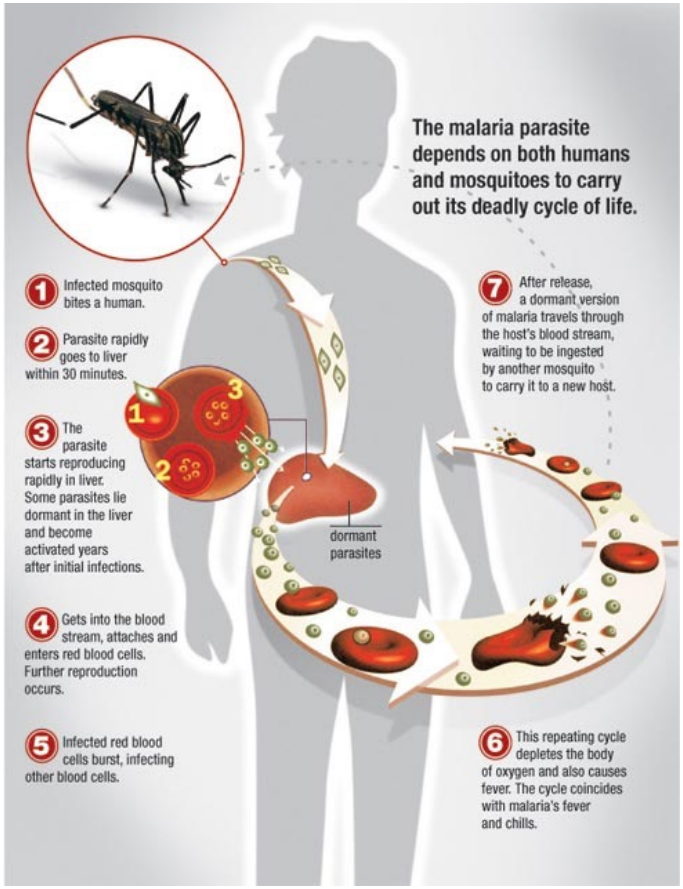
Statistical approaches to antibody data analysis for populations on the path of malaria elimination

Nuno Sepúlveda, nuno.sepulveda@lshtm.ac.uk

LSHTM, 28th April 2017

Malaria



Plasmodium parasite species

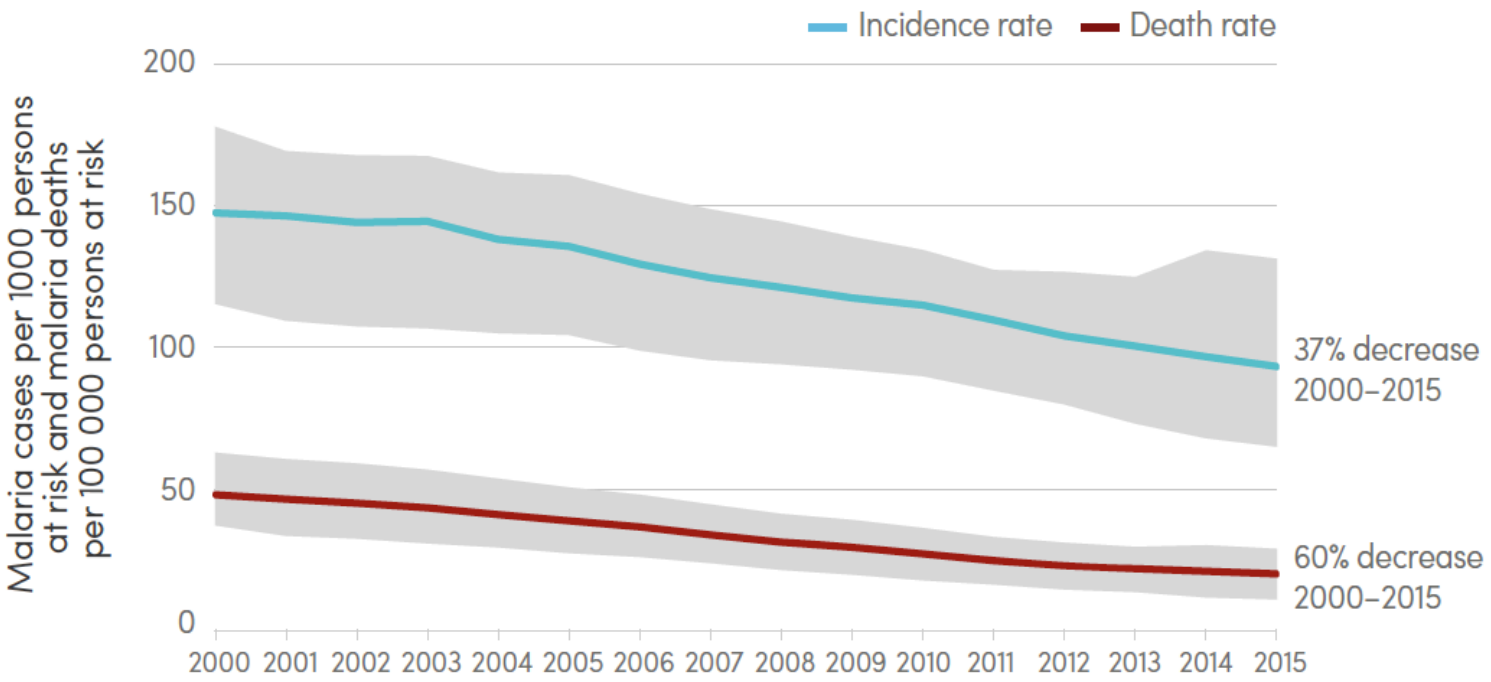
Plasmodium falciparum
Plasmodium vivax
Plasmodium ovale
Plasmodium malariae
Plasmodium knowlesi
....

Anophele mosquito species

Anopheles gambiae
Anopheles funestus
Anopheles arabiensis
Anopheles darlingi
...

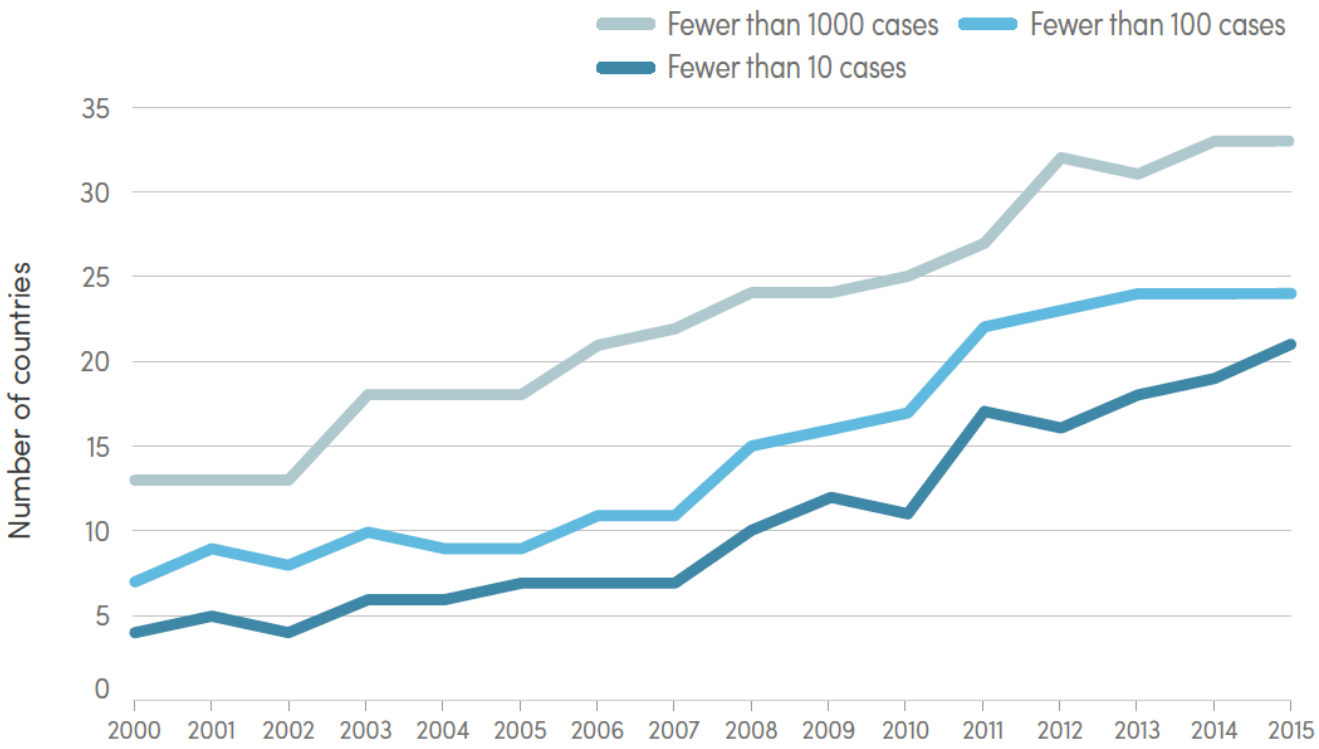
World malaria report 2015

Great achievements in the last 15 years



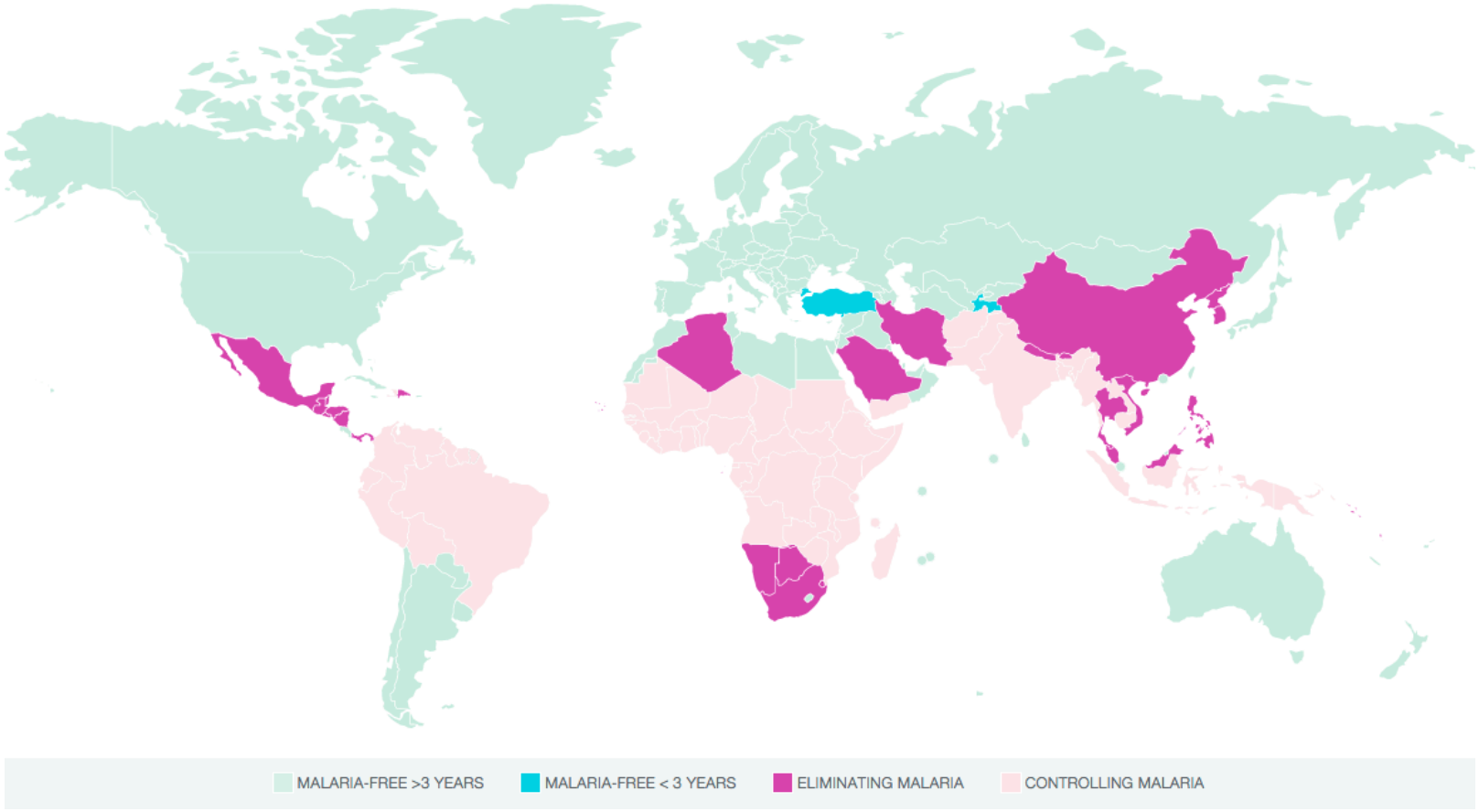
World malaria report 2015

Many countries are on route to malaria elimination



World malaria report 2015

These countries are spread over the world



Source: http://www.shrinkingthemalariamap.org/elimination-progress

How to assess progress of a country towards malaria elimination ?

Passive detection of infection

Annual parasite index
Number of official cases reported in a region /year

Problem: asymptomatic individuals are not included in this statistic.

