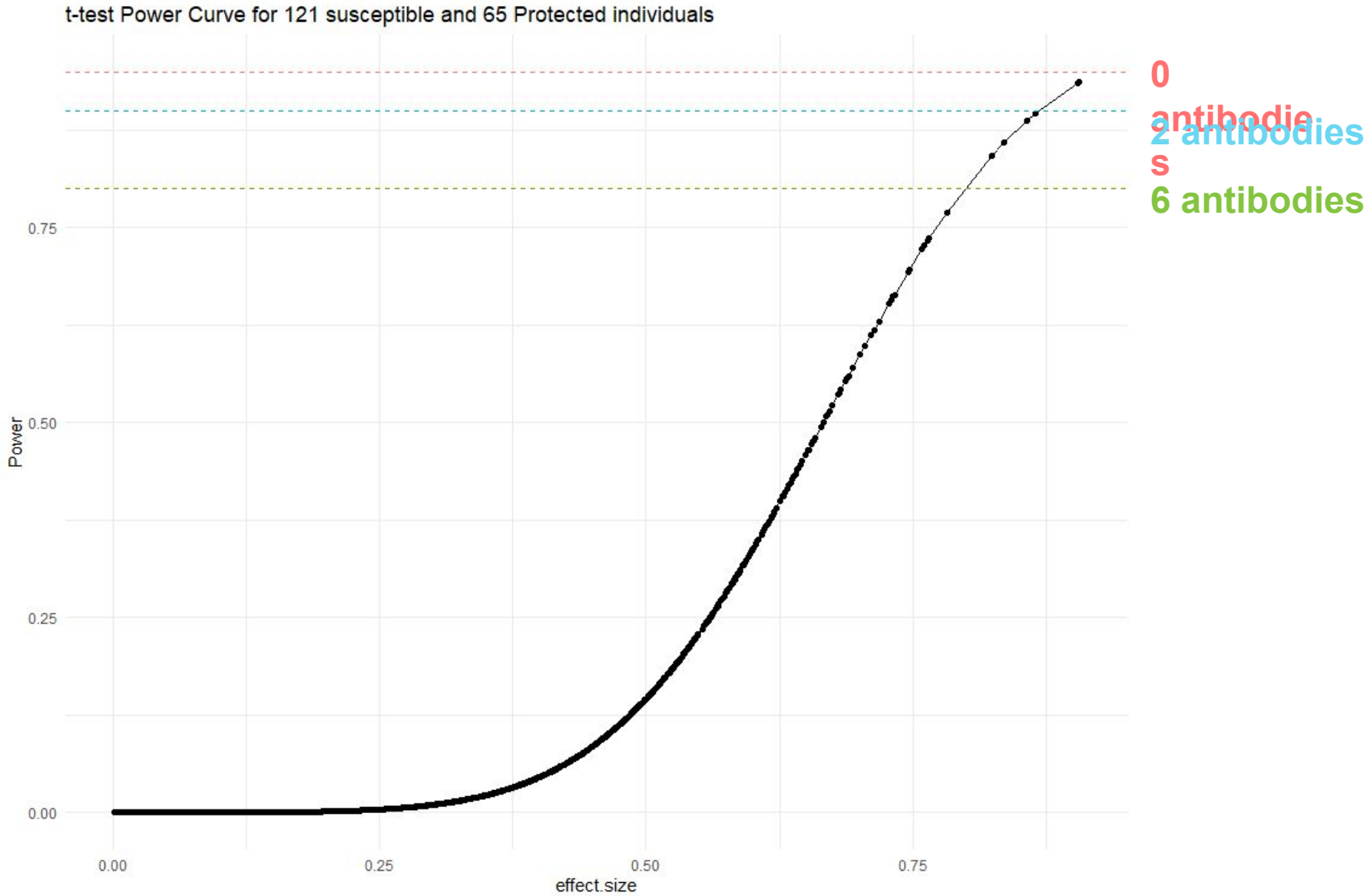
n1 = number of susceptibles (121)
n2= number of protected (65)

$$d = \frac{\text{mean}(x1) - \text{mean}(x2)}{\sqrt{\frac{(n1-1).sd1^2 + (n2-1).sd2^2}{n1+n2-2}}}$$

$$\text{sig.level} = 2.38 * 10^{-5}$$

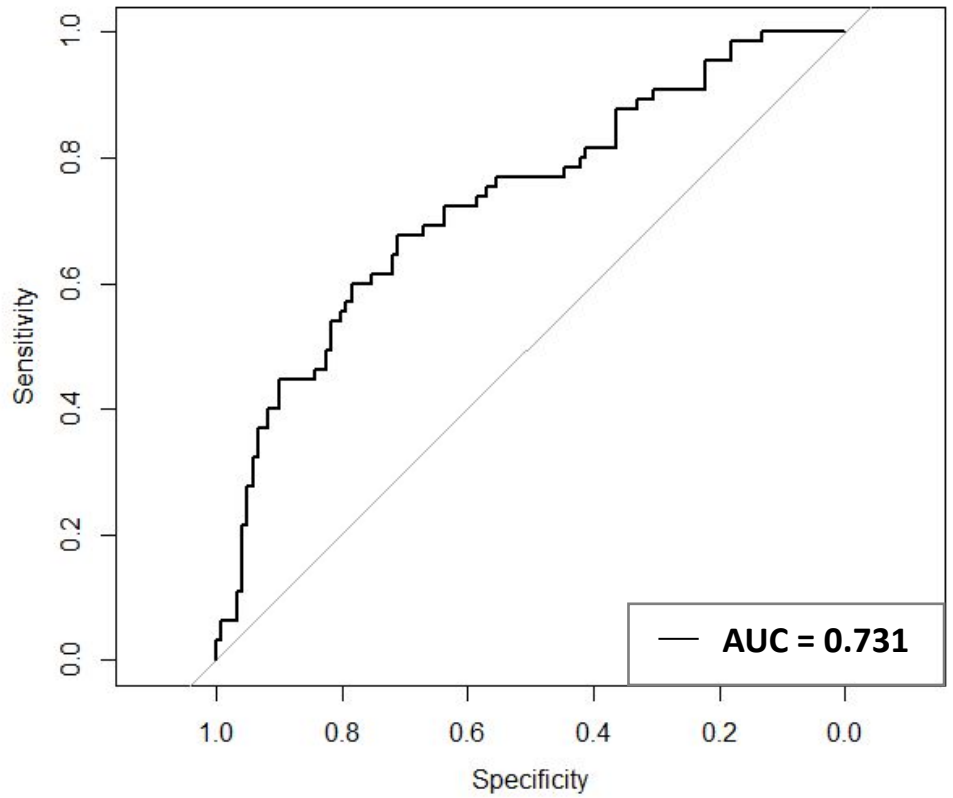
alternative = “two-sided”

