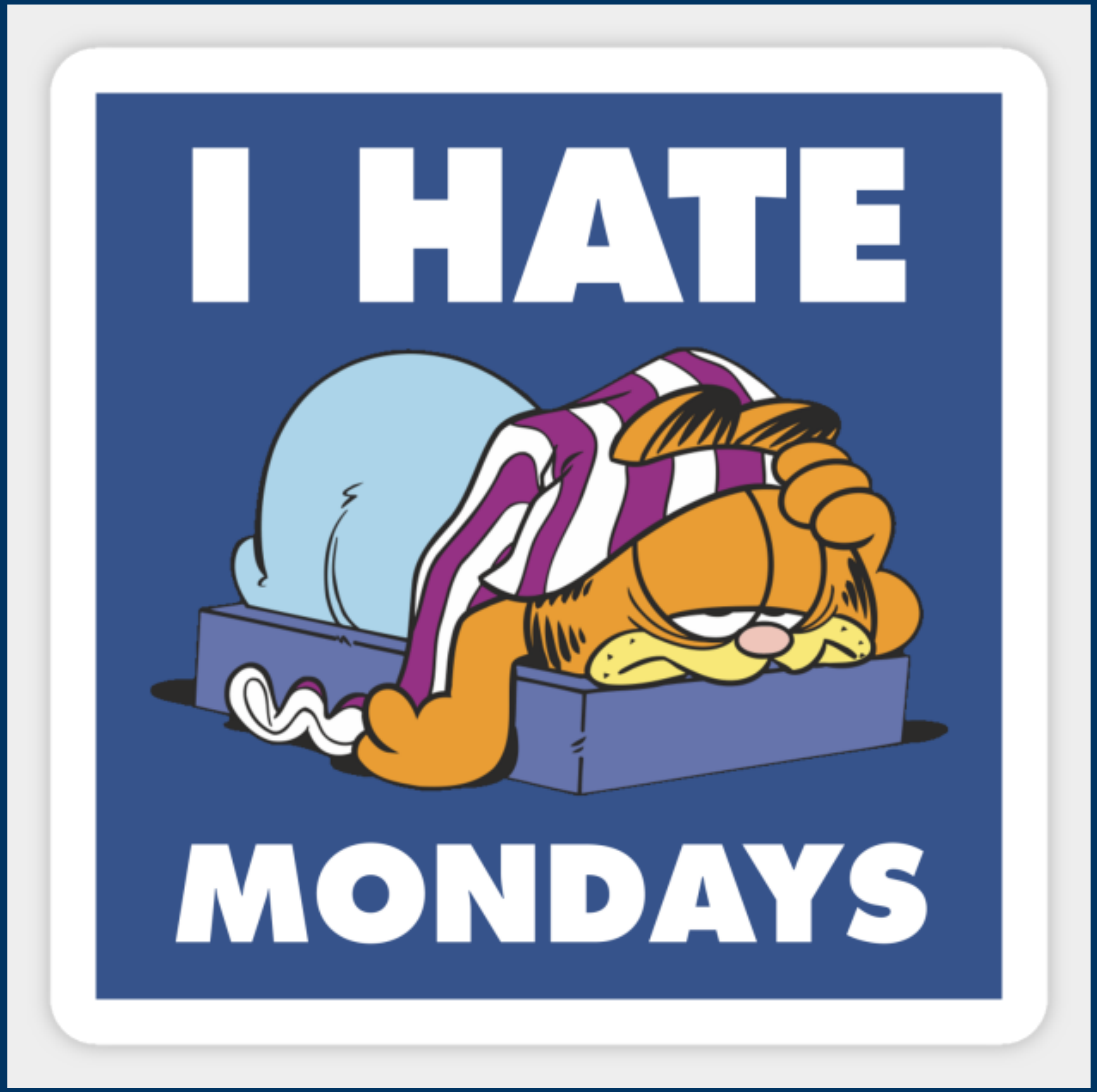
Biostatistics

Applications in Medicine

Nuno Sepúlveda, 23.10.2023


I HATE
MONDAYS


CLASSIC 80s GARFIELD
I HATE MONDAYS
JIM DAVIS


I HATE MONDAY

Syllabus

1. General review
    a. What is Biostatistics?
    b. Population/Sample/Sample size
    c. Type of Data – quantitative and qualitative variables
    d. Common probability distributions
    e. Work example – Malaria in Tanzania
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 Genetics, Genomics, and other 'omics 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, estimation of biological age
    c. Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative-Binomial
4. Other Topics
    a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
    b. Serological analysis – Gaussian (skew-normal) mixture models
    c. Advanced sample size and power calculations

Prevent

Diagnose

Medicine

Improve

Treat

Develop

Diagnosis

Negative / Positive

What is the type of random variable?

Diagnosis

Negative / Positive

What is the type of random variable?

Binary

$X=0$ $X=1$

Diagnosis

Negative / Positive

What is the type of random variable? **Binary**

What is the probability distribution associated with this random variable?

Diagnosis

Negative / Positive

What is the type of random variable? **Binary**

What is the probability distribution associated with this random variable?

Bernoulli

$$P\left[X = x \mid \pi\right] = \pi^{x}\left(1 - \pi\right)^{1-x} I_{\{0,1\}}\left(x\right)$$

Diagnosis

Number of Positive Tests in a Sample of n Individuals

What is the type of random variable?

