Biostatistics

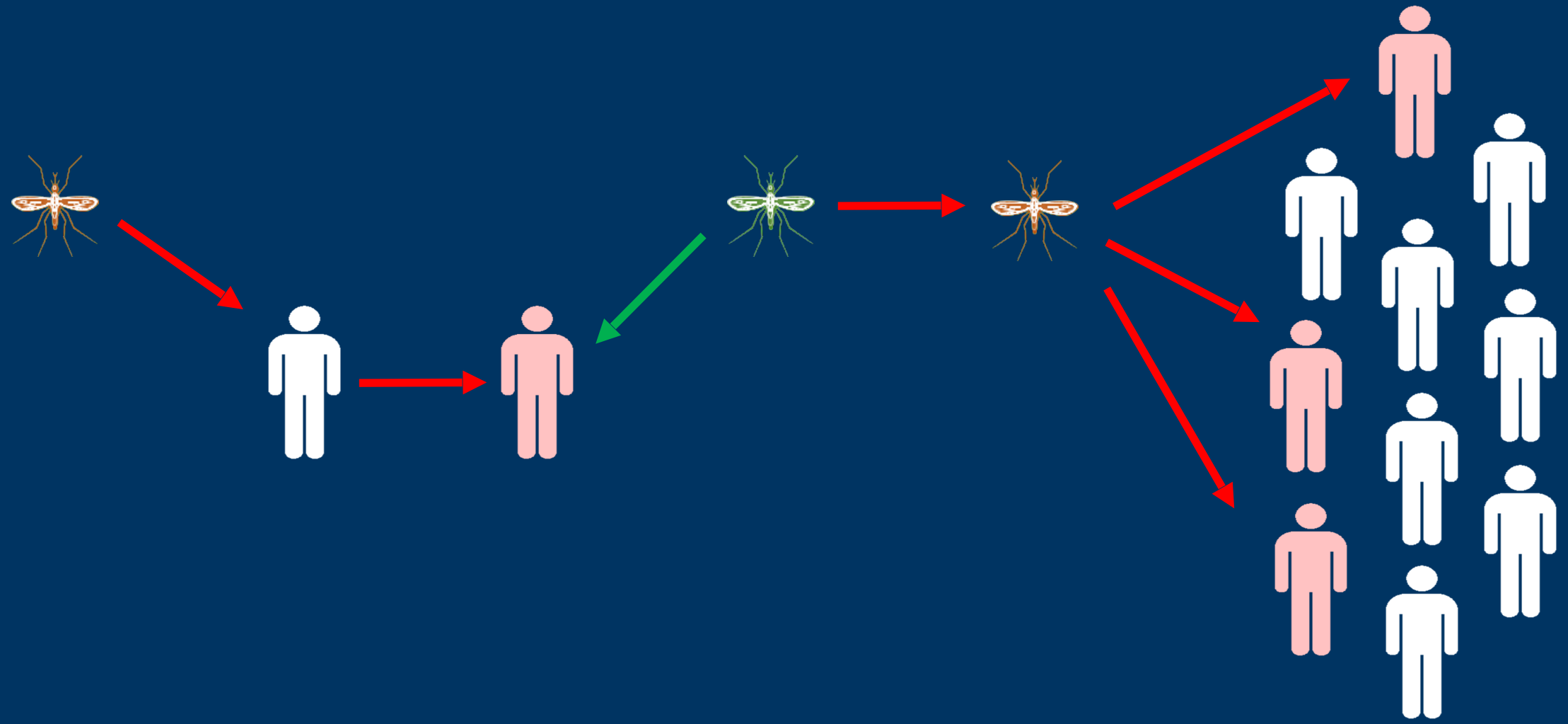
Applications in Serological Data Analysis

Nuno Sepúlveda, 12.01.2026

Syllabus

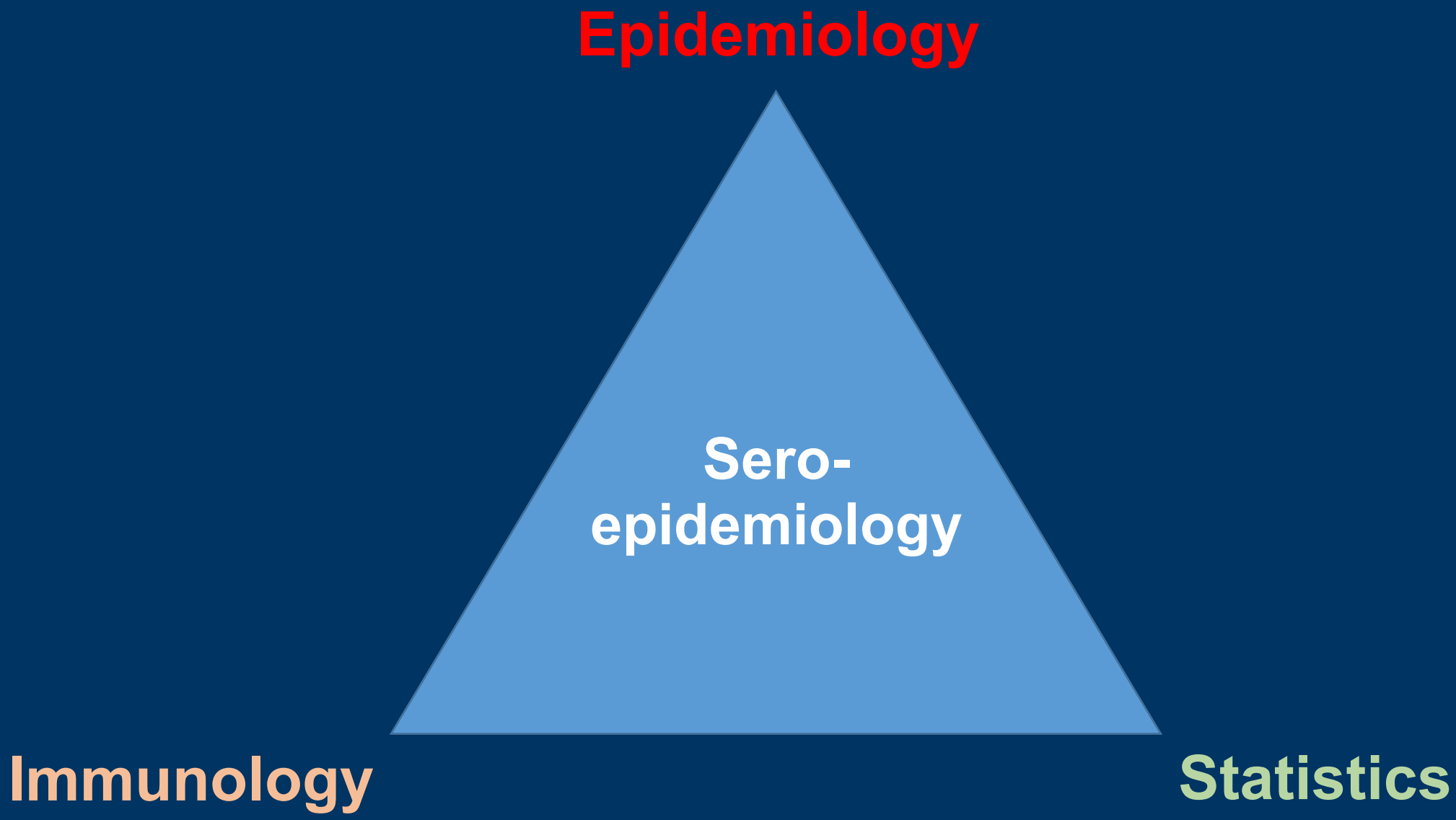
1. General review
    a. Population/Sample/Sample size
    b. Type of Data – quantitative and qualitative variables
    c. Common probability distributions/popular tests
2. Applications in Medicine
    a. Construction and analysis of diagnostic tools – Binomial distribution, ROC curve, sensitivity, specificity, Rogal-Gladen estimator
    b. Estimation of treatment effects - generalized linear models
    c. Survival analysis - Kaplan-Meier curve, log-rank test, Cox's proportional hazards model
3. Applications in Genetic and Epigenetic Data
    a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
    b. Methylation association studies – M versus beta values
4. Applications in Serological Data Analysis
    a. Determination of seropositivity using Gaussian mixture models
    b. Reversible catalytic models for estimating seroconversion rate
    c. Sample size calculation for estimating seroconversion rate

Motivation:
How to measure malaria transmission?



