Statistical approaches for dealing with imperfect reference standards

Nandini Dendukuri

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

Technology Assessment Unit, McGill University Health Centre

nandini.dendukuri@mcgill.ca

Advanced TB diagnostics course, Montreal, July 2012

Evaluating Diagnostic Tests in the Absence of a Gold Standard

- Remains a challenging area, particularly relevant to TB diagnostics
- A number of statistical methods have been proposed to get around the problem
- I will review the pros and cons of some of these methods

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

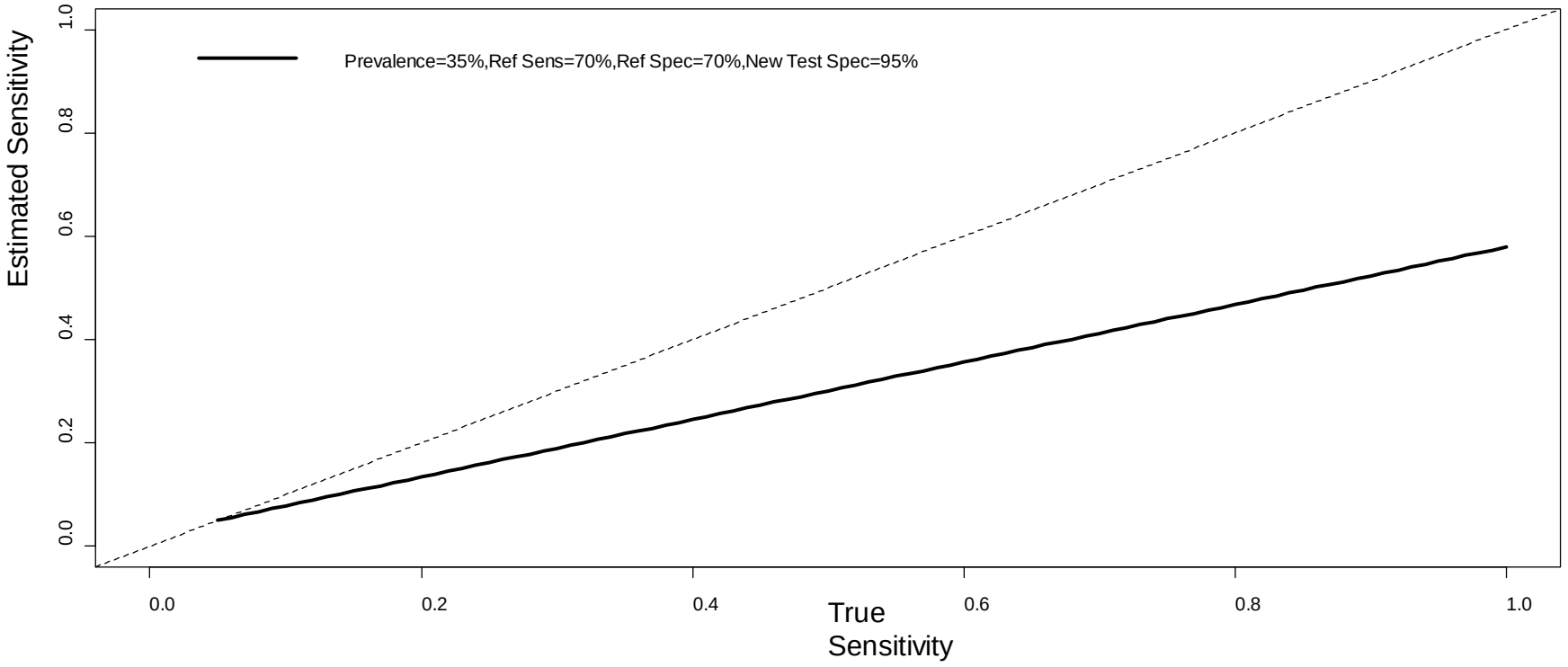
Compare new test to existing standard

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

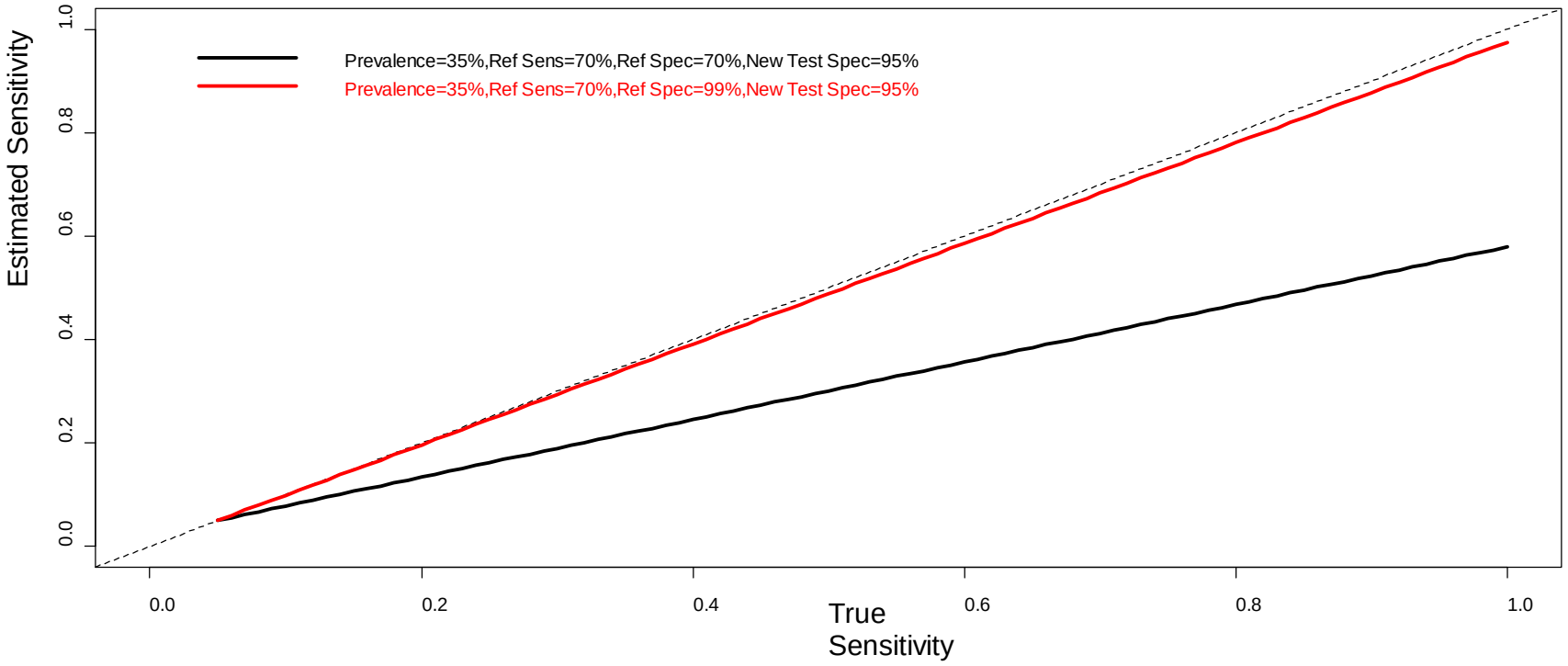
Sensitivity of new test = A/(A+C)