Active detection of infection

Entomological inoculation rate
Number of infectious mosquito bites / person /year

Parasite rate
Proportion of infected individuals in active surveillance studies

Problems with measures based on current infection

Technical

- Sensitivity and specificity of diagnostic test

Variation in transmission

- Time of sampling is important when transmission is seasonal

Logistics

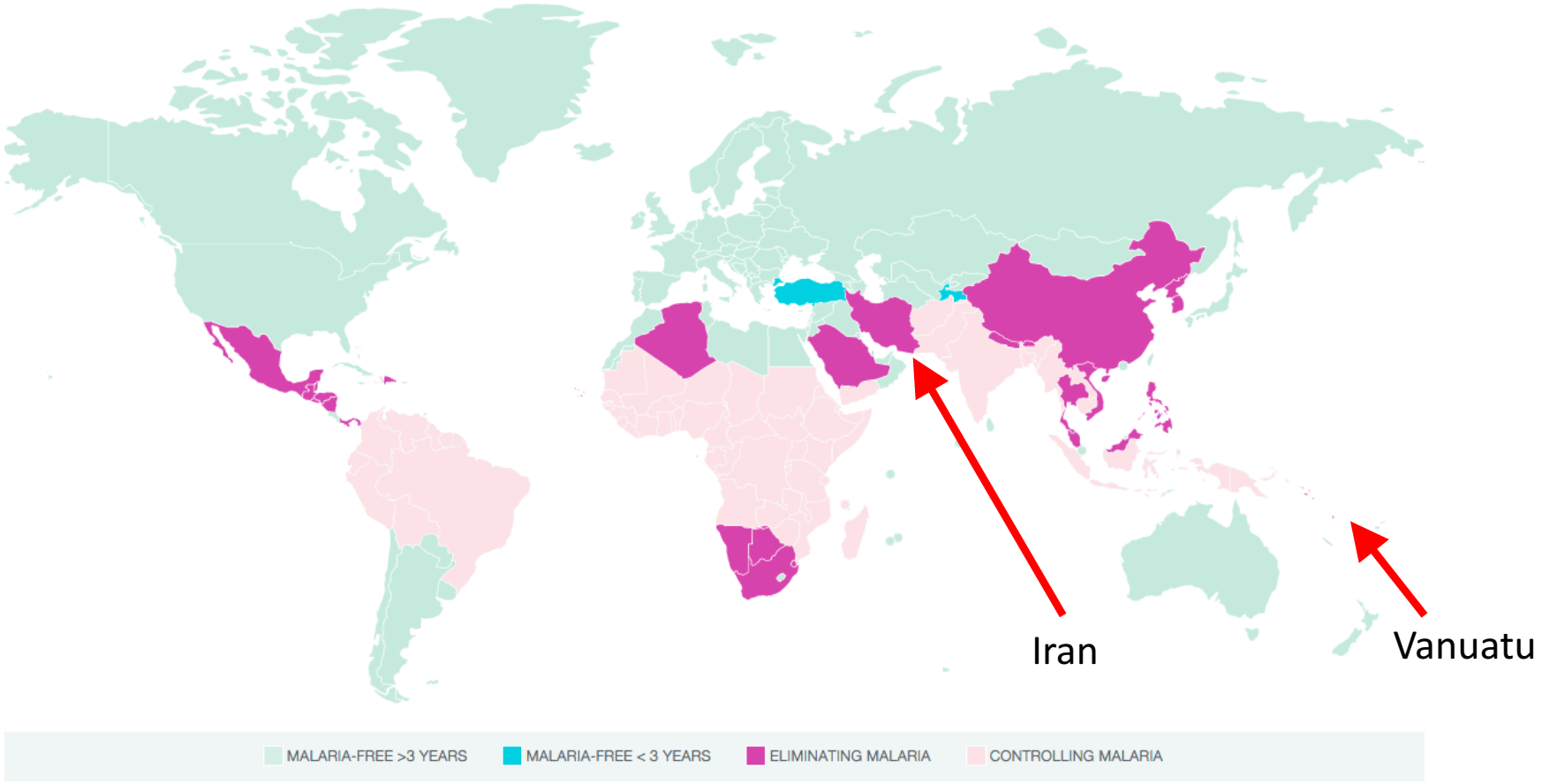
- Requirement of large sample sizes to obtain good estimation precision

Statistical

- Estimation methods do not usually work well close to ‘zero’

Illustration of common statistical problems

Data from two malaria elimination settings



Source: http://www.shrinkingthemalariamap.org/elimination-progress

What is the uncertainty association with each study?

Malaria infection	Iran	Vanuatu
Sample size (n)	1500	3009
Positive to any diagnostic test	0	0
Parasite rate (%)	0.000	0.000

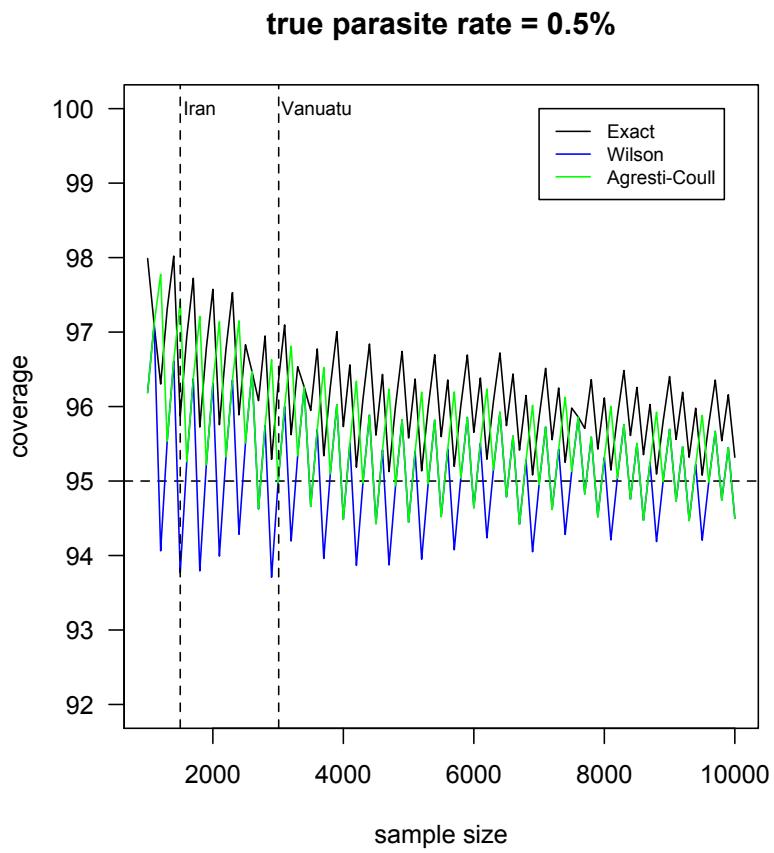
Iran: Zakeri et al. Malar J 15:382, 2016.
Vanuatu: Chan et al. Epidemiol Infect 145:41-45, 2017.

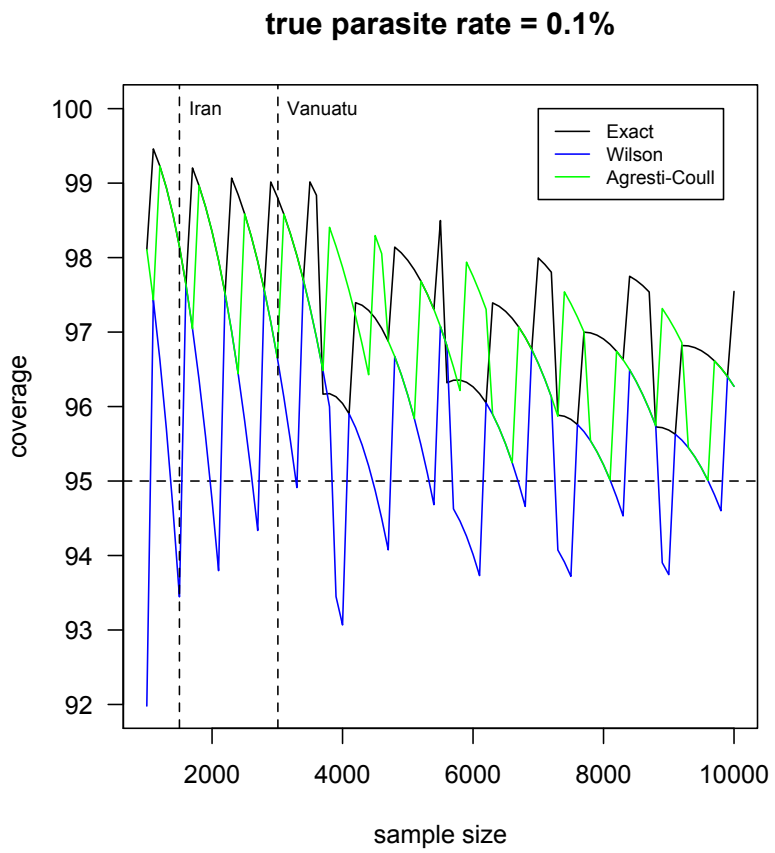
95% Conf/Cred. intervals for parasite rate (%)

Study	Method	Lower bound	Upper bound	
Iran	Wald	0.000	0.000	Degeneracy
	Exact	0.000	0.246	
	Wilson	0.000	0.247	
	Agresti-Coull	-0.005	0.298	Overshooting
	Bayesian (Jeffreys)	0.000	0.128	Reduced uncertainty
	Bayesian (Uniform)	0.000	0.199	
Vanuatu	Wald	0.000	0.000	Degeneracy
	Exact	0.000	0.123	
	Wilson	0.000	0.128	
	Agresti-Coull	-0.026	0.153	Overshooting
	Bayesian (Jeffreys)	0.000	0.064	Reduced uncertainty
	Bayesian (Uniform)	0.000	0.099	

Sepulveda & Drakeley (2016). Proceedings of the Portuguese Society for Statistics.

