

Presenting results and summary of findings tables

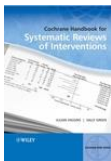
Yemisi Takwoingi

October 2015

Based on slides developed by Jon Deeks and Mariska Leeflang

Learning objectives

- Understand how estimate of test accuracy can be used to describe the practical implications of using a test
- Use the RevMan DTA calculator tool to compute likelihood ratios, predictive values, and normalised frequencies
- Understand the importance of prevalence



See Chapter 11 of the DTA Handbook
available at dta.cochrane.org

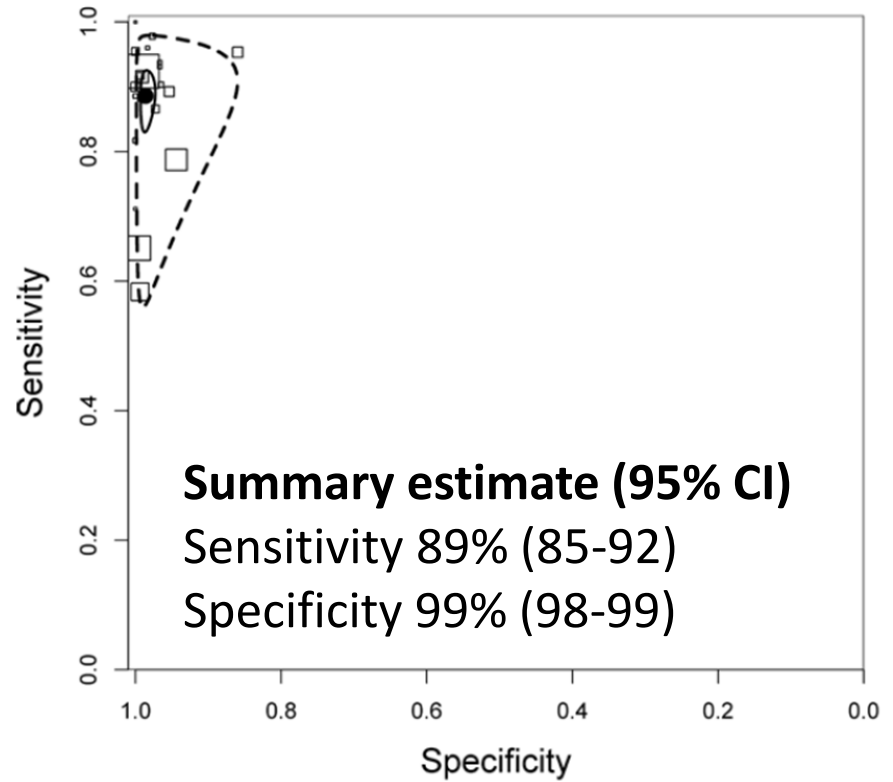
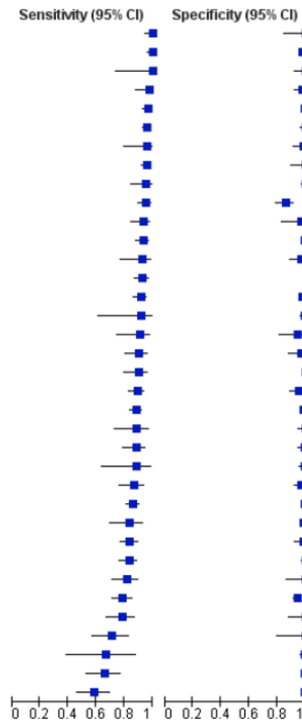
Outline

- Review measures of test accuracy
- Apply the RevMan calculator to compute:
 - likelihood ratios
 - positive and negative predictive values
 - normalised frequencies
- Application for comparisons of tests
- Simple summaries from SROC curves
- To consider the challenges of summarising the findings of a DTA review
- To be able to define the key components of a summary of findings table for Cochrane DTA reviews

Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults (Review)