Discrete

$$0, 1, \ldots, n$$

Diagnosis

Number of Positive Tests in a Sample of n Individuals

What is the type of random variable? **Discrete**

What is the probability distribution associated with this random variable?

Diagnosis

Number of Positive Tests in a Sample of n Individuals

What is the type of random variable? **Discrete**

What is the probability distribution associated with this random variable?

Hypergeometric

$$P[X = x \mid N, M, n] = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

N is the population size

M is the size of population with a positive test

Estimation of the proportion of positive tests

$x \mid N, M, n \rightsquigarrow$ Hypergeometric $(N; M; n)$

N is typically known

$$P[X = x \mid N, M, n] = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$$\hat{M} = \,?$$

Exercise

$x \,|\, N, M, n \rightsquigarrow$ Hypergeometric $(N; M; n)$

N is typically known

$$P[X = x \,|\, N, M, n] = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$N = 700$ - Aneityum Island (Mistery Island)

$n = 50$ individuals

$x = 2$ positive malaria tests

$\hat{M} = ?$

Can you estimate this parameter by the maximum likelihood method using R?

Can you estimate this parameter by the method of moments?

Diagnosis

Number of Positive Tests in a Sample of n Individuals

What is the type of random variable? **Discrete**

What is the probability distribution associated with this random variable?

Hypergeometric

$$P[X = x \,|\, N, M, n] = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$N \to \infty$

Binomial

$$P[X = x \,|\, n, \pi] = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

Estimation of proportion of positive tests

$x \,|\, n, \pi \rightsquigarrow$ Binomial $(n; \pi)$

$$P[X = x \,|\, n, \pi] = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

$$\hat{\pi} = ?$$

$$IC_{95\%}(\pi) = ?$$

Estimation of proportion of positive tests

$x \mid n, \pi \rightsquigarrow$ Binomial $(n; \pi)$

$$P[X = x \mid n, \pi] = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

$$\hat{\pi}_{mle} = \frac{X}{n}$$

Maximum likelihood estimator

$$\hat{\pi}_{mle} = \frac{x}{n}$$

Maximum likelihood estimate

Estimation of proportion of positive tests

$$x \mid n, \pi \rightsquigarrow \text{Binomial}\,(n; \pi)$$

$$P[X = x \mid n, \pi] = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

$$IC_{95\%}\,(\pi) = \hat{\pi}_{mle} \pm 1.96 \times se\left(\hat{\pi}_{mle}\right)$$

Wald's confidence interval

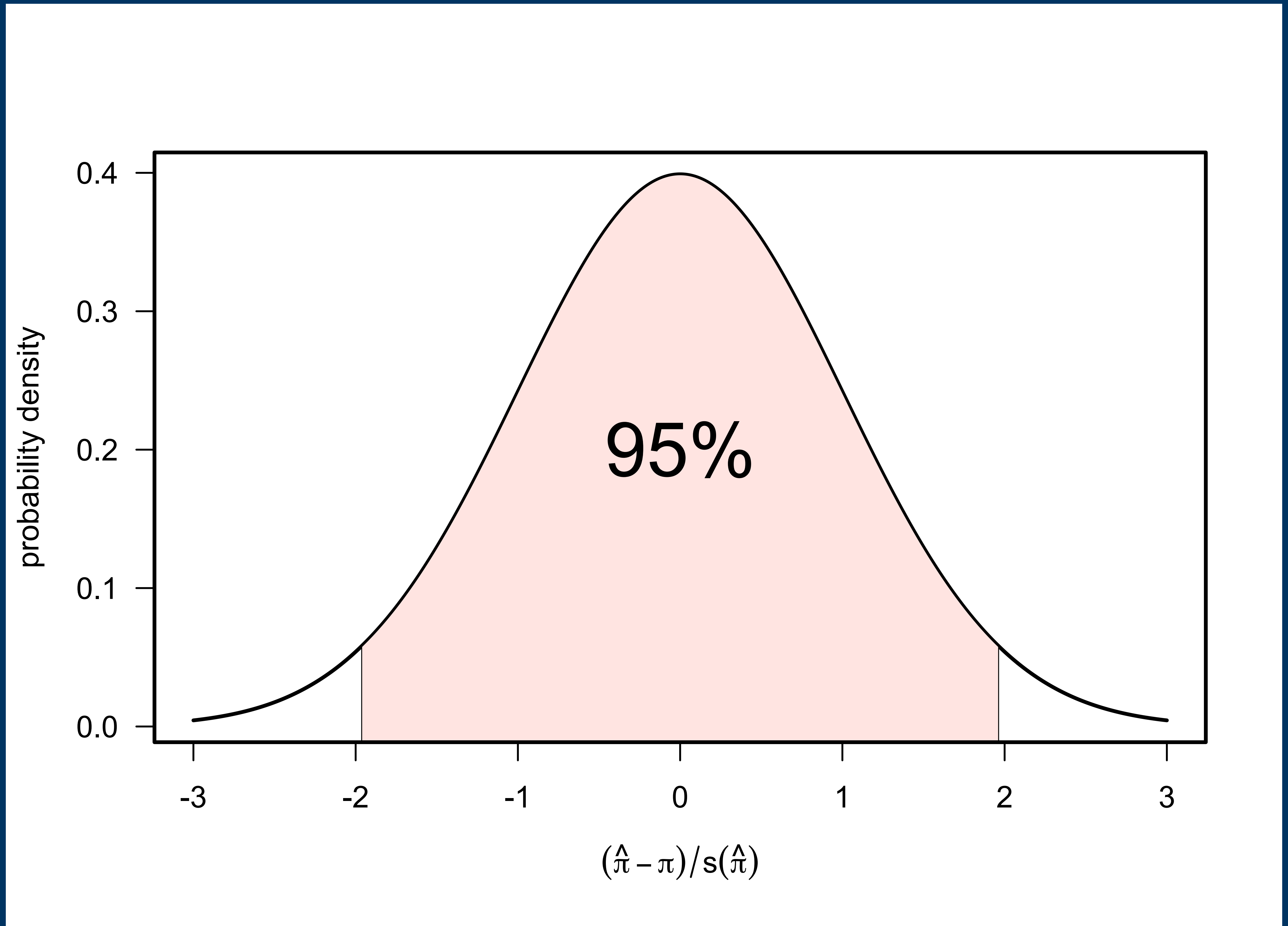
$$s\left(\hat{\pi}_{mle}\right) = \sqrt{\frac{\pi\,(1 - \pi)}{n}}$$