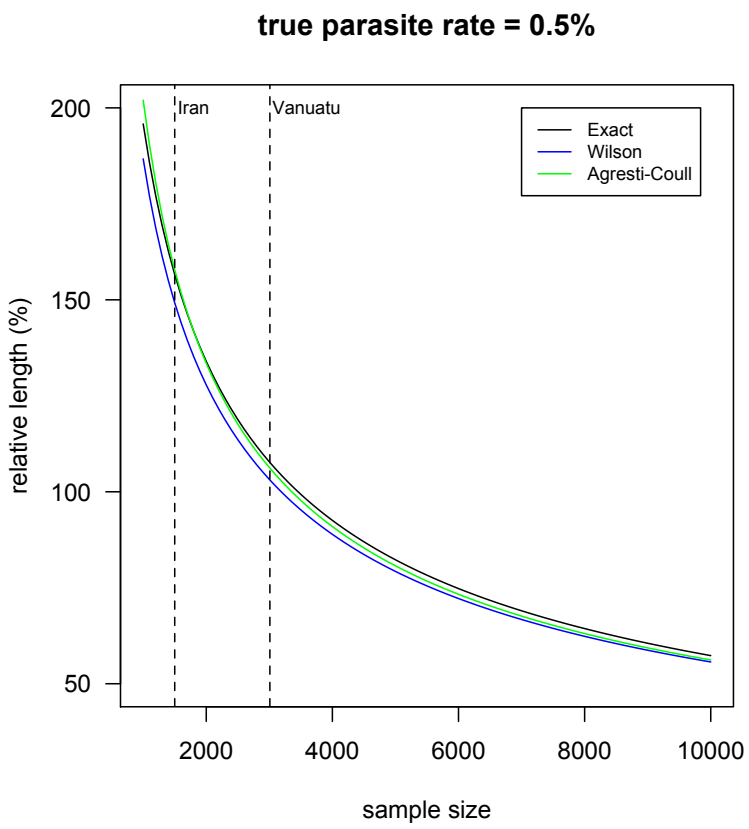
Approximate coverage at 95%

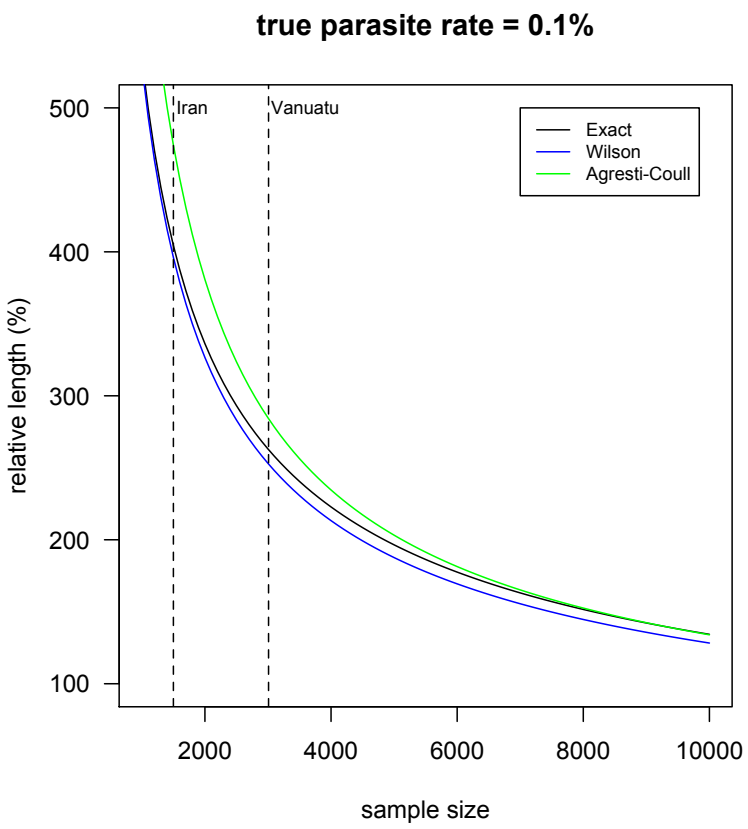




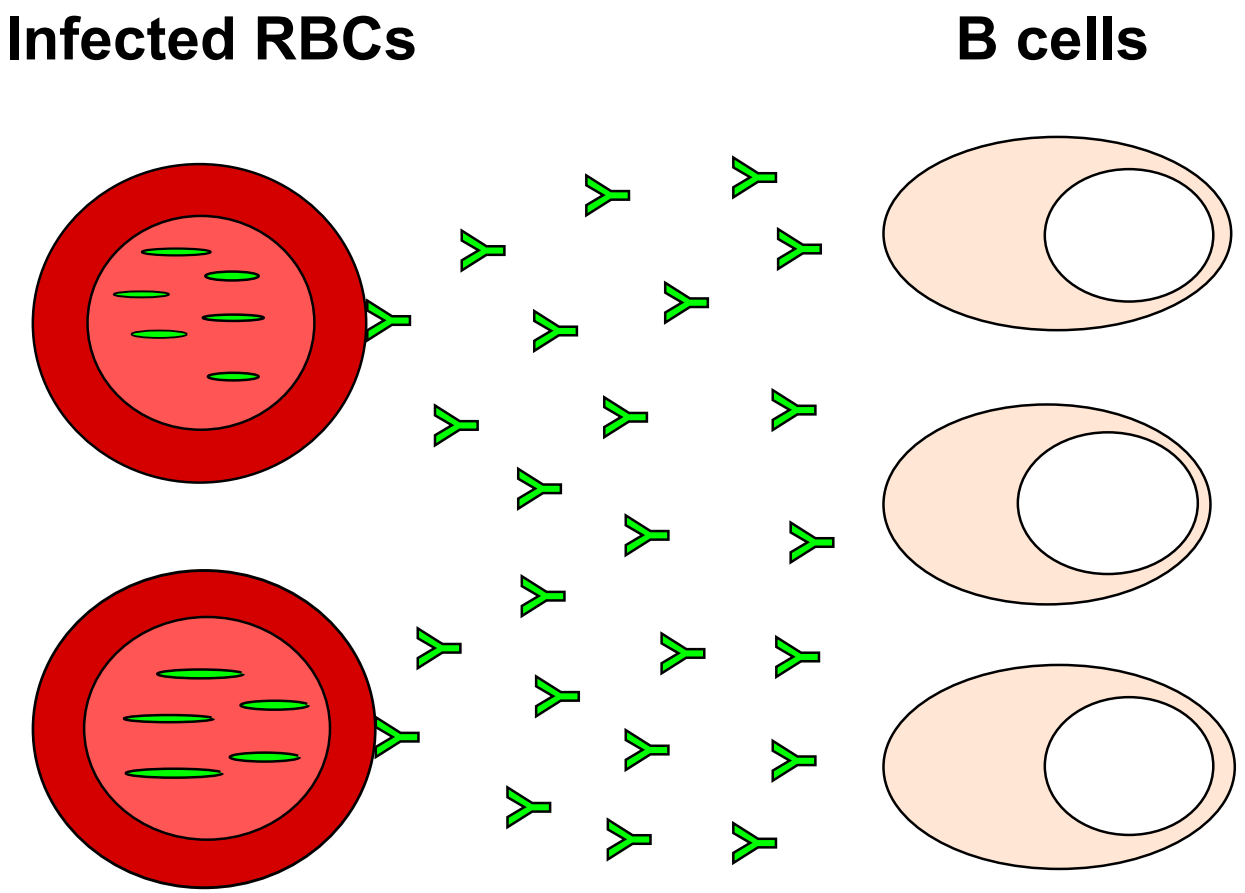
Sepulveda & Drakeley (2016). Proceedings of the Portuguese Society for Statistics.

Approximate length at 95%





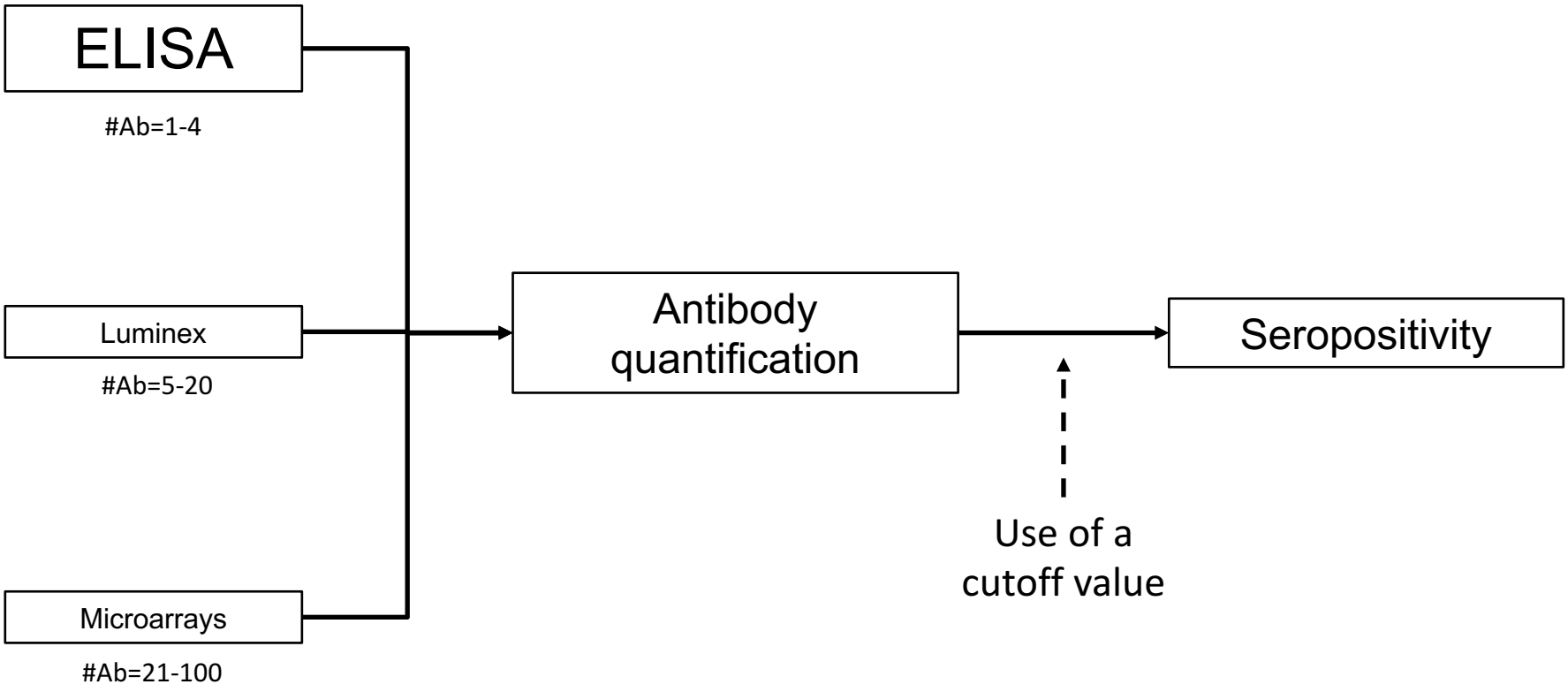
Time to be smart



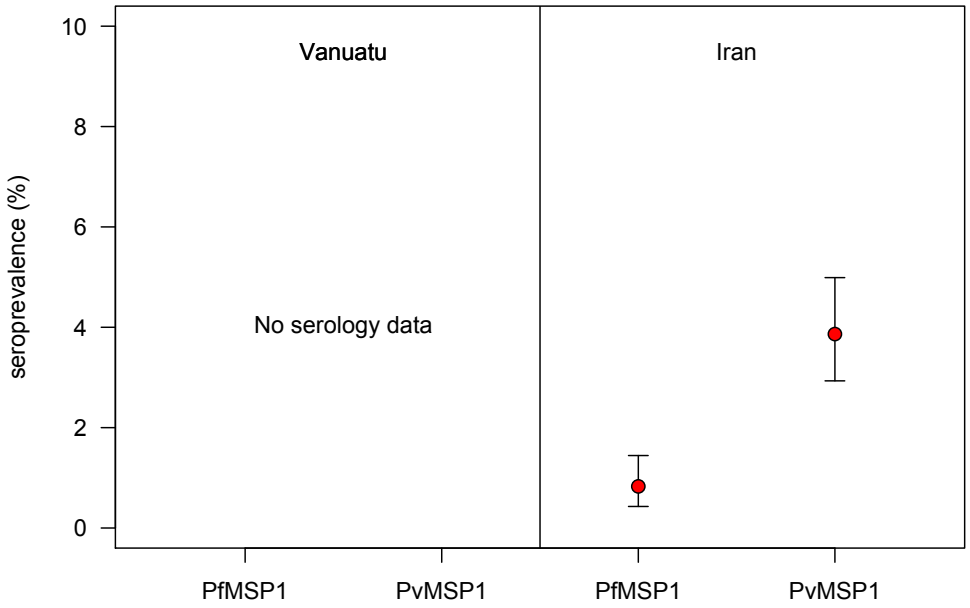
B cells produce antibodies specific to malaria antigens

Serology data analysis

Commonly studied antigens:
AMA-1 – Apical membrane antigen-1
MSP-1 – Membrane surface protein 1



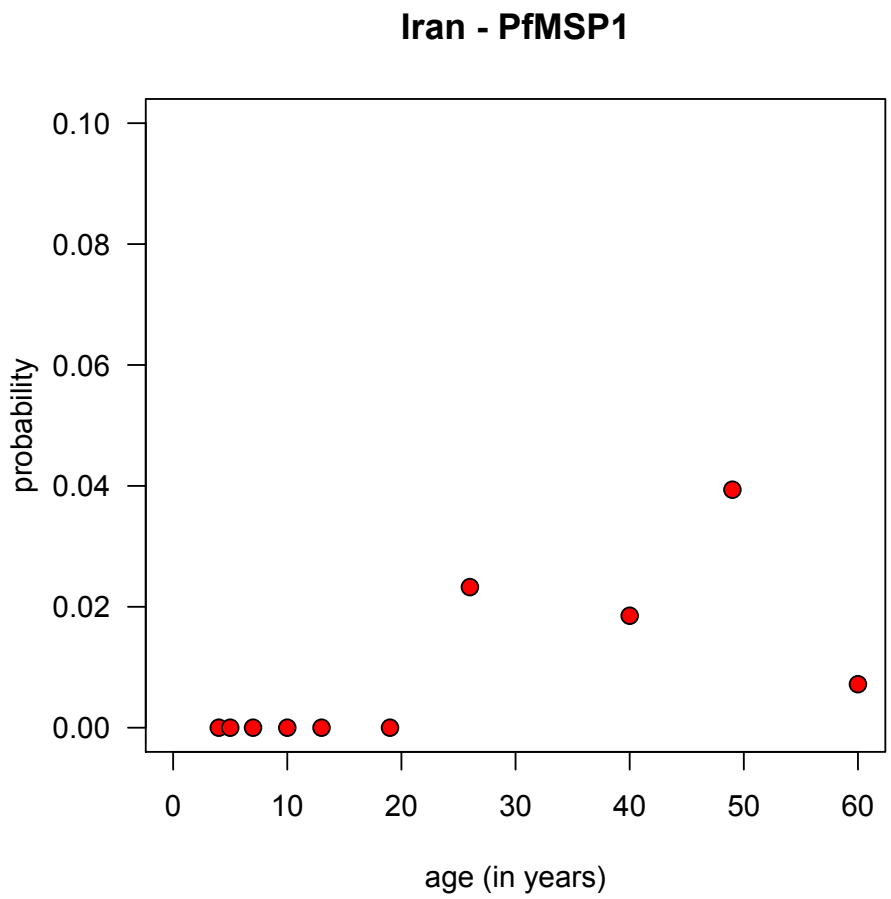
Seroprevalence as a measure of malaria exposure