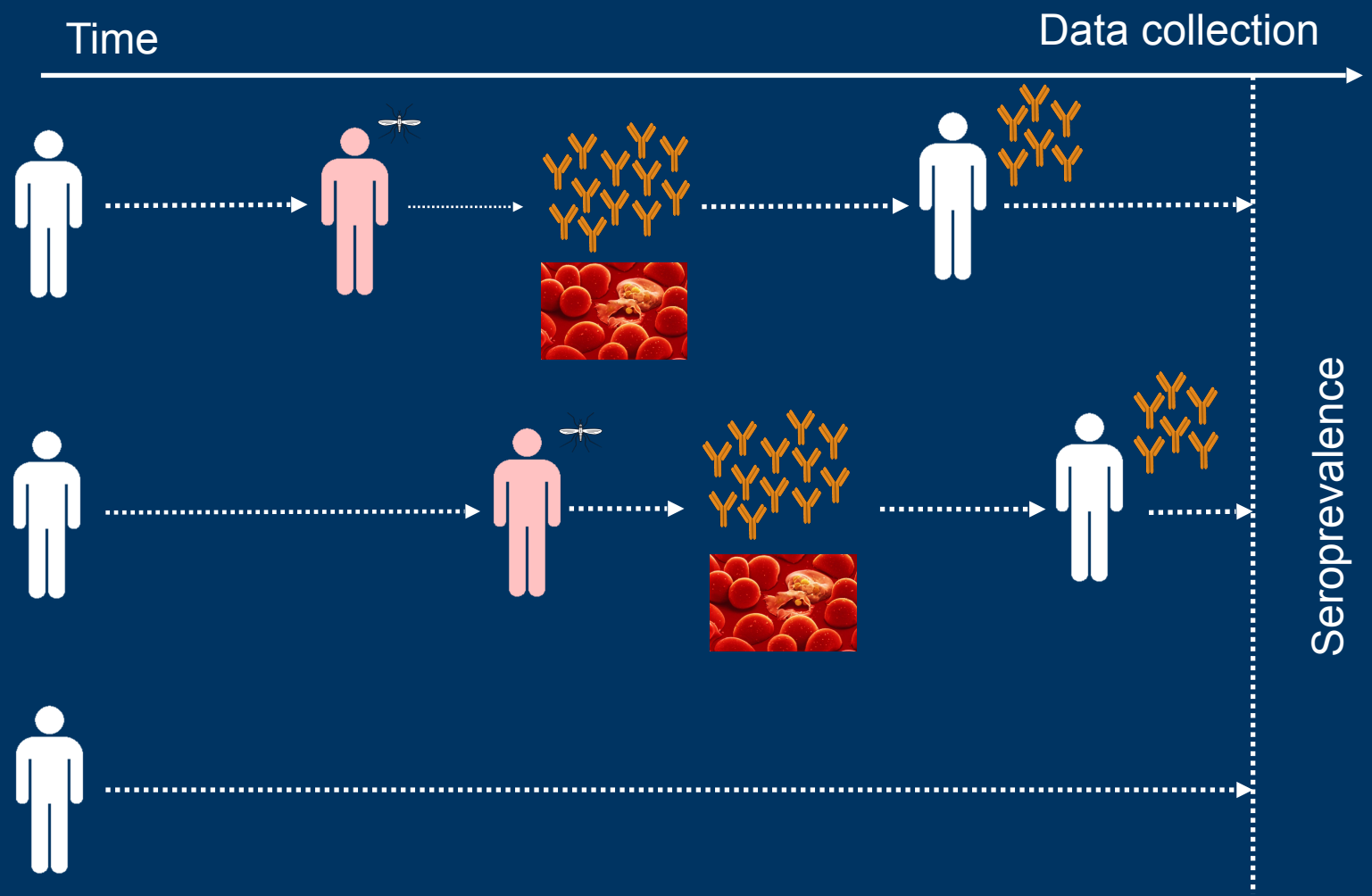
How to measure malaria transmission?

1. Prevalence of infection or parasite rate (non-informative when disease transmission intensity is low)
2. Entomological inoculation rate (trick to estimate)
3. Seroprevalence (prevalence of exposure)
4. Seroconversion rate (proxy of transmission intensity)

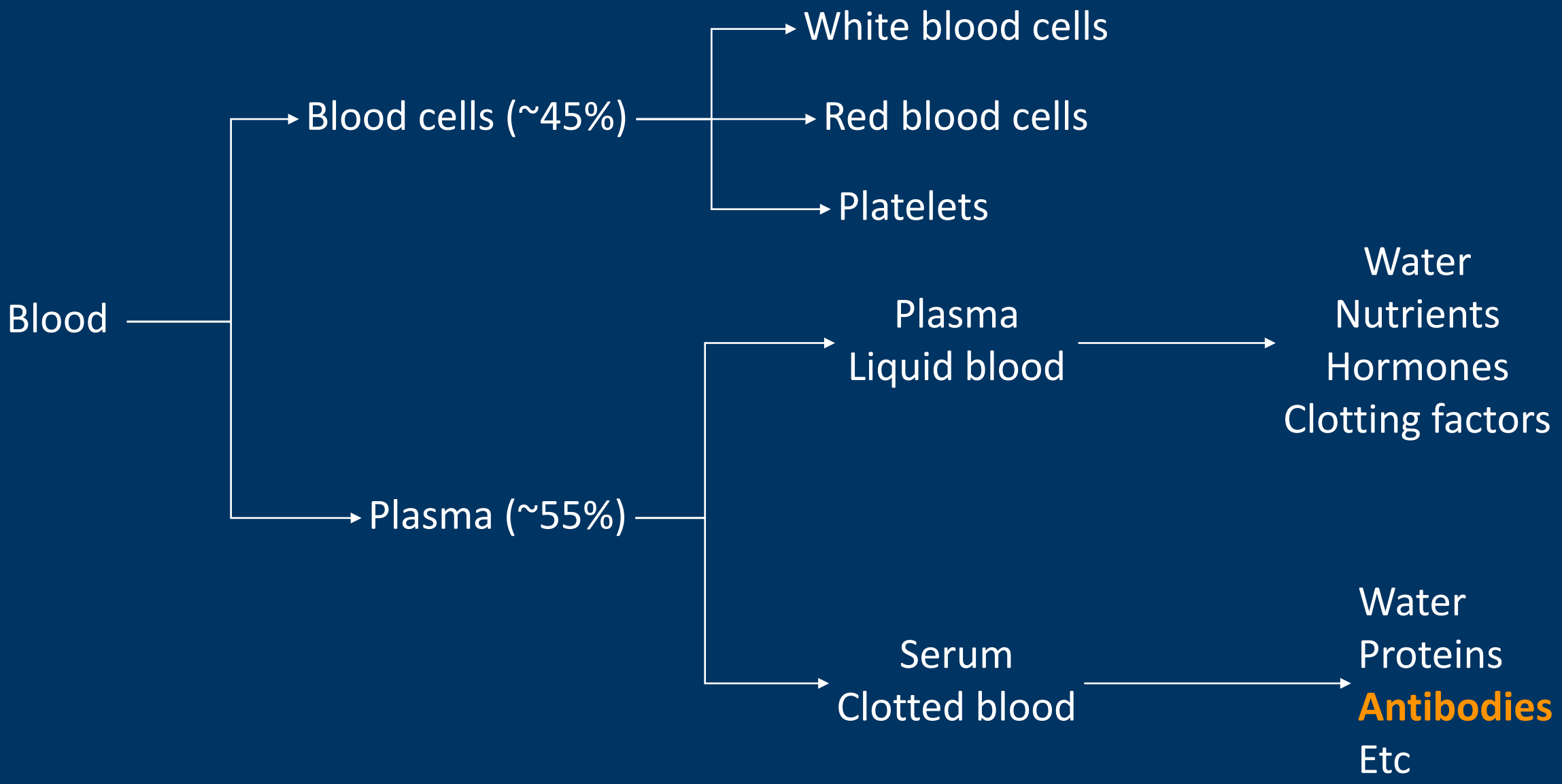

Epidemiology
Sero-
epidemiology
Immunology
Statistics

Defining seropositivity (using two-Gaussian mixture models) and estimating seroprevalence