$$s\left(\hat{\pi}_{mle}\right) = \sqrt{\frac{\hat{\pi}_{mle}\left(1 - \hat{\pi}_{mle}\right)}{n}}$$

Two-tail Wald's confidence interval

$$Y = \frac{\hat{\pi}_{mle} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \mid \pi, n \rightsquigarrow \text{Normal}\left(\mu = 0; \sigma = 1\right)$$

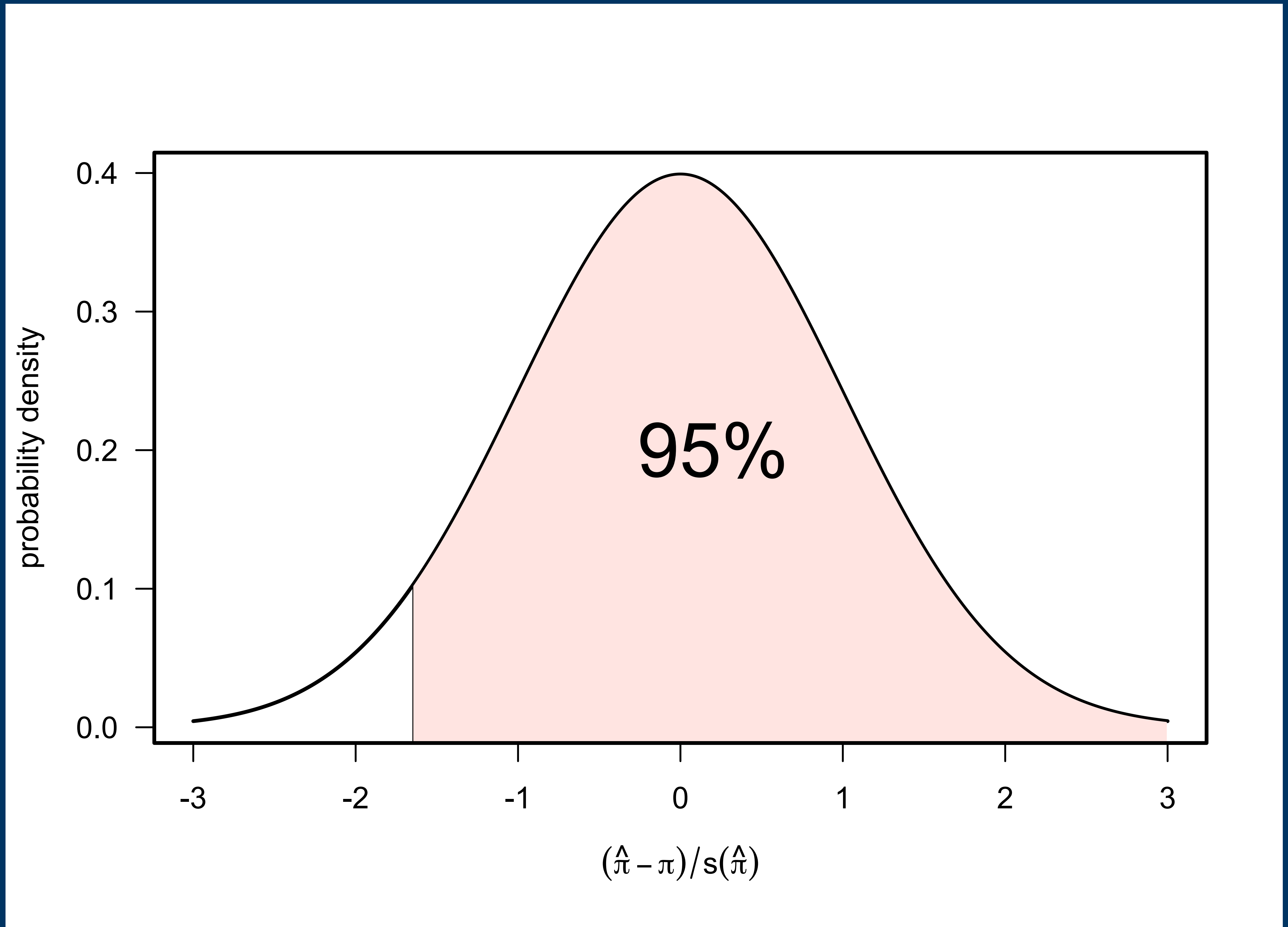
For large samples



One-tail Wald's confidence interval

$$Y = \frac{\hat{\pi}_{mle} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \mid \pi, n \rightsquigarrow \text{Normal}\left(\mu = 0; \sigma = 1\right)$$