Steingart KR, Schiller I, Horne DJ, Pai M, Boehme CC, Dendukuri N

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Williamson 2012	67	0	0	22	1.00 [0.95, 1.00]	1.00 [0.85, 1.00]		
Boehme 2011e	101	16	0	671	1.00 [0.96, 1.00]	0.98 [0.96, 0.99]		
Malbruny 2011	12	0	0	46	1.00 [0.74, 1.00]	1.00 [0.92, 1.00]		
Carriquiry 2012	44	2	1	84	0.98 [0.88, 1.00]	0.98 [0.92, 1.00]		
Boehme 2011b	171	3	6	825	0.97 [0.93, 0.99]	1.00 [0.99, 1.00]		
Boehme 2010b	201	0	8	101	0.96 [0.93, 0.98]	1.00 [0.96, 1.00]		
Ciftci 2011	24	1	1	59	0.96 [0.80, 1.00]	0.98 [0.91, 1.00]		
Boehme 2010e	179	0	8	35	0.96 [0.92, 0.98]	1.00 [0.90, 1.00]		
Al-Ateah 2012	42	0	2	128	0.95 [0.85, 0.99]	1.00 [0.97, 1.00]		
Kurbatova 2013	102	17	5	104	0.95 [0.89, 0.98]	0.86 [0.78, 0.92]		
Bowles 2011	60	1	4	29	0.94 [0.85, 0.98]	0.97 [0.83, 1.00]		
Boehme 2010c	136	1	10	185	0.93 [0.88, 0.97]	0.99 [0.97, 1.00]		
Miller 2011	27	2	2	58	0.93 [0.77, 0.99]	0.97 [0.88, 1.00]		
Friedrich 2011	117	0	9	0	0.93 [0.87, 0.97]	Not estimable		
Boehme 2011f	136	5	12	234	0.92 [0.86, 0.96]	0.98 [0.95, 0.99]		
Balcells 2012	11	1	1	147	0.92 [0.62, 1.00]	0.99 [0.96, 1.00]		
Ioannidis 2011	29	2	3	33	0.91 [0.75, 0.98]	0.94 [0.81, 0.99]		
Teo 2011	56	2	6	55	0.90 [0.80, 0.96]	0.96 [0.88, 1.00]		
Hanif 2011	54	0	6	146	0.90 [0.79, 0.96]	1.00 [0.98, 1.00]		
Marlowe 2011	116	4	14	82	0.89 [0.83, 0.94]	0.95 [0.89, 0.99]		
Boehme 2011a	203	4	26	303	0.89 [0.84, 0.92]	0.99 [0.97, 1.00]		
Zeka 2011	31	0	4	68	0.89 [0.73, 0.97]	1.00 [0.95, 1.00]		
Rachow 2011	61	1	8	102	0.88 [0.78, 0.95]	0.99 [0.95, 1.00]		
Safianowska 2012	15	1	2	127	0.88 [0.64, 0.99]	0.99 [0.96, 1.00]		
Scott 2011	58	3	9	107	0.87 [0.76, 0.94]	0.97 [0.92, 0.99]		
Boehme 2011c	201	2	32	669	0.86 [0.81, 0.90]	1.00 [0.99, 1.00]		
Boehme 2010d	36	3	7	215	0.84 [0.69, 0.93]	0.99 [0.96, 1.00]		
Boehme 2010a	123	1	24	68	0.84 [0.77, 0.89]	0.99 [0.92, 1.00]		
Boehme 2011d	121	0	24	144	0.83 [0.76, 0.89]	1.00 [0.97, 1.00]		
Helb 2010	67	0	15	25	0.82 [0.72, 0.89]	1.00 [0.86, 1.00]		
Theron 2011	111	19	30	320	0.79 [0.71, 0.85]	0.94 [0.91, 0.97]		
Moure 2011	61	0	17	29	0.78 [0.67, 0.87]	1.00 [0.88, 1.00]		
Barnard 2012	37	0	15	16	0.71 [0.57, 0.83]	1.00 [0.79, 1.00]		
Van Rie 2013	10	1	5	145	0.67 [0.38, 0.88]	0.99 [0.96, 1.00]		
Hanrahan 2013	42	2	22	487	0.66 [0.53, 0.77]	1.00 [0.99, 1.00]		
Lawn 2011	42	2	30	320	0.58 [0.46, 0.70]	0.99 [0.98, 1.00]		



Sensitivity and specificity

Reference standard

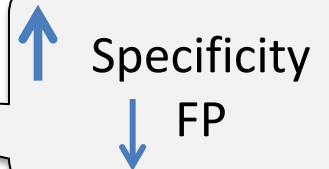
	(+)	(-)
Index (+)	TP	FP
Index (-)	FN	TN

Sensitivity Specificity

➤ Sensitivity: what proportion of those with the disease does the test detect?

