Evaluating Diagnostic Tests in the Absence of a Gold Standard

Nandini Dendukuri

Departments of Medicine & Epidemiology, Biostatistics and Occupational Health, McGill University;

Technology Assessment Unit, McGill University Health Centre

Advanced TB diagnostics course, Montreal, July 2011

Evaluating Diagnostic Tests in the Absence of a Gold Standard

- An area where we are still awaiting a solution
  - We are still working on methods for Phase 0 studies

- A number of methods have appeared in the pipeline
  - Some have been completely discredited (e.g. Discrepant Analysis)
  - Some are more realistic but have had scale-up problems due to mathematical complexity and lack of software (e.g. Latent Class Analysis)
  - Some in-between solutions seem easy to use, but their inherent biases are poorly understood (e.g. Composite reference standards)

- Not yet at the stage where you can stick your data into a machine and get accurate results in 2 hours

No gold-standard for many types of TB

- Example 1: TB pleuritis:
  - Conventional tests have less than perfect sensitivity*

Microscopy of the pleural fluid	<5%
Culture of pleural fluid	24 to 58%
Biopsy of pleural tissue + culture of biopsy material	~ 86%

  - Most conventional tests have good, though not perfect specificity ranging from 90-100%

* Source: Pai et al., BMC Infectious Diseases, 2004

No gold-standard for many types of TB

- Example 2: Latent TB Screening/Diagnosis:
  - Traditionally based on Tuberculin Skin Test (TST)
    - TST has poor specificity* due to cross-reactivity with BCG vaccination and infection with non-TB mycobacteria

TST Sensitivity	75-90
TST Specificity	70-90

Menzies et al., Ann Int Med, 2007

Usual approach to diagnostic test evaluation

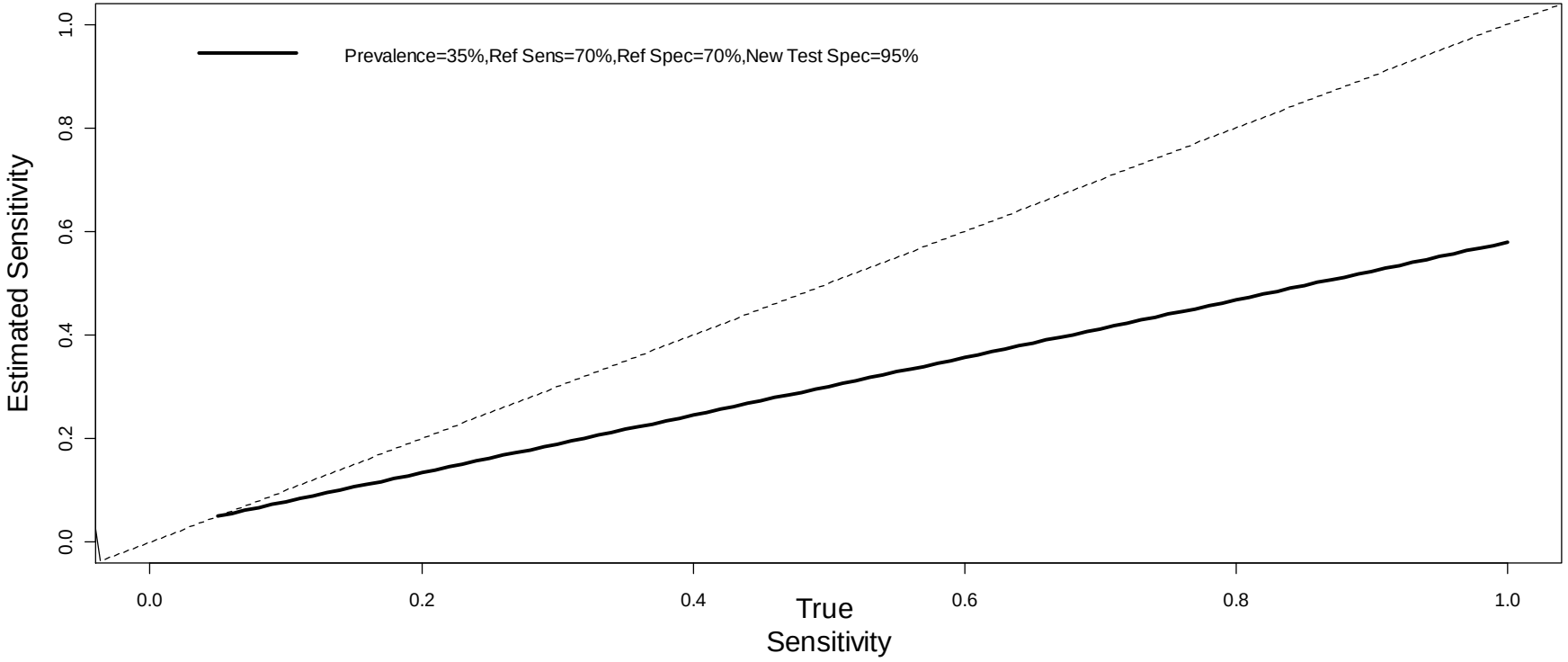
Compare new test to existing standard

	Standard Test+	Standard Test-
New Test+	A	B
New Test-	C	D

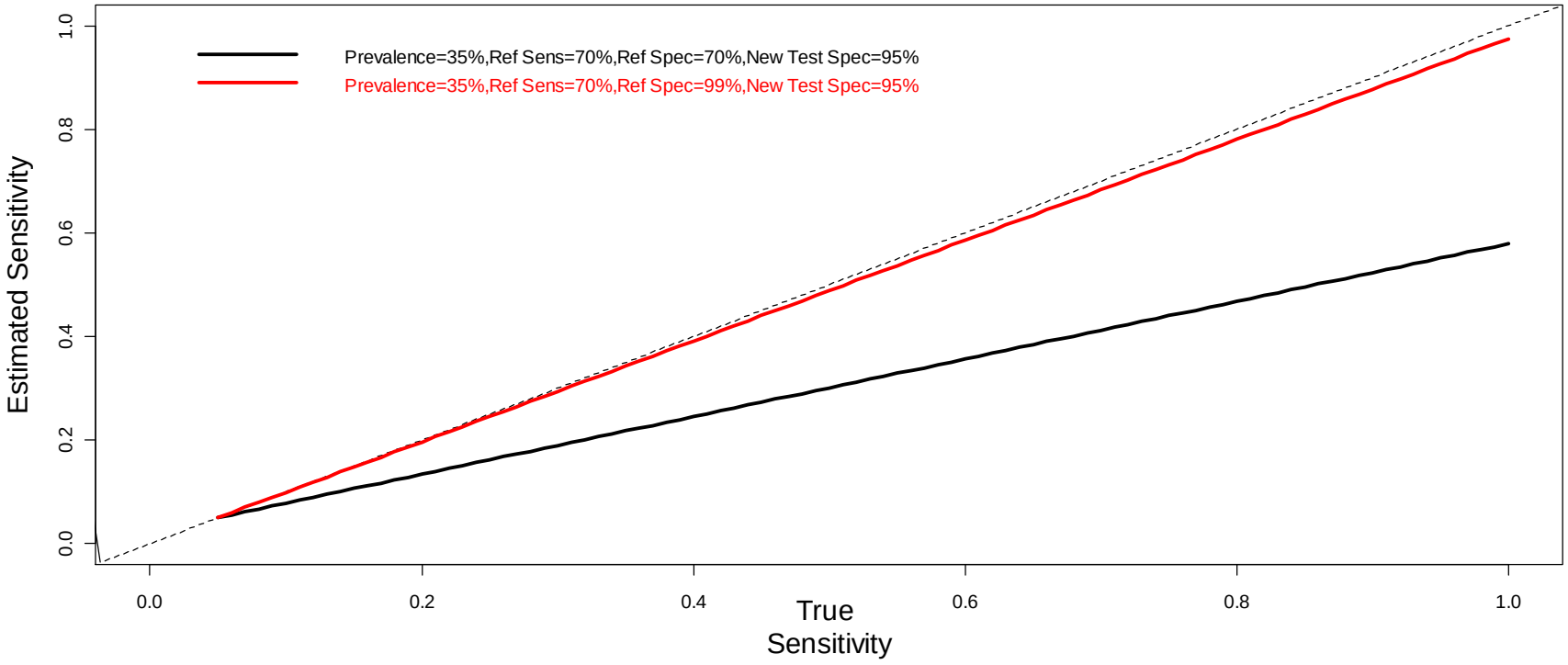
Sensitivity of new test = A/(A+C)

Specificity of new test = D/(B+D)