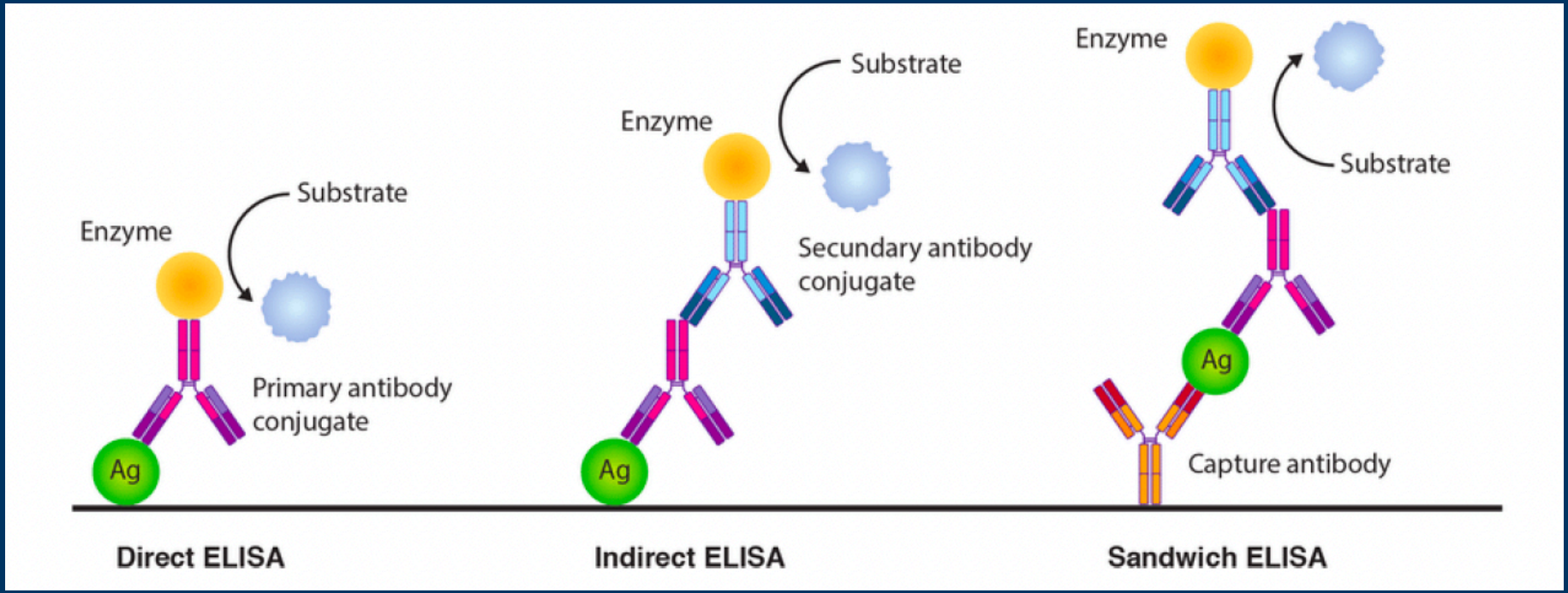
Basic principle



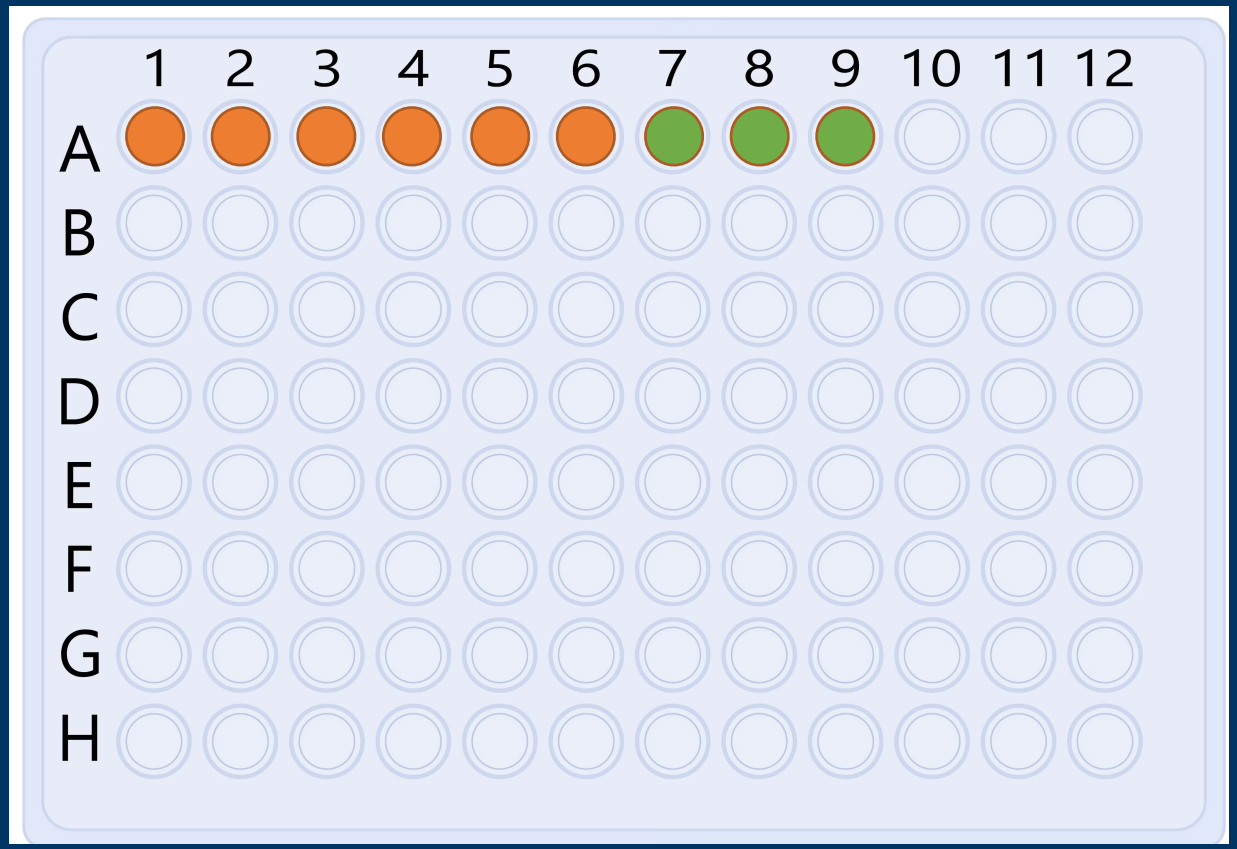
Blood


Blood
Blood cells (~45%)
White blood cells
Red blood cells
Platelets
Plasma (~55%)
Plasma
Liquid blood
Water
Nutrients
Hormones
Clotting factors
Serum
Clotted blood
Water
Proteins
Antibodies
Etc

Serology data - ELISA



Serology data - 96-well plate

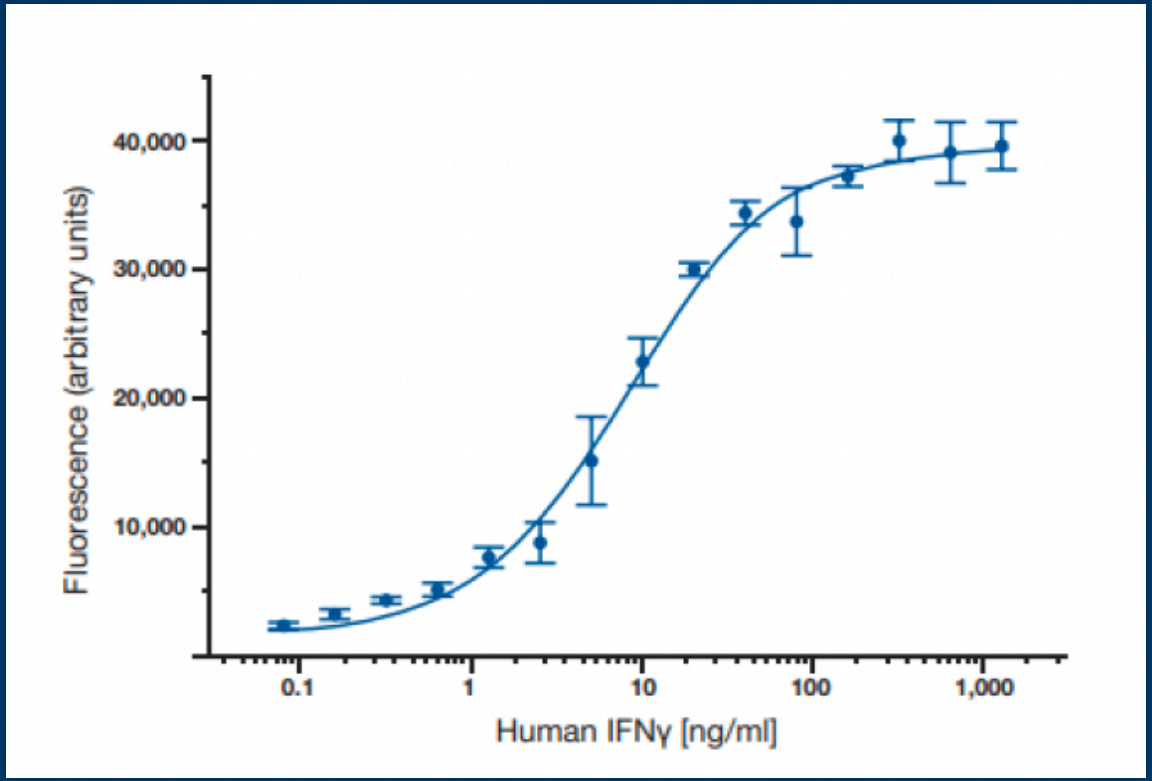


Positive controls (different dilutions)

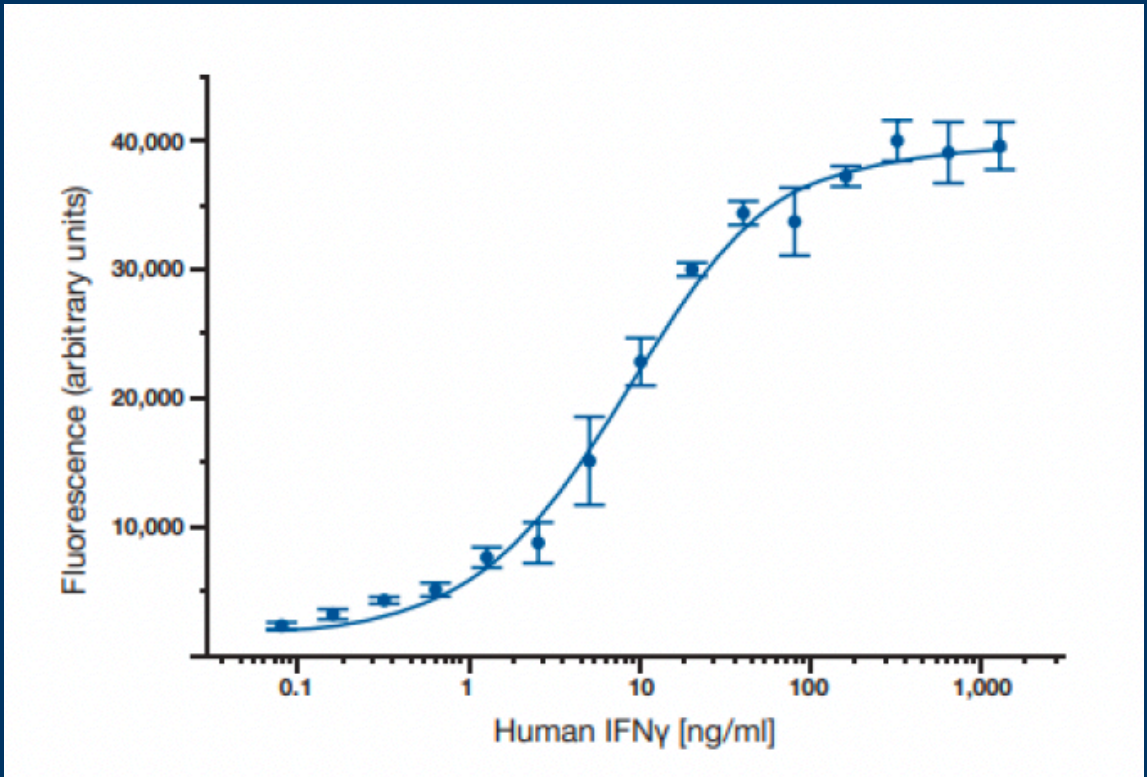
Blank (negative control)

Field samples

Serology data - Calibration curve



Serology data - Calibration curve



$$y_j = a + \frac{(d-a)}{1+(x_j/c)^b},$$

where y_j is the response at concentration x_j, a is the upper asymptote, d is the lower asymptote, c is the concentration at the inflection point of the curve and b is the growth factor (Azadeh et al., 2017).