For large samples



Estimation of proportion of positive tests

$$x \,|\, n, \pi \rightsquigarrow \text{Binomial}\,(n; \pi)$$

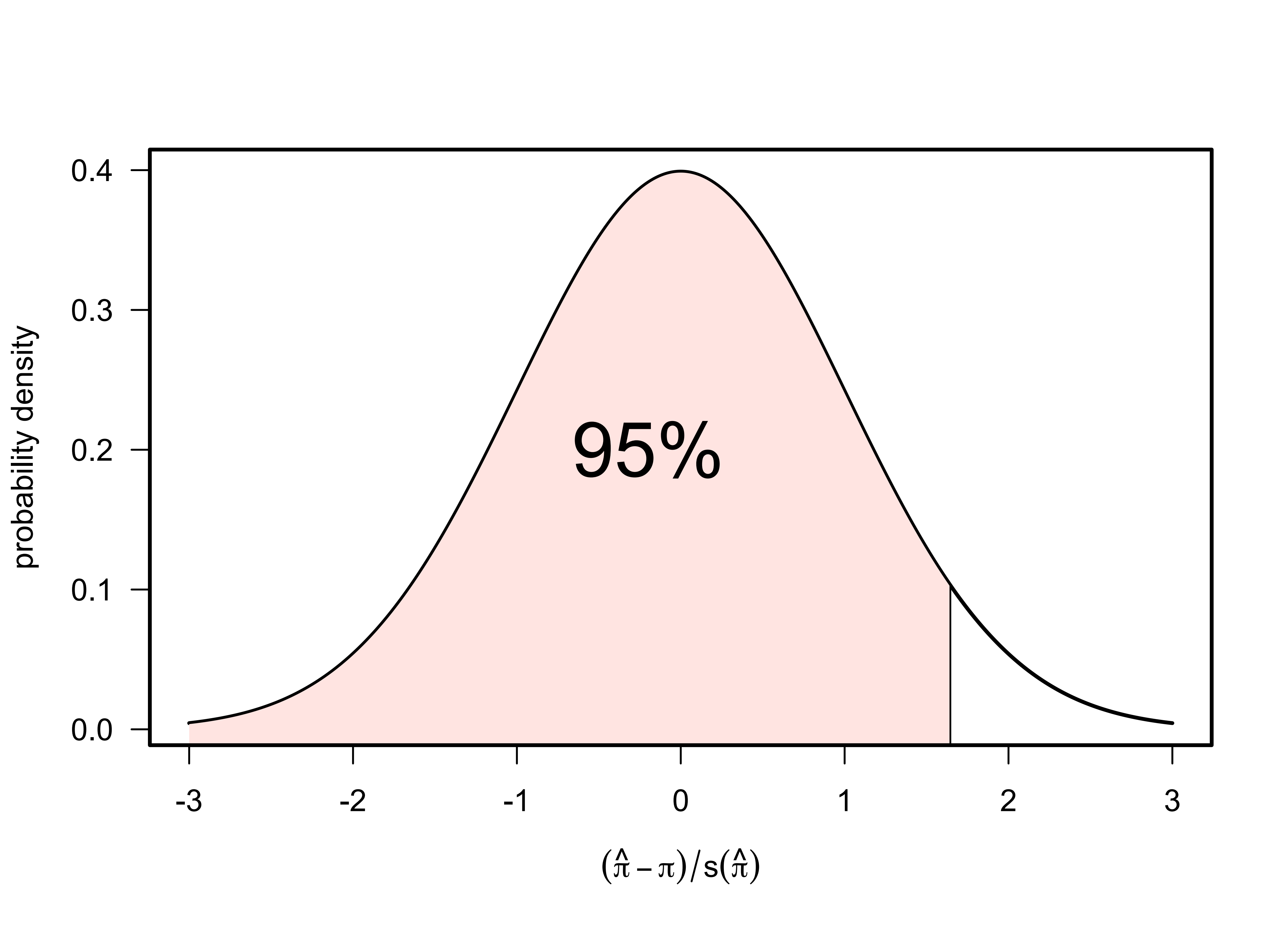
$$P[X = x \,|\, n, \pi] = \binom{n}{x} \pi^x \,(1 - \pi)^{n-x}$$

$$IC_{95\%}\,(\pi) = \left(\hat{\pi}_{mle} - 1.64 \times s\,\left(\hat{\pi}_{mle}\right)\,,1\right)$$

One-tail Wald's confidence interval

$$Y = \frac{\hat{\pi}_{mle} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \mid \pi, n \rightsquigarrow \text{Normal}\left(\mu = 0; \sigma = 1\right)$$