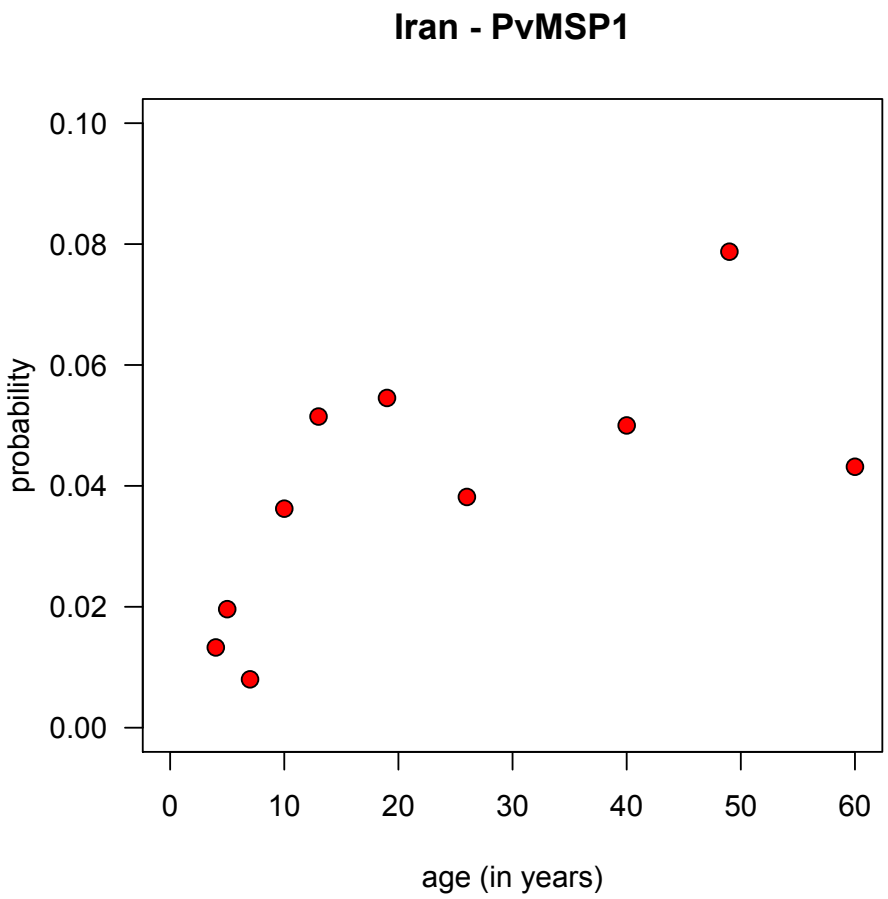


Last 10 years history	Estimate
Any malaria	16.8%
Pf malaria	1.5%
Pv malaria	6.0%

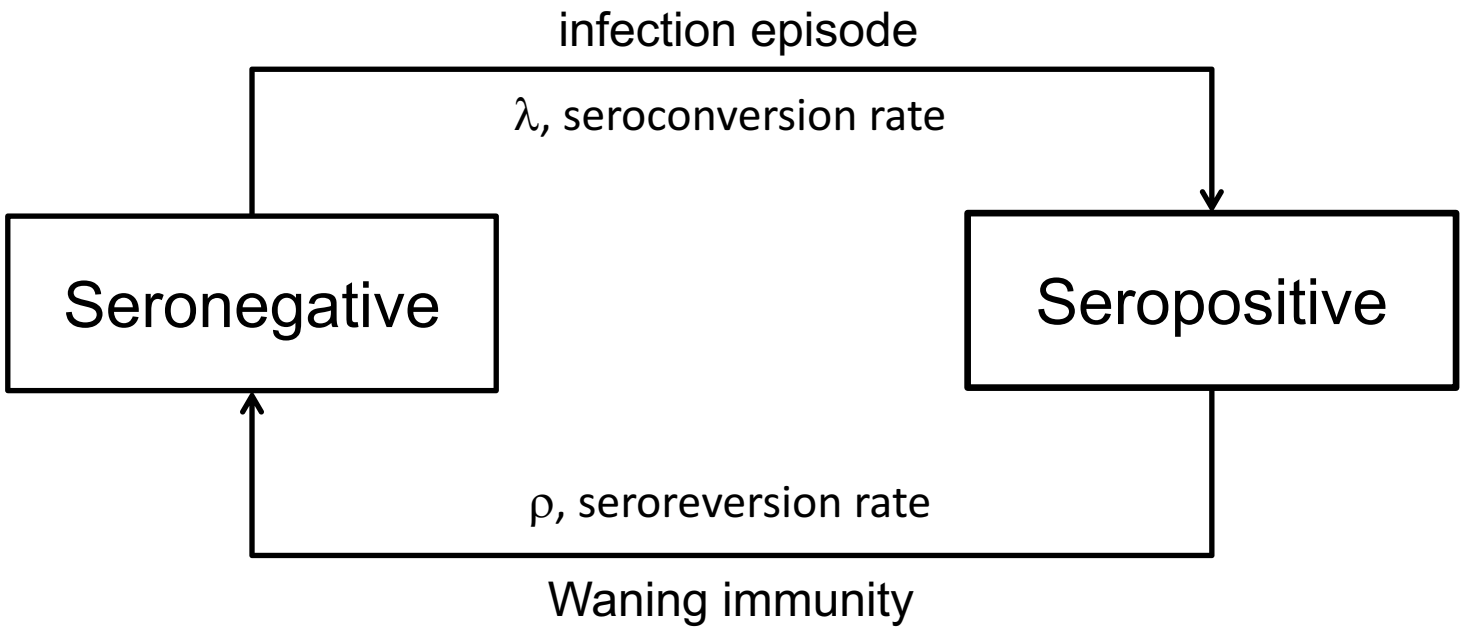
Zakeri et al. Malar J 15:382, 2016.