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

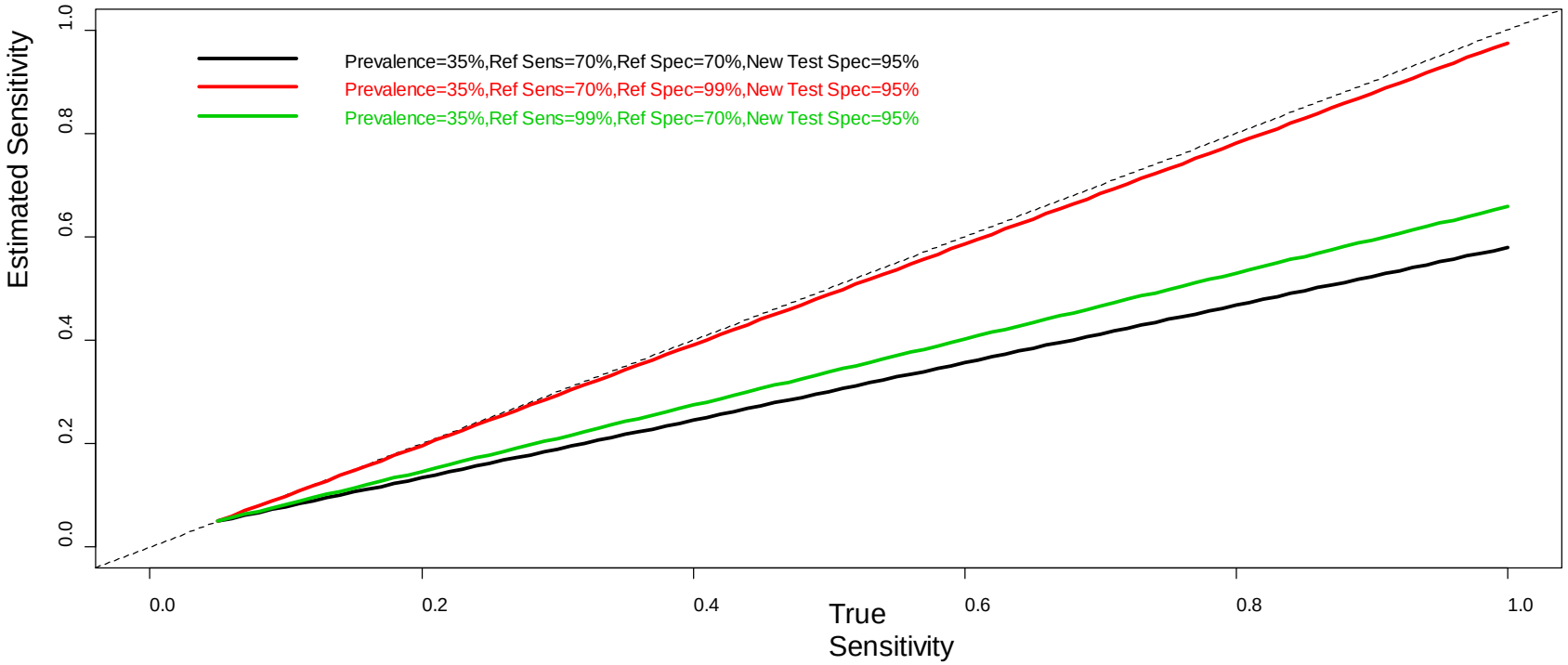
Bias due to assuming reference test is perfect: Impact on sensitivity



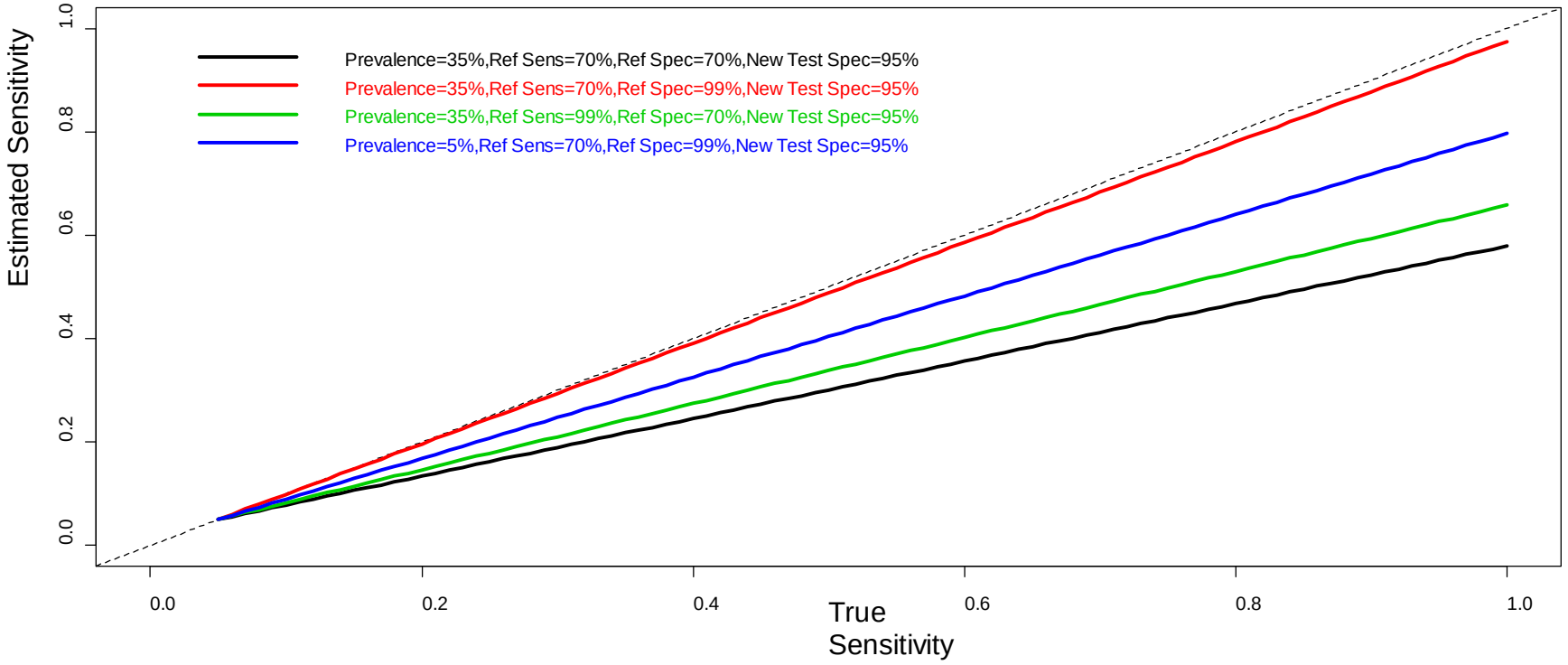
Bias due to assuming reference test is perfect: Impact on sensitivity