For large samples



Estimation of proportion of positive tests

$$x \mid n, \pi \rightsquigarrow \text{Binomial}\,(n; \pi)$$

$$P[X = x \mid n, \pi] = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

$$IC_{95\%}(\pi) = \left(0, \hat{\pi}_{mle} + 1.64 \times se\left(\hat{\pi}_{mle}\right)\right)$$

Exercise (in the R software)

Table 1 Comparison of screening results for blood samples from community mass blood surveys and passive case detection in the Thai–Myanmar border area

	qPCR (reference)	Expert light microscopy				
	Number of samples	*P. falciparum*	*P. vivax*	*P. malariae*	Mixed *Pf* + *Pv*	Negative
Community mass blood survey						
P. vivax	21	–	2	–	–	19
P. falciparum	10	–	–	–	–	10
Mixed *Pf* + *Pv*	6	–	1	–	–	5
Mixed *Pf* + *P. ovale*	2	–	–	–	–	2
Mixed *Pf* + *Pv* + *Po*	1	–	–	–	–	1
Mixed *Pf* + *Pv* + *Po* + *P. malariae*	1	–	–	–	–	1
Negative	1306	–	–	–	–	1306
Total *n*	1347	–	3	–	–	1344
Hospital and malaria clinic PCD						
P. falciparum	5	5	–	–	–	–
P. vivax	4	–	1	–	–	3
P. malariae	1	–	–	1	–	–
Mixed *Pf* + *Pv*	22	5	14	–	–	3
Negative	265	–	–	–	–	265
Total *n*	297	10	15	1	–	271

Exercise (in the R software)

Estimate the number of positive tests by qPCR

Community

Hospital

Estimate the number of positive tests by light expert microscopy

Community

Hospital

Use binom.test function

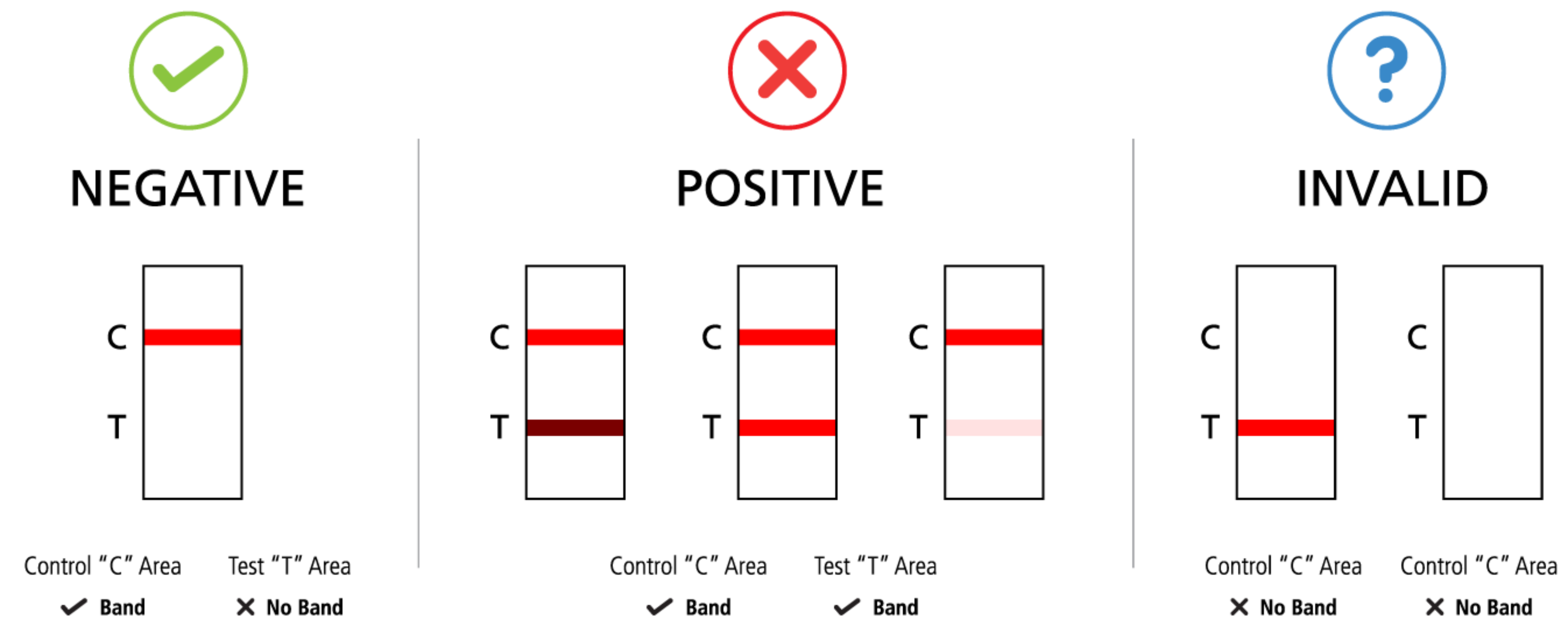
Exercise (in the R software)