Seroprevalence is a function of age



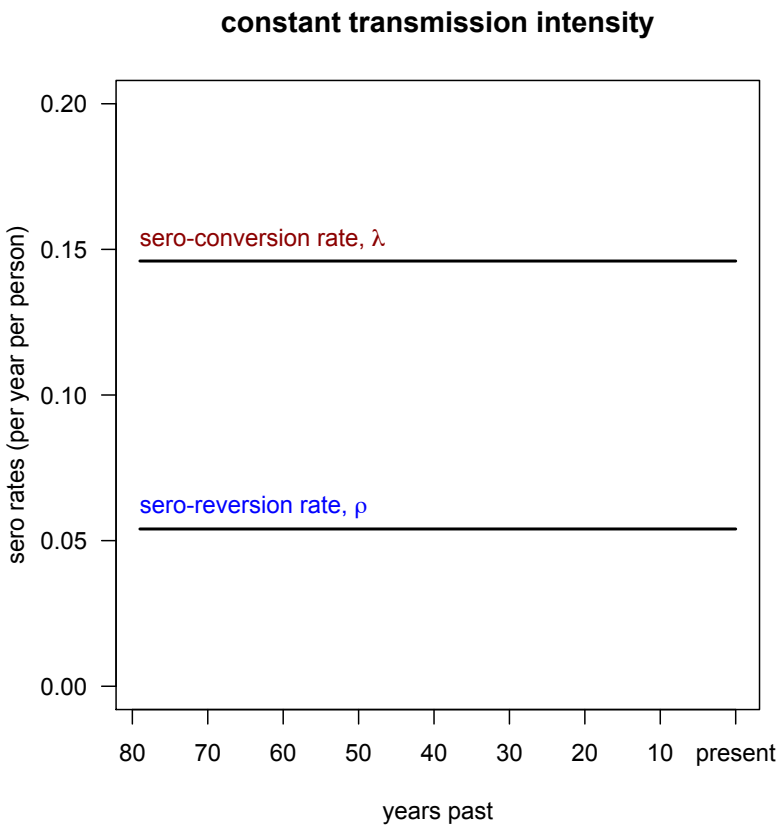


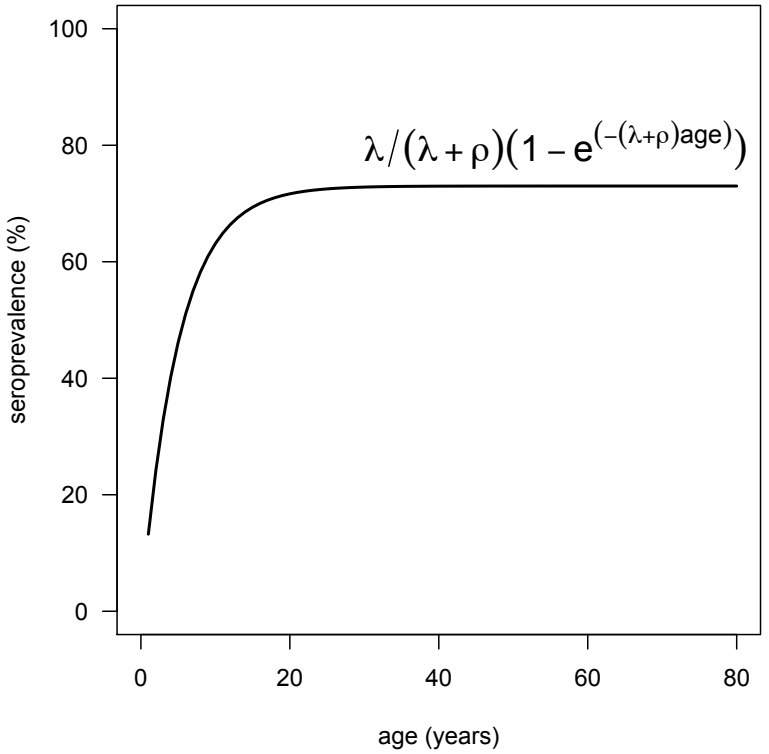
The reversible catalytic models



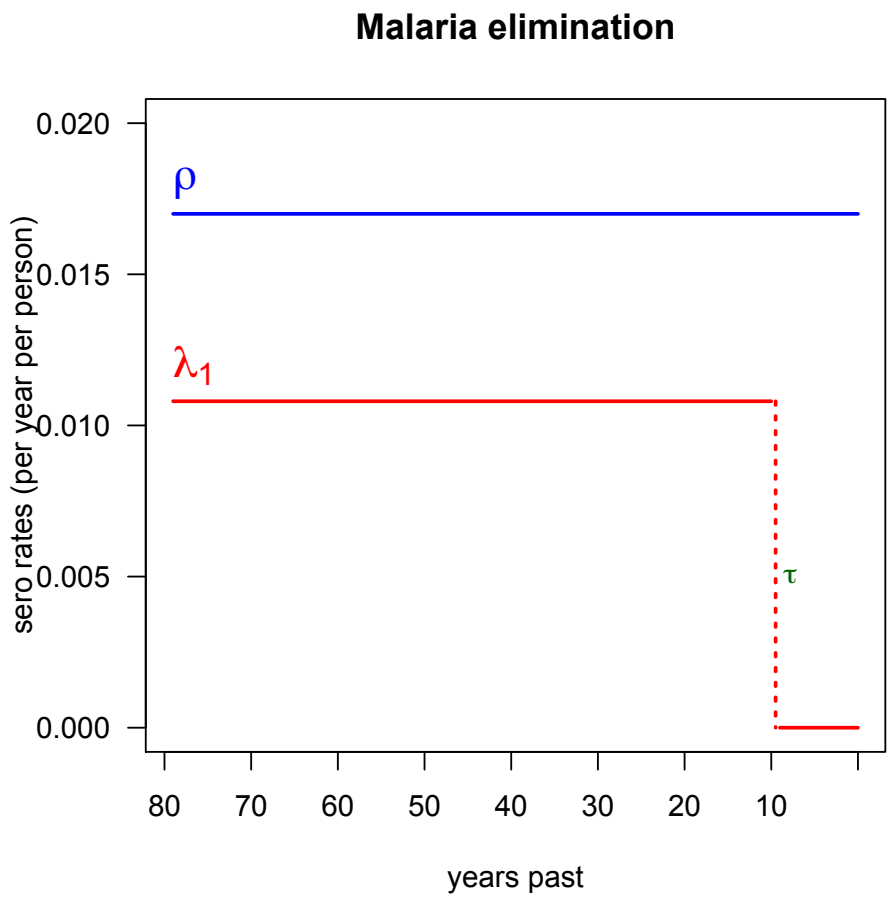
Markov chain in continuous time

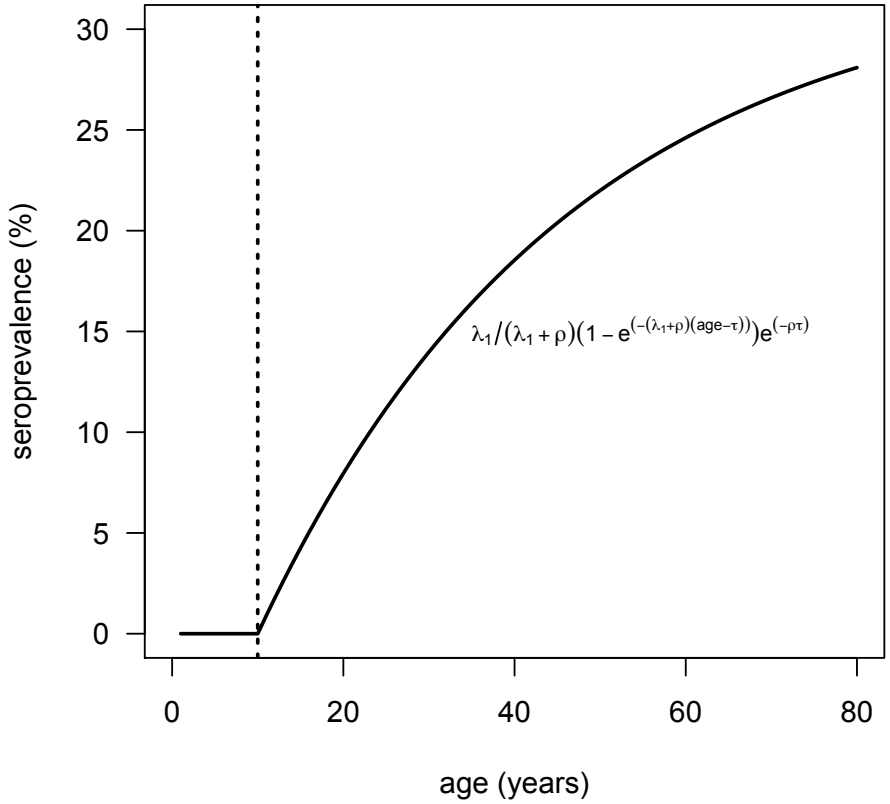
The most simple situation



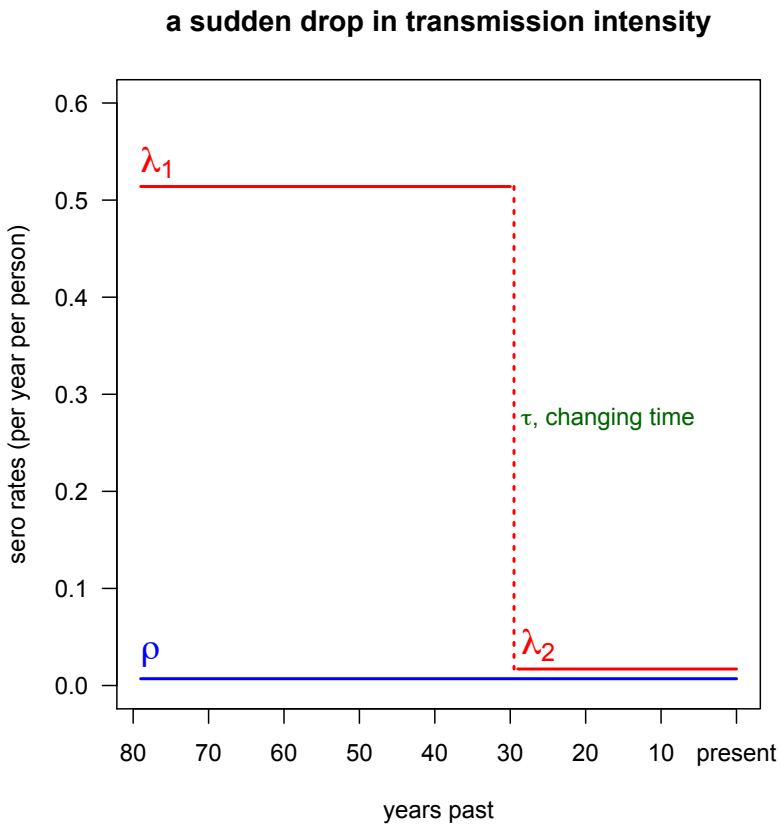


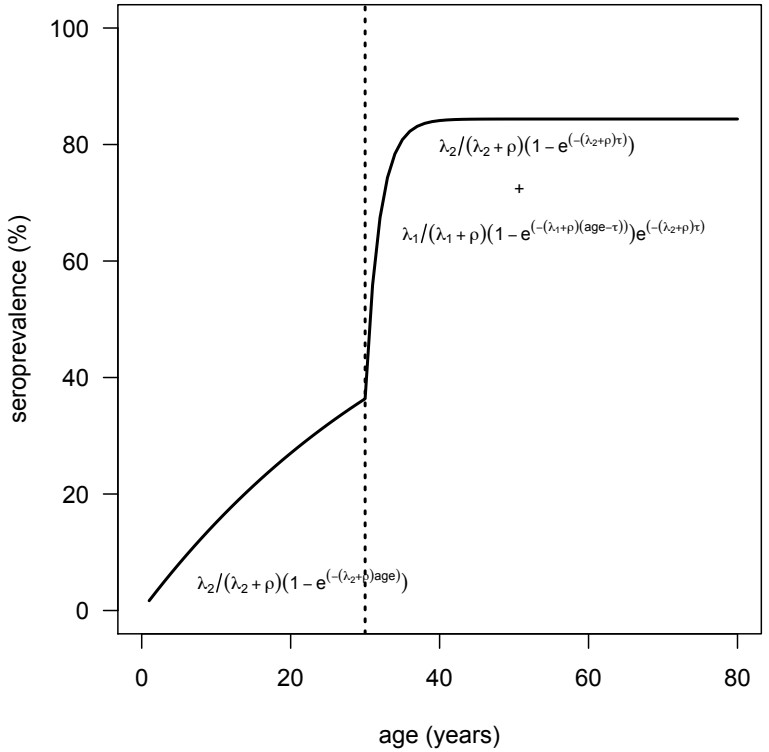
Modelling malaria elimination



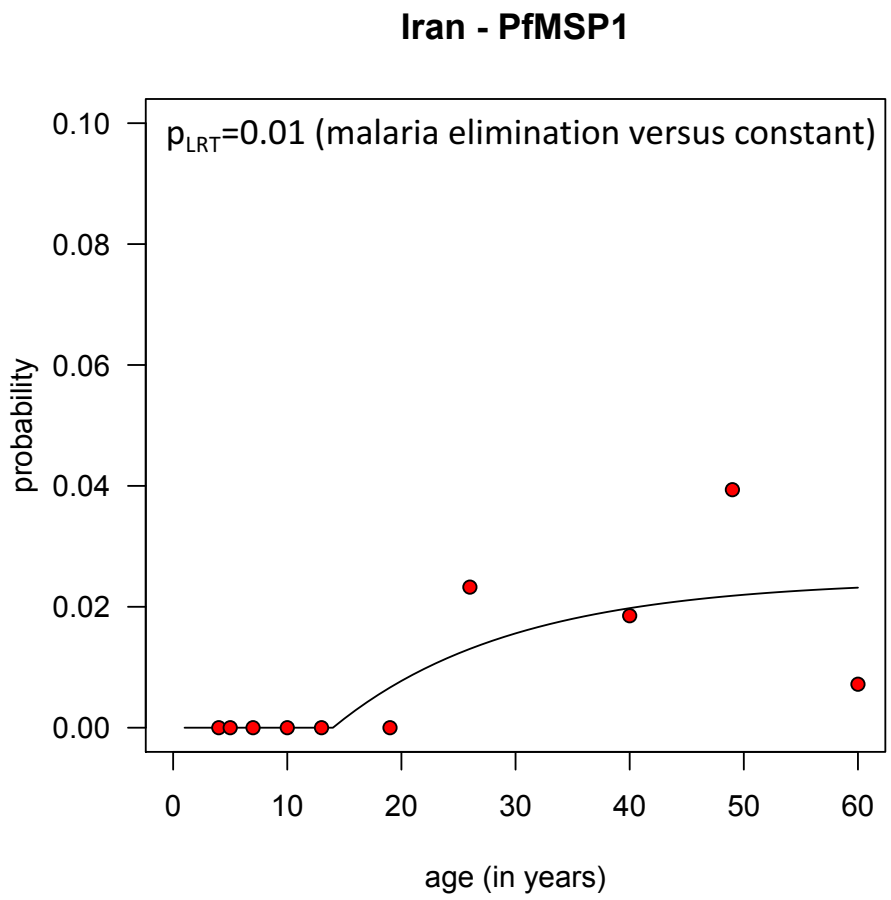


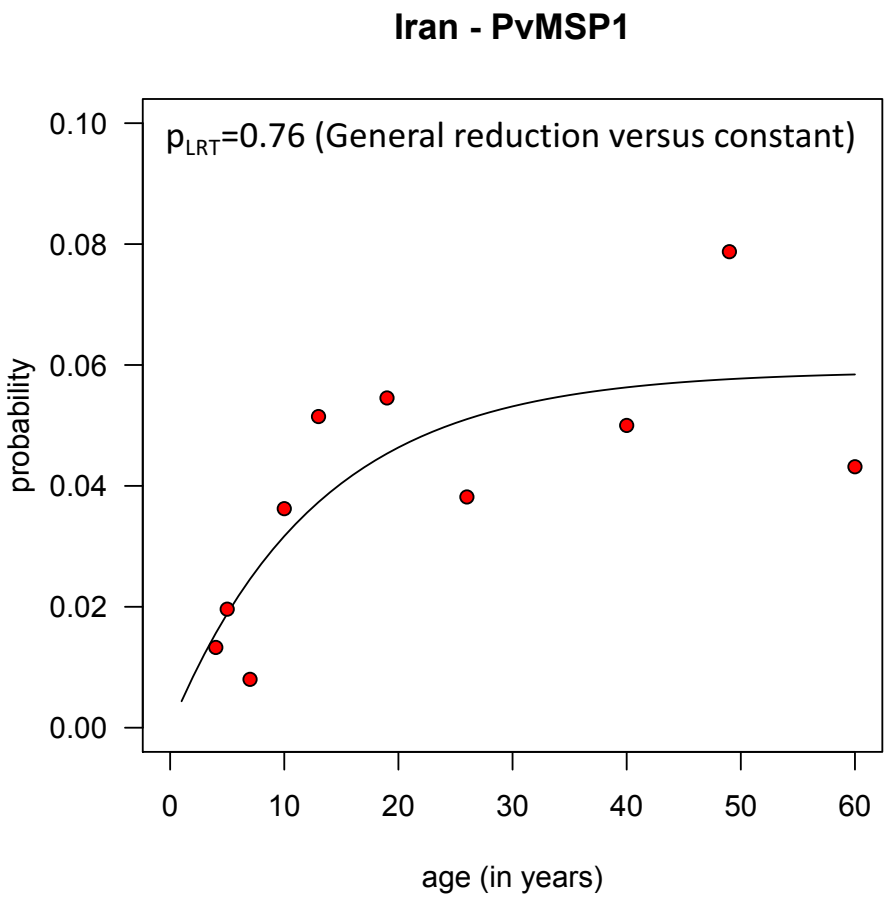
General situation





Best fits for the Iran data

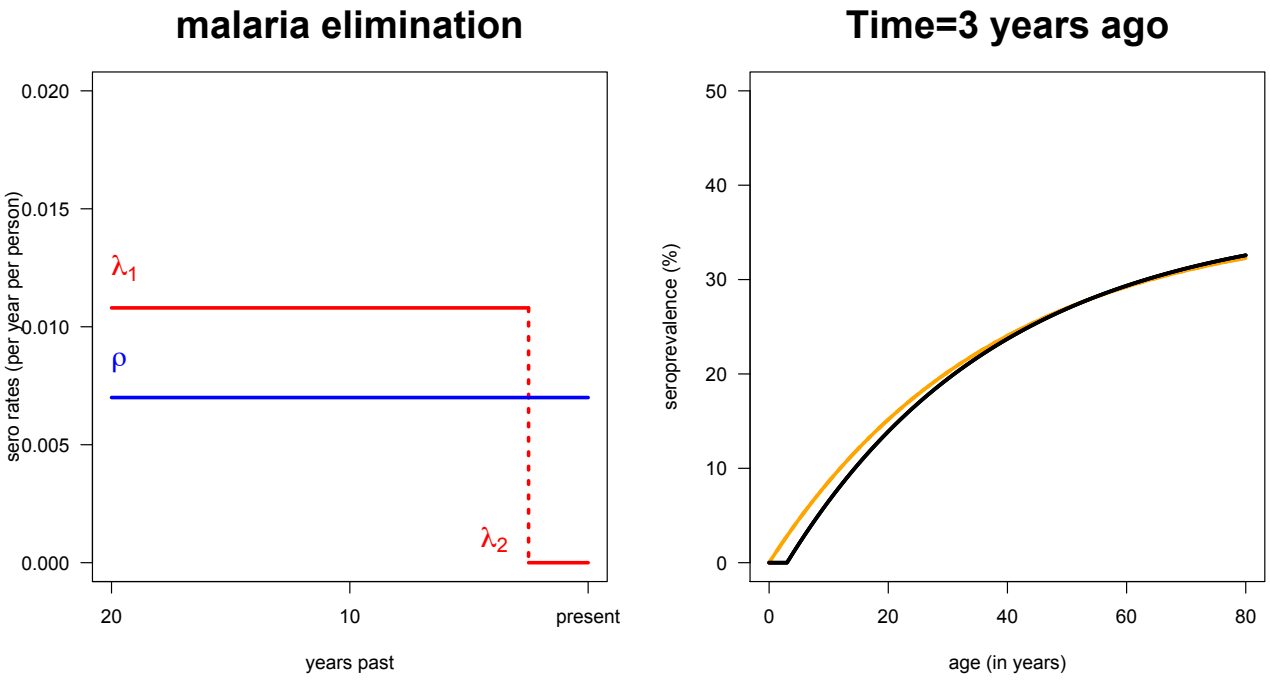




Challenge I:
The problem of model discrimination

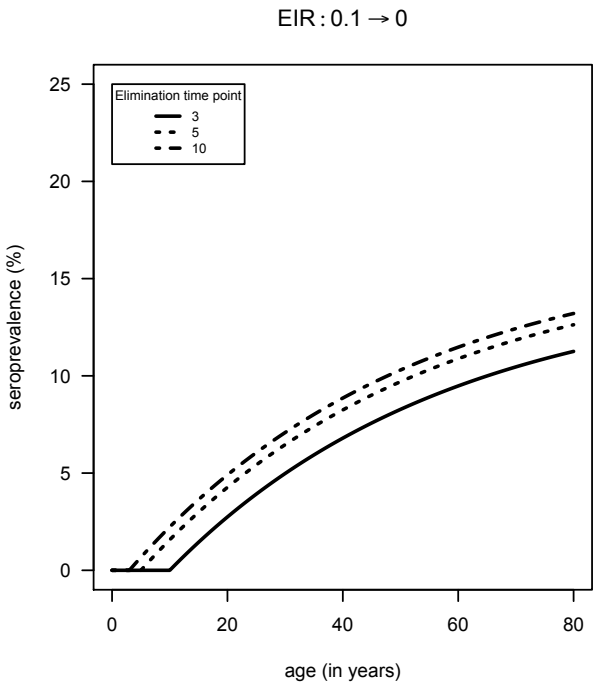