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

- Sensitivity and specificity of the reference, as well as prevalence play a role in determining magnitude of bias
  - Specificity (rather than sensitivity) of reference has greater impact on sensitivity of new test
- Similar results can be derived for specificity of the new test
  - 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
   - Estimation of sensitivity/specificity/prevalence
   - Estimation of incremental value
   - Meta-analysis

Discrepant 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

Composite Reference Standard (CRS)*

- Proposed with the aim of developing a reference standard that
  1. Does not involve the test under evaluation
  2. Has higher sensitivity/specificity than individual 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
+	+	+
+	-	-
-	+	-
-	-	-

Patient Infection Status Algorithm (PISA)

- PISA is a type of CRS that has been used to produce sensitivity/specificity estimates in test kits cleared by FDA

- PISA is typically based on two tests carried out on two specimens
  - e.g. Two different NAATs both carried out on urine and cervical specimens
  - Once again, different definitions of PISA are possible

Patient Infection Status Algorithm (PISA)*

TABLE 1. Infection Status Based on the Patient-infected-status Algorithm (PISA) and PISA2

	Test 1		Test 2			
Profile	Specimen 1 Comparator1	Specimen 2 Comparator2	Specimen 1 Comparator3	Specimen 2 Comparator4	PISA	PISA2
1	+	+	+	+	Infected	Infected
2	+	+	+	−	Infected	Infected
3	+	+	−	+	Infected	Infected
4	+	−	+	+	Infected	Infected
5	+	−	+	−	Infected	Infected
6	+	−	−	+	Infected	Infected
7	−	+	+	+	Infected	Infected
8	−	+	+	−	Infected	Infected
9	−	+	−	+	Infected	Infected
10	+	+	−	−	Infected	Uninfected
11	−	−	+	+	Infected	Uninfected
12	+	−	−	−	Uninfected	Uninfected
13	−	+	−	−	Uninfected	Uninfected
14	−	−	+	−	Uninfected	Uninfected
15	−	−	−	+	Uninfected	Uninfected
16	−	−	−	−	Uninfected	Uninfected