Exercise: data_serology.csv

Multiplex assays for the identification of serological signatures of SARS-CoV-2 infection: an antibody-based diagnostic and machine learning study

Jason Rosado, Stéphane Pelleau, Charlotte Cockram, Sarah Hélène Merkling, Narimane Nekkab, Caroline Demeret, Annalisa Meola, Solen Kerneis, Benjamin Terrier, Samira Fafi-Kremer, Jerome de Seze, Timothée Bruel, François Dejardin, Stéphane Petres, Rhea Longley, Arnaud Fontanet, Marija Backovic, Ivo Mueller, Michael T White

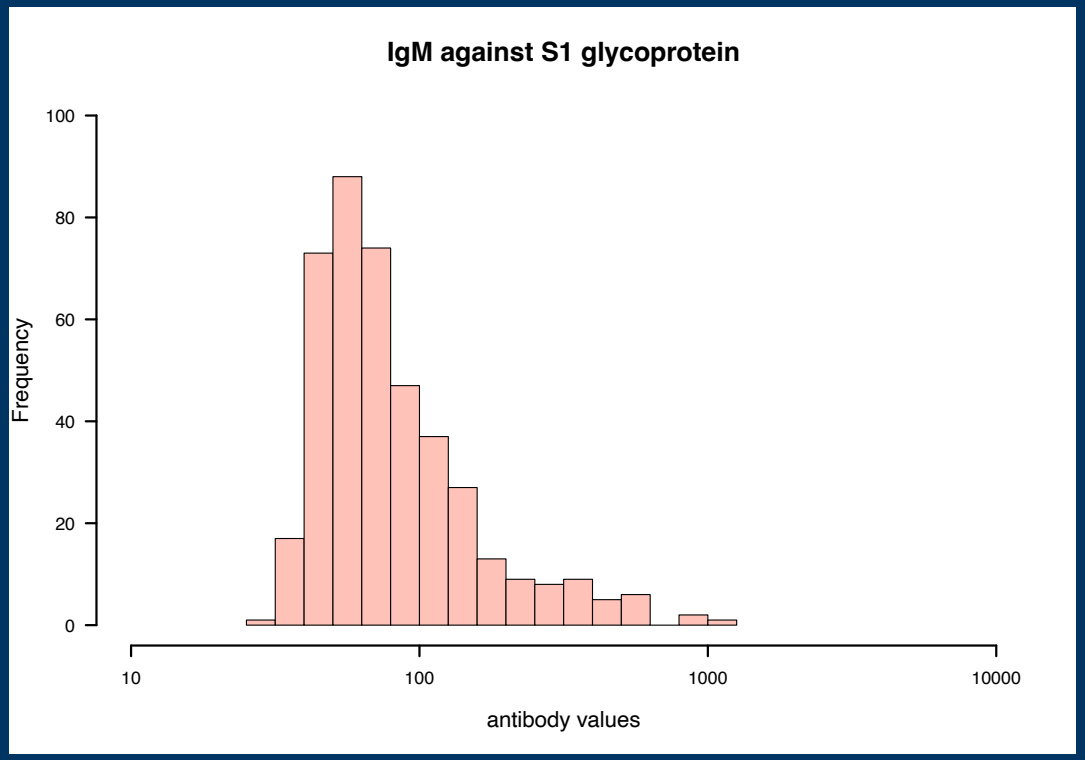
Focus on data from participant with ID B01-R002

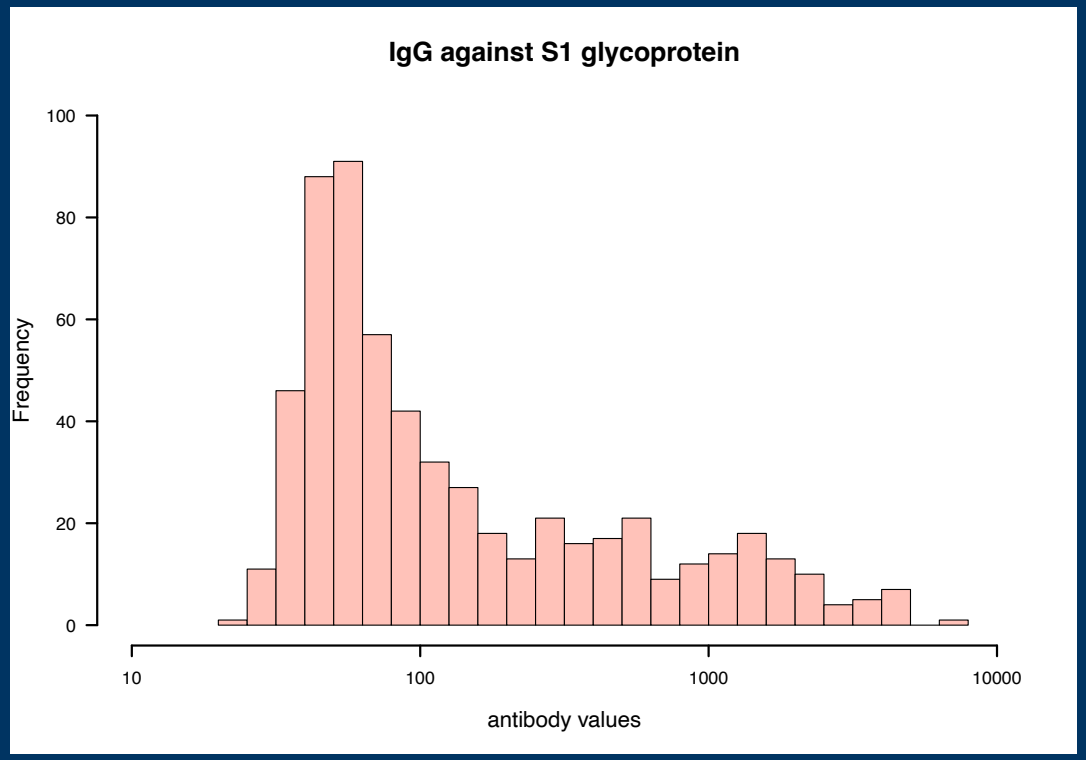
Plot S1RBD_NA_IgG_dil versus S1RBD_NA_IgG_MFI

Estimate the calibration curve of these data and estimate the concentrations expected from an MFI of 3000, 6000, and 8000

Use nplr package

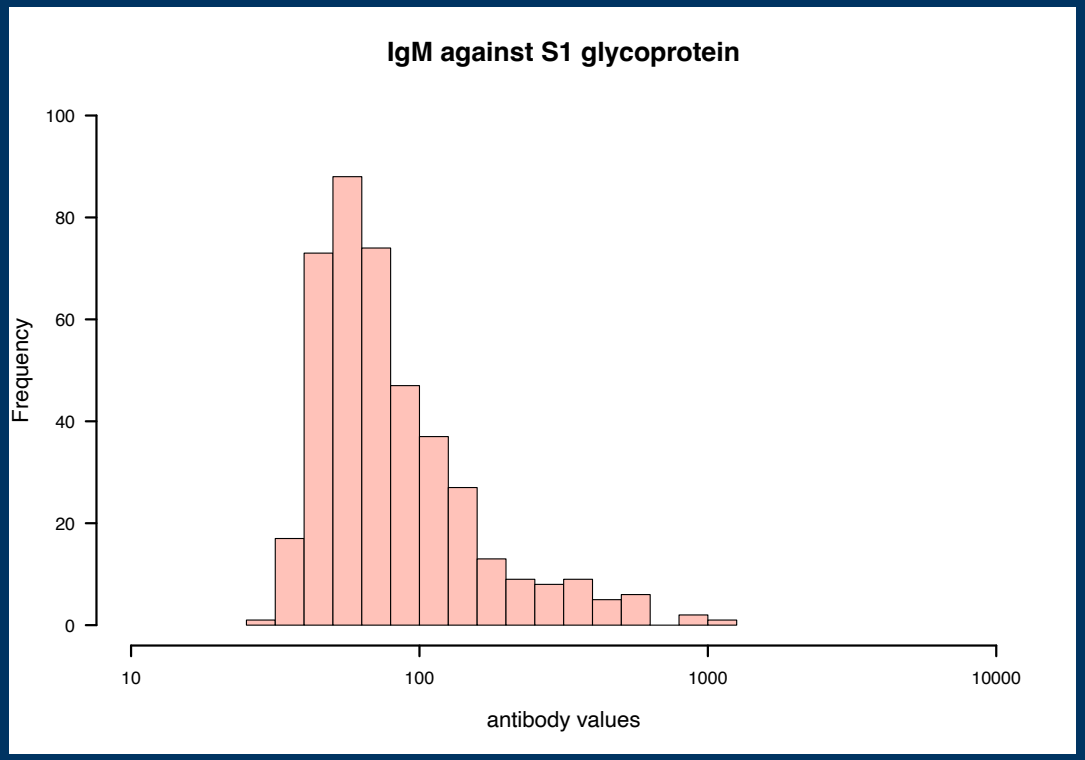
Antibody data are intrinsically quantitative