Concept

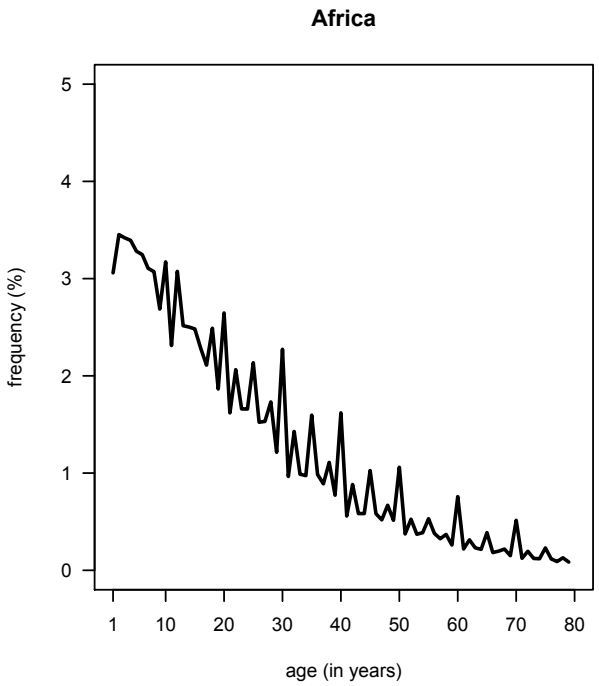
Short term effects of transmission interruption

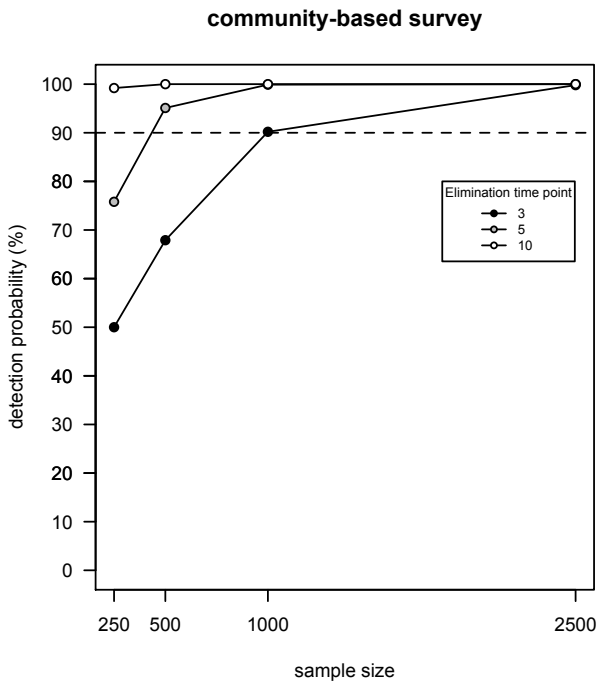


Find the optimal study design for discriminating each model

Sample size calculations for cross-sectional surveys



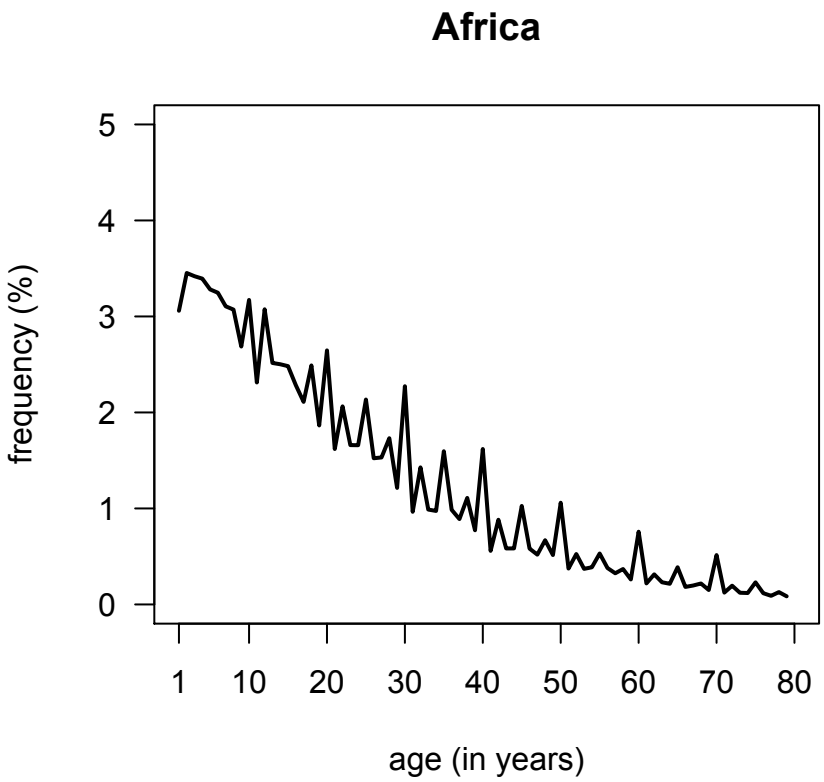


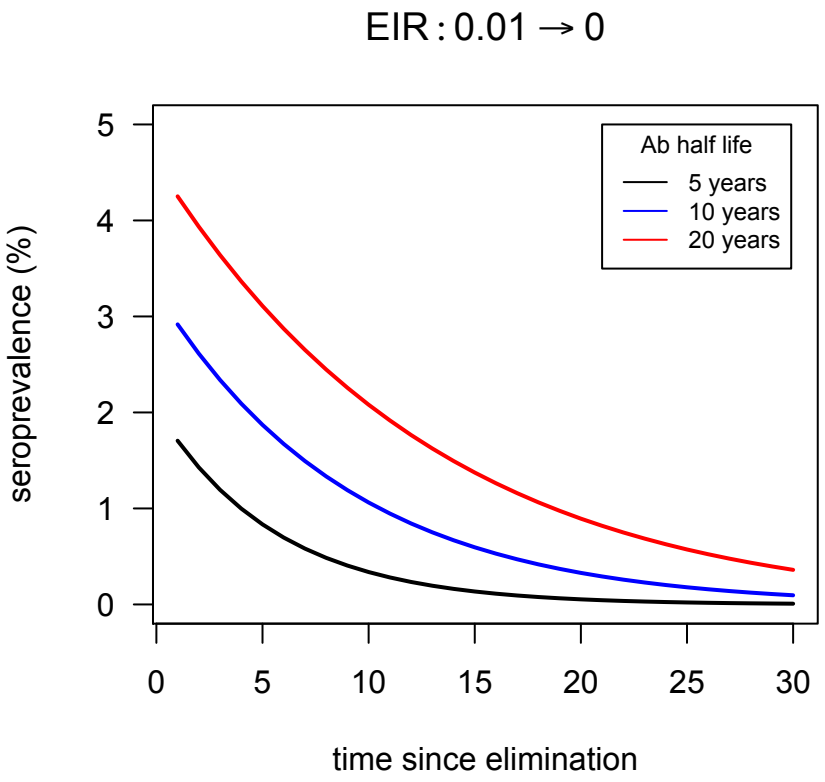


Sepúlveda et al (2015). Malaria J, 14, 529. Sepúlveda et al (2015). J Immunol Res, 2015, 738030.

Challenge II:
The problem of detecting seropositive individuals

Concept

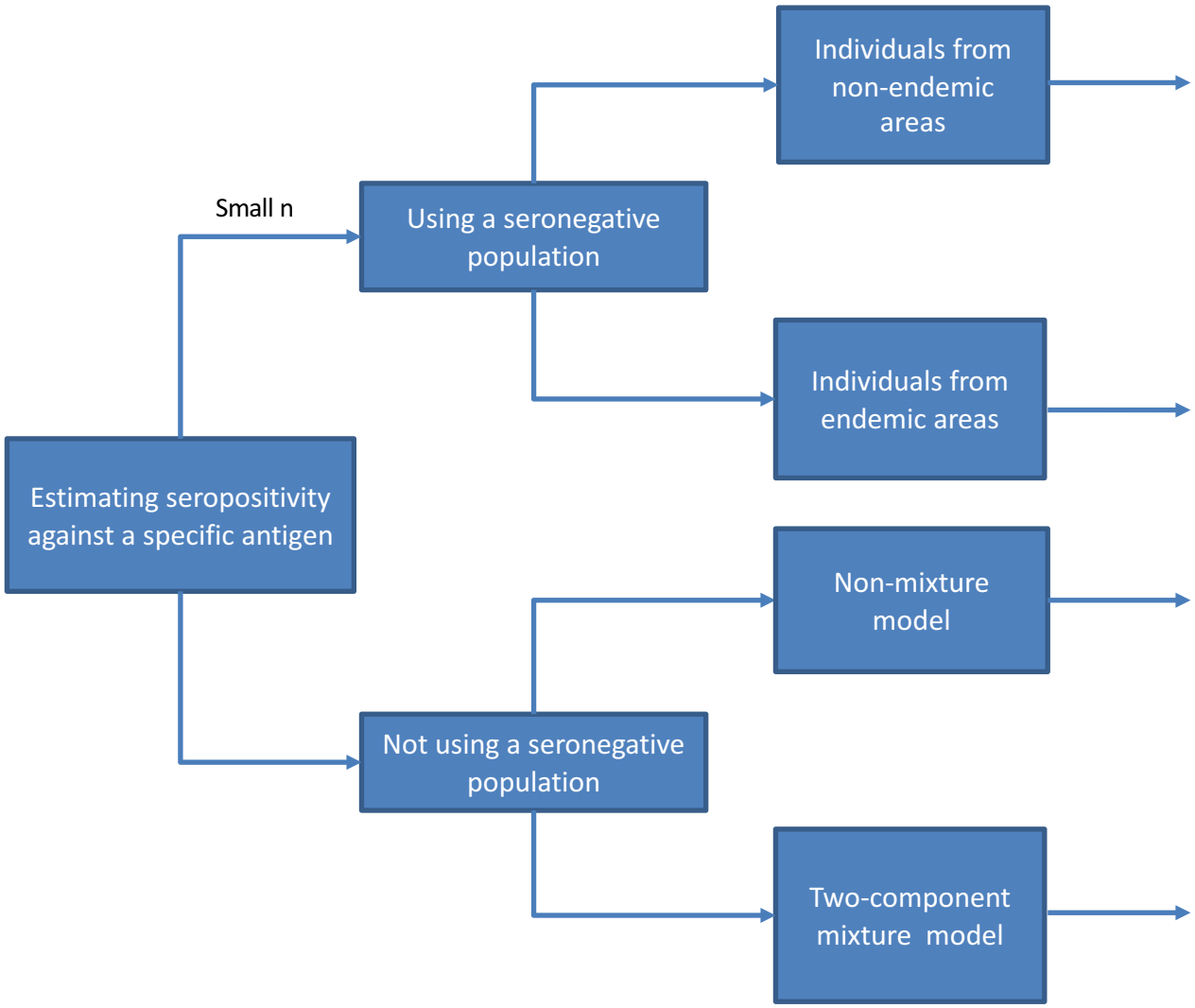




Statistical challenge(s):

How to accurately detect seropositive individuals over time?