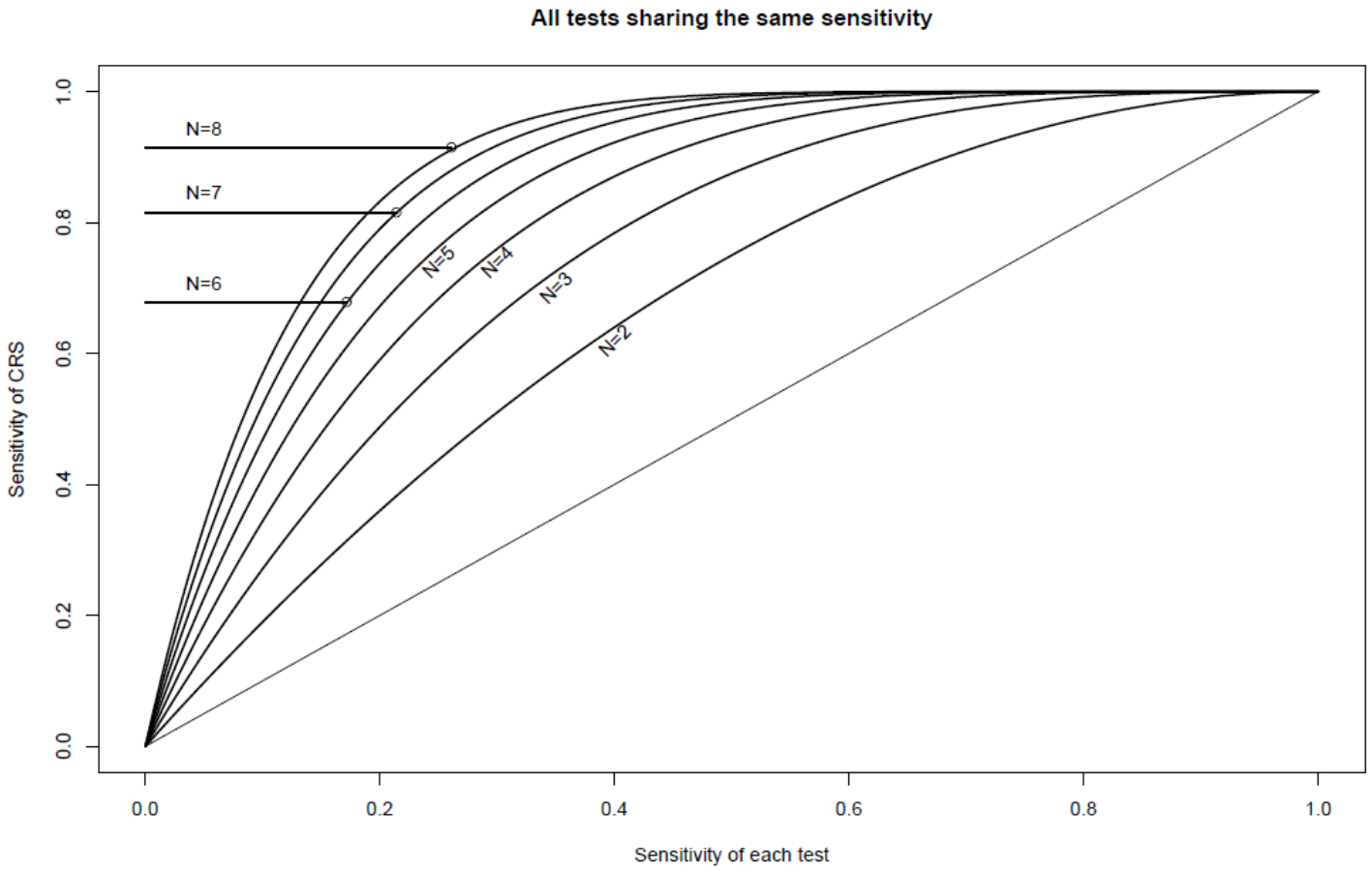
Hadgu et al., Epidemiology, 2012

Composite Reference Standard

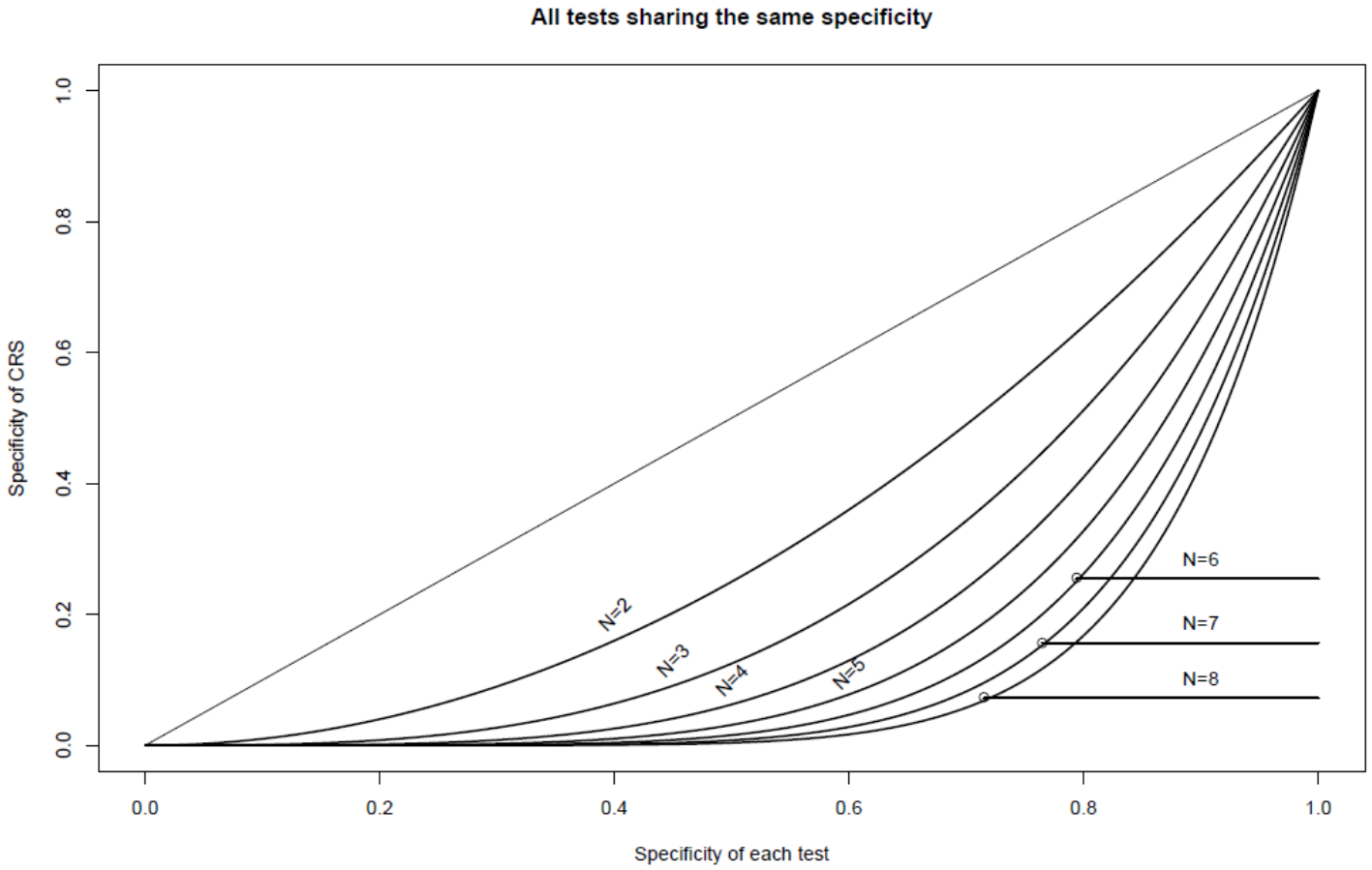
- Drawbacks *
  - Same problems that arise when treating a single, imperfect standard test as perfect, i.e. underestimation of test properties
  - Operating characteristics poorly understood
    - e.g. Liberal CRS assumes that both T1 and T2 have perfect specificity, which may not be the case
  - When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

* Dendukuri et al, SBR, 2011

Sensitivity of CRS as a function of sensitivity of individual tests and number of tests*



Specificity of CRS as a function of specificity of individual tests and number of tests*



Biased estimates possible even with high sensitivity and specificity of PISA*

TABLE 4. Sensitivity and Specificity Estimates and Their Associated 95% Confidence/Credible Intervals (CIs) Obtained by Using 4 Estimation Approaches

Test	True Value	PISA[a] Estimate (95% CI)	PISA2[b] Estimate (95% CI)	Imperf GS Estimate (95% CI)	LCM Estimate (95% CI)
			Sensitivity		
Comparator1	95.0			75.1 (71.1–78.7)	94.6 (92.6–96.4)
Comparator2	95.0			75.3 (71.3–78.9)	93.6 (91.4–95.6)
Comparator3	95.0			76.5 (72.6–80.0)	94.6 (92.4–96.4)
Comparator4	95.0			76.5 (72.6–80.0)	95.8 (93.9–97.3)
New Test1	95.0	75.5 (72.0–78.7)	80.8 (77.4–83.8)	75.3 (71.3–78.9)	93.7 (91.5–95.7)
New Test2	75.0	62.3 (58.4–65.9)	67.2 (63.3–70.9)	60.4 (56.1–64.6)	77.2 (73.5–80.7)
Imperfect GS	80.0	62.1 (58.3–65.8)	66.7 (62.8–70.4)	—	77.7 (74.0–81.1)
			Specificity		
Comparator1	95.0			93.5 (93.0–94.0)	94.6 (94.2–95.0)
Comparator2	95.0			94.3 (93.8–94.7)	95.3 (94.9–95.7)
Comparator3	95.0			93.7 (93.2–94.2)	94.7 (94.3–95.1)
Comparator4	95.0			94.1 (93.6–94.5)	95.1 (94.7–95.6)
New Test1	95.0	94.7 (94.2–95.1)	94.7 (94.2–95.1)	93.7 (93.1–94.1)	94.7 (94.2–95.1)
New Test2	97.5	97.5 (97.2–97.8)	97.5 (97.2–97.8)	96.5 (96.2–96.9)	97.5 (97.1–97.8)
Imperfect GS	99.0	98.9 (98.7–99.1)	98.9 (98.7–99.1)	—	98.9 (98.7–99.1)
Prevalence	5.0	6.4 (5.9–6.9)	5.9 (5.4–6.4)	5.0 (4.6–5.4)	5.0 (4.6–5.5)

The data were simulated assuming conditional independence and that the comparator tests have good true sensitivity and specificity values (95% sensitivity and 95% specificity).
[a]Sensitivity of PISA = 99.9; specificity of PISA = 98.6.
[b]Sensitivity of PISA2 = 99.5; specificity of PISA2 = 99.0.