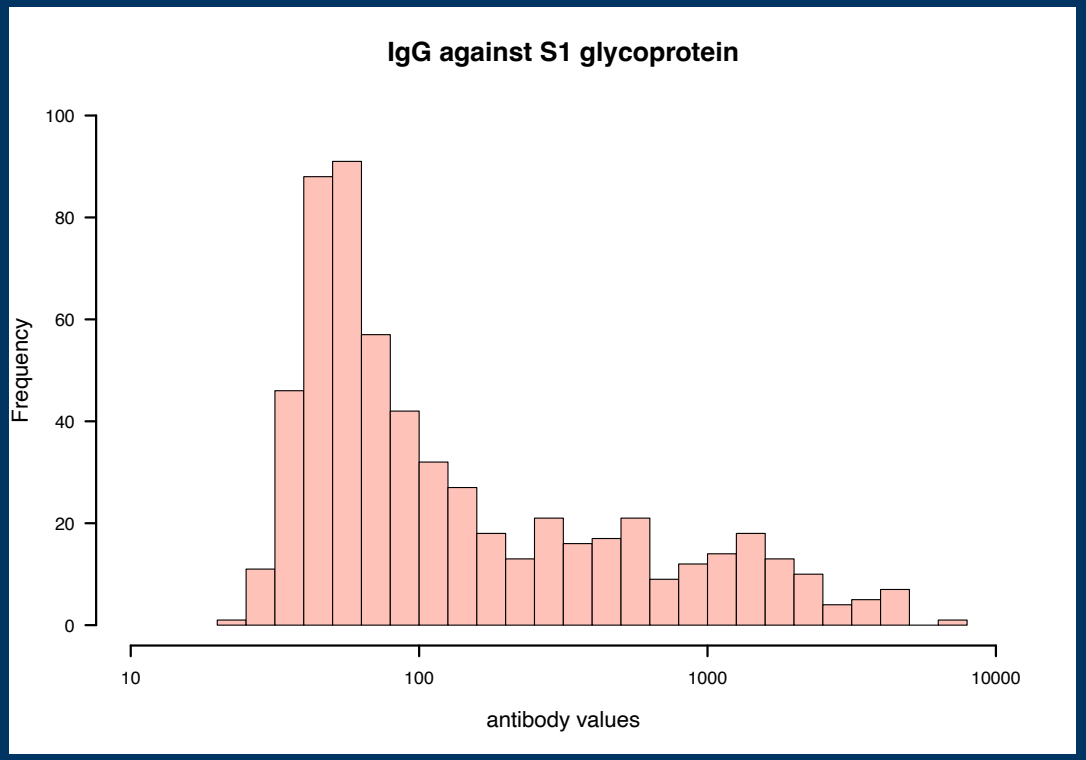




Rosado et al (2020). Serological signatures of SARS-CoV-2 infection: Implications for antibody-based diagnostics. medRxiv 2020.05.07.20093963.

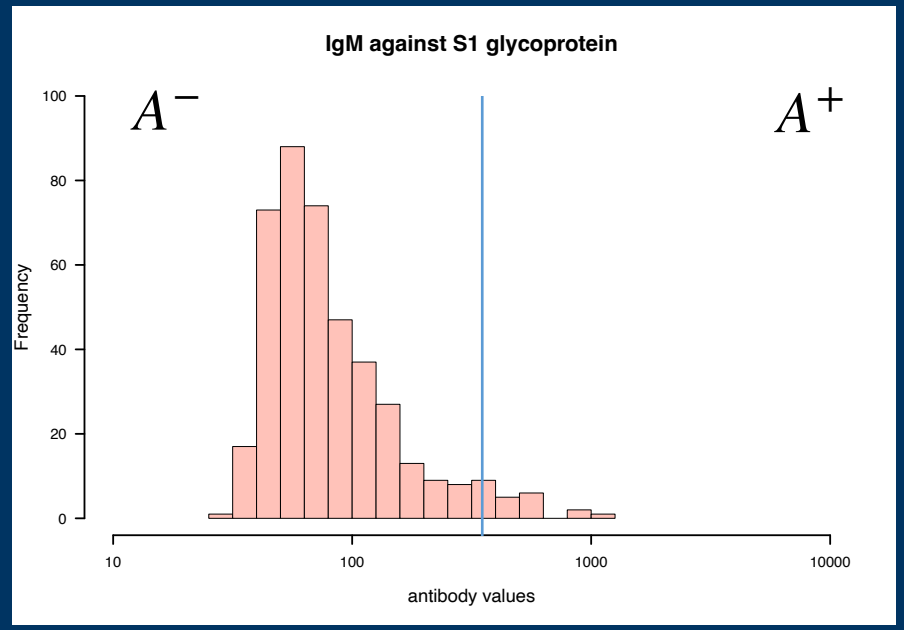
Who are the seropositive individuals?