Statistical approaches for seropositivity



Approach 1
Normal, Lognormal, Gamma, Weibull, Gumbel, Rayleigh, etc

Haiti (n=284) - Rogier et al (2015)
Normal
Haiti (n=60-70) - Arnold et al (2014)

Approach 2
Normal distribution
Sri Lanka (n=16) - Dewasurendra et al (2017)

Approach 3
Normal distribution
Haiti (n=814) Weppelmen et al, 2016; von Fricken et al, 2014

Approach 4
Mixture of two Gaussian distributions
Iran (n=1500) - Zakeri et al (2016)
Aneityum (n=516) - Cook et al (2010)
Haiti (n=580) - Rogier et al, 2015

Defining seropositivity using a reference sample

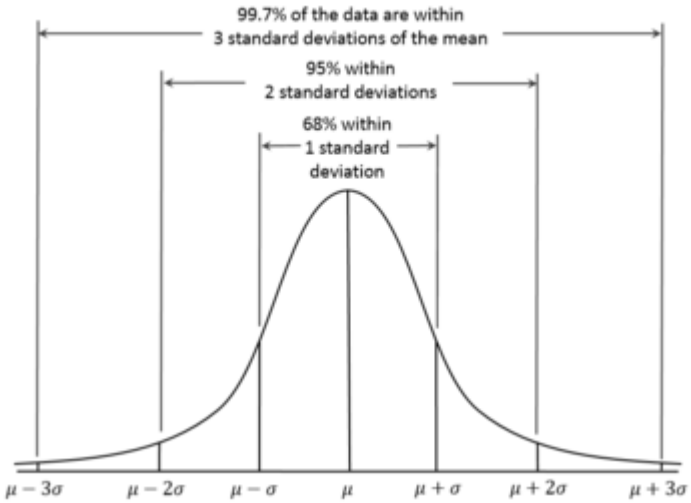
Theoretical Cutoff

$\mu_0+i*\sigma_0$

μ_0 = mean of the seronegative population
σ_0 = standard deviation of the seronegative population

i=2 → P(X>cutoff)=0.0228
i=3 → P(X>cutoff)=0.0013
i=5 → P(X>cutoff)=$2.9x10^{-7}$

These results hold true for any values of m_0 and s_0.



Estimated cutoff

m_0+i*s_0

m_0 = sampled mean of the seronegative population
s_0 = sampled standard deviation of the seronegative population

Current cutoff estimates are slightly biased

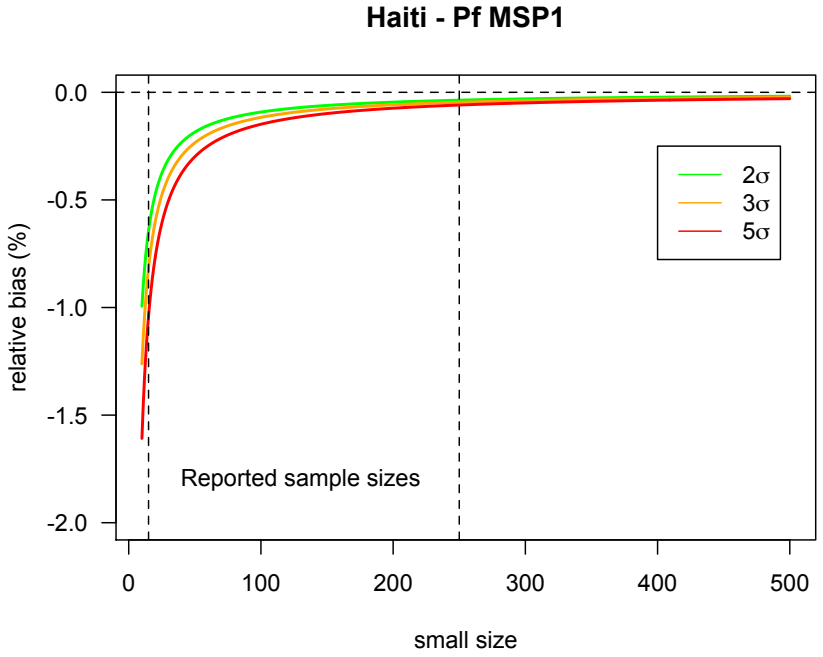
$$E[m_0 + i \times s_0] \neq \mu_0 + i \times \sigma_0$$

$$E[m_0 + i \times s_0^*] = \mu_0 + i \times \sigma_0$$

where

$$s_0^* = k_n s_0$$

$$k_n = \sqrt{\frac{n-1}{2}}\, e^{\ln \Gamma\left(\frac{n-1}{2}\right) - \ln \Gamma\left(\frac{n}{2}\right)}$$



Current estimated cutoffs do not have high precision

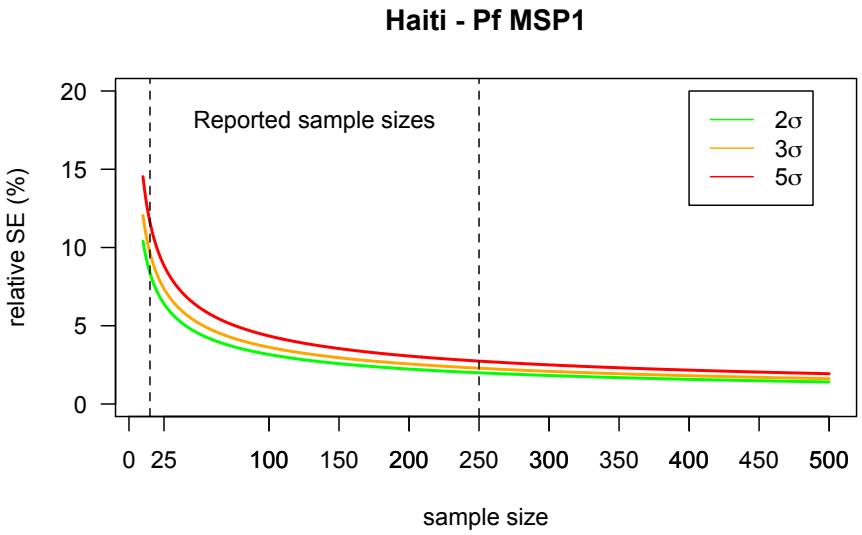
$$Var[m_0 + is_0^*] \approx Var[m_0] + i^2 Var[s_0^*]$$

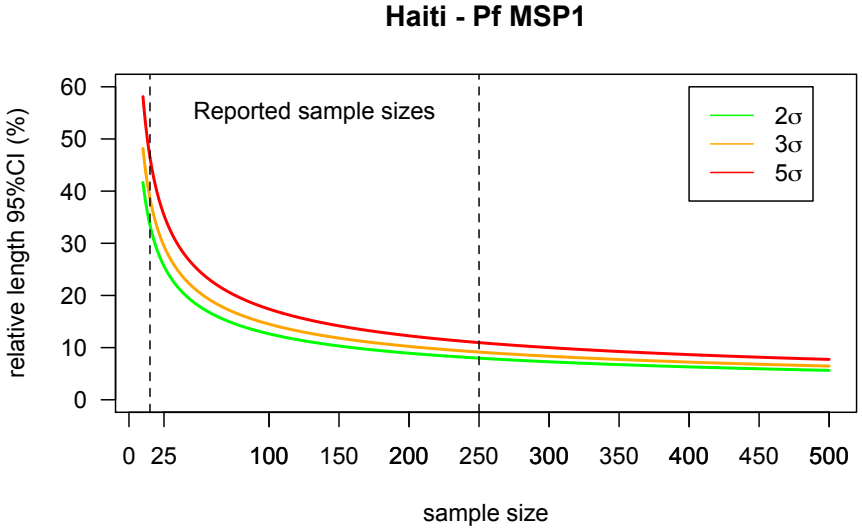
$$Var[m_0] = \frac{\sigma^2}{n}$$

$$Var[s_0^*] = \sigma^2 k_n^2 \frac{V_n}{n-1}$$

$$k_n = \sqrt{\frac{n-1}{2}} e^{\ln \Gamma\left(\frac{n-1}{2}\right) - \ln \Gamma\left(\frac{n}{2}\right)}$$

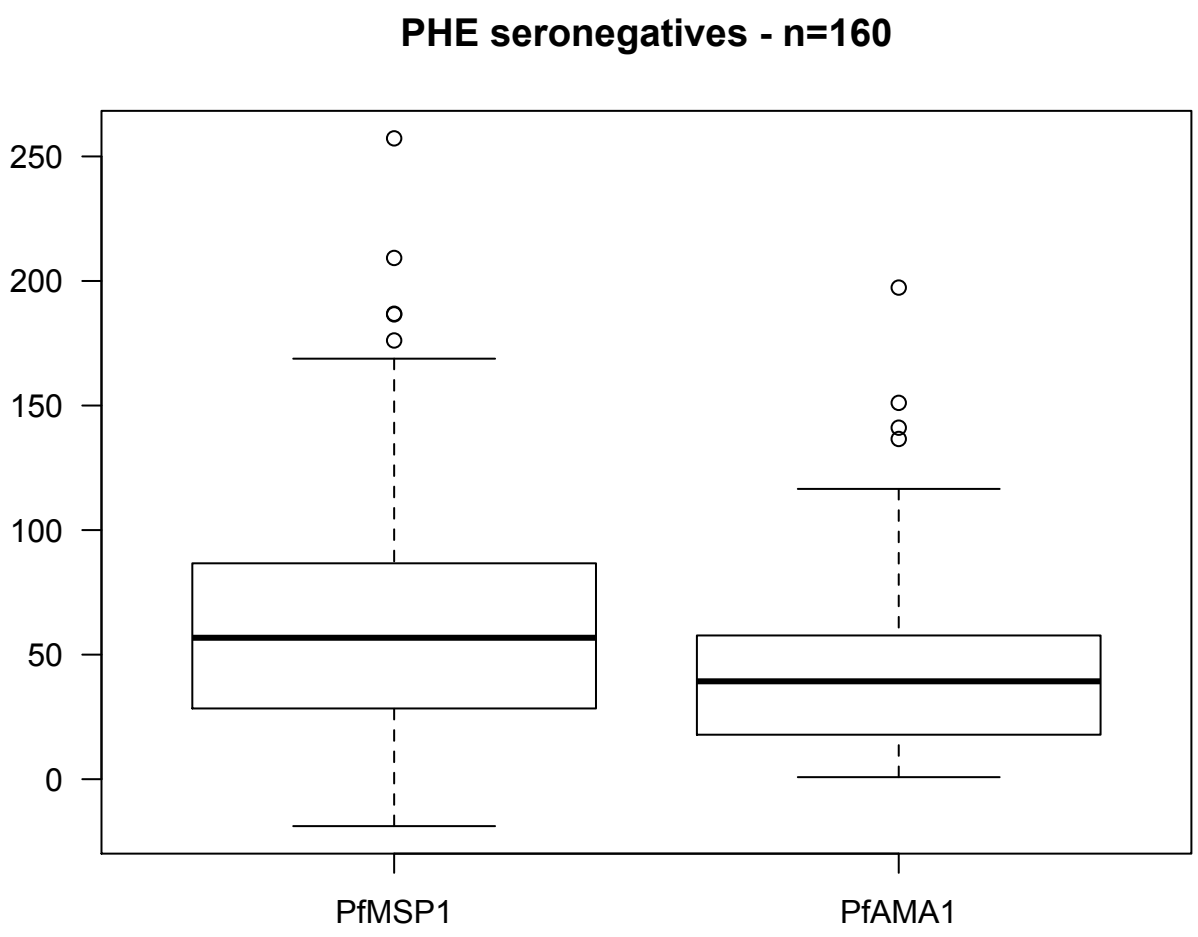
$$V_n = 2\left[\frac{n-1}{2} - \frac{\Gamma^2(n/2)}{\Gamma^2((n-1)/2)}\right]$$





Analyzing a UK seronegative sample

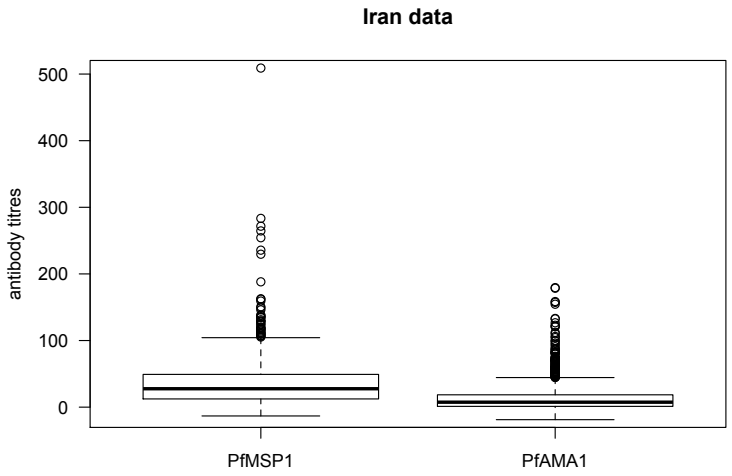
Public Health England



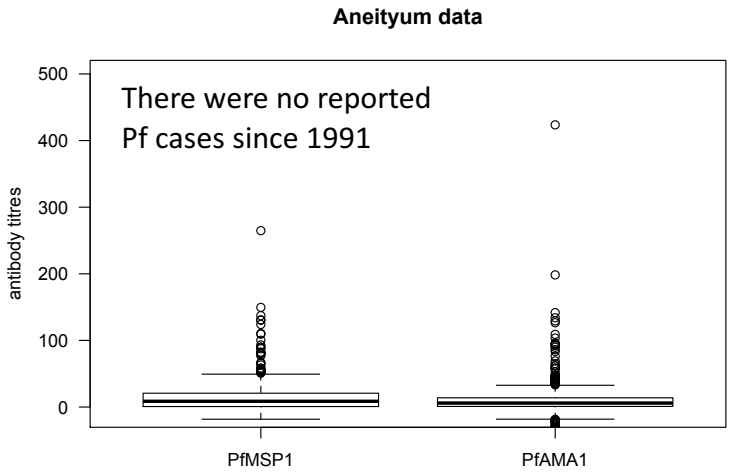
Variation in cutoff values from equally good models