Results


t-test Power Curve for 121 susceptible and 65 Protected individuals
0 antibodies
2 antibodies
6 antibodies
Power
0.00
0.25
0.50
0.75
effect.size

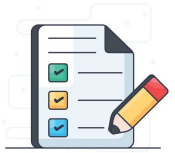
Results

Logistic regression



Setbacks

What about the other antibodies that did not follow a normal distribution (p=221) ?

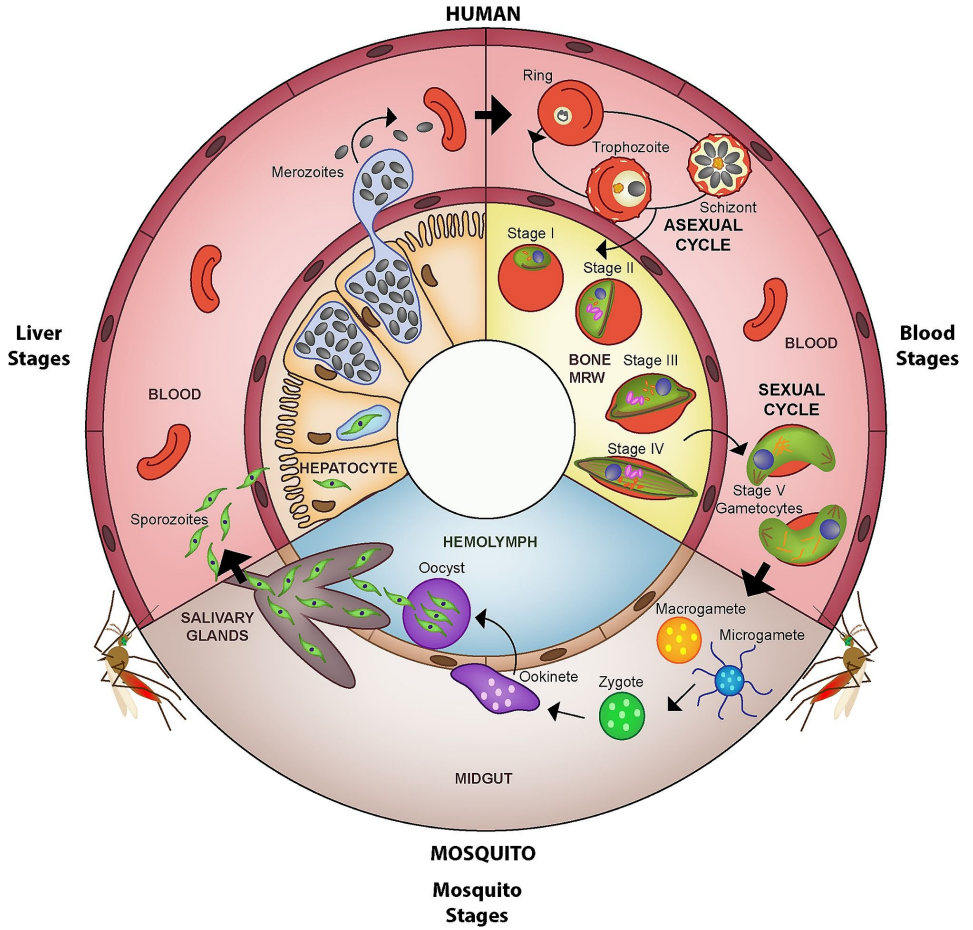


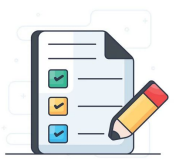
Discussion

Our pipeline needs some tweaking to be more suitable for higher dimension datasets

Discussion

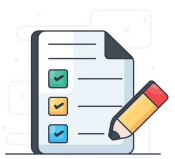
- Our pipeline needs some tweaking to be more suitable for higher dimension datasets
- High dimension datasets difficult the identification of important antibodies. An alternative could be to analyse antibodies in a more biological oriented approach (e.g. MSP)





Discussion

- Our pipeline needs some tweaking to be more suitable for higher dimension datasets
- High dimension datasets difficult the identification of important antibodies. An alternative could be to analyse antibodies in a more biological oriented approach (e.g. MSP)
- Pipelines such as these may help to increase reproducibility among studies



Discussion

- Our pipeline needs some tweaking to be more suitable for higher dimension datasets
- High dimension datasets difficult the identification of important antibodies. An alternative could be to analyse antibodies in a more biological oriented approach (e.g. MSP)
- Pipelines such as these may help to increase reproducibility among studies
- Since antibody data transcends the field of malaria these pipelines may provide promising tools to be applied in other settings