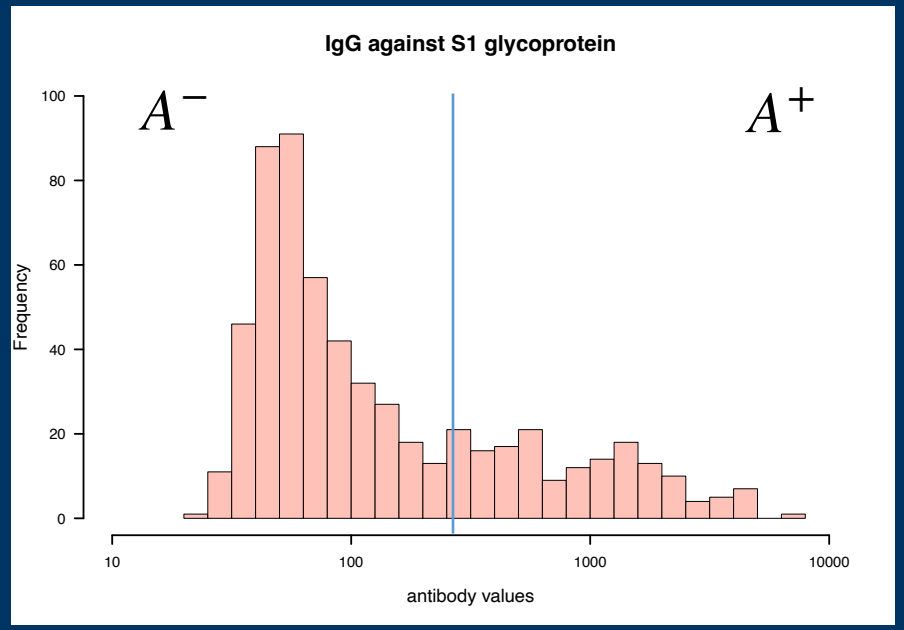


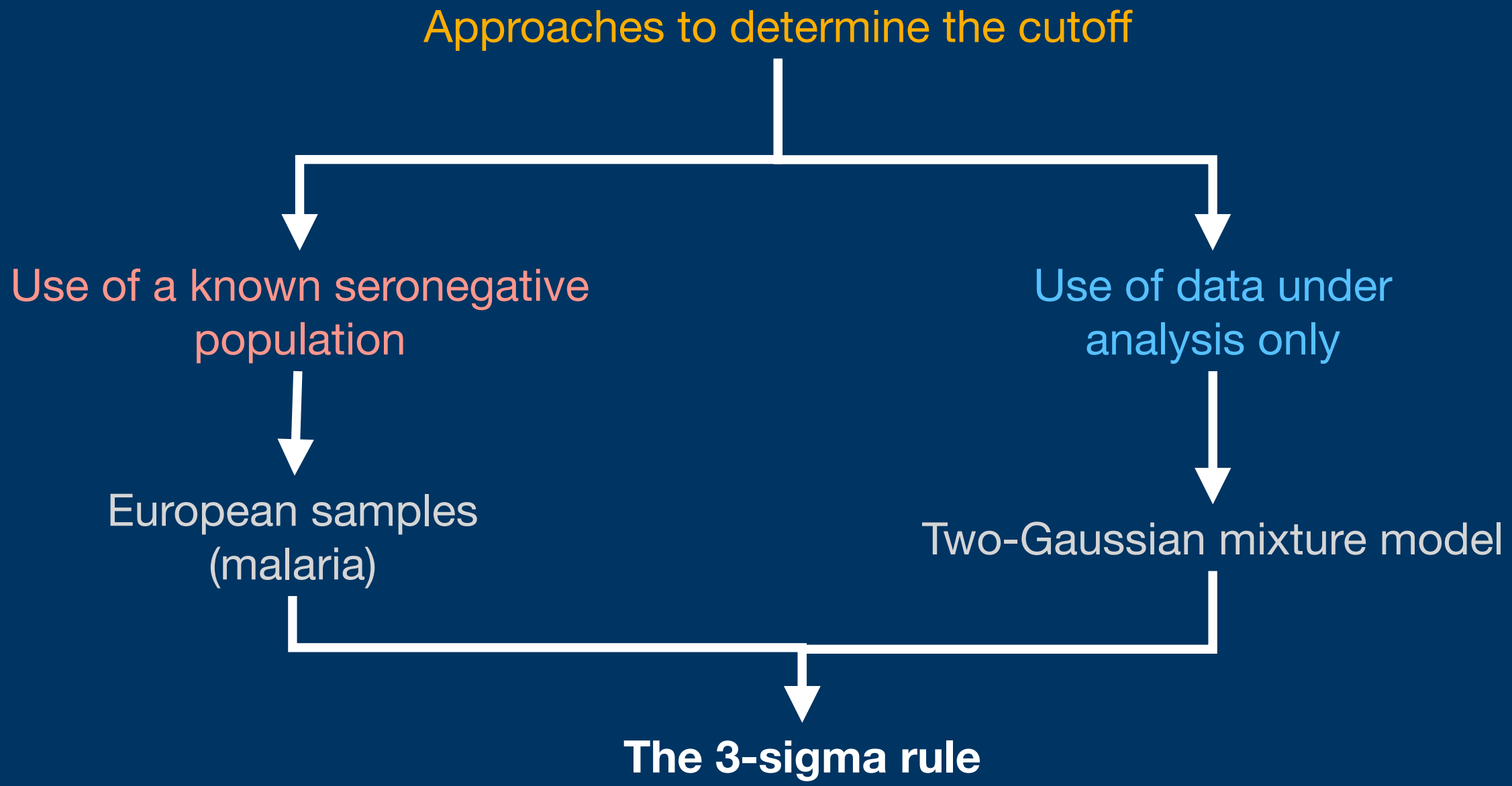


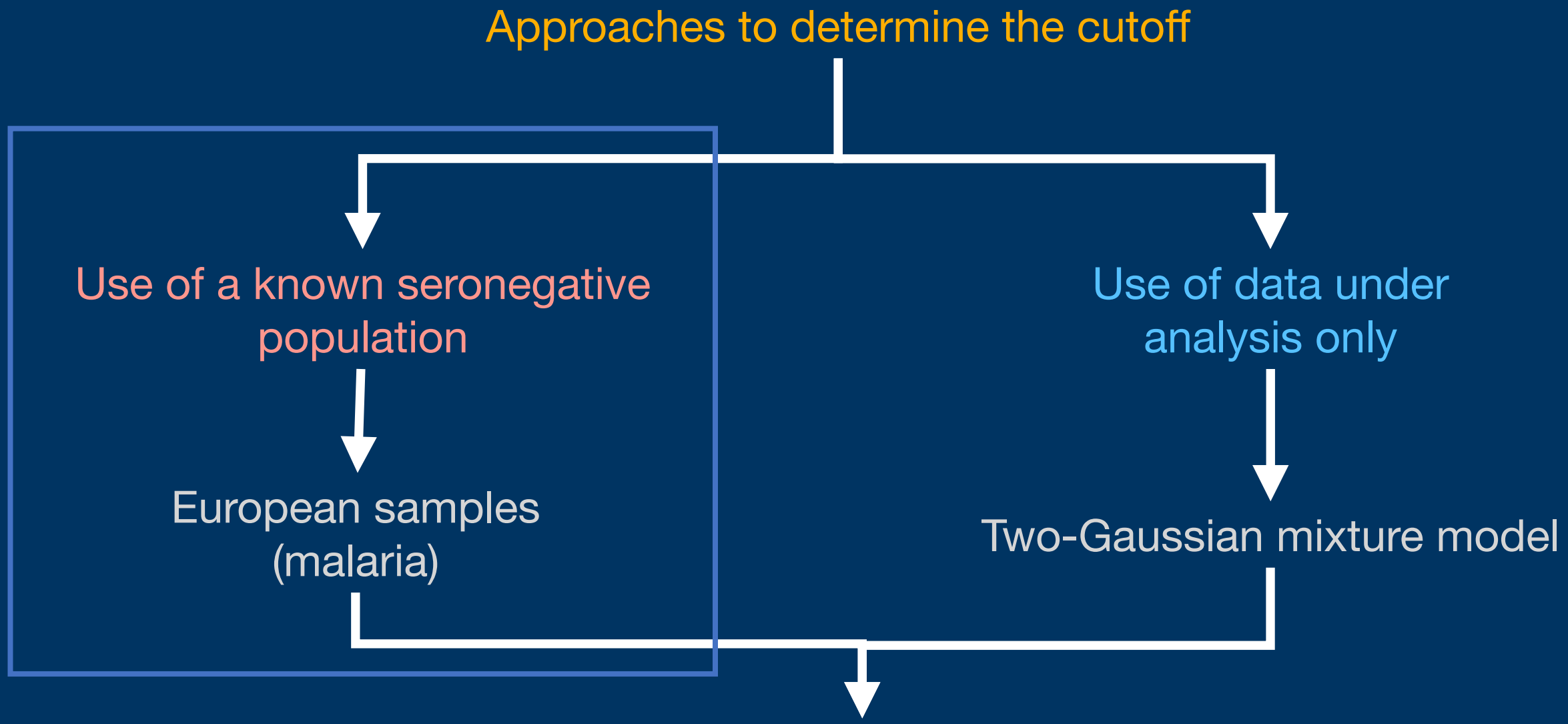
Rosado et al (2020). Serological signatures of SARS-CoV-2 infection: Implications for antibody-based diagnostics. medRxiv 2020.05.07.20093963.

How to determine the cut-off?






Approaches to determine the cutoff
Use of a known seronegative population
Use of data under analysis only
European samples (malaria)
Two-Gaussian mixture model
The 3-sigma rule


Approaches to determine the cutoff
Use of a known seronegative population
European samples (malaria)
Use of data under analysis only
Two-Gaussian mixture model
The 3-sigma rule

Theoretical 3-sigma rule

$$\mu_{A^-} = E\left[X \mid A^-\right]$$

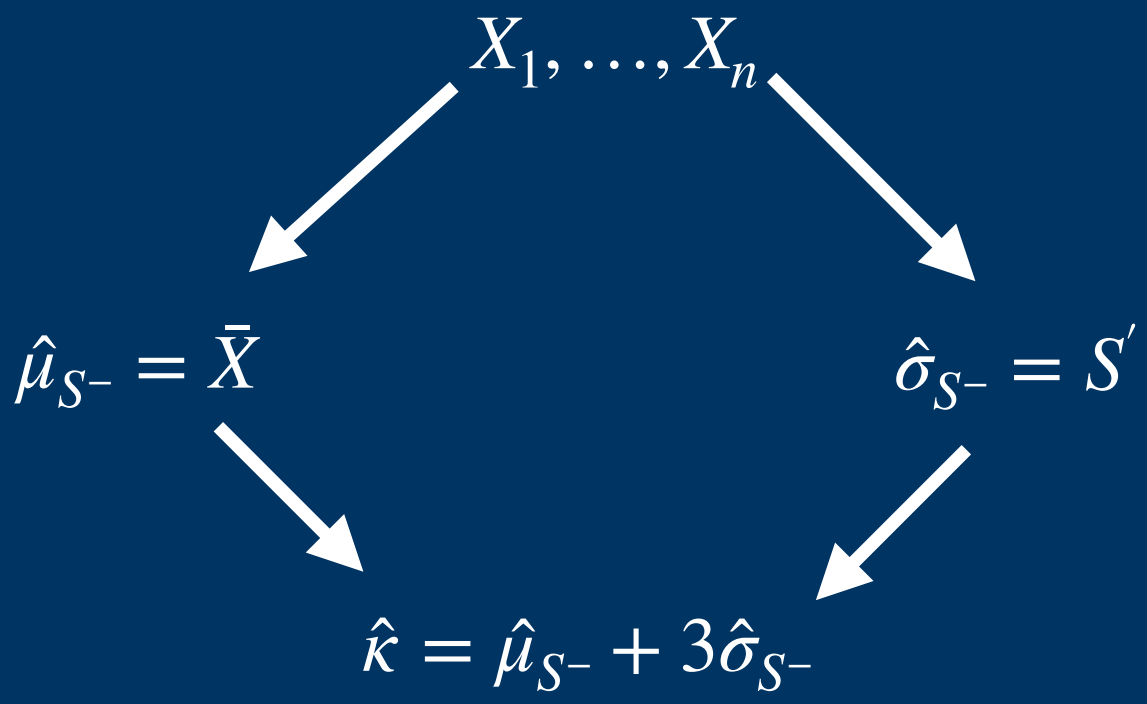
$$\sigma_{A^-} = \sqrt{Var\left[X \mid A^-\right]}$$