Table 1 Comparison of screening results for blood samples from community mass blood surveys and passive case detection in the Thai–Myanmar border area

	qPCR (reference)	Expert light microscopy				
	Number of samples	*P. falciparum*	*P. vivax*	*P. malariae*	Mixed *Pf* + *Pv*	Negative
Community mass blood survey						
P. vivax	21	–	2	–	–	19
P. falciparum	10	–	–	–	–	10
Mixed *Pf* + *Pv*	6	–	1	–	–	5
Mixed *Pf* + *P. ovale*	2	–	–	–	–	2
Mixed *Pf* + *Pv* + *Po*	1	–	–	–	–	1
Mixed *Pf* + *Pv* + *Po* + *P. malariae*	1	–	–	–	–	1
Negative	1306	–	–	–	–	1306
Total *n*	1347	–	3	–	–	1344
Hospital and malaria clinic PCD						
P. falciparum	5	5	–	–	–	–
P. vivax	4	–	1	–	–	3
P. malariae	1	–	–	1	–	–
Mixed *Pf* + *Pv*	22	5	14	–	–	3
Negative	265	–	–	–	–	265
Total *n*	297	10	15	1	–	271

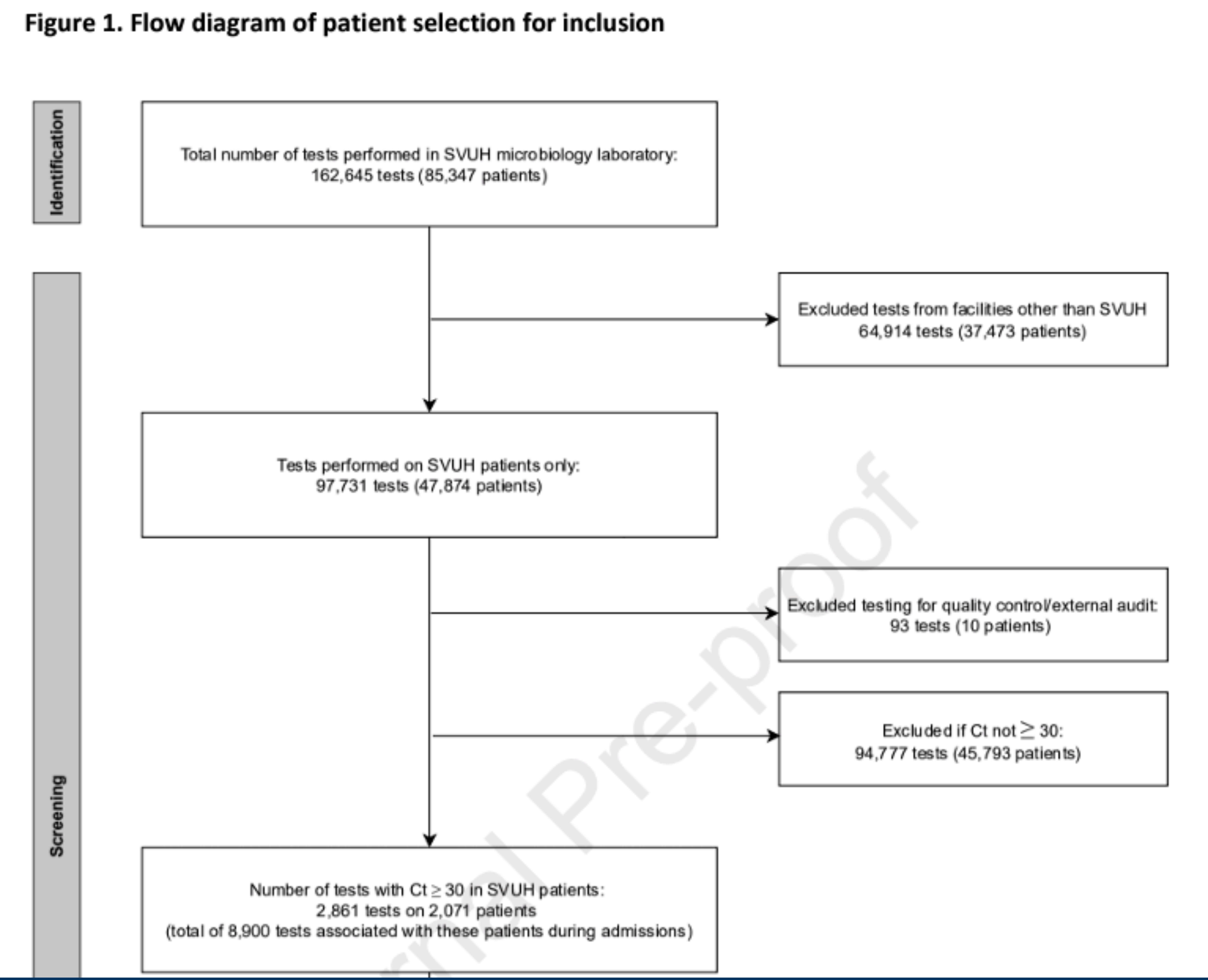
Break

Diagnosis (more complex situation)



Rapid diagnostic test for SARS-CoV-2

Diagnosis (more complex situation)



Molecular test for SARS-CoV-2 detection

Invalid

Positive

Indeterminate

Dolan & Fitzgerald (2023). Journal of Hospital Infection

Diagnosis (more complex situation)

Invalid / Indetermine / Negative / Positive

What is the type of random variable?