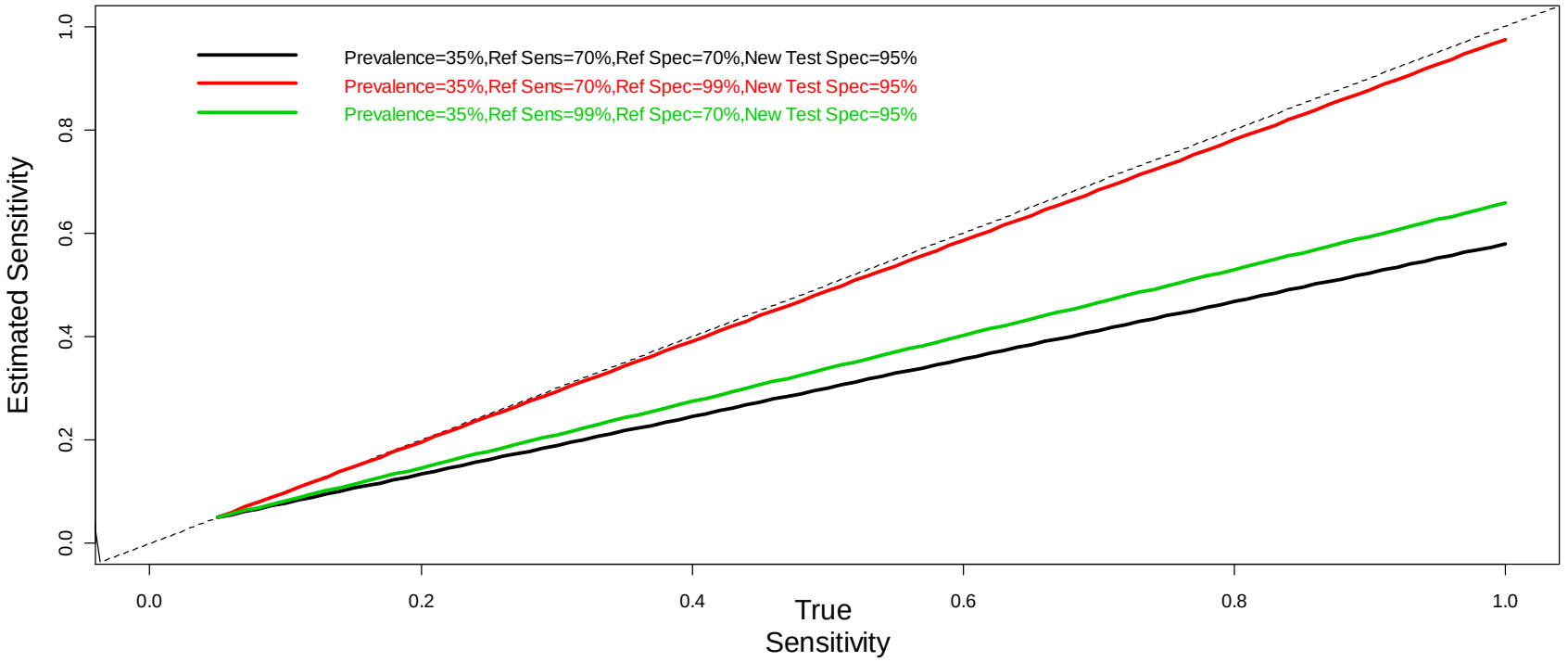
Bias due to assuming reference test is perfect: Impact on sensitivity



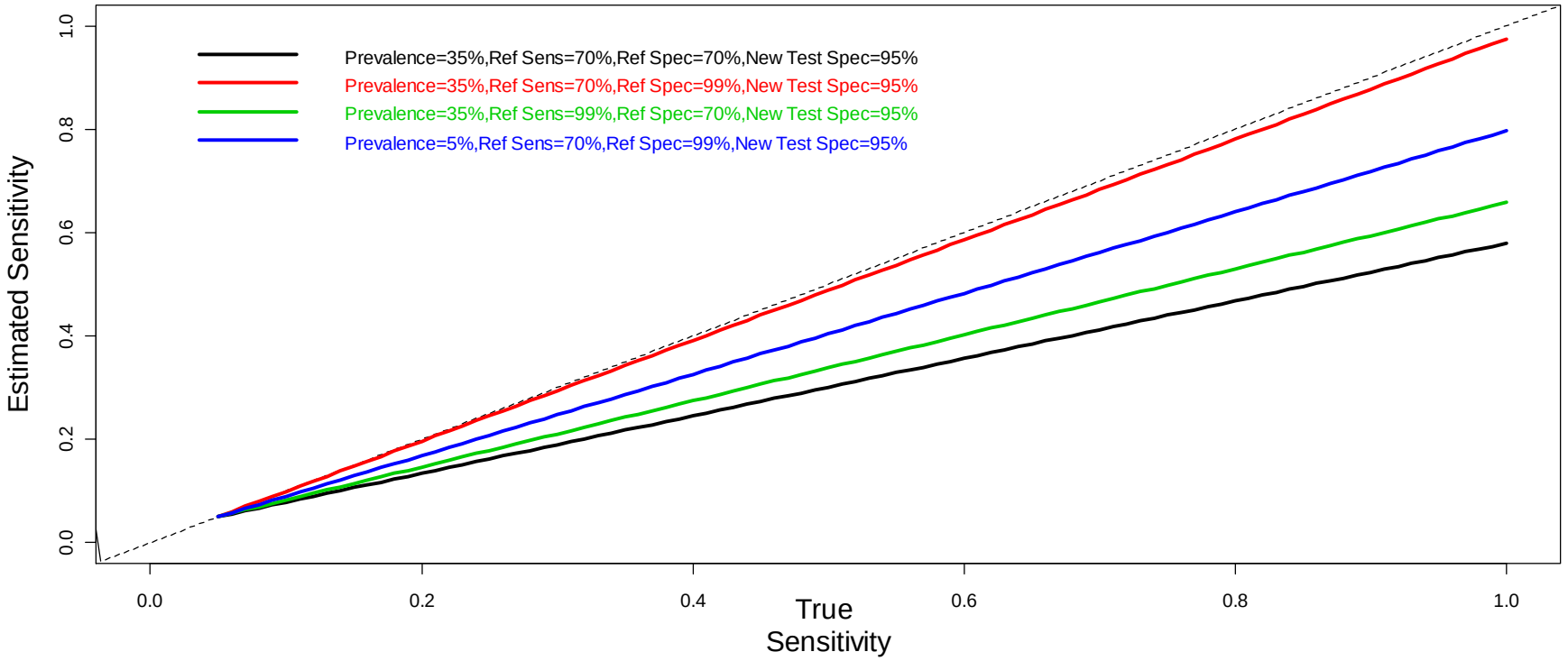
Bias due to assuming reference test is perfect: Impact on sensitivity



Bias due to assuming reference test is perfect: Impact on sensitivity



Bias due to assuming reference test is perfect: Impact on sensitivity



Bias due to assuming reference test is perfect

- Thus, sensitivity and specificity of the reference, as well as prevalence play a role in determining the magnitude of the bias
  - Specificity rather than sensitivity of the reference has greater impact on sensitivity of new test
- Similar results can be derived for specificity of the new test
  - The sensitivity of the reference will have a greater impact there
- Since we do not have accurate measures of these quantities, our subjective knowledge of them is needed to make meaningful inferences in these problems

Solutions that have been proposed to adjust reference standard bias

1. Discrepant analysis
2. Composite reference standard
3. Plug in values for sensitivity and specificity
4. Latent class analysis

Discrepant Analysis

- Arose in the area of *C. Trachomatis* tests when the standard test, culture, was found inadequate for evaluating NAATs
  - Culture has high specificity, but poor sensitivity

- Involves a two-stage design
  - First patients were tested by both the NAAT under evaluation and culture
  - Then, those NAAT+, culture- individuals were re-tested with a resolver test that was typically also an NAAT. The result of the resolver test was used to classify patients as ‘infected’ or not

Discrepant Analysis: Example*

TABLE 1. Comparison of LCR and Cell Culture Assays for *C. trachomatis* in Urine Collected From 237 Women Attending an STD Clinic (adopted from van Doornum et al[16])

	Cell Culture (Cervix)			Discrepant Analysis by MOMP-LCR	
Plasmid-LCR Results	**Positive**	**Negative**	**Total**	**Positive**	**Negative**
Positive	13 (cell A)	12 (cell B)	25	25 (13+12)	0 (12−12)
Negative	2 (cell C)	210 (cell D)	212	2	210
Total	15	222	237	27	210

Culture-based sensitivity of LCR = (13/15) = 86.7%, specificity = (210/222) = 94.6%.
Discrepant analysis-based estimates of sensitivity = (25/27) = 92.6% and specificity = (210/210) = 100%.
STD indicates sexually transmitted disease.

* Hadgu et al, Epidemiology, 2005

Discrepant Analysis discredited

- Several papers* showed the method to be biased due to:
  - Selective selection of patients for the second stage of the design
  - Use of the NAAT under evaluation in the definition of the reference standard.

* Hadgu, Stats in Med and Lancet, 2007

Composite Reference Standard (CRS)*

- Proposed with the aim of
  - Developing a reference standard that did not involve the test under evaluation
  - That increased the overall accuracy of the reference tests

- Approach:
  - CRS defines a decision rule to classify patients as 'infected' or not based on observed results of 2 or more standard, imperfect tests
    - e.g. A CRS based on culture and biopsy may assume that a positive result on either test is equivalent to 'infected'

* Alonzo and Pepe, Stats in Med, 1999

Composite Reference Standard (CRS)

- Liberal definition of CRS will result in an increase in sensitivity, but a loss of specificity