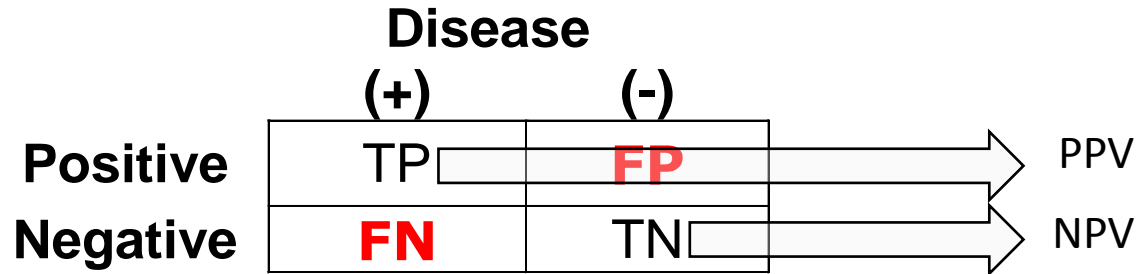


➤ Specificity: what proportion of those without the disease get negative test results?

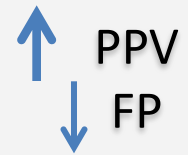


➤ Sensitivity and specificity depend on the patient spectrum recruited to the study

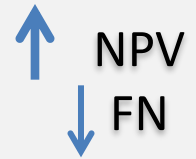
Predictive values



➤ PPV: What proportion of those who test positive with the index test really have disease?



➤ NPV: What proportion of those who test negative with the index test really do not have disease?



➤ PPV and NPV mathematically depend on prevalence.

Likelihood ratios

- Positive and negative likelihood ratios describe by how much the chances of disease increase and decrease with positive and negative test results
- Values can be computed from summary sensitivity and specificity estimates
- They are used in Bayesian updating to compute post-test probabilities of disease
- Most commonly encountered in studies of signs and symptoms

Red flags to screen for vertebral fracture in patients presenting with low-back pain

Christopher M Williams¹, Nicholas Henschke², Christopher G. Maher¹, Maurits W van Tulder³, Bart W Koes⁴, Petra Macaskill⁵, Les Irwig⁶

¹The George Institute for Global Health, University of Sydney, Sydney, Australia. ²Institute of Public Health, University of Heidelberg, Heidelberg, Germany. ³Department of Health Sciences, Faculty of Earth and Life Sciences, VU University, Amsterdam, Netherlands. ⁴Department of General Practice, Erasmus Medical Center, Rotterdam, Netherlands. ⁵Screening and Test Evaluation Program (STEP), School of Public Health, Sydney, Australia. ⁶School of Public Health, University of Sydney, Sydney, Australia

Main results

Eight studies set in primary (four), secondary (one) and tertiary care (accident and emergency = three) were included in the review. Overall, the risk of bias of studies was moderate with high risk of selection and verification bias the predominant flaws. Reporting of index and reference tests was poor. The prevalence of vertebral fracture in accident and emergency settings ranged from 6.5% to 11% and in primary care from 0.7% to 4.5%. There were 29 groups of index tests investigated however, only two featured in more than two studies. Descriptive analyses revealed that three red flags in primary care were potentially useful with meaningful positive likelihood ratios (LR+) but mostly imprecise estimates (significant trauma, older age, corticosteroid use; LR+ point estimate ranging 3.42 to 12.85, 3.69 to 9.39, 3.97 to 48.50 respectively). One red flag in tertiary care appeared informative (contusion/abrasion; LR+ 31.09, 95% CI 18.25 to 52.96). The results of combined tests appeared more informative than individual red flags with LR+ estimates generally greater in magnitude and precision.