Categorical

Diagnosis (more complex situation)

Invalid or Indetermine / Negative / Positive

What is the type of random variable? **Categorical**

What is the probability distribution associated with this random variable?

Multivariate Bernoulli

$$P\left[\mathbf{X} = (x_1, x_2, x_3) \,|\, (\pi_1, \pi_2, \pi_3)\right] = \Pi_{i=1}^{3} \pi_i^{x_i}$$

with the restrictions: $x_i \in \{0,1\}, \sum_{i=1}^{3} x_i = 1$ and $\pi \in (0,1), \sum_{i=1}^{3} \pi_i = 1$

Diagnosis (more complex situation)

Number of Invalid or Indetermine / Negative / Positive

What is the type of random variable?

Diagnosis (more complex situation/less common)

Number of Invalid or Indetermine / Negative / Positive

What is the type of random variable? **Multivariate Categorical**

Diagnosis (more complex situation/less common)

Number of Invalid / Indetermine / Negative / Positive

What is the type of random variable? **Multivariate Categorical**

What is the probability distribution associated with this random variable?

Diagnosis (more complex situation/less common)

Number of Invalid or Indetermine / Negative / Positive

What is the type of random variable? **Multivariate Categorical**

What is the probability distribution associated with this random variable?

Multivariate Hypergeometric (small population sizes)

$$P[(n_1, \ldots, n_4) \,|\, n, N, (M_1, M_2, M_3)] = \frac{\binom{M_1}{n_1}\binom{M_2}{n_2}\binom{M_3}{n_3}}{\binom{N}{n}}$$

with $\sum_{i=1}^{3} n_i = n$ and $\sum_{i=1}^{3} M_i = N$

Diagnosis (more complex situation/less common)

Number of Invalid or Indetermine / Negative / Positive

What is the type of random variable? **Multivariate Categorical**

What is the probability distribution associated with this random variable?

Multinomial (large population sizes)

$$P[(n_1, \ldots, n_3) \,|\, n, (\pi_1, \pi_2, \pi_3)] = \frac{n!}{n_1!n_2!n_3!} \pi_1^{n_1} \pi_2^{n_2} \pi_3^{n_3} \text{ with } \sum_{i=1}^{3} n_i = n \text{ and } \sum_{i=1}^{3} \pi_i = 1$$

Estimation of the proportions

$$\hat{\pi}_1 = ?$$

$$\hat{\pi}_2 = ?$$

$$\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2$$

$$IC_{95\%}(\pi_1) = ?$$

$$IC_{95\%}(\pi_2) = ?$$

$IC_{95\%}(\pi_3)$ – no need

Estimation of the proportions

$$\hat{\pi}_1 = \frac{n_1}{n}$$

$$\hat{\pi}_2 = \frac{n_2}{n}$$

$$\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2$$

Estimation of the proportions

$$\hat{\pi}_1 = \frac{n_1}{n}$$

$$IC_{95\%}(\pi_1) = ?$$

$$\hat{\pi}_2 = \frac{n_2}{n}$$

$$IC_{95\%}(\pi_2) = ?$$

$$\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2$$

Estimation of the proportions

$$\hat{\pi}_1 = \frac{n_1}{n} \qquad IC_{95\%}\left(\pi_1\right) = \hat{\pi}_1 \pm 2.24 \times se\left(\hat{\pi}_1\right)$$

$$\hat{\pi}_2 = \frac{n_2}{n} \qquad IC_{95\%}\left(\pi_2\right) = \hat{\pi}_2 \pm 2.24 \times se\left(\hat{\pi}_2\right)$$

$$\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2 \qquad 2.24 = \Phi^{-}1\left(\frac{0.975}{2}\right)$$

$$IC_{\gamma\%}\left(\pi_1\right) = \hat{\pi}_1 \pm \Phi^{-1}\left(1 - \frac{\gamma}{2}\right) \times se\left(\hat{\pi}_1\right) \qquad \Phi^{-}1\left(\frac{1-\gamma}{2p}\right)$$

Bonferroni's method

p is the number of estimated parameters

$$P\left[\cup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} P\left[A_i\right]$$

Exercise (in the R software)

Cliff et al (2019). Frontiers in Medicine

Herpesvirus	Seronegative	Indeterminate	Seropositive
Cytomegalovirus	254	7	133
Epstein-Barr virus (VCA)	46	4	344
Epstein-Barr virus (EBNA1)	83	15	296
Herpesvirus simplex 1	195	20	179
Herpesvirus simplex 2	232	12	150

Estimate the proportion of positive and indeterminate tests and calculate the respective confidence intervals