Liberal definition of CRS		
T1	T2	CRS
+	+	+
+	-	+
-	+	+
-	-	-

- Vice-versa for the conservative definition of the CRS

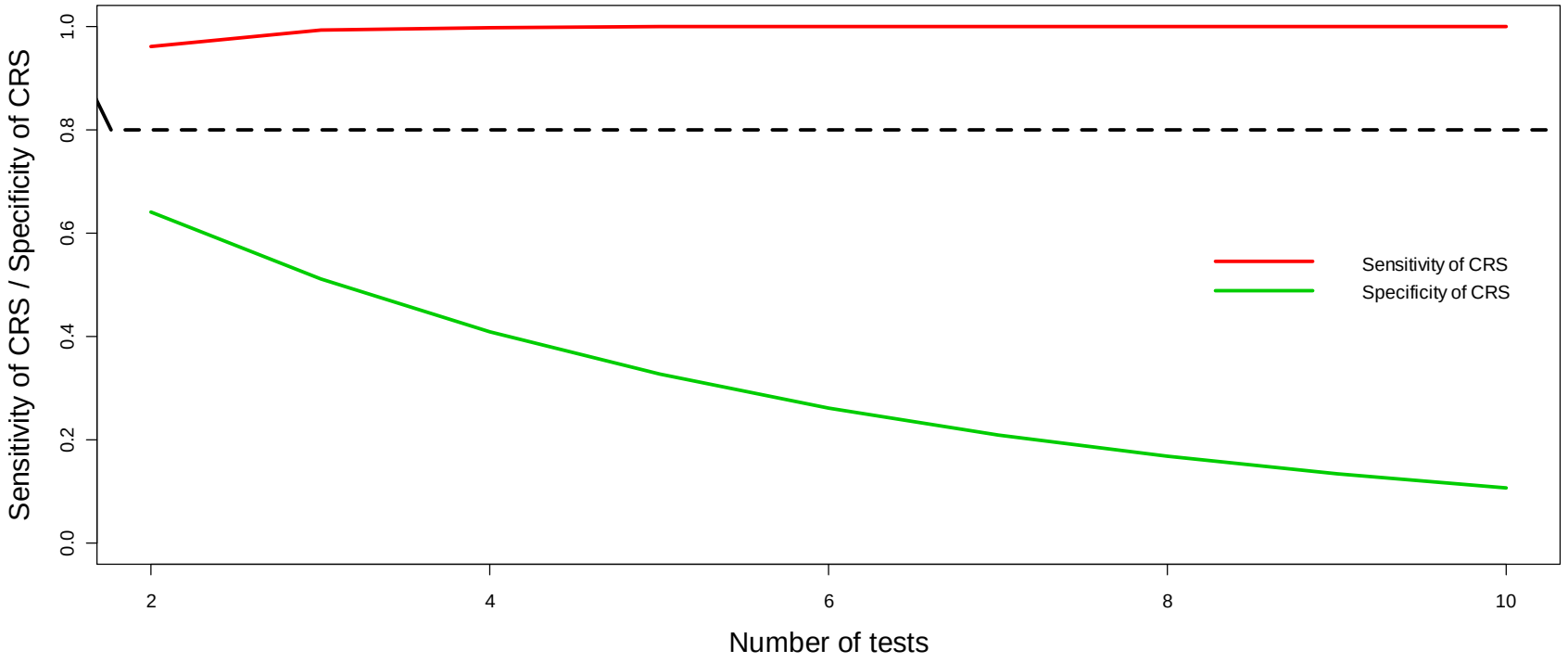
Conservative definition of CRS		
T1	T2	CRS
+	+	+
+	-	-
-	+	-
-	-	-

Composite Reference Standard

- Drawbacks *
  - Same problems that arise when treating a single, imperfect standard test as perfect, i.e. underestimation of test properties
  - Operating characteristics poorly understood
    - e.g. Liberal CRS assumes that both T1 and T2 have perfect specificity, which may not be the case
  - When several standard tests are available it is unclear which combination to use as the reference standard

* Dendukuri et al, SBR, 2011

Relating performance of liberal CRS to number of tests*



* Assuming all tests in CRS have same sensitivity of 80% and same specificity of 80%

Latent Class Models

- Example on: “Estimation of latent TB infection prevalence using mixture models”
  - Though the focus of this article was estimation of disease prevalence, the models used can also be used to estimate the sensitivity and specificity of the observed tests

- Article:
  - Pai et al., Intl. J of TB and Lung Disease. 12(8): 1-8. 2008

Latent TB infection (LTBI)

- Latent Tuberculosis Infection
  - Definition: Patient carries live, dormant *Mycobacterium TB* organisms but does not have clinically apparent disease
  - Risk of developing full-blown TB about 10%
  - As high as 50% prevalence in health care workers in endemic countries

- LTBI Screening/Diagnosis:
  - traditionally based on Tuberculin Skin Test (TST)
  - TST has poor specificity due to cross-reactivity with BCG vaccination and infection with non-TB mycobacteria
  - T-cell based interferron-gamma release assays (IGRAs) more specific alternative to TST

Study on LTBI prevalence at MGIMS, Sevagram

Original study:	Pai et al. (JAMA 2005)
Participants:	719 health care workers in rural India
Data:	All participants were tested on both TST and QFT-G (a commercial IGRA)
	TST: Score from 0mm-30mm QFT-G: Score from 0-10 IU/mL

Traditional approach to LTBI prevalence estimation

- Prevalence reported as % >= cut-off separately for each test
  - e.g. >5mm, >10mm or >15mm for TST
  - > 0.35 IU/mL for QFT-G

- Problems
  - Amounts to assuming test is perfect at cut-off
  - Wastes information on continuous scores
  - Not straightforward to combine results of TST and QFT-G

Cross-tabulation of TST and QFT-G

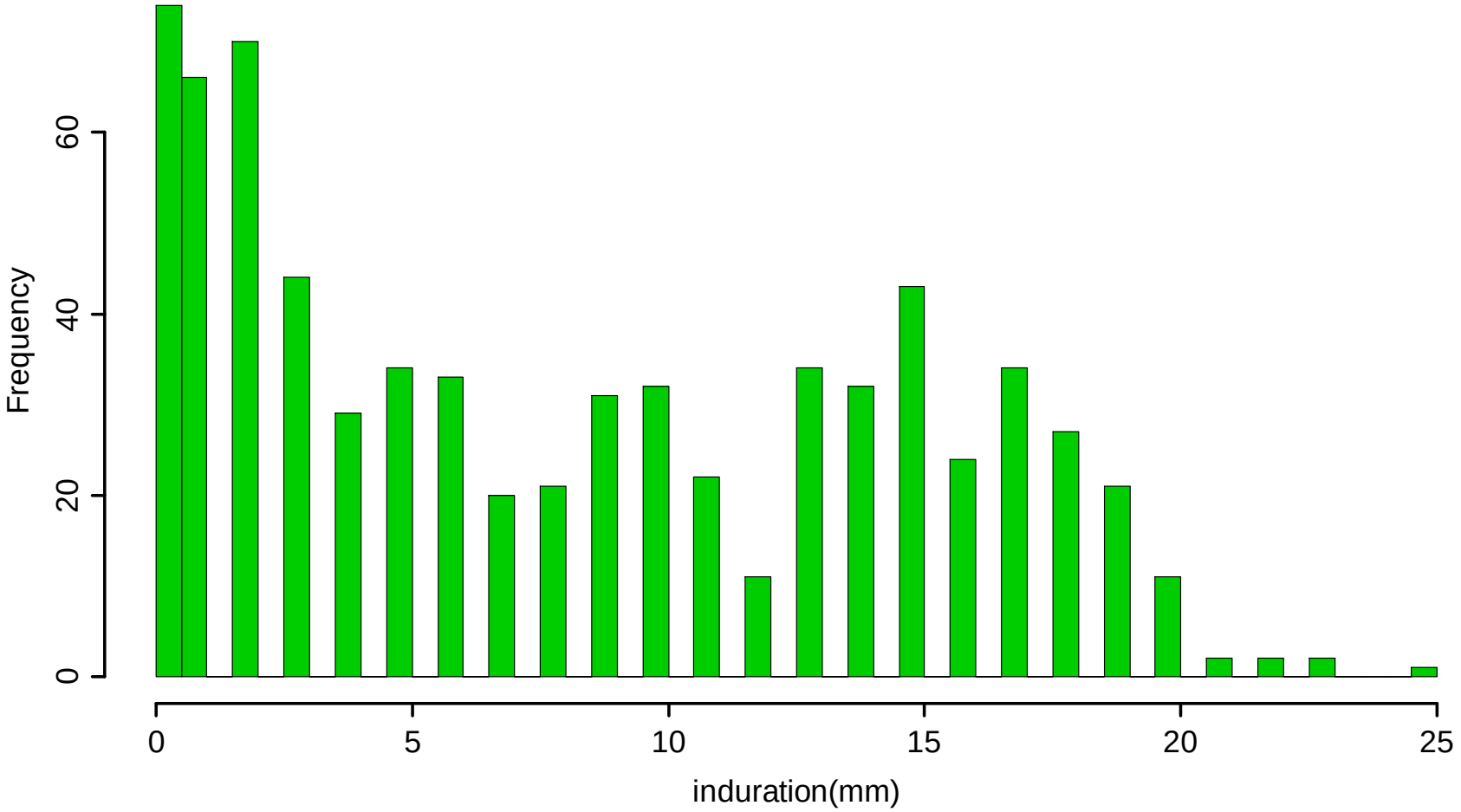
	TST+*	TST-
QFT-G+	226	62
QFT-G-	72	359

- Prevalence
  - TST+: 41% (95% CI: 38%-45%)
  - QFT-G+: 40% (95% CI: 37%-43%)

- What is the probability of LTBI in each cell, particularly discordant cells?
- Can our prior knowledge of the tests' properties help?