Using the RevMan calculator tool

The image shows the RevMan calculator tool interface. At the top, there are two horizontal axes for Sensitivity (95% CI) and Specificity (95% CI), both ranging from 0 to 1. Below these is a 'Calculator' dialog box. The dialog box is divided into two main sections: 'Reference standard' and 'Index test'.

Reference standard:

	+	-	Total
+	TP <input type="text"/>	FP <input type="text"/>	Test+ <input type="text"/>
-	FN <input type="text"/>	TN <input type="text"/>	Test- <input type="text"/>
Total	D+ <input type="text"/>	D- <input type="text"/>	N <input type="text"/>

Index test:

Sensitivity	0.89
Specificity	0.99
PPV	<input type="text"/>
NPV	<input type="text"/>
LR+	89.0000
LR-	0.1111
Prevalence	<input type="text"/>

The 'Calculator' dialog box also includes a 'Reset' button and 'OK' and 'Cancel' buttons. The 'Sensitivity' and 'Specificity' values are highlighted with a yellow oval, and the 'LR+' and 'LR-' values are also highlighted with a yellow oval.

Computing PPV and NPV

Calculator -

		Reference standard		Total
		+	-	
Index test	+	TP 0	FP 0	Test+ 0
	-	FN 0	TN 1	Test- 1
Total	D+ 0	D- 1	N 1	

Sensitivity	Specificity
0.89	0.99
PPV 0.6953	NPV 0.9972
LR+ 89.0000	LR- 0.1111
Prevalence 0.025	

Buttons: ? [Icon] Reset OK Cancel

Impact of prevalence

Calculator -

		Reference standard		Total
		+	-	
Index test	+	TP 0	FP 0	Test+ 0
	-	FN 0	TN 1	Test- 1
Total		D+ 0	D- 1	N 1

Sensitivity	0.89	Specificity	0.99
PPV	0.9674	NPV	0.9643
LR+	89.0000	LR-	0.1111
Prevalence	0.25		

Buttons: ? [Icon] Reset OK Cancel

Computing normalised frequencies

The screenshot shows a 'Calculator' window with a 2x2 contingency table for a diagnostic test. The table is structured as follows:

		Reference standard		Total
		+	-	
Index test	+	TP: 89	FP: 1	Test+: 90
	-	FN: 11	TN: 99	Test-: 110
Total		D+: 100	D-: 100	N: 200

Diagnostic metrics displayed on the right:

- Sensitivity: 0.89
- Specificity: 0.99
- PPV: 0.9889
- NPV: 0.9000
- LR+: 89.0000
- LR-: 0.1111
- Prevalence: 0.5000

Buttons at the bottom: ? (help), [icon], Reset, OK, Cancel.

11 out of every 100 with TB will be missed

1 out of every 100 without TB will be falsely positive

Normalised frequencies using prevalence

The screenshot shows a 'Calculator' window with a 2x2 contingency table for a diagnostic test. The table is structured as follows:

		Reference standard		Total
		+	-	
Index test	+	TP: 22	FP: 10	Test+: 32
	-	FN: 3	TN: 965	Test-: 968
Total		D+: 25	D-: 975	N: 1000

Key metrics displayed on the right side of the calculator:

- Sensitivity: 0.89
- Specificity: 0.99
- PPV: 0.6953
- NPV: 0.9972
- LR+: 89.0000
- LR-: 0.1111
- Prevalence: 0.025

At the bottom of the window, there are buttons for '?', a folder icon, 'Reset', 'OK', and 'Cancel'.

Out of every 1000 tested:

3 with disease will be missed

10 without disease will be falsely positive

Presentation for a single test

Summary of findings

Review question: What is the diagnostic accuracy of Xpert MTB/RIF assay for detection of pulmonary TB?