Hadgu et al, Epidemiology, 2012

Latent Class Analysis: Estimating sensitivity, specificity and prevalence

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

Cross-tabulation of TST and QFT-G

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

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

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

* >=10mm induration

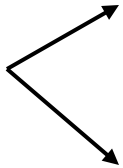
Mixture models

- 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 applied when one or more test results are observed
- 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

Latent Class Model: TST alone

TST+	TST-
288	431



	TST+	TST-	
LTBI+	X	Y	X+Y
LTBI -	288-X	431-Y	719-(X+Y)

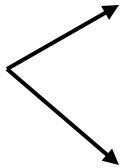
- Note that if we knew X and Y we can estimate:
  - Prevalence of LTBI = (X+Y)/719
  - Sensitivity of TST = X/(X+Y)
  - Specificity of TST = (431-Y)/(719-(X+Y))

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% = 431/(719-X). Therefore, X=144
  - Knowing, X and Y we can determine the third parameter. Prevalence = 144/719 = 20%

X and Y in terms of sens/spec/prev

TST+	TST-
288	431



	TST+	TST-	
LTBI+	X	Y	X+Y
LTBI -	288-X	431-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 -	288-X = 719(1-p)(1-C)	431-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)

- $\Rightarrow 288 = 719 \times (ps + (1-p)(1-c))$
  - $\Rightarrow$ 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 (Pai et al., Ann Int Med 2008):
  - $0.70 < S < 0.80$
  - $0.96 < C < 0.99$
- Different values in these ranges would yield different prevalence estimates.
  - If s=0.70 and c=0.96 then p=0.57
  - If s=0.80 and c=0.99 then p=0.51
- 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 = $\alpha + \beta$ Contact