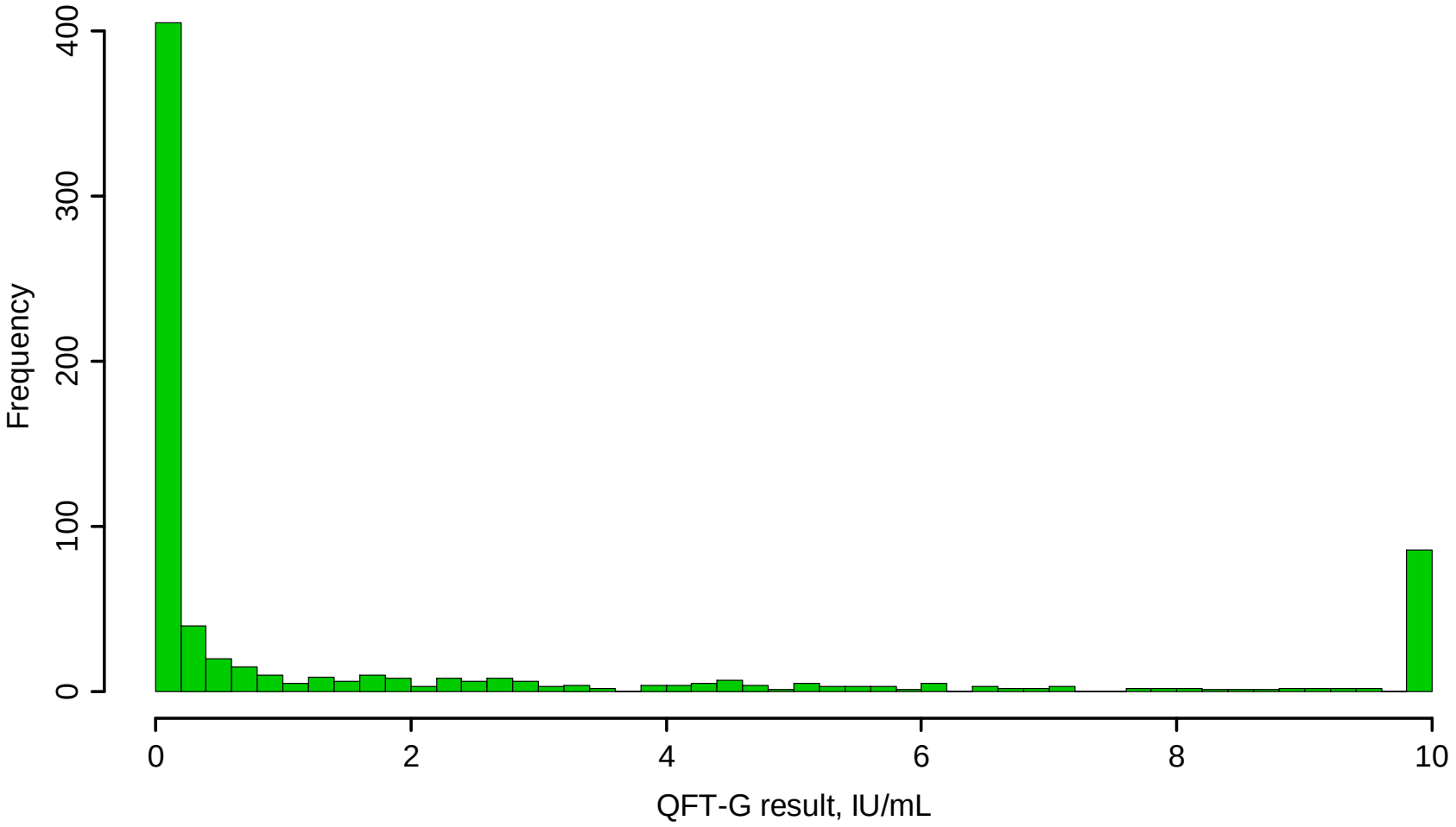
* >=10mm induration

Histogram of TST results



How does the P(LTBI) change with increasing TST?

Histogram of QFT-G results



Mixture models

- Assume the observed data arise from mixture of true LTBI+ and LTBI- groups
- Can be applied to either continuous or categorical test results
- Can be estimated using software packages such as SAS, WinBUGS or specialized programs such as BLCM* or the LCMR package in R

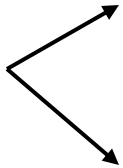
* See software page at http://www.nandinidendukuri.com

We considered two types of mixture models for LTBI prevalence estimation

- Latent class model
  - Data on one or both of TST and QFT-G dichotomized at standard cut-off values
- Continuous mixture model
  - For TST alone

Latent Class Model: TST alone

TST+	TST-
298	421



	TST+	TST-	
LTBI+	X	Y	X+Y
LTBI -	298-X	421-Y	719-(X+Y)

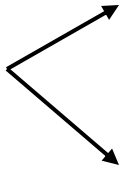
- Note that if we knew X and Y we can estimate:
  - Prevalence of LTBI = (X+Y)/719
  - Sensitivity of TST = X/(X+Y)
  - Specificity of TST = (421-Y)/(719-(X+Y))

How do we determine X and Y?

- By using information external to the data on the sensitivity and specificity of TST or on the prevalence
  - e.g. Just for illustration, say sensitivity of TST=100% and specificity of TST=75%.
  - This means Y=0.
  - And, 75% = 421/(719-X). Therefore, X=158
  - Knowing, X and Y we can determine the third parameter. Prevalence = 158/719 = 22%

X and Y in terms of sens/spec/prev

TST+	TST-
298	421



	TST+	TST-	
LTBI+	X	Y	X+Y
LTBI -	298-X	421-Y	719-(X+Y)

- X = # of true positives

  = 719 × P(TST+, LTBI+)

  $= 719 \times \frac{X+Y}{719} \times \frac{X}{X+Y}$

  = 719 × prevalence × sensitivity

Latent Class Model: TST alone

	TST+	TST-	
LTBI+	X = 719 pS	Y = 719 p(1-S)	719 p
LTBI -	298-X = 719(1-p)(1-C)	421-Y = 719(1-p)C	719 (1-p)

- Each cell in the 2 X 2 table can be written in terms of the prevalence (p), sensitivity (S) and specificity (C)

- $\Rightarrow 298 = 719 \times (ps + (1-p)(1-c))$
  - $\Rightarrow$ 1 equation and 3 unknown parameters
  - $\Rightarrow$ Problem is not identifiable!

Using valid external (prior) information

- Based on meta-analyses of studies evaluating TST we have (Menzies et al. 2007):
  - $0.75 < S < 0.90$
  - $0.70 < C < 0.90$
- Different values in these ranges would yield different prevalence estimates.
  - If s=0.75 and c=0.70 then p=0.24
  - If s=0.90 and c=0.90 then p=0.39
- Considering all possible combinations of sens/spec would mean repeating this infinite times!
  - How do we pick amongst these infinite possible results?

Bayesian vs. Frequentist estimation

Added step in Bayesian Analysis:

Prior Distribution: Summarizes any information on unobserved population parameters that is external to the observed data

e.g. For Linear Regression Model
Vague (non-informative) priors on α and β

e.g. For Latent Class Model
Informative priors on LTBI prevalence, Sens/Spec of both tests

Model: Relates unobserved population parameters to observed values

e.g. Linear Regression
TST result = α + β Contact

e.g. Latent Class Model
TST and QFT results = f(LTBI prevalence, Sens/Spec of both tests)

Observed Data x, y: Collected on a sample

e.g. Y = TST result
X = contact with TB patients

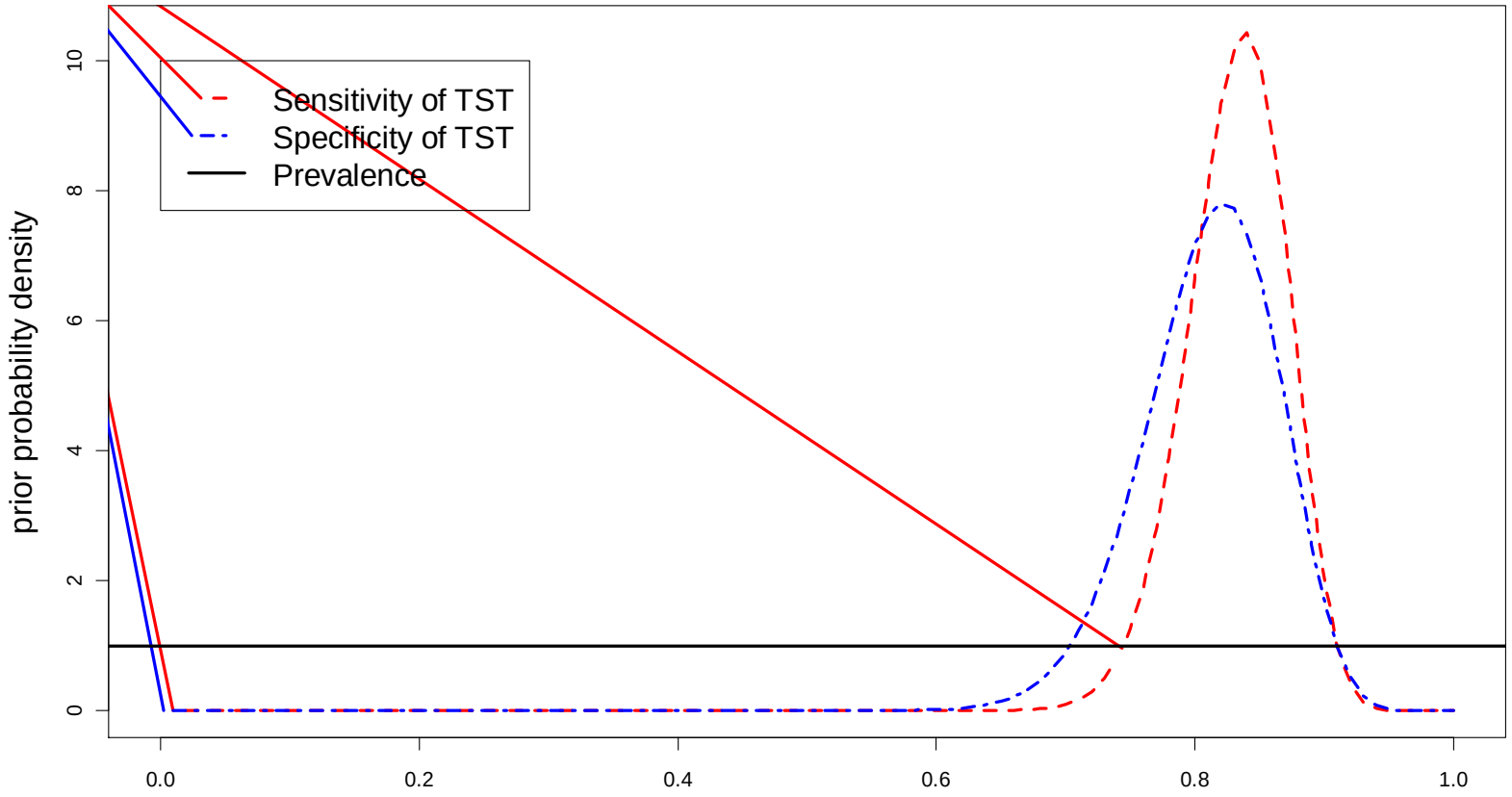
e.g. Y = TST result
X = QFT result

Bayesian approach to estimating a latent class model

- A natural updating method that can simultaneously adjust for uncertainty in all parameters involved in a problem