Break

Diagnosis

Detection of disease / Identification of cases

Infectious diseases

Detection of the pathogen genetic material or antigen

positive

negative

Invalid or Indetermine

Non-communicable diseases

Use of a biomarker

not ill

ill

Diagnosis of infectious diseases

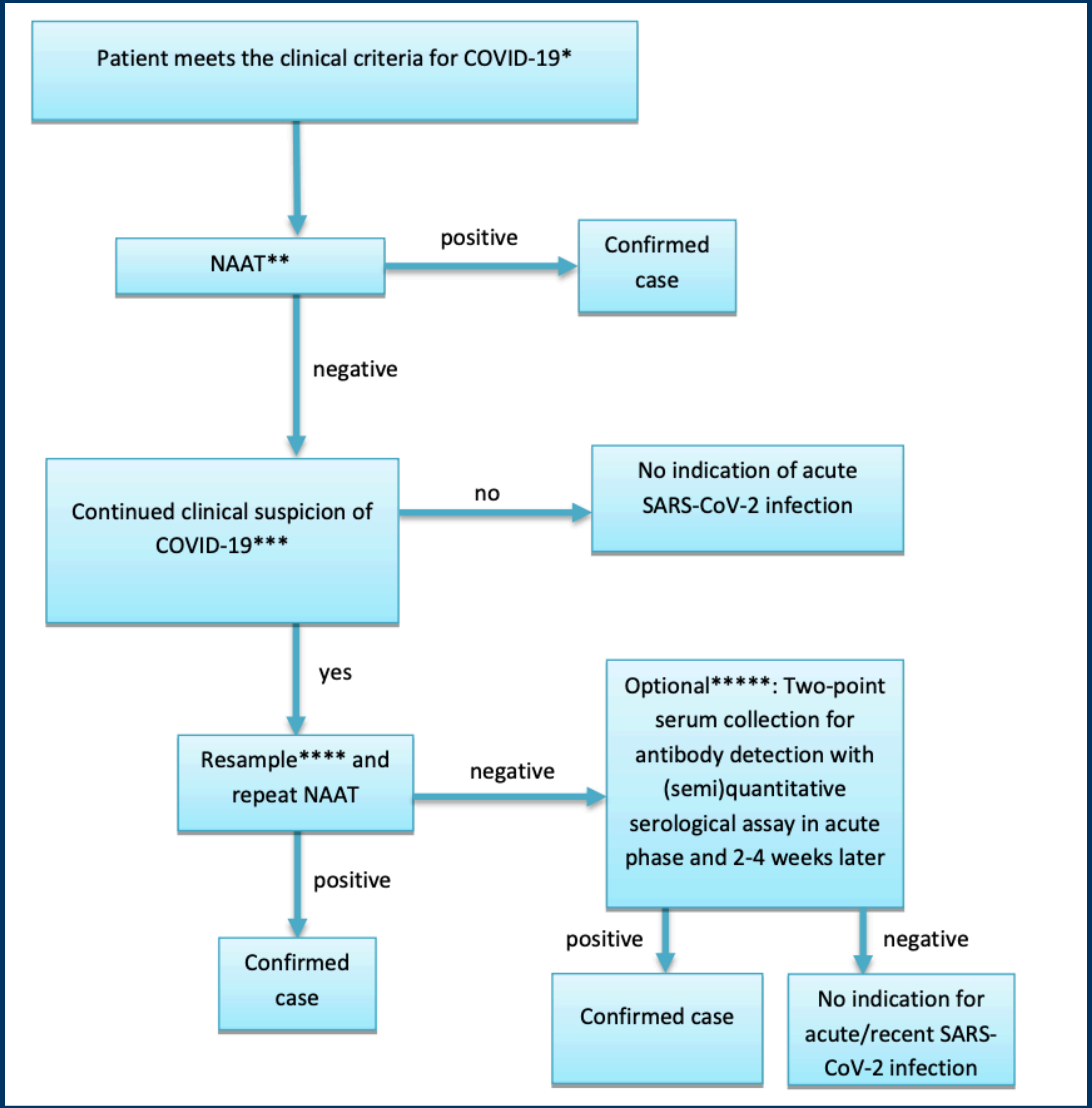
Malaria

Lab test (PCR)
Microscopy
Rapid diagnostic test

COVID-19

Lab test (PCR)
Antigen test
Rapid diagnostic test
Serological test

Diagnosis of suspected COVID-19 cases



Biomarker

A biomarker is usually measurement or a substance that indicates important facts about a living organism, usually a patient.

It provides information about:

- The biological state of the organism;
- Disease risk;
- Disease diagnosis;
- Disease progression;
- Treatments of choice;
- Monitoring responses to treatment;
- Endpoints for treatment efficacy.

Little quiz
Do you know the associated biomarker?

Anemia

Diabetes Mellitus

Multiple sclerosis

Little quiz
Do you know the associated biomarker?

Anemia

Hemoglobulin

Diabetes Mellitus

Fasting glucose levels

Multiple sclerosis

Antibodies against Myelin basic protein

Imperfect diagnosis

0<Sensitivity<1

0<Specificity<1

Statistical inference

Hypothetical data (2 x 2 contingency table)

True status	Positive	Negative
Cases	x_1	$n_1 - x_1$
Non cases	$n_2 - x_2$	x_1

Example: malaria testing

Estimate Se / Sp (with confidence intervals) of microscopy testing

True status (PCR)	Positive	Negative
Cases	1695	1057
Non cases	216	4169

Doctor, S.M., Liu, Y., Anderson, O.G. et al. Low prevalence of Plasmodium malariae and Plasmodium ovale mono-infections among children in the Democratic Republic of the Congo: a population-based, cross-sectional study. Malar J 15, 350 (2016).

Results

Binomial test

Se: 0.6159 CI (0.5974, 0.6341)
Sp: 0.9507 CI (0.9439, 0.9569)

Prop test

Se: 0.6159 CI (0.5974, 0.6341)
Sp: 0.9507 CI (0.9439, 0.9569)