Patients/population: Adults with presumed pulmonary TB

Role: Xpert MTB/RIF assay used as an initial test replacing microscopy and used as an add-on test following a negative smear microscopy result

Index test: Xpert MTB/RIF assay

Reference standards: Solid or liquid culture

Studies: Cross-sectional

Setting: Mainly intermediate level laboratories

Type of analysis	Effect (95% credible interval)	No. of participants (studies)	Test result	Number of results per 1000 patients tested (95% CrI) ¹		
				Prevalence 2.5%	Prevalence 5%	Prevalence 10%
TB detection, Xpert MTB/RIF used as an initial test replacing smear microscopy	Pooled median sensitivity 89% (85, 92) and pooled median specificity 99% (98, 99)	8998 (22)	True Positives	22 (21, 23)	45 (43, 46)	89 (85, 92)
			False Negatives	3 (2, 4)	6 (4, 8)	11 (8, 15)
			False Positives	10 (10, 20)	10 (10, 19)	9 (9, 18)
			True Negatives	965 (956, 965)	941 (931, 941)	891 (882, 891)

Values for lower 95% limit

Calculator -

		Reference standard		Total
		+	-	
Index test	+	TP 21	FP 20	Test+ 41
	-	FN 4	TN 956	Test- 959
Total		D+ 25	D- 975	N 1000

Sensitivity	0.85	Specificity	0.98
PPV	0.5215	NPV	0.9961
LR+	42.5000	LR-	0.1531
Prevalence	0.025		

? Reset OK Cancel

Values for upper 95% limit

Calculator -

		Reference standard		Total
		+	-	
Index test	+	TP 23	FP 10	Test+ 33
	-	FN 2	TN 965	Test- 967
Total		D+ 25	D- 975	N 1000

Sensitivity	0.92	Specificity	0.99
PPV	0.7023	NPV	0.9979
LR+	92.0000	LR-	0.0808
Prevalence	0.025		

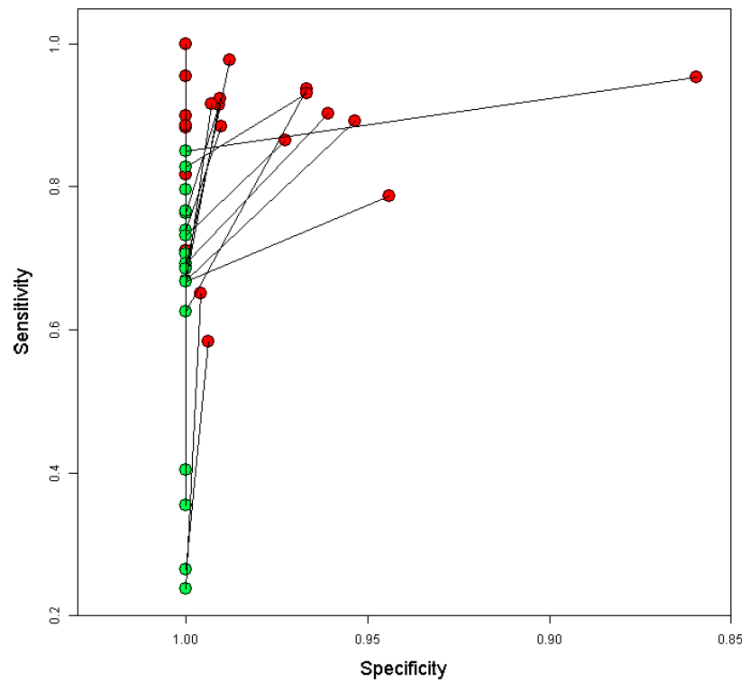
? Reset OK Cancel

Sources of estimates of prevalence

- Observed prevalence in the systematic review
 - Median prevalence
 - Dependent on studies providing representative estimates
 - Exclude case-control studies as prevalence estimates are artefactual
- Other sources
 - Disease registries and routine data sources
 - Audits and epidemiological studies
 - Professional opinion
- Evaluate across a range of plausible values

Comparison of tests

Figure 7. Study results of smear microscopy (green circle) versus Xpert MTB/RIF (red circle) plotted in ROC space. The specificity of smear was assumed to be 100%.



TB detection, Xpert MTB/RIF compared with smear microscopy

Twenty-one studies (8880 participants) provided data from which to compare the sensitivity of Xpert MTB/RIF and smear microscopy. Figure 7 displays results of smear microscopy versus Xpert MTB/RIF for the individual studies. In the meta-analysis, the sensitivity estimate for Xpert MTB/RIF was the same as the estimate in the meta-analysis in I. A., the difference in the number of studies and participants being due to use of the subset of studies that also reported results by smear status. For smear microscopy, the pooled sensitivity was 65% (95% CrI 57% to 72%). For Xpert MTB/RIF, the pooled sensitivity was 88% (95% CrI 84% to 92%). Therefore, in comparison with smear microscopy, Xpert MTB/RIF increased TB detection among culture-confirmed cases by 23% (95% CrI 15% to 32%).

Presentation for test comparison

Test result	Number of TB positives per 1000 culture-positive individuals tested (95% CrI)							
	Prevalence 2.5%		Prevalence 5%		Prevalence 10%		Prevalence 30%	
	Smear Microscopy	Xpert MTB/RIF	Smear Microscopy	Xpert MTB/RIF	Smear Microscopy	Xpert MTB/RIF	Smear Microscopy	Xpert MTB/RIF
True positives	16 (14, 18)	22 (21, 23)	33 (29, 36)	44 (42, 46)	65 (57, 72)	88 (84, 92)	195 (171, 216)	264 (252, 276)
Mean absolute difference in true positives	6 more		11 more		23 more		69 more	
False negatives	9 (7, 11)	3 (2, 4)	18 (14, 22)	6 (4, 8)	35 (28, 43)	12 (8, 16)	105 (84, 129)	36 (24, 48)
Mean absolute difference in false negatives	6 less		12 less		23 less		69 less	

Comparison with smear microscopy

In comparison with smear microscopy, Xpert® MTB/RIF increased TB detection among culture-confirmed cases by 23% (95% CrI 15% to 32%; 21 studies, 8880 participants).

For TB detection, if pooled sensitivity estimates for Xpert® MTB/RIF and smear microscopy are applied to a hypothetical cohort of 1000 patients where 10% of those with symptoms have TB, Xpert® MTB/RIF will diagnose 88 cases and miss 12 cases, whereas sputum microscopy will diagnose 65 cases and miss 35 cases.

Using SROC curves

- Reviews may estimate an average SROC where included studies do not have a common threshold
- Summary summary statistics available include:
 - Diagnostic odds ratios
 - Areas under the SROC curvewhich are not amenable to easy explanation
- One approach is to quote sensitivity for a fixed false positive rate (specificity)

Second trimester serum tests for Down's Syndrome screening

S Kate Alldred¹, Jonathan J Deeks², Boliang Guo³, James P Neilson¹, Zarko Alfirevic¹

Figure 4. Studies evaluating combination of maternal age, Total hCG, AFP and uE3 showing summary ROC curve

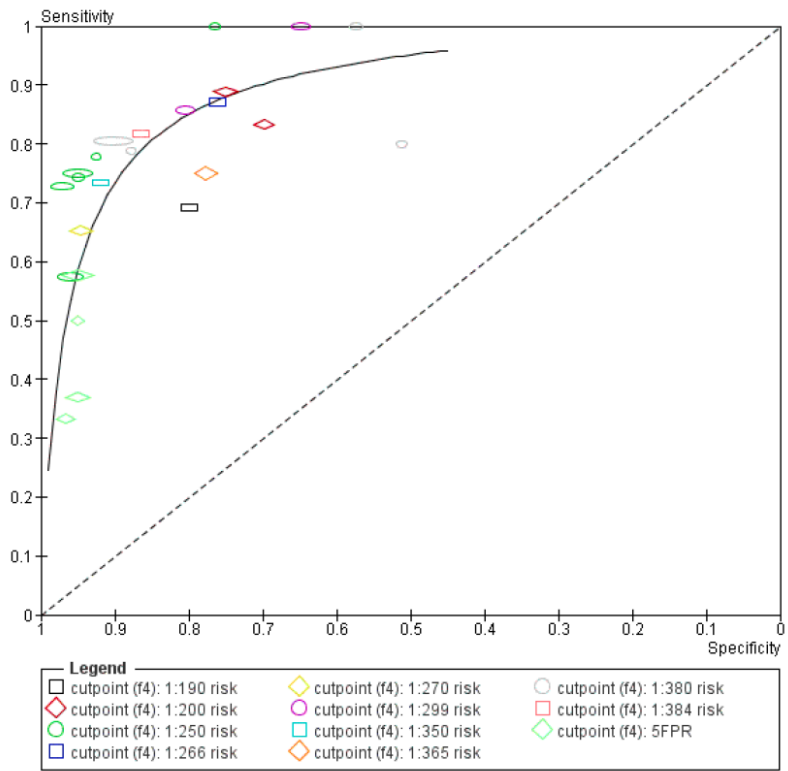
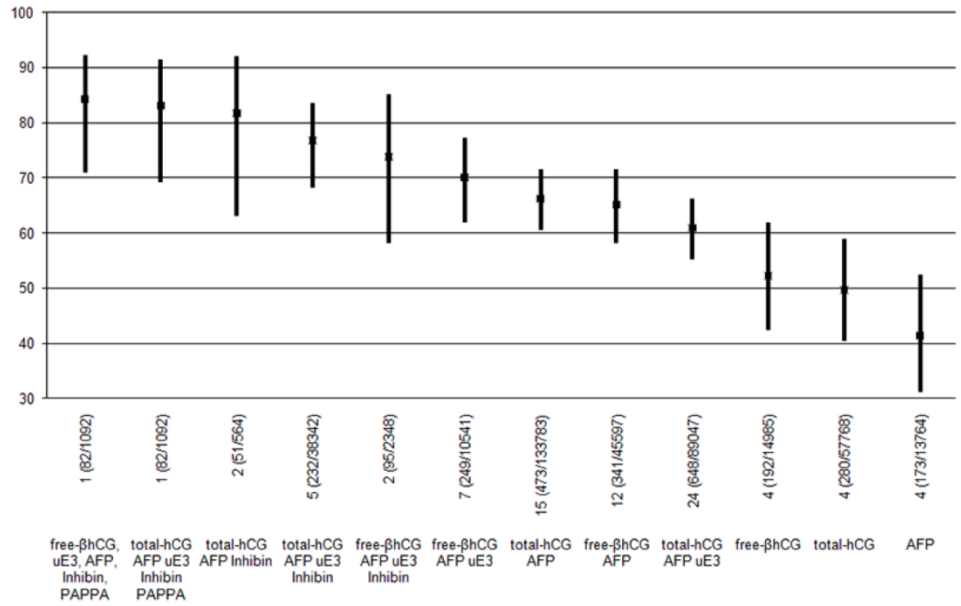


Figure 11. Detection rates (% sensitivity) at a false positive rate of 5% for the 12 selected test combinations (estimates from summary ROC curves)



free-βhCG, uE3, AFP, Inhibin, PAPPA total-hCG AFP uE3 Inhibin PAPPA total-hCG AFP Inhibin total-hCG AFP uE3 Inhibin free-βhCG AFP uE3 Inhibin free-βhCG AFP uE3 total-hCG AFP free-βhCG AFP total-hCG AFP uE3 free-βhCG total-hCG AFP

Prevalence of 25 Down's per 10,000

		Reference standard		Total
		+	-	
Index test	+	TP 15	FP 499	Test+ 514
	-	FN 10	TN 9476	Test- 9486
Total	D+ 25	D- 9975	N 10000	

Sensitivity	0.61
Specificity	0.95
PPV	0.0297
NPV	0.9990
LR+	12.2000
LR-	0.4105
Prevalence	0.0025

Out of every 10,000 tested:

15 pregnancies with Down's will be detected

10 pregnancies with Down's will be missed

499 pregnancies will have unnecessary invasive testing


Summary of findings table (SDTMG)

- Mandatory table for Cochrane reviews
- No prescribed format yet, but
 - Review question must be stated in full as an integral part of table
 - One table for each main question addressed any particular review
 - Each row represents one index test, version of the tests, or use of a test in a particular sub-population/setting
- GRADE Working Group in process of developing SoF template but...

Summary of findings table (SDTMG)

- **Outcomes** of a DTA review:
 - sensitivity and specificity of a test at a specific threshold;
 - the summary ROC curve and its parameters;
 - comparative outcomes (accuracy of one test relative to another)
- Interpretation of accuracy outcomes in relation to patient outcomes CONSEQUENCES..... (**remember role of the test**)
- **Numbers** (included studies; number with and without target condition)
- Risk of Bias and Concerns regarding Applicability (**QUADAS-2**)
- **Heterogeneity and precision**
- Inconsistency if comparative questions were considered
- **Normalised frequencies** (eg: hypothetical population of 1000 persons)

How does this relate to **GRADE**?

- **G**radings of **R**ecommendations **A**ssessment, **D**evelopment and **E**valuation (GRADE) Working Group
- <http://www.gradeworkinggroup.org/> The logo for GRADE, consisting of the word "GRADE" in bold, red, uppercase letters inside a red rectangular border with a slightly distressed or stamped appearance.
- The **GRADE** working group has developed an approach to grading quality of evidence and strength of recommendations. They have ‘invented’ the summary of findings tables.

Concepts of quality

Cochrane DTA reviews

“Quality” is defined as a combination of applicability of the results and low risk of bias.

Overlap with some domains (e.g. risk of bias, imprecision, indirectness).

No formal downgrading or overall ‘credibility’ judgment

GRADE

“Quality” is defined as a combination of:

- Low risk of bias
- Low degree of indirectness
- Low degree of inconsistency
- Low degree of imprecision
- No publication bias

Formal downgrading from high quality to very low quality

Implementation of items may be difficult in DTA reviews

Summary of findings table – Cochrane review example

Rapid diagnostic tests for diagnosing uncomplicated *P. falciparum* malaria in endemic countries (Review)

Abba K, Deeks JJ, Olliaro PL, Naing CM, Jackson SM, Takwoingi Y, Donegan S, Garner P

What is the diagnostic accuracy of Rapid Diagnostic Tests for detecting malaria? What are the best types of tests?	
Patients/populations	People presenting with symptoms suggestive of uncomplicated malaria
Prior testing	None
Settings	Ambulatory healthcare settings in <i>P. falciparum</i> malaria endemic areas in Asia, Africa and South America
Index tests	Immunochromatography-based rapid diagnostic tests for <i>P. falciparum</i> malaria
Reference standard	Conventional microscopy or PCR
Importance	Accurate and fast diagnosis allows appropriate and quick treatment for malaria to be provided
Studies	Consecutive series of patients; 74 studies presented 111 test evaluations based on 60,396 patient test results
Quality concerns	Poor reporting of patient characteristics, sampling method and reference standard methods were common concerns

Summary of findings table – Cochrane review example

Rapid diagnostic tests for diagnosing uncomplicated *P. falciparum* malaria in endemic countries (Review)

Abba K, Deeks JJ, Olliaro PL, Naing CM, Jackson SM, Takwoingi Y, Donegan S, Garner P

Test types	Quantity of evidence	Brands (studies)	Average pooled results	Consequences in a cohort of 1000		
				<i>P. falciparum</i> prevalence	Missed cases	Overtreated non-cases
HRP-2 antibody-based tests compared with microscopy						
Type 1 HRP-2 (<i>P. falciparum</i> specific)	71 evaluations 40,062 participants 11,966 malaria cases	Paracheck-Pf (27), ParaSight (17), ICT Malaria Pf (16), ParaHIT-F (4), PATH (2), Determine Malaria Pf (1), Rapid Test Malaria (1), Diaspot Malaria (1), New mini-Pf (1), and Hexagon Malaria (1)	sens = 94.8% (93.1% to 96.1%)	30%	16	34

Take home message (1)

- Results of reviews will be more accessible if they are presented as normalised frequencies for realistic scenarios
- Normalised frequencies can be computed from summary estimates of test accuracy, as can predictive values and likelihood ratios
- The RevMan calculator can be used for all computations
- Interpretation requires consideration of the prevalence
- Helpful to present test comparisons as absolute difference in the numbers of false negatives and false positives
- Results should be clearly summarised and presented

Take home message (2)

- **‘Summary of findings’** tables present the main findings of a review in a transparent and simple tabular format.
- They contain key information about:
 - review question
 - accuracy estimates
 - sum of available data
 - quality of evidence
 - practical implications

References

- DTA Handbook Chapter 11 contains relevant information for this chapter. This is available at dta.cochrane.org
- See <http://dta.cochrane.org/dta-author-training-online-learning> for additional training materials

ACKNOWLEDGEMENTS

Materials for this presentation are based in part on material adapted from members of the Cochrane Screening and Diagnostic Test Methods Group