- Bayesian estimation is a three-step process:
  - Summarize prior information as prior probability distributions for unknown parameters
  - Combine prior information with observed data using Bayes theorem
  - Use resulting posterior distributions to make inferences about unknown parameters

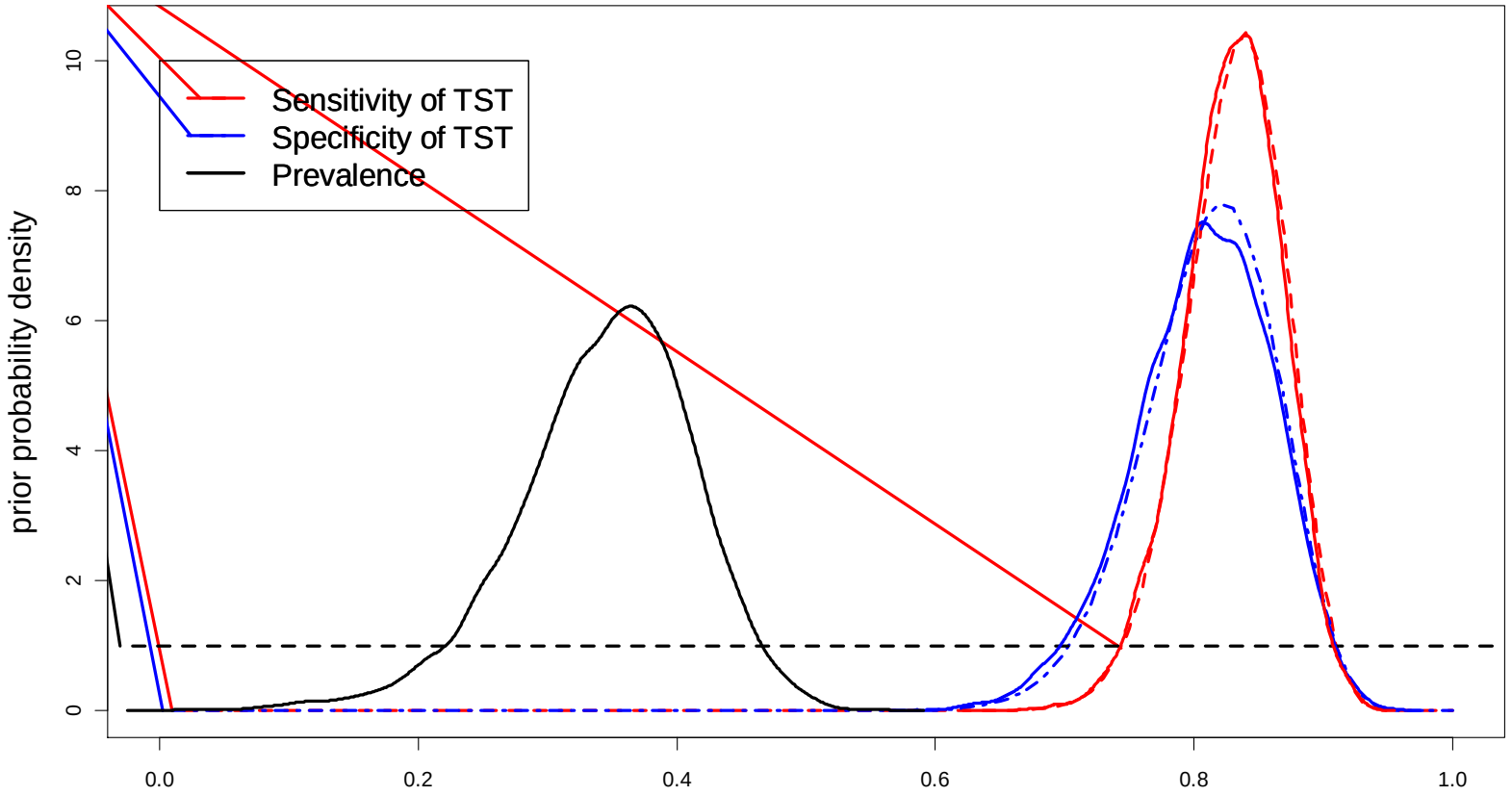
Prior probability distributions



Results of Latent Class Model for TST data alone

	Posterior distribution	
Variable	Median	95 % Credible Interval
P(LTBI+|TST+)	70.5%	40.1% – 86.4%
P(LTBI-|TST-)	90.1%	81.4% – 95.8%
Sensitivity of TST	83.1%	74.9% – 89.7%
Specificity of TST	81.0%	69.3% – 89.9%
Prevalence of LTBI	35.1%	19.3% - 46.3%

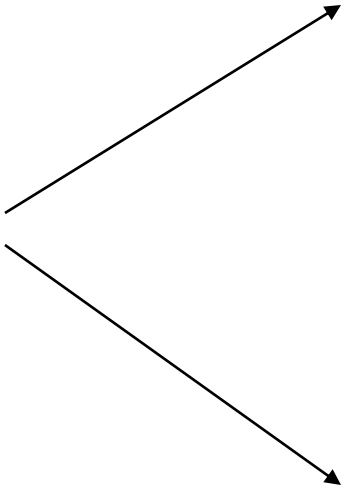
Prior and posterior distributions



Dashed lines: prior distributions; Solid lines: posterior distributions

Latent Class Model: QFT and TST

Observed data		
	QFT-G +	QFT-G -
TST +	226	62
TST -	72	359



Truly infected		
	QFT-G +	QFT-G -
TST +	y11	y10
TST -	y01	y00

Truly non-infected		
	QFT-G +	QFT-G -
TST +	226-y11	62-y10
TST -	72-y01	359-y00

Latent Class Model: QFT and TST

- 5 unknown parameters (S and C of each test, and p) but 3 degrees of freedom
  - Need prior information on atleast 2 parameters

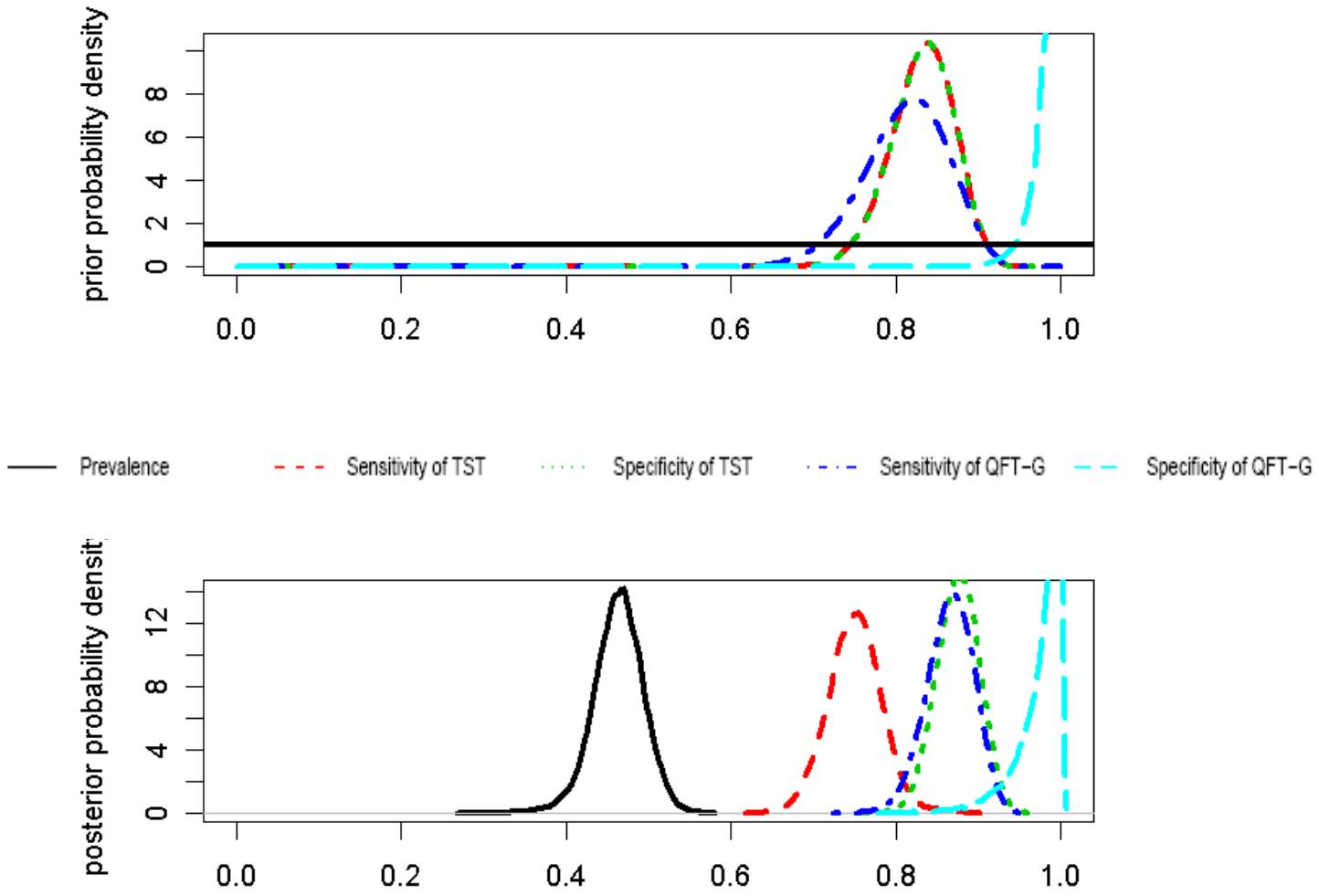
- Prior information on S and C of QFT-G also available:
  - $0.75 < S < 0.90$; $0.95 < C < 1$

- If our focus had been estimation of S and C of QFT-G, we could have used uniform distributions over these parameters instead.