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

Observed Data x, y: Collected on a sample

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

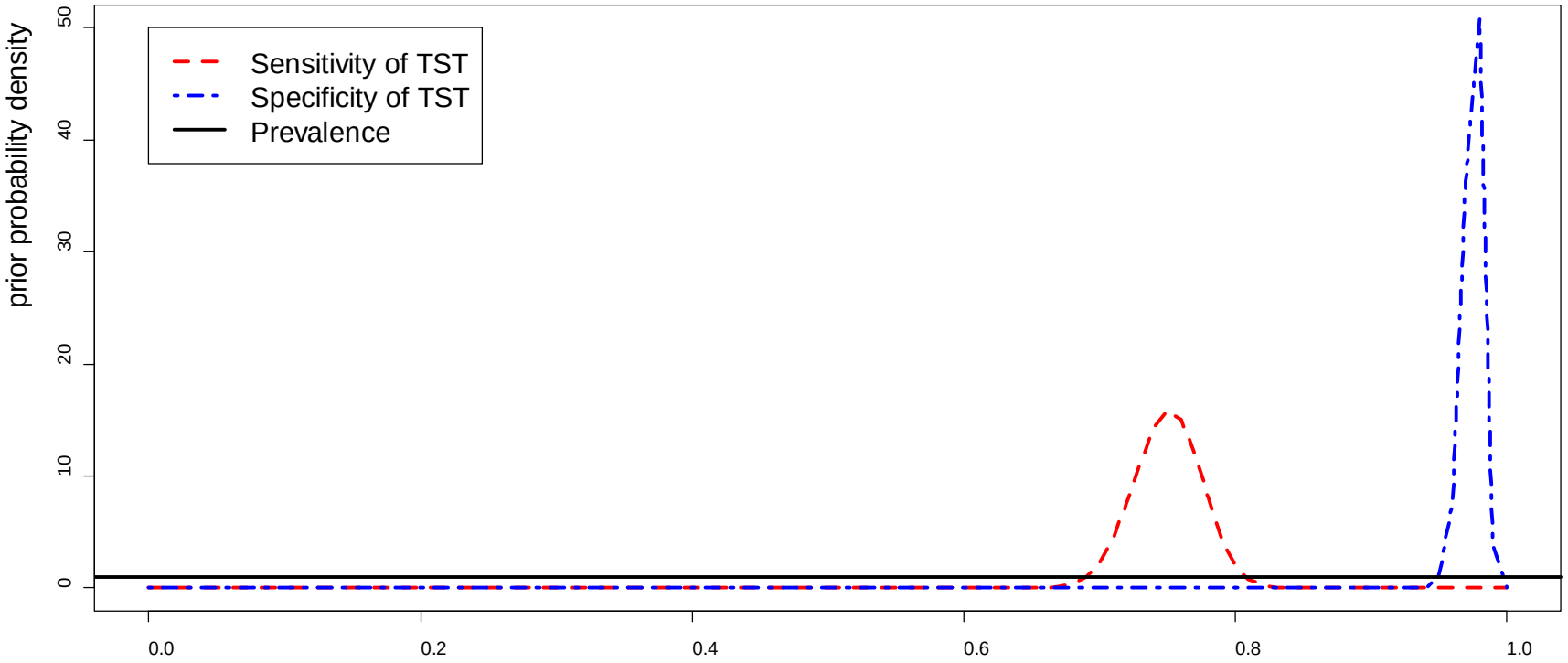
e.g. Y = TST result
X = QFT result

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

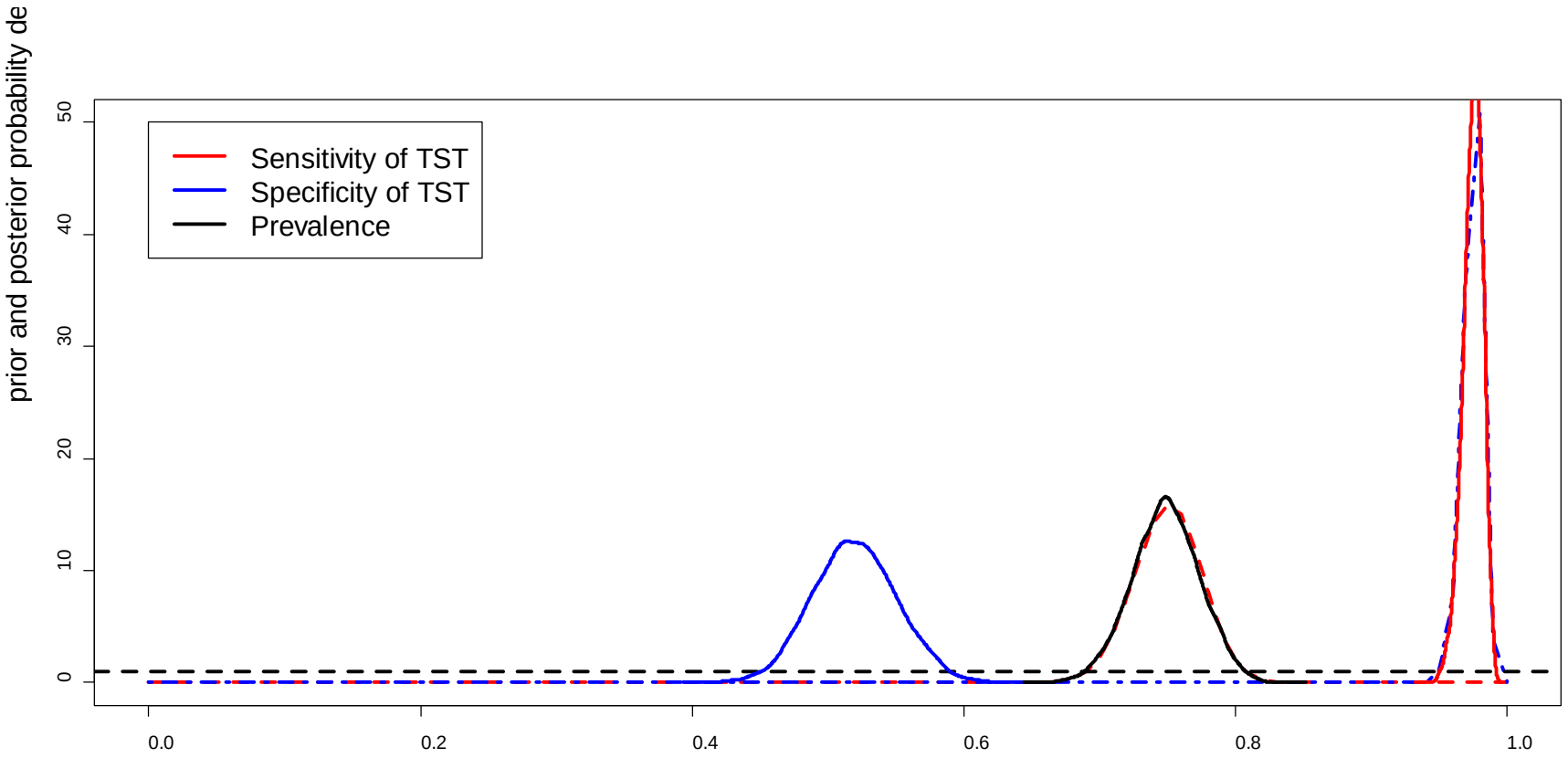
Prior probability distributions



Results of Latent Class Model for TST data alone

	Posterior distribution	
Variable	Median	95 % Credible Interval
P(LTBI+|TST+)	97.1%	94.6% – 98.6%
P(LTBI-|TST-)	78.3%	70.3% – 84.2%
Sensitivity of TST	74.9%	69.9% – 79.7%
Specificity of TST	97.6%	95.8% – 98.8%
Prevalence of LTBI	51.9%	46.0% - 58.1%

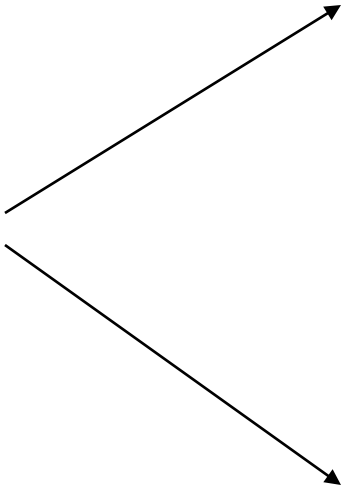
Prior and posterior distributions



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 at least 2 parameters

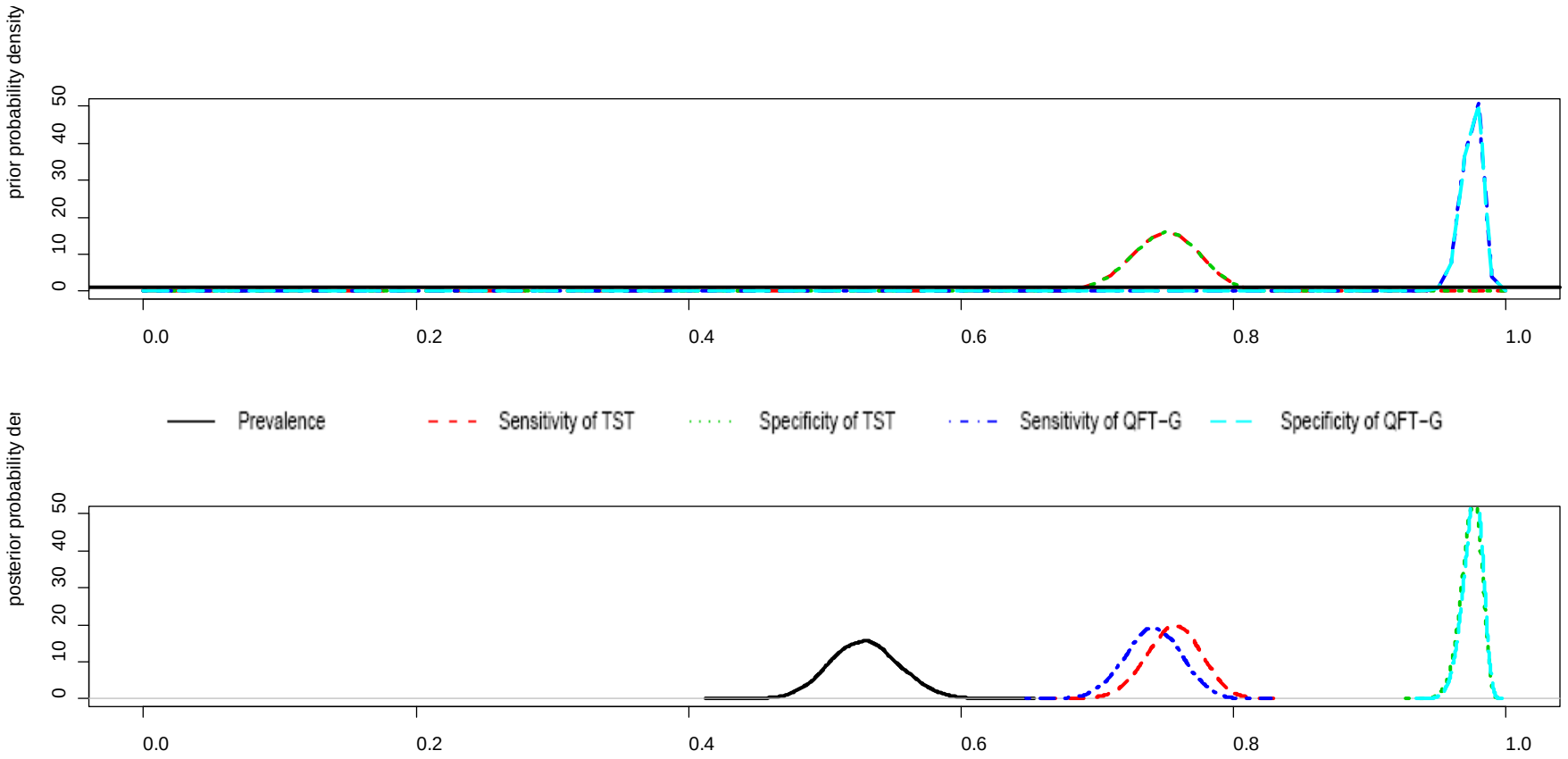
- Prior information on S and C of QFT-G also available:
  - $0.7 < S < 0.8$; $0.96 < C < 0.99$