Seronegative, if $X_i \leq \mu_{A^-} + 3\sigma_{A^-}$

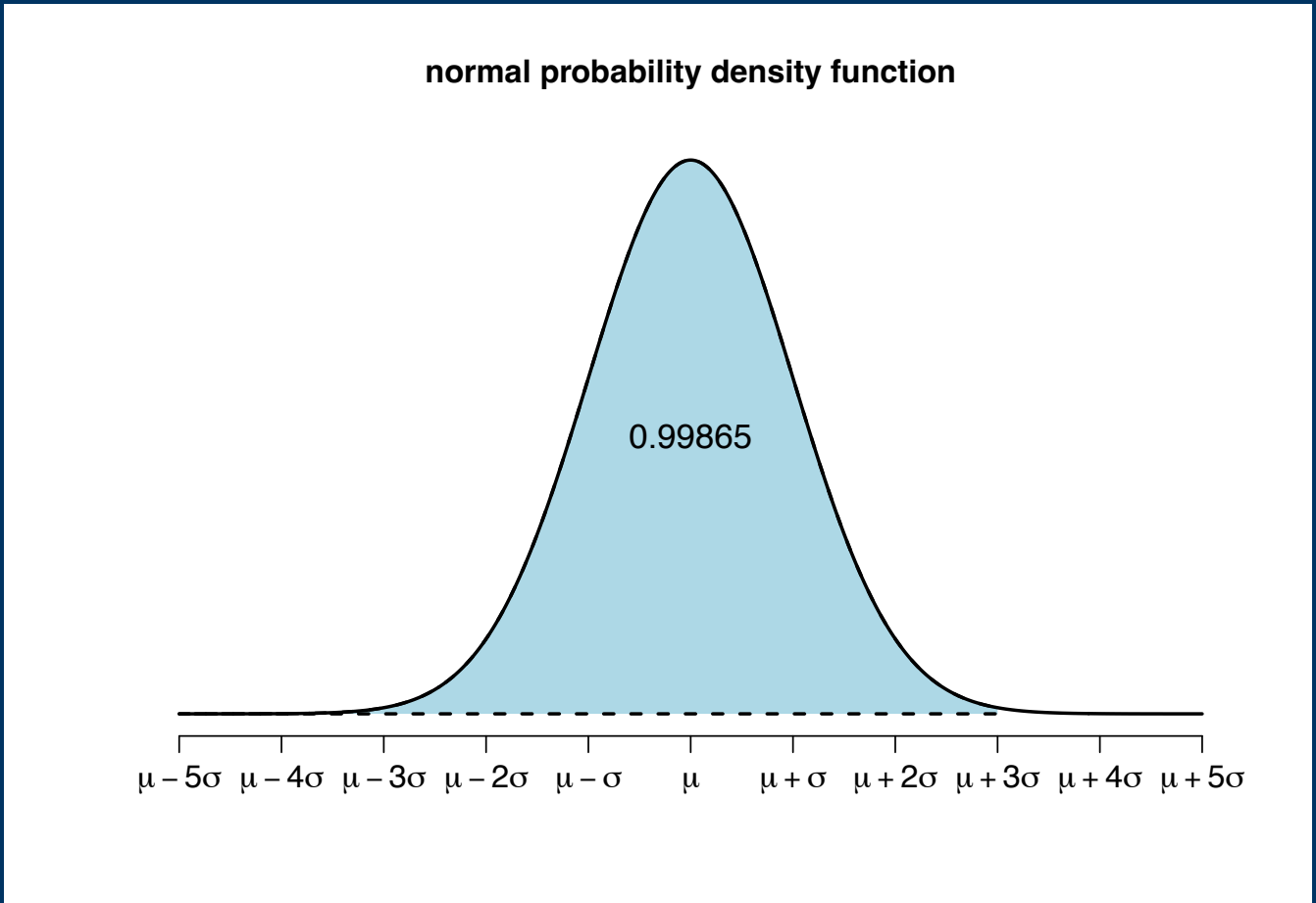
Seropositive, otherwise

Estimated 3σ rule

X_i = r. v. that describes background noise in the i-th seronegative individual



The link to the Normal distribution



Theoretical results

Cantelli-Chebyshev inequality

$$P\left[X \geq \mu + \lambda\right] \leq \frac{\sigma^2}{\sigma^2 + \lambda^2}, \text{ if } \lambda > 0$$

$$\mu = E\left[X\right] \qquad \sigma^2 = Var\left[X\right] < \infty$$

Application to $\lambda = 3\sigma_{S^-}$

$$P\left[X \geq \mu_{S^-} + 3\sigma_{S^-}\right] \leq \frac{1}{10} \equiv 0.1$$

Specificity > 0.90

Theoretical results

One-sided Vysochanskii-Petunin inequality

$$P\left[X-\mu \geq r\right] \leq \begin{cases} \frac{4}{9}\frac{\sigma^2}{r^2+\sigma^2}, & \text{for } r^2 \geq \frac{5}{3}\sigma^2 \\ \frac{4}{3}\frac{\sigma^2}{r^2+\sigma^2}-\frac{1}{3}, & \text{otherwise} \end{cases}$$

$$\mu = E\left[X\right] \qquad \sigma^2 = Var\left[X\right] < \infty$$

Unimodal distributions

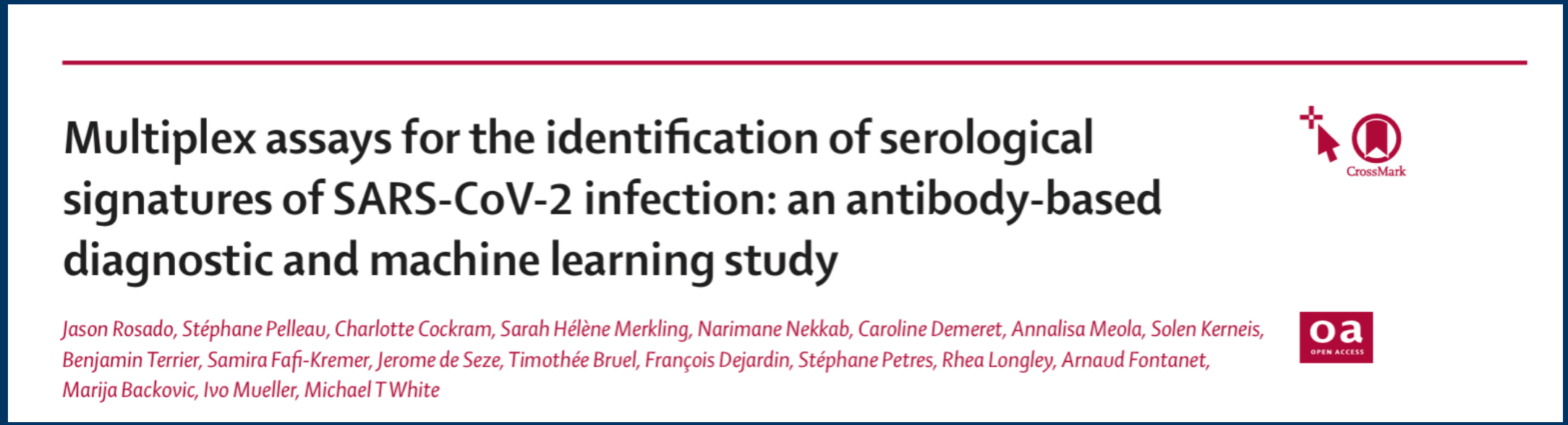
Application to $r = 3\sigma_{S^-}$

$$P\left[X-\mu_{S^-} \geq 3\sigma_{S^-}\right] \leq 0.044$$

Specificity > 0.956

Mercadier, Mathieu & Strobel, Frank, 2021. "A one-sided Vysochanskii-Petunin inequality with financial applications," European Journal of Operational Research, vol. 295(1), pages 374-377.

Exercise: data_serology.csv



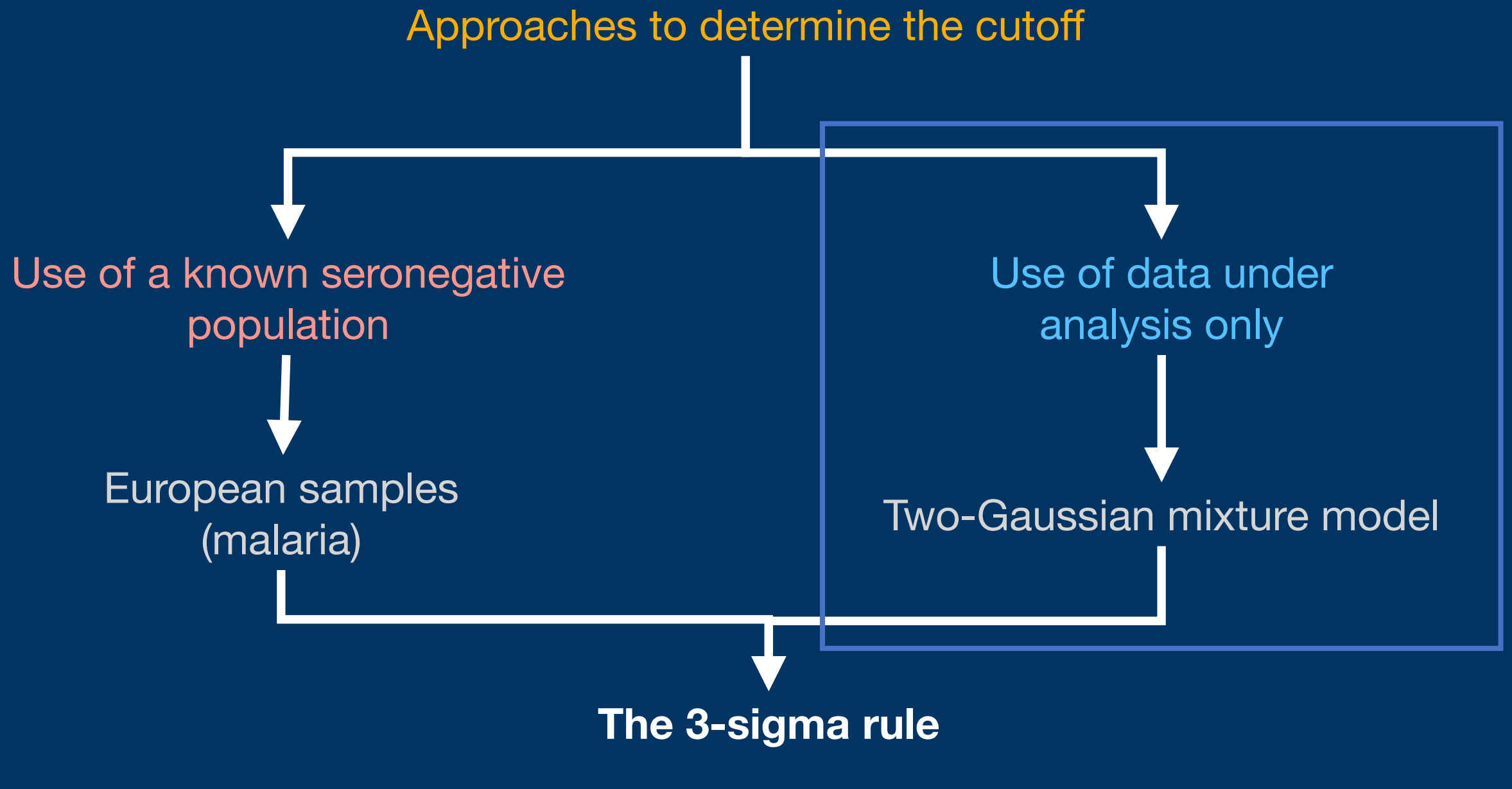
Multiplex assays for the identification of serological signatures of SARS-CoV-2 infection: an antibody-based diagnostic and machine learning study