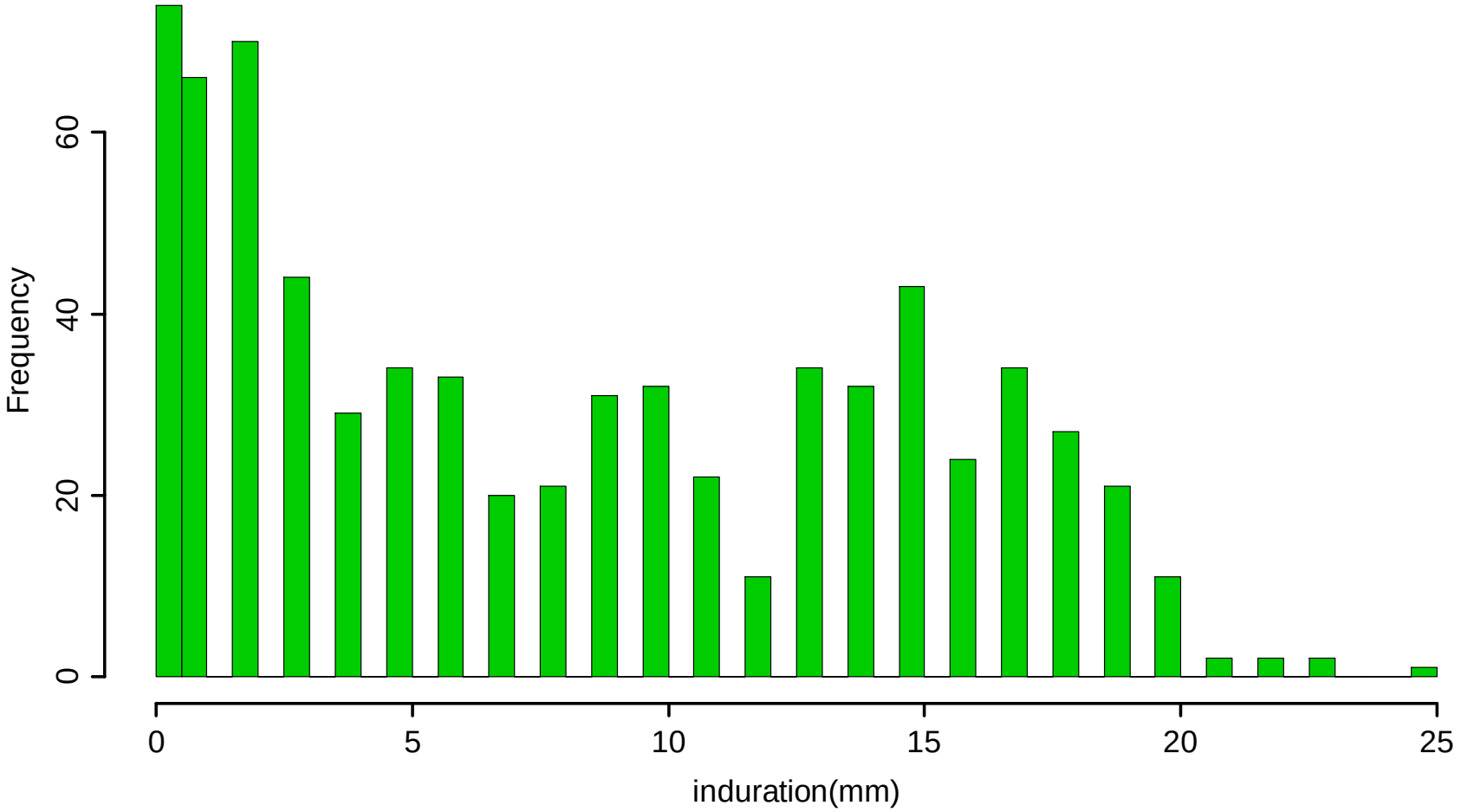
Results of LCA

	Posterior distribution	
Variable	Median	95 % Credible Interval
P(LTBI+|TST+, QFT-G+)	99.2%	99.0% – 100.0%
P(LTBI+|TST+, QFT-G-)	46.0%	29.0% – 65.0%
P(LTBI+|TST-, QFT-G+)	85.0%	69.0% – 94.0%
P(LTBI+|TST-, QFT-G-)	2.0%	1.0% – 4.0%
Sensitivity of TST	79.5%	74.9% – 84.4%
Specificity of TST	87.4%	82.3% – 91.8%
Sensitivity of QFT	89.9%	86.1% – 93.7%
Specificity of QFT	97.4%	94.2% – 98.9%
Prevalence of LTBI	45.4%	39.5% - 51.1%