- 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 for both tests

	Posterior distribution	
Variable	Median	95 % Credible Interval
P(LTBI+|TST+, QFT-G+)	98.6%	96.2% – 99.9%
P(LTBI+|TST+, QFT-G-)	92.9%	82.6% – 99.5%
P(LTBI+|TST-, QFT-G+)	92.8%	81.9% – 99.5%
P(LTBI+|TST-, QFT-G-)	9.3%	5.5% – 15.5%
Sensitivity of TST	75.7%	71.6% – 79.5%
Specificity of TST	97.5%	95.8% – 98.7%
Sensitivity of QFT	74.1%	70.0% – 78.1%
Specificity of QFT	97.6%	95.9% – 98.7%
Prevalence of LTBI	52.9%	48.1% - 58.1%

Prior and posterior distributions



Latent Class Analysis: Estimating incremental value

Estimation of incremental value

- We have recently shown† how to estimate incremental value in the absence of a gold-standard
  - If the predictive value of the test (s) is used in decision making, then statistics like difference in AUC* or IDI** may be useful
  - If decisions are based on the observed test results then the incremental value may be determined by comparing the predictive values of the decision rules based on using 2 tests vs. 1 test

† Ling et al, under review* AUC: Area Under the Curve

**IDI: Integrated Discrimination Index

How do we define the true disease status?

- We have argued that the best estimate of the true disease status is obtained by using all available information, i.e. results of both tests and prior information

- One way to think of it is that at each iteration of the Gibbs sampler, each patient is classified as D+ or D-. Incremental value is obtained by averaging across iterations

Illustration of calculation of IDI

(assuming sens1=0.7, sens2=0.8, spec1=spec2=0.9)

T1, T2, D	P(D|T1,T2)	P(D|T1)	Difference	Weight (P(T1,T2|D))	Contribution to IDI (weight×difference)
+++	0.96	0.75	0.21	0.56	0.12
+-+	0.41	0.75	-0.34	0.14	-0.05
-++	0.54	0.13	0.41	0.24	0.10
--+	0.03	0.13	-0.1	0.06	-0.001
	Incremental value among D+ (Σweight×difference)				0.17
++-	0.04	0.25	-0.21	0.01	-0.002
+--	0.59	0.25	0.34	0.09	0.03
-+-	0.46	0.87	-0.41	0.09	-0.04
---	0.97	0.87	0.1	0.81	0.08
	Incremental value among D- (Σweight×difference))				0.07
				Overall incremental value	0.24

Median incremental value of second test vs. its sens & spec*

Accuracy of T2 vs T1		AUC difference	IDI in events	IDI in non events	IDI [b]
1) higher sens	S_2=80, C_2=90	0.13	0.17	0.07	0.24
2) higher spec	S_2=70, C_2=100	0.12	0.20	0.08	0.28
3) lower sens	S_2=60, C_2=90	0.09	0.10	0.04	0.14
4) lower spec	S_2=70, C_2=80	0.09	0.08	0.03	0.11
5) both better	S_2=80, C_2=100	0.14	0.24	0.10	0.34
6) both worse	S_2=60, C_2=80	0.07	0.05	0.02	0.07
7) No better	S_2=70, C_2=90	0.10	0.13	0.06	0.19
8) No value	S_2=70, C_2=30	0.008	<0.001	<0.001	0.001

sens1=0.7 and spec1=0.9

Applied example: Incremental value of IFN-γ over TST

- IFN-γ is a promising alternative to TST for screening latent TB infection due to its supposedly superior specificity

- Experience over the last decade has shown that its performance may vary according to whether it is used in a setting where BCG vaccination was given once (e.g. India) or multiple times (e.g. Portugal)

- We estimated incremental value separately in datasets from two different studies of health care workers, one from India and one from Portugal

Cross-tabulation of TST and QFT-G in data from India and Portugal

India (Pai et al, JAMA, 2004)		
	TST+*	TST-
QFT-G+	226	72
QFT-G-	62	359

Portugal (Torres et al, Eur Res J, 2009)		
	TST+*	TST-
QFT-G+	371	26
QFT-G-	532	289

* >=10mm induration

Range of prior distributions for India and Portugal data*

	India	Portugal
Sensitivity of TST	70-80%	70-80%
Specificity of TST	96-99%	55-65%
Sensitivity of QFT	70-80%	70-80%
Specificity of QFT	96-99%	96-99%