Antigen	Distribution	P-value (gof)	2σ	3σ	5σ
Pf MSP1	Normal	0.174	157.4	205.1	300.5
	Skew Normal	0.665	174.5	243.8	387.8
	Non-central T	0.314	156.1	242.3	776.6
Pf AMA1	Normal	$<10^{-5}$	108.7	141.1	205.9
	Lognormal	$<10^{-3}$	169.1	382.5	1955.6
	Skew Normal	0.058	119.4	165.6	261.6
	Gamma	0.338	126.5	199.2	405.8
	Weibull	0.183	122.6	180.8	321.2

Seroprevalence of two elimination settings



Antigen	Distribution	2σ (2.3%)	3σ (0.1%)	5σ (<0.0001%)
Pf MSP1	Normal	0.8%	0.6%	0.1%
	Skew Normal	0.6%	0.4%	0.1%
	Non-central T	0.8%	0.4%	0.1%
Pf AMA1	Skew Normal	0.8%	0.1%	0.0%
	Gamma	0.6%	0.0%	0.0%
	Weibull	0.6%	0.0%	0.0%



Antigen	Distribution	2σ (2.3%)	3σ (0.1%)	5σ (<0.0001%)
Pf MSP1	Normal	0.2%	0.2%	0.0%
	Skew Normal	0.2%	0.2%	0.0%
	Non-central T	0.2%	0.2%	0.0%
Pf AMA1	Skew Normal	1.2%	0.4%	0.2%
	Gamma	1.2%	0.2%	0.2%
	Weibull	1.2%	0.4%	0.2%

Defining seropositivity not using a reference sample

Assume two latent Gaussian (or normal) populations for the data:

Seronegative (never or at least not recently exposed):

$Ab^{-} \sim N(\mu_0, \sigma_0)$

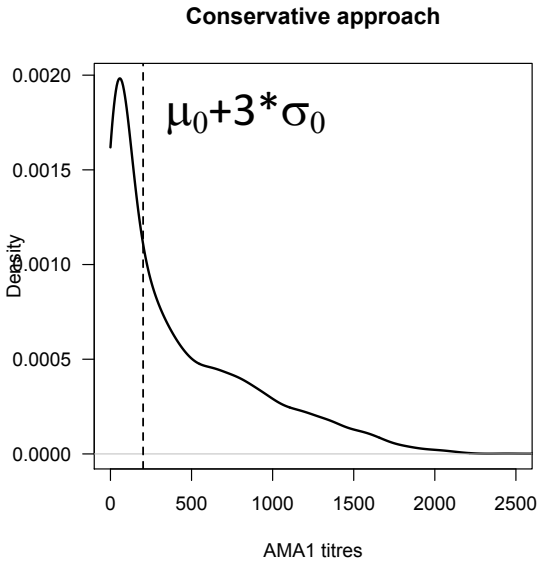
Seropositive (recently exposed):

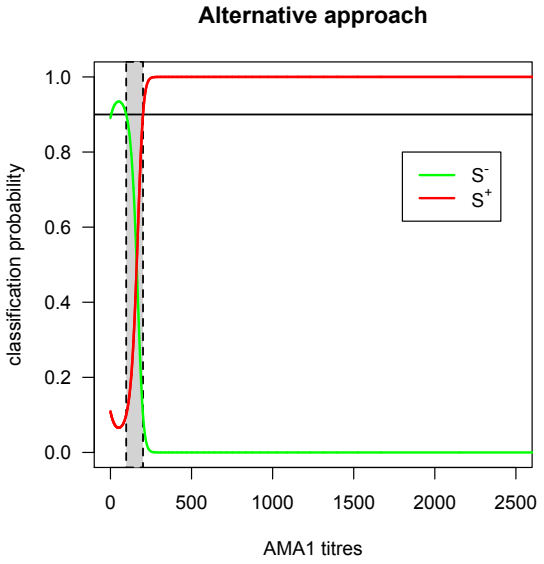
$Ab^{+} \sim N(\mu_1, \sigma_1)$

Estimation:

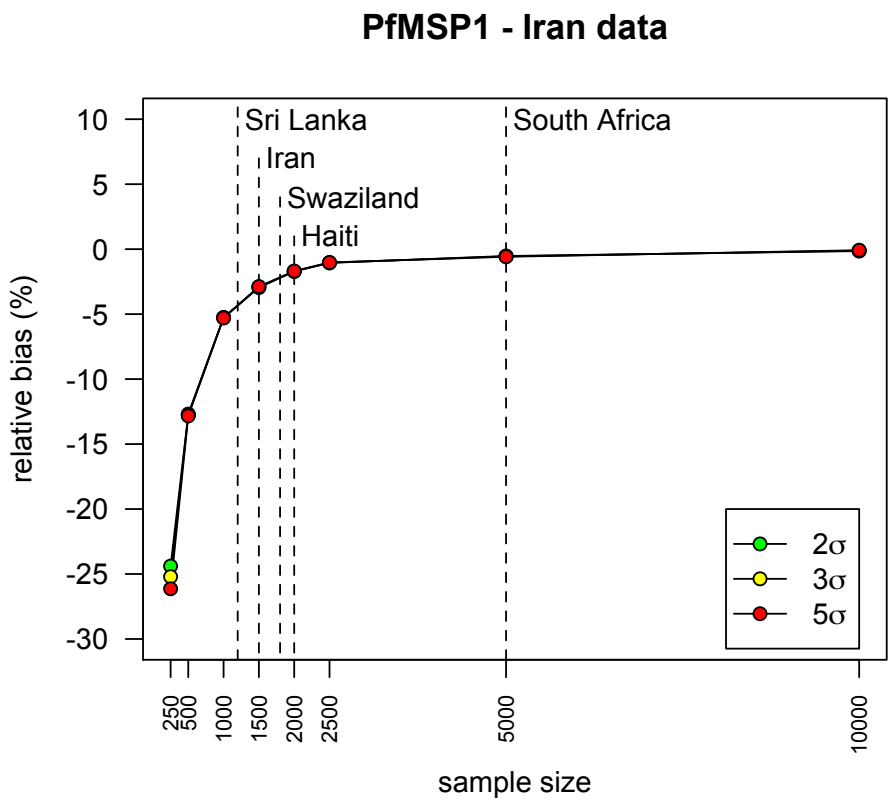
Expectation-maximization algorithm

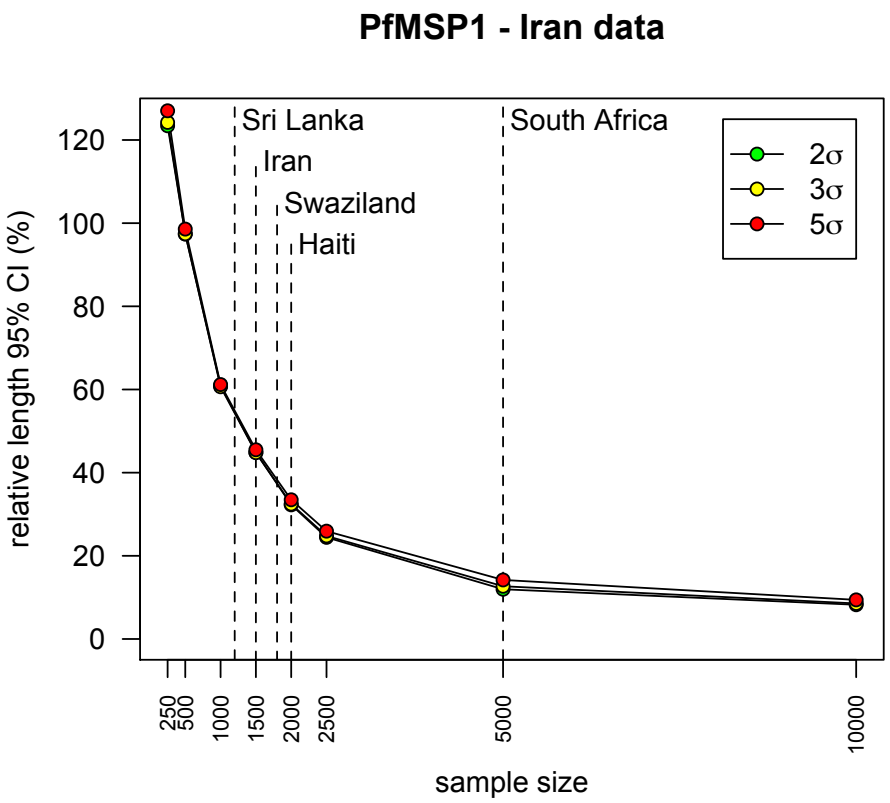
Calculate threshold dividing the seronegative and seropositive populations.





Cutoff values are biases and low precision





Conclusions

Estimating parasite rate near zero leads to poor performance of the classical inference methods.

Serological data can be used to estimate the degree of malaria exposure and detect malaria elimination.

- Having the correct sample size is essential in malaria elimination settings.
- Reduce bias, increase precision and discriminate models

- Normal or lognormal distributions are the statistical dogma whilst analyzing the seronegative population. However, other models can also describe the data well, implying different seropositivity cutoff values.

Ongoing work

Box-Cox transformation

Find the data transformation that best describes a Gaussian distribution or any other probability for the data.

Generalized Tukey's distribution

Very flexible distribution that can describe data from the following probability distribution: Gaussian, Lognormal, Uniform, Weibull, etc.

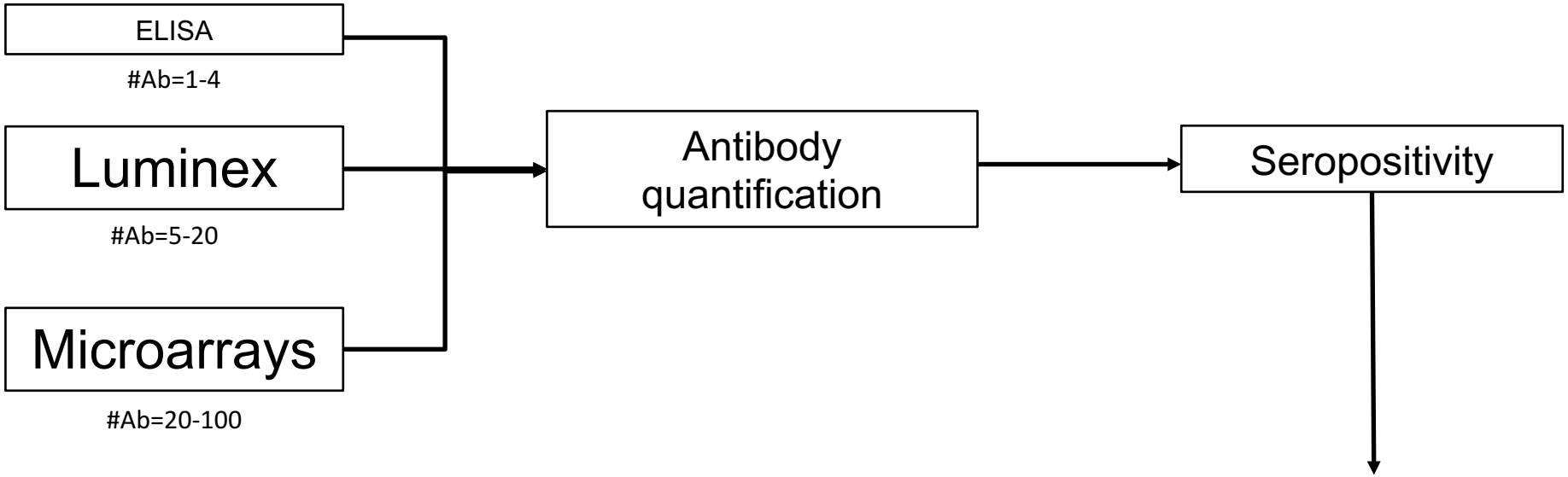
Harrell-Davis quantile estimator

Non-parametric estimator to calculate the cutoff values based on the quantiles of the observed data.

Bayes model averaging

Fit different probability models and average the respective cutoff values according to the plausibility of each model (unbiased estimates if the true model is included in the analysis). Computationally intensive.

The future



Seroprevalence	Half-life of antibodies			
Sites	**<1 year**	**1-5 years**	**6-10 years**	**>10 years**
A	5%	15%	40%	60%
B	0%	4%	14%	20%
C	0%	0%	5%	10%

Summer project of Steve Tate, Med Stat MSc student

Acknowledgements

LSHTM, UK

Chris Drakeley et al
Lotus van Hoogen (Iran and PHE data)
Jackie Cook (Aneityum data)

University of Lisbon, Portugal
Carlos Daniel Paulino

Univ. Santiago of Campostela, Spain
Jose Ameijeiras-Alonso
Rosa Crujeiras

Riga University, Latvia
Maksims Čistjakovs
Modra Murovska

Funding:

wellcome trust

BILL & MELINDA GATES foundation

MALARIA ZERO
THE ALLIANCE FOR A MALARIA-FREE HAITI

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