Prior and posterior distributions

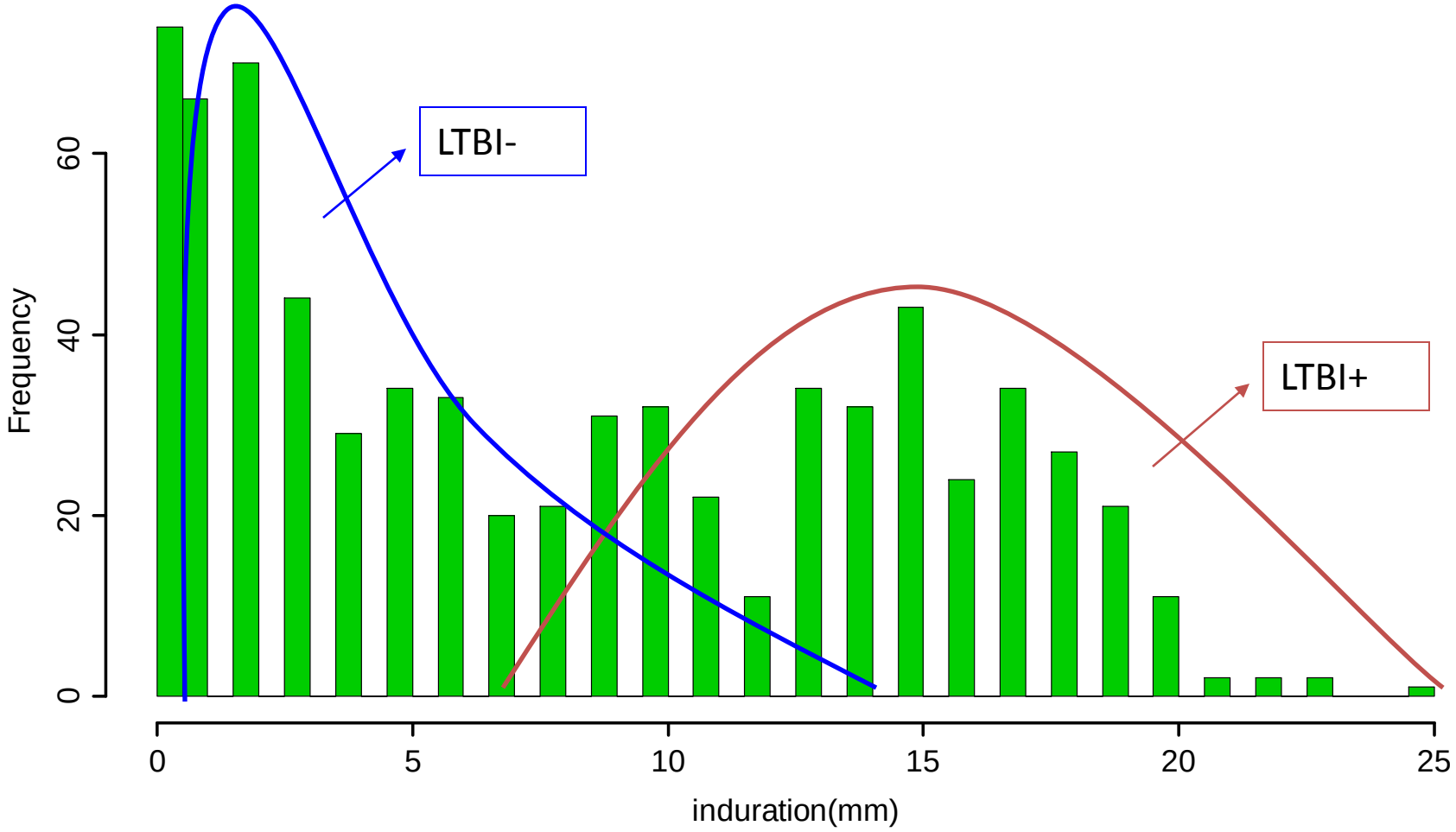


Histogram of TST results

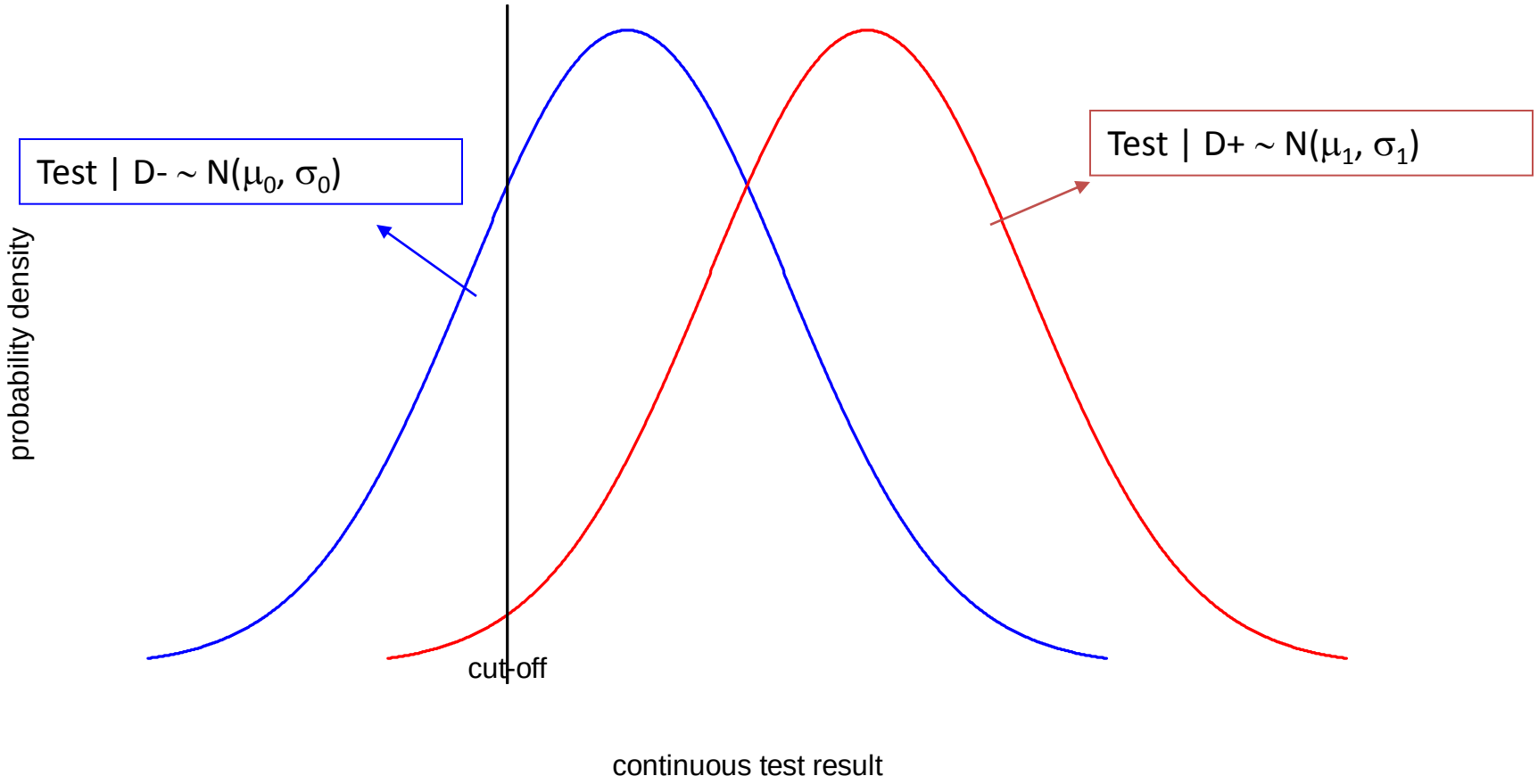


How does the P(LTBI) change with increasing TST?

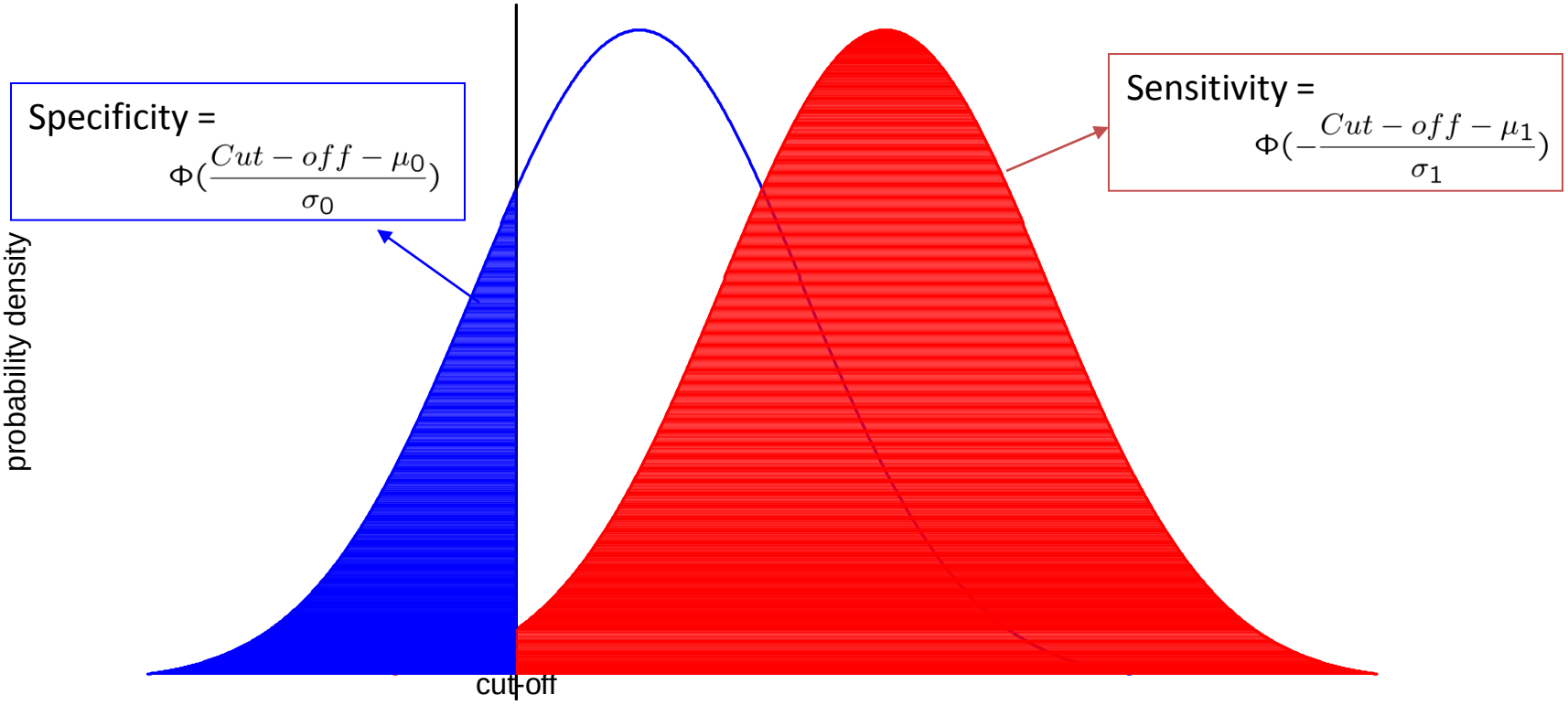
Continuous mixture model



Results from continuous test



Results from continuous test



Continuous mixture model

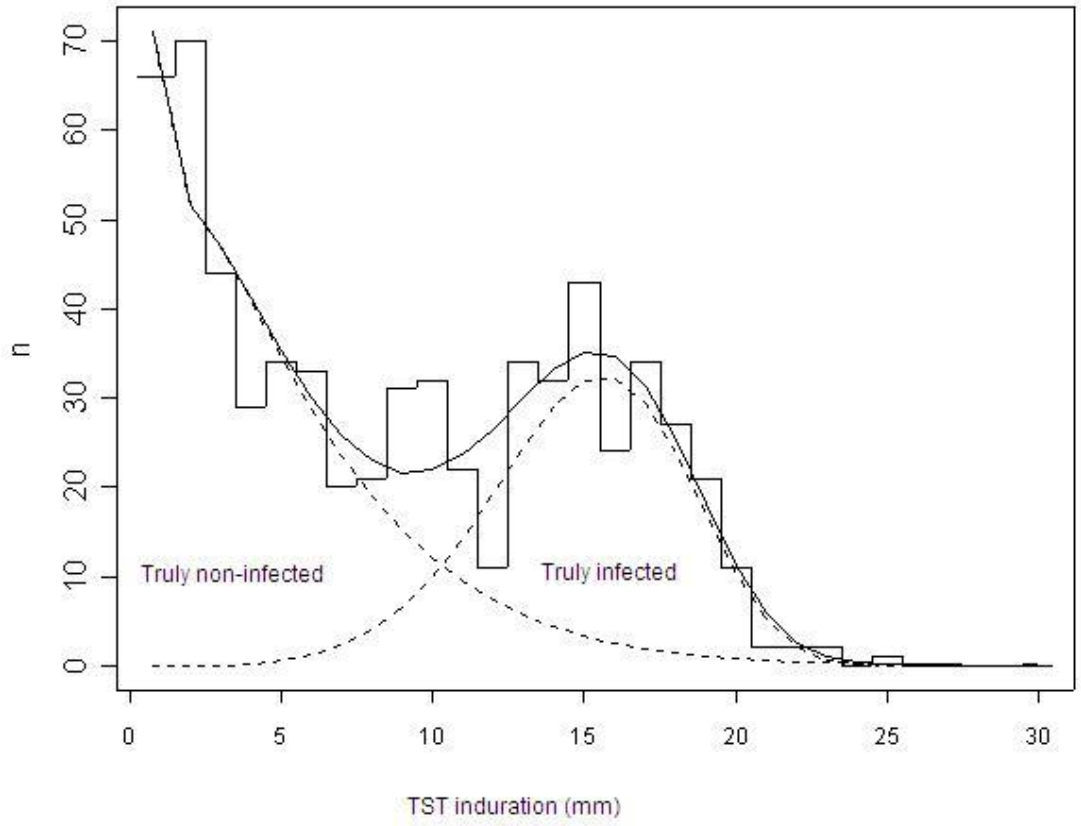
- Unknown parameters are mean and variance of each distribution, and prevalence

- Prior information on these parameters can be used when using a Bayesian approach