Jason Rosado, Stéphane Pelleau, Charlotte Cockram, Sarah Hélène Merkling, Narimane Nekkab, Caroline Demeret, Annalisa Meola, Solen Kerneis, Benjamin Terrier, Samira Fafi-Kremer, Jerome de Seze, Timothée Bruel, François Dejardin, Stéphane Petres, Rhea Longley, Arnaud Fontanet, Marija Backovic, Ivo Mueller, Michael T White

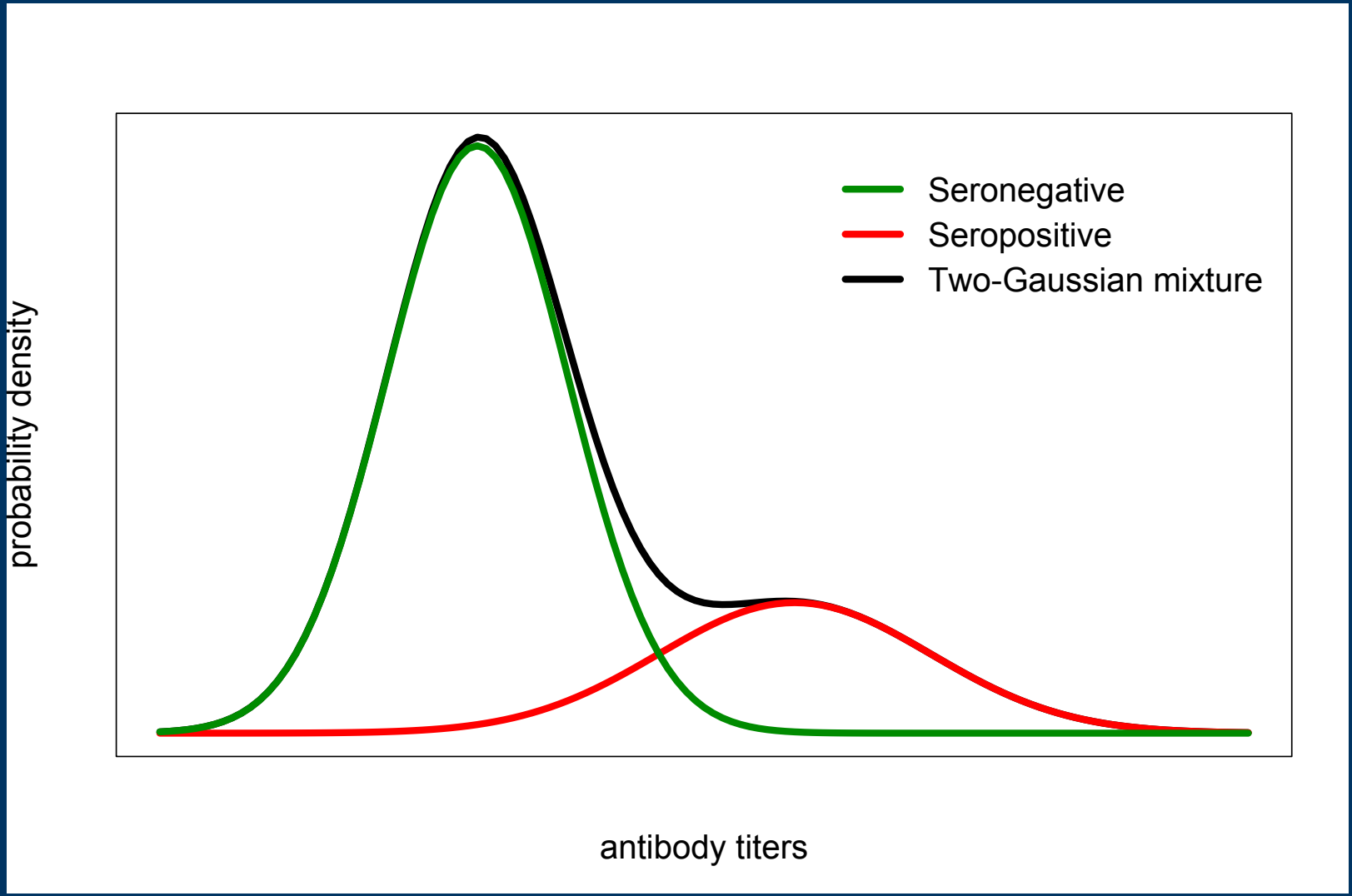
Focus on data from participants with status = negative (samples collected before the pandemic)

Estimate the cut-off using either S1RBD_NA_IgG_dil or S1RBD_NA_IgG_MFI. Estimate the respective specificity.
Is the estimated specificity in agreement with theoretical results?

Estimate the seroprevalence associated with that antibody for participants with status = positive (samples during the pandemic).


Approaches to determine the cutoff
Use of a known seronegative population
European samples (malaria)
Use of data under analysis only
Two-Gaussian mixture model
The 3-sigma rule

Two-Gaussian mixture model



Two-Gaussian mixture models

$$f_X(x) = (1-\pi) f_{N(\mu_{S^-},\sigma_{S^-})}(x) + \pi f_{N(\mu_{S^+},\sigma_{S^+})}(x)$$

Definition of $S^- \Rightarrow \mu_{S^-} < \mu_{S^-}$

In general:

$$f_X(x) = \sum_{i=1}^{k} \pi_i f_{N(\mu_i,\sigma_i)}(x) \qquad \text{where} \sum_{i=1}^{k} \pi_i = 1$$

Estimation of the model by maximum likelihood method

EM (Expectation-Maximization) Algorithm

Package mixtools

1. Start with initial estimates for the parameters
2. E-Step - calculate the probability of each individual belonging to a given subpopulation according to estimates at 1.
3. M-Step - re-estimate the parameters using these probabilities and repeat the E-step with these new estimates
4. Stop with the increment in the log-likelihood is below a given tolerance error.

Calculate the cutoff for seropositivity, $\hat{k} = \hat{\mu}_{S^-} + 3\hat{\sigma}_{S^-}$

Estimate the "raw" seroprevalence using this cutoff ($\hat{\pi} = \frac{\sum_{i=1}^{n} I_{x>\hat{k}}(x_i)}{n}$)

Estimating specificity and sensitivity of the serological classification

cutoff for seropositivity

$$\hat{k} = \hat{\mu}_{S^-} + 3\hat{\sigma}_{S^-}$$

Sensitivity (detecting true seropositive individuals)

$$Se = 1 - F_{\mathcal{N}(\hat{\mu}_{S^+}),\hat{\sigma}_{S^+})}(\hat{k})$$

Specificity (detecting true seronegative individuals)

$$Sp = F_{\mathcal{N}(\hat{\mu}_{S^-}),\hat{\sigma}_{S^-})}(\hat{k})$$

Estimating corrected seroprevalence

Rogan-Gladen estimator (lecture 3)

$$\hat{\pi}_{+} = \frac{\hat{\pi} + \hat{\pi}_{Sp} - 1}{\hat{\pi}_{Se} + \hat{\pi}_{Sp} - 1}$$

π = estimated "raw" seroprevalence

$\hat{\pi}_{Se}$ = Estimated sensitivity

$\hat{\pi}_{Sp}$ = Estimated specificity

Exercise: data_serology.csv

Multiplex assays for the identification of serological signatures of SARS-CoV-2 infection: an antibody-based diagnostic and machine learning study

Jason Rosado, Stéphane Pelleau, Charlotte Cockram, Sarah Hélène Merkling, Narimane Nekkab, Caroline Demeret, Annalisa Meola, Solen Kerneis, Benjamin Terrier, Samira Fafi-Kremer, Jerome de Seze, Timothée Bruel, François Dejardin, Stéphane Petres, Rhea Longley, Arnaud Fontanet, Marija Backovic, Ivo Mueller, Michael T White

Focus on data from participants with status = positive (samples collected during the pandemic)

Estimate a two-Gaussian mixture model for S1RBD_NA_IgG_dil or S1RBD_NA_IgG_MFI using package mixtools. Estimate the 3-sigma cutoff and the respective sensitivity and specificity. Estimate the raw seroprevalence and corrected seroprevalence using the Rogan-Gladen Estimator.

Do you think the two-Gaussian mixture model is adequate for the data?