Pai et al, Ann Int Med, 2008

Results of latent class analysis

	TST Sensitivity (95% CrI)	**TST Specificity (95% CrI)**	**QFT Sensitivity (95% CrI)**	**QFT Specificity (95% CrI)**	**Prevalence (95% CrI)**
India study (n=719)	0.74 (0.70, 0.78)	0.98 (0.96, 0.99)	0.76 (0.72, 0.80)	0.98 (0.96, 0.99)	0.53 (0.48, 0.58)
Portugal study (n=1218)	0.84 (0.81, 0.87)	0.46 (0.42, 0.51)	0.69 (0.62, 0.75)	0.98 (0.97, 0.99)	0.47 (0.41, 0.55)

Results of latent class analysis

	AUC difference (95% CrI)	**IDI (95% CrI)**
India study (n=719)	0.08 (0.06, 0.11)	0.23 (0.16, 0.29)
Portugal study (n=1218)	0.21 (0.17, 0.25)	0.40 (0.29, 0.51)

Incremental value of decision rules based on observed data

Decision Rule	n (%) Classified correctly	Incremental value of QFT (%)
India study (N=719)		
LTBI+ if TST+	611 (85)	
LTBI+ if TST+ and QFT+	558 (78)	-7%
LTBI+ if TST+ or QFT+	673 (94)	9%
Portugal study (N=1218)		
LTBI+ if TST+	749 (61)	
LTBI+ if TST+ and QFT+	995 (82)	21%
LTBI+ if TST+ or QFT+	771 (63)	2%

Latent Class Analysis: Meta-analysis setting

Biometrics DOI: 10.1111/j.1541-0420.2012.01773.x

Bayesian Meta-Analysis of the Accuracy of a Test for Tuberculous Pleuritis in the Absence of a Gold Standard Reference

Nandini Dendukuri,[1,*] Ian Schiller,[2] Lawrence Joseph,[1] and Madhukar Pai[1]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal H3A 1A2, Canada

[2]Division of Clinical Epidemiology, McGill University Health Centre, Montreal H3A 1A1, Canada

[*]*email:* nandini.dendukuri@mcgill.ca

Reference standard bias in TB diagnostic meta-analyses

- As previously discussed, reference standard bias may arise in individual studies due to an imperfect reference test

- In a meta-analysis setting, the problem is worsened because each study may use a different reference standard
  - Thus the diagnostic meta-analyses may not be pooling the same quantity across studies!

Reference standard bias in TB diagnostic meta-analyses

Table 1

Studies included in meta-analysis of in-house nucleic acid amplification tests for tuberculous pleuritis (Source: Pai et al. 2004).

Study	Author (Year)	Index (T_1) and reference (T_2) test results				Reference test	Sensitivity of reference test
		$T_1 = 1$ $T_2 = 1$	$T_1 = 1$ $T_2 = 0$	$T_1 = 0$ $T_2 = 1$	$T_1 = 0$ $T_2 = 0$		
1	Chan (1996)	11	1	14	75	Culture	20–60%
2	Gunisha (2001)	1	1	3	25	Culture	20–60%
3	Almeda (2000)	8	0	1	16	Culture/Clinical data	20–70%
4	Tan (1997)	16	6	0	43	Culture/Clinical data	20–70%
5	Portillo-Gomez (2000)	16	0	1	56	Culture/Biopsy	70–90%
6	De Lassence (1992)	9	0	6	10	Culture/Biopsy	70–90%
7	Mangiapan (1996)	13	0	4	25	Culture/Biopsy	70–90%
8	Querol (1995)	17	2	4	84	Culture/Biopsy	70–90%
9	Tan, Jama (1995)	7	0	3	13	Culture/Biopsy	70–90%
10	Villena (1998)	14	1	19	97	Culture/Biopsy	70–90%
11	Villegas (2000)	31	7	11	63	Culture/Biopsy	70–90%

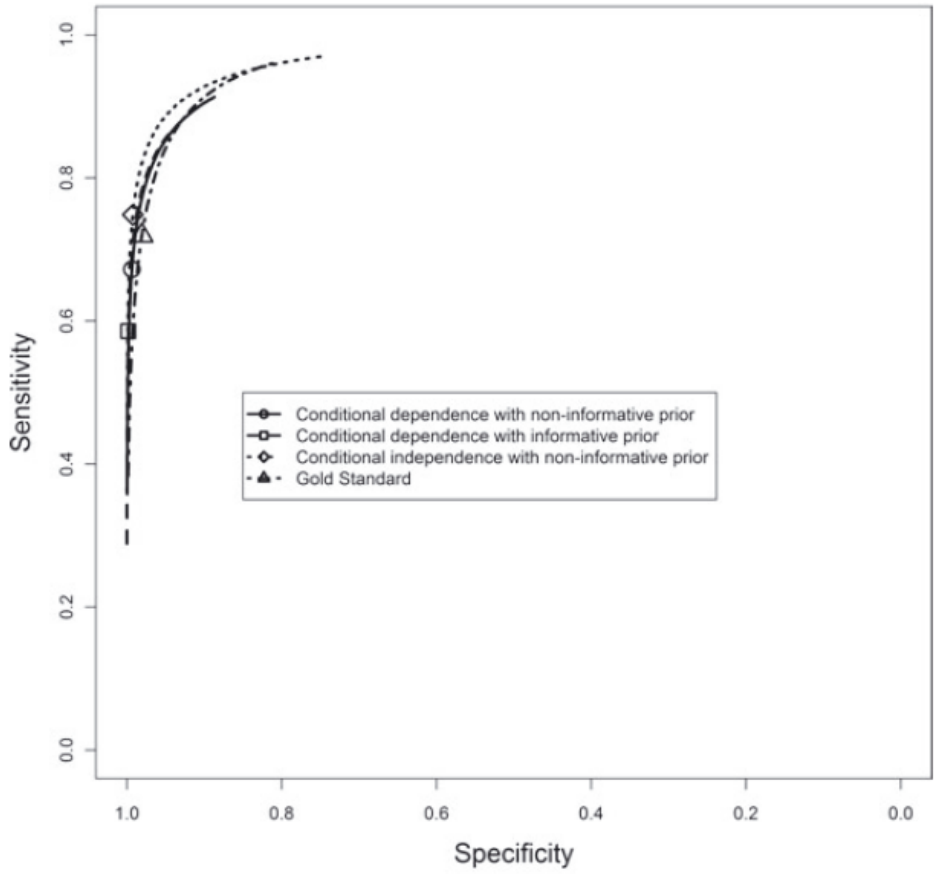
Latent class analysis in a meta-analytic setting*

- We recently extended the well known HSROC model to include a latent class framework*

- We have developed a number of programs in R, SAS and WinBUGS to support these models:
  - See http://www.nandinidendukuri.com under Software

Dendukuri et al, Biometrics, 2012

Meta-analysis of in-house NAATs for TB pleuritis



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