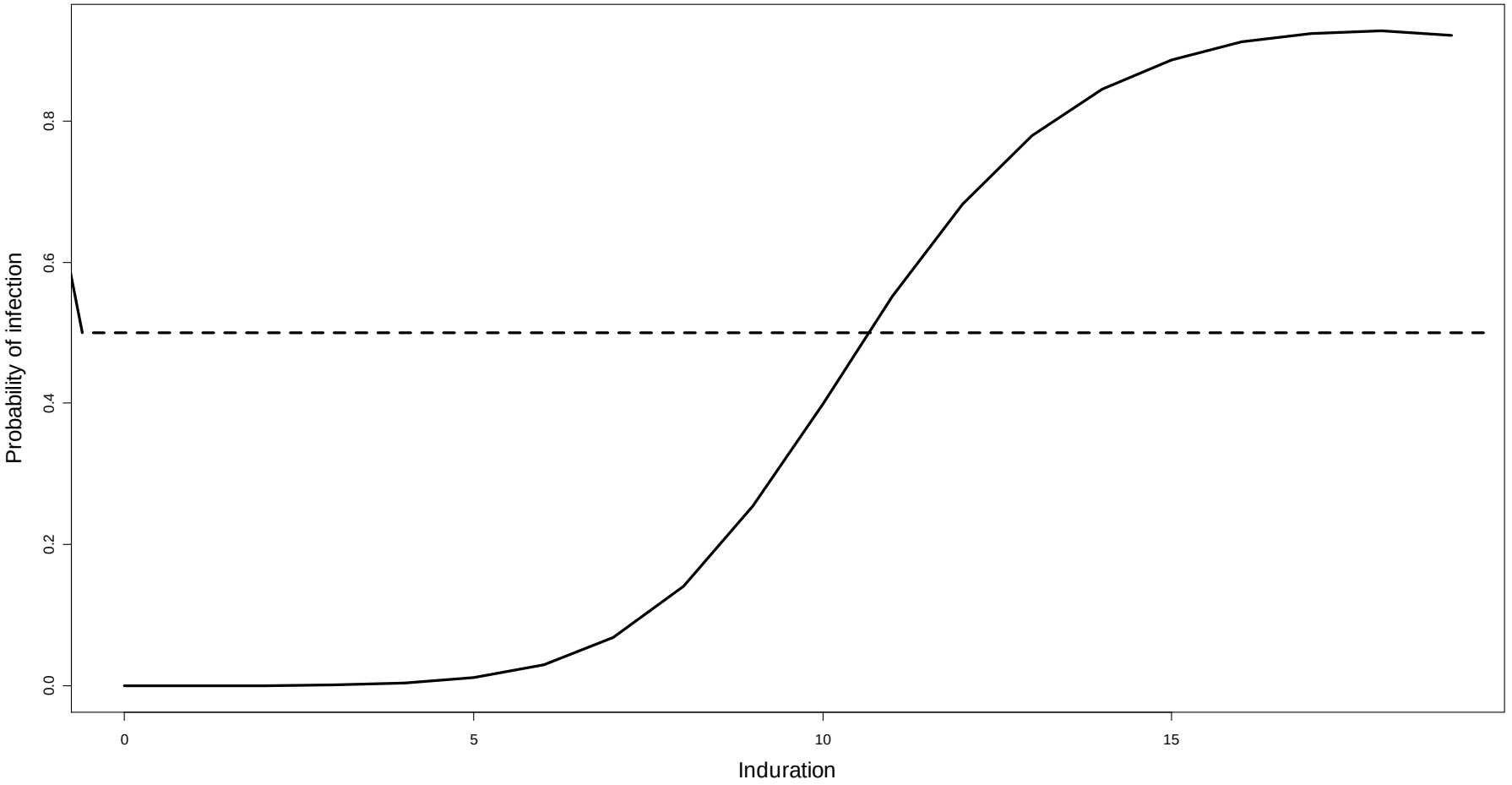
Continuous mixture model for TST

- Fit with R program from IUATLD
  - Program assumes those with induration 0 are true LTBI-
  - Need to provide the density function of TST among cross-reactors and among LTBI+ (i.e. normal, weibull, log-normal)
- Best fitting model for our data was Weibull in both groups
  - Prevalence estimated as 36.5% (28.5%, 47.0%)
  - 0mm induration 10.4%
  - Cross-reactors 53.1%

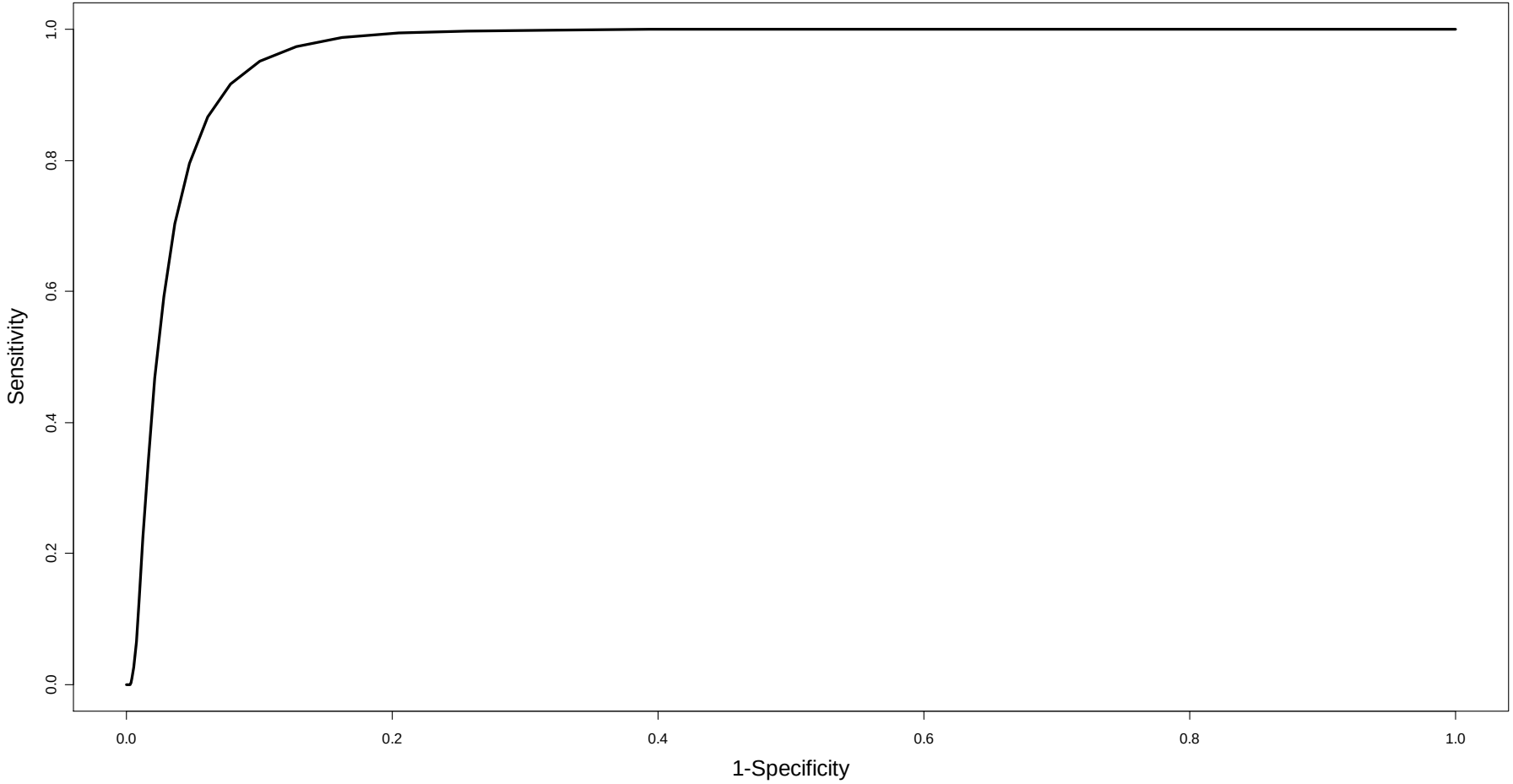
Continuous mixture model for TST



Variation in P(LTBI) with induration



ROC curve



Different estimates from the two mixture models

- At 10mm induration:
  - Continuous mixture: sens=spec=92%
  - LCA: sens=79.5%, spec=89.9%

- Reflects different assumptions of the two models

Prevalence estimates from the various approaches

Method used to estimate prevalence	LTBI Prevalence (%)	95% confidence interval or credible interval (%)
Cut-off point based analysis of TST data		
TST (≥5 mm cut-point)	60.7	57.1 – 64.2
TST (≥10 mm cut-point)	41.4	37.7 – 44.9
TST (≥15 mm cut-point)	23.2	20.1 – 26.3
Cut-off point based analysis of QFT-G data		
QFT-G (IFN-γ ≥0.35 IU/mL, manufacturer's cut-point)	40.1	36.6 – 43.7
Mixture analysis of continuous TST data		
Mixture model of TST (assuming Weibull distributions for both infected and cross-reacting subgroups)	36.5	28.5 - 47.0
Latent class analysis of TST and QFT-G data		
LCA (using prior information on TST and QFT-G)	45.4	40.1 – 49.7

Pros and cons of mixture modeling

- Pros:
  - More realistic
  - Incorporate prior information
  - Extend easily to multiple tests

- Cons:
  - Need specialized software
  - Inferences depend heavily on assumptions

Sample sizes needed for diagnostic studies in the absence of a gold-standard

- Much larger sample sizes are needed to estimate prevalence/sensitivity/specificity in the absence of a gold-standard*
  - In some cases even an infinite sample size may be insufficient

- Falsely assuming the reference standard is perfect in sample size calculations will lead to underestimation of the required sample size

* Dendukuri et al., Biometrics 2004 and Stats in Med 2010