Methodological Challenges in Complex Reviews

Cochrane UK & Ireland Symposium 2016

Outline

- Current challenges to complex reviews
- Direct and indirect evidence
- Diagnostic test accuracy data
- Supporting complex reviews

Why are reviews increasingly complex?

- Increasingly complex clinical and policy questions
- More interests in complex interventions
- Existing evidence often limited and heterogeneous
- Multiple treatment/intervention options with no head-to-head evidence
- Outcomes of interest have complex data structure

Aims

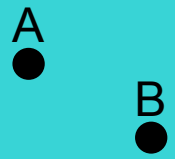
1. Raise awareness of the challenges of conducting complex reviews with multiple comparators and diagnostic test accuracy (DTA) data
2. Offer potential (simple to more complex) solutions to some but not all of the challenges
3. Provoke discussion regarding how to ensure complex reviews answer clinically-relevant questions

Direct and Indirect Evidence

Olivia Wu
Neil Hawkins

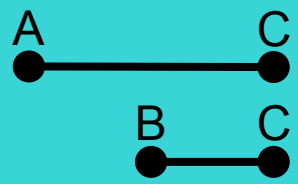
A Taxonomy of Comparisons

Direct Comparison (head to head)



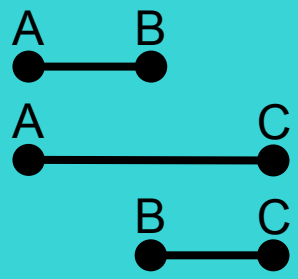
‘Naïve’ or ‘Unadjusted’ Indirect Comparison

Absolute effect estimates from individual trial arms



‘Adjusted’ Indirect Comparison

Relative effect estimates between treatments



Mixed Treatment Comparison or ’Network’ Meta-Analysis

‘Adjusted’ indirect comparison extended to more complex networks of trial evidence

(i.e. head to head and indirect evidence)

An Example

Early thrombolysis for the treatment of acute myocardial infarction: a systematic review and economic evaluation

A Boland
Y Dundar
A Bagust
A Haycox
R Hill
R Mujica Mota
T Walley
R Dickson*

Liverpool Reviews and Implementation Group, New Medical School, Liverpool, UK

Executive summary

Health Technology Assessment 2003; Vol. 7: No. 15

Health Technology Assessment
NHS R&D HTA Programme

Multiple treatments and trial comparisons

TABLE 2 *Summary of included clinical studies*

Alteplase/streptokinase	Alteplase/ tenecteplase	Alteplase/ reteplase	Streptokinase/ reteplase	Dose-ranging and mixed regimes
GUSTO I[18] Central Illinois[43] Cherng *et al.*[44] ECSG[45] GISSI-2/ISG[46,47] ISIS-3[48] KAMIT[49] PAIMS[50] TIMI 1[51] White *et al.*[59]	ASSENT-2[20] *	GUSTO III[19] * RAPID-2[17] *	INJECT[52]	COBALT[57] (t-PA)* Xu *et al.*[58] (SK) Six *et al.*[53] (SK) ASSENT-1[54] (TNK) TIMI 10B[55] (TNK)* RAPID-1[56] (r-PA)

* *Involved accelerated alteplase*

One of the comparisons summarised in a pairwise meta-analysis

TABLE 8 *Alteplase, excluding accelerated alteplase, versus streptokinase (GUSTO I[18] omitted)*

Outcome	Study	Alteplase	Streptokinase	OR random effect (95% CI)
Mortality up to 35 days	Central Illinois[43]	6/123	9/130	0.69 (0.24 to 2.00)
	Cherng *et al.*[44]	2/59	5/63	0.41 (0.08 to 2.18)
	ECSG[45]	3/64	3/65	1.02 (0.20 to 5.23)
	GISSI-2/ISG[46,47]	929/10,372	887/10,396	1.05 (0.96 to 1.16)
	GUSTO I[18]	1,418/13,746	1,455/13,780	0.97 (0.90 to 1.05)
	ISIS-3[48]	4/86	7/85	0.54 (0.15 to 1.93)
	PAIMS[50]	7/143	12/147	0.58 (0.22 to 1.52)
	TIMI 1[51]	5/135	10/135	0.48 (0.16 to 1.45)
	Total	**2,374/24,728**	**2,388/224,801**	**1.00 (0.94 to 1.06)**
				Test for heterogeneity $\chi^2 = 6.87$, df = 7, $p = 0.44$

Full data set summarised in four separate meta-analyses

1. Alteplase vs Streptokinase

Outcome	Study	Alteplase	Streptokinase	OR random effect (95% CI)
Mortality up to 35 days	Central Illinois[43]	6/123	9/130	0.69 (0.24 to 2.00)
	Cherng *et al.*[44]	2/59	5/63	0.41 (0.08 to 2.18)
	ECSG[45]	3/64	3/65	1.02 (0.20 to 5.23)
	GISSI-2/ISG[46,47]	929/10,372	887/10,396	1.05 (0.96 to 1.16)
	GUSTO I[18]	652/10,344	1,472/20,173	0.85 (0.78 to 0.94)
	ISIS-3[48]	1,418/13,746	1,455/13,780	0.97 (0.90 to 1.05)
	PAIMS[50]	4/86	7/85	0.54 (0.15 to 1.93)
	TIMI 1[51]	7/143	12/147	0.58 (0.22 to 1.52)
	White *et al.*[59]	5/135	10/135	0.48 (0.16 to 1.45)
	Total	**3,026/35,072**	**3,860/44,974**	**0.94 (0.85 to 1.04)**
				Test for heterogeneity $\chi^2 = 13.96$, df = 8, $p = 0.083$

2. Acc Alteplase vs Tenecteplase

Outcome	Study	Accelerated alteplase	Tenecteplase	OR random effect (95% CI)
Mortality up to 35 days	ASSENT-2[20]	522/8,488	523/8,461	0.99 (0.88 to 1.13)

3. Acc Alteplase vs Reteplace

Outcome	Study	Accelerated alteplase	Reteplase	OR random effect (95% CI)
Mortality up to 35 days	GUSTO III[19]	356/4,921	757/10,138	0.97 (0.85 to 1.10)
	RAPID-2[17]	13/155	7/169	2.12 (0.82 to 5.46)
	Total	**369/5,076**	**764/10,307**	**1.24 (0.61 to 2.53)**
				Test for heterogeneity $\chi^2 = 2.60$, df = 1, $p = 0.11$

4. Reteplace vs Streptokinase

Outcome	Study	Reteplase	Streptokinase	OR random effect (95% CI)
Mortality up to 35 days	INJECT[52]	270/2,994	285/2,992	0.94 (0.79 to 1.12)

This is difficult to summarise…

"Definitive conclusions on efficacy are that streptokinase is as effective as non-accelerated alteplase, that tenecteplase is as effective as accelerated alteplase, and that reteplase is at least as effective as streptokinase.

Some conclusions require interpretation of data, i.e. whether streptokinase is as effective as, or inferior to accelerated alteplase; and whether reteplase is as effective as accelerated alteplase or not.

Depending on these, two further conclusions on indirect comparisons arise, whether tenecteplase is superior to streptokinase or not, and whether reteplase is as effective as tenecteplase or not."

From Boland A, Dundar Y, Bagust A, Haycox A, Hill R, Mujica Mota R, *et al*. Early thrombolysis for the treatment of acute myocardial infarction: a systematic review and economic evaluation. *Health Technol Assess* 2003;**7**(15).

Acknowledgements to Julian Higgins

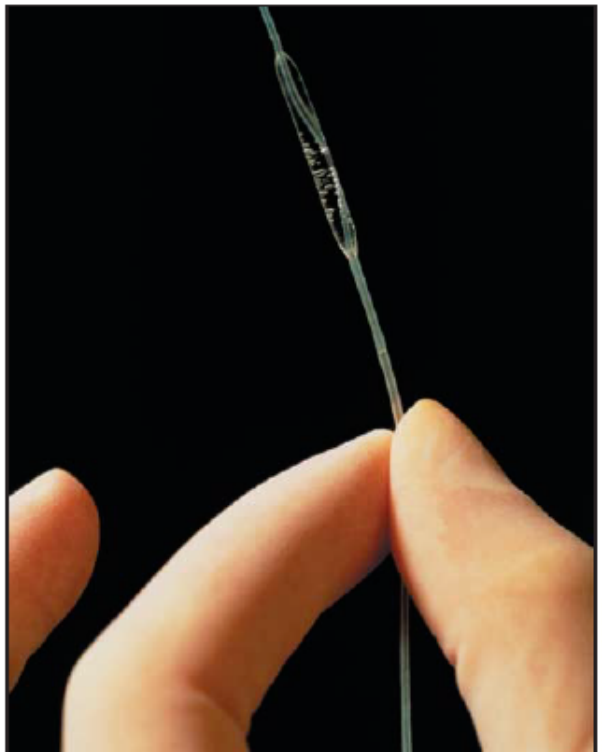
An alternative approach – network meta-analysis

Simultaneous comparison of multiple treatments: combining direct and indirect evidence

Deborah M Caldwell, A E Ades, J P T Higgins

How can policy makers decide which of five treatments is the best? Standard meta-analysis provides little help but evidence based decisions are possible

Several possible treatments are often available to treat patients with the same condition. Decisions about optimal care, and the clinical practice guidelines that inform these decisions, rely on evidence based evaluation of the different treatment options.[1 2] Systematic reviews and meta-analyses of randomised controlled trials are the main sources of evidence. However, most systematic reviews focus on pair-wise, direct comparisons of treatments (often with the comparator being a placebo or control group), which can make it difficult to determine the best treatment. In the absence of a collection of large, high quality, randomised trials comparing all eligible treatments (which is invariably the situation), we have to rely on indirect comparisons of multiple treatments. For example, an indirect estimate of the benefit of A over B can be obtained by comparing trials of A *v* C with trials of B *v* C,[3–5] even though indirect comparisons produce relatively imprecise estimates.[6] We describe comparisons of three or more treatments, based on pair-wise or multi-arm comparative studies, as a multiple treatment comparison evidence structure.



Medical Research Council Health Services Research Collaboration, Department of Social Medicine, University of Bristol, Bristol BS8 2PR

Deborah M Caldwell
research associate

A E Ades
professor of public health science

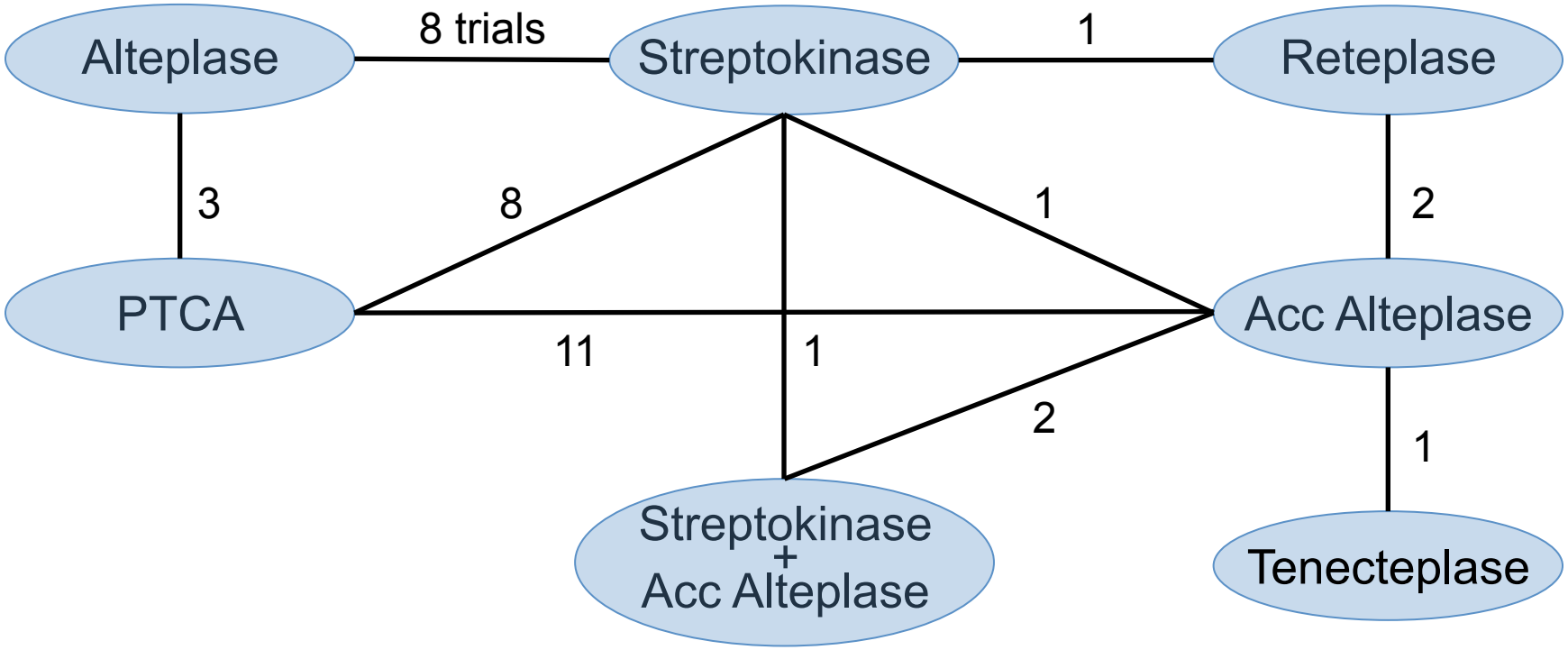
Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR

J P T Higgins
senior research associate

Correspondence to:
D M Caldwell
d.m.caldwell@bristol.ac.uk

BMJ 2005;331:897–900

The network of trial evidence is analysed as a ‘whole’



Network meta-analysis provides comparable estimates of effectiveness for all treatments

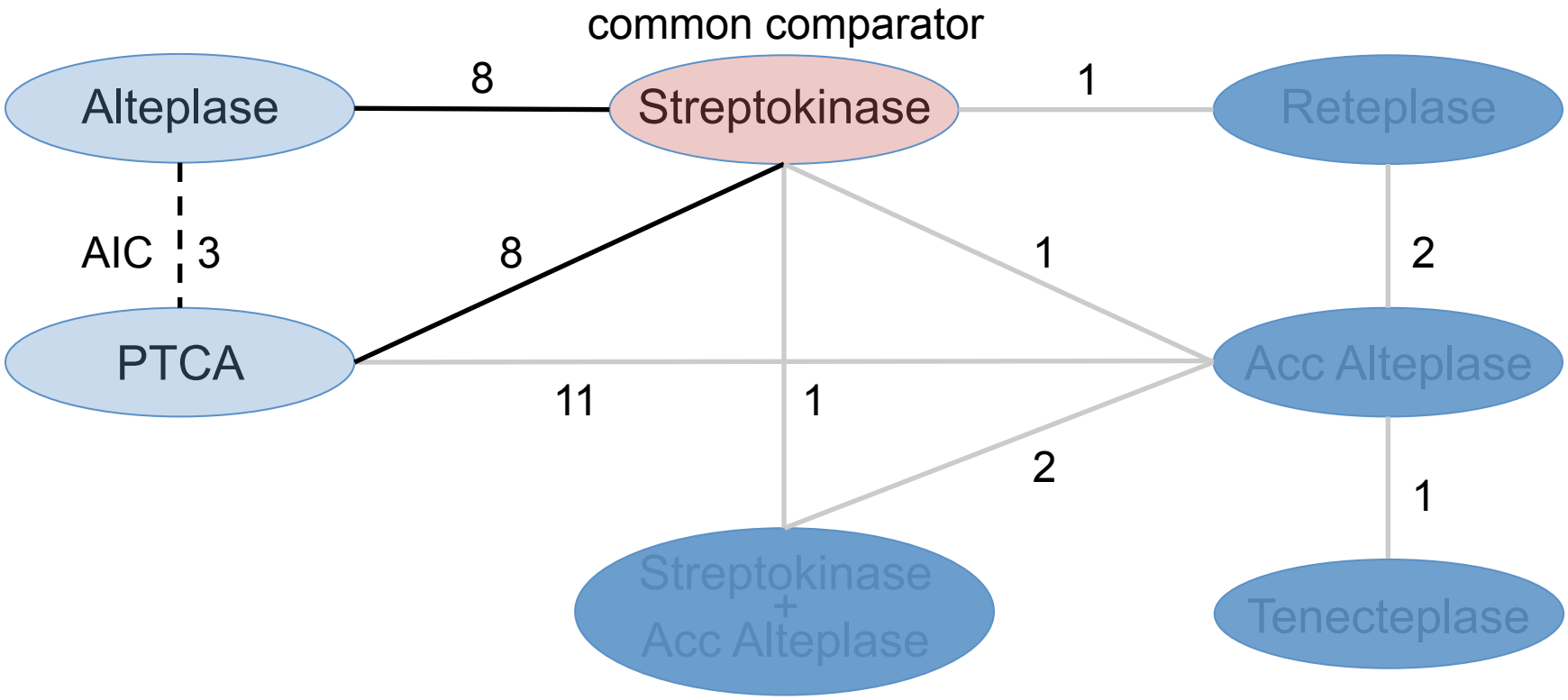
30-day Mortality Treatment	Odds Ratio (Mean (95% CrI))
Streptokinase	1.04 (0.91 to 1.35)
Alteplase	1 (Reference Treatment)
Acc. Alteplase	0.88 (0.70 to 1.19)
Streptokinase+Alteplase	1.02 (0.78 to 1.51)
Reteplase	0.92 (0.70 to 1.24)
Tenecteplase	0.90 (0.61 to 1.35)
PCTA	0.65 (0.49 to 0.86)

(Bayesian) network meta-analysis can provide useful summaries of uncertainty

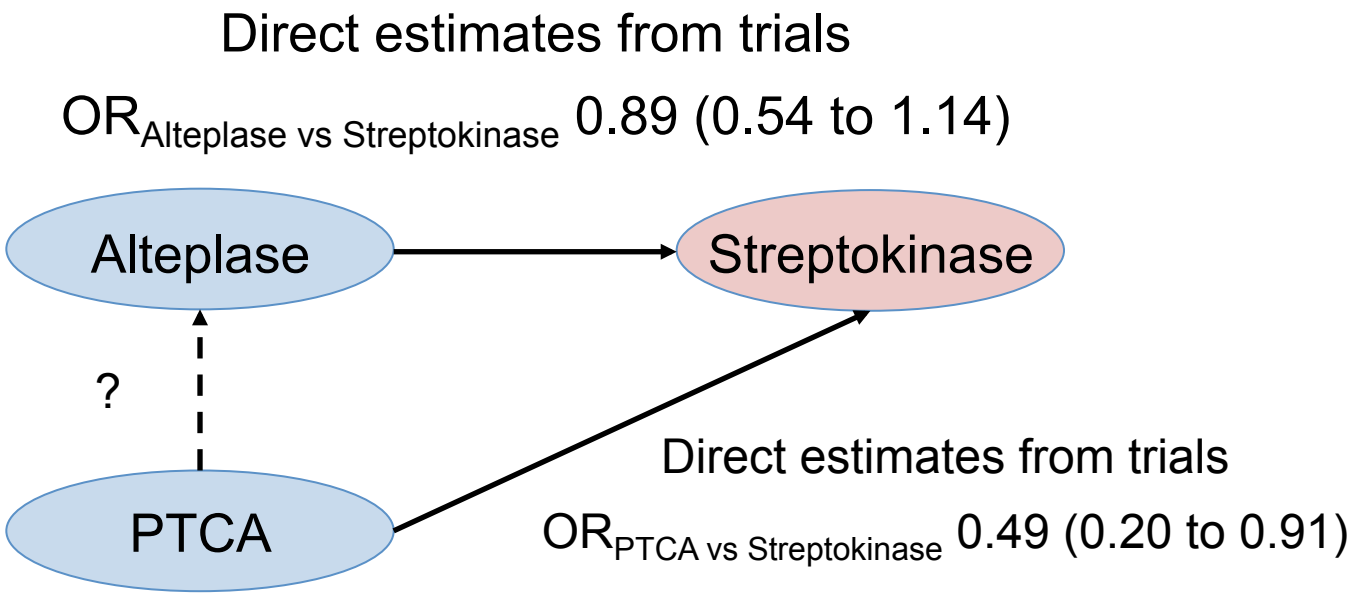
Table 3 Percentage mortality at 35 days and the probability that each treatment is best (lowest mortality) in multiple treatment comparison analysis*

	Fixed effect model		Random effects model	
	35 day Mortality %	Probability best	35 day Mortality %	Probability best
Streptokinase	6.7	0	6.8	0
Alteplase	6.7	0	6.5	0.003
Accelerated alteplase	5.8	0	5.8	0.001
Streptokinase + alteplase	6.5	0	6.6	0.002
Reteplase	6.1	0	6.0	0.01
Tenecteplase	5.8	0.004	5.8	0.03
Percutaneous transluminal coronary angioplasty	4.4	0.995	4.3	0.95

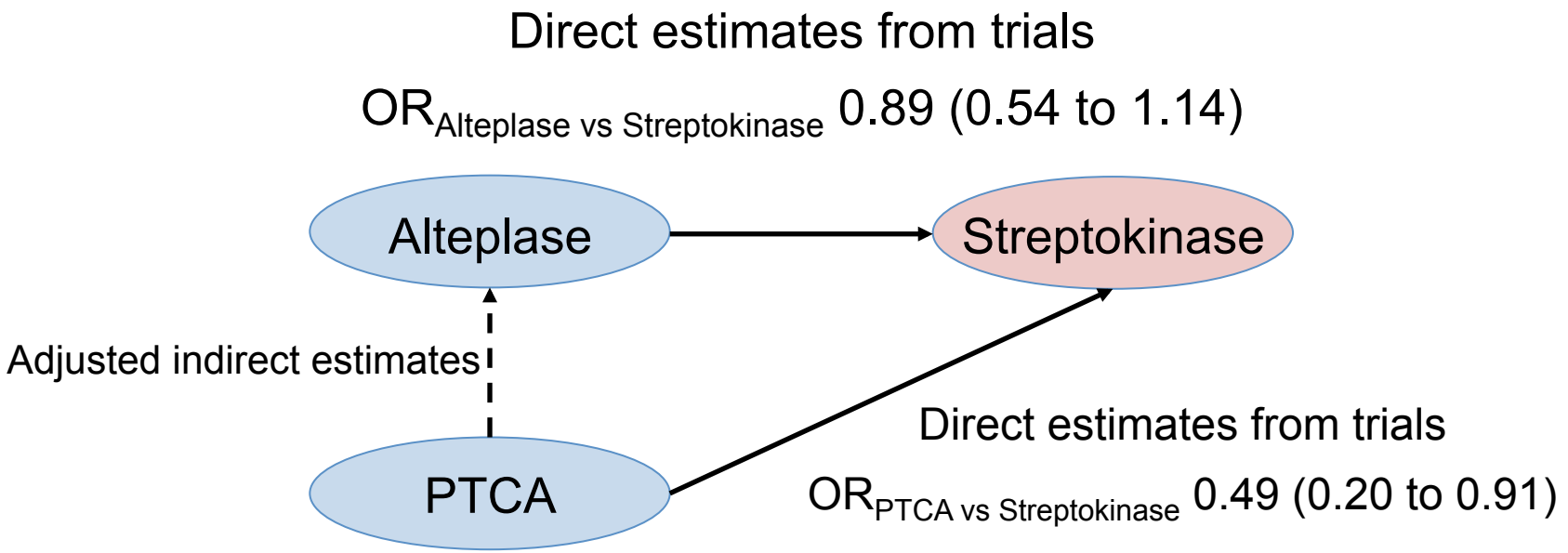
The basic building block – adjusted indirect comparison (AIC)



Indirect Comparison: PTCA vs Alteplase



Indirect Comparison: PTCA vs Alteplase



Adjusted indirect estimates

$OR_{PTCA\ vs\ Alteplace} = OR_{PTCA\ vs\ Streptokinase} / OR_{Alteplase\ vs\ Streptokinase}$

$= 0.49 / 0.89$

$= 0.55$

Generic Assumption

$$OR_{AB} = {OR_{AC}}/{OR_{BC}}$$

$$\log(OR_{AB}) = \log(OR_{AC}) - \log(OR_{BC})$$

$$\partial_{AB} = \partial_{AC} - \partial_{BC}$$

The generic assumption of transitivity

Basic Assumption

- Similarity
  - Trials are clinically and methodologically similar and comparable
- Exchangeability
  - If patients in one trial were substituted in another, the observed treatment estimates would be expected to be the same (allowing for random variation)
- Transitivity
  - $\partial_{AB} = \partial_{AC} - \partial_{BC} \qquad \partial_{AC} = \partial_{AB} - \partial_{CB}$
- Consistency
  - Indirect and direct estimates are consistent

Network Meta-Analysis

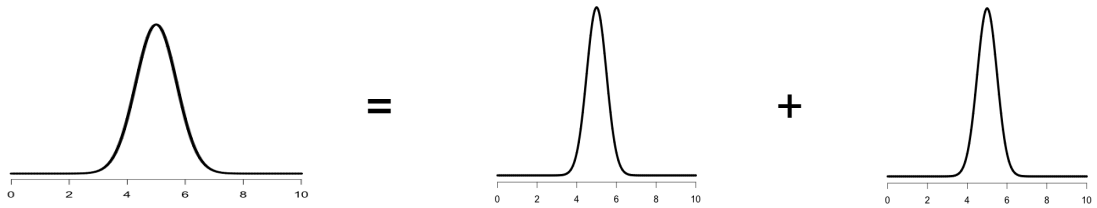
- Extension of the basic indirect comparison to more complex networks
- Estimates treatment effects that best 'fit' the network of trial comparisons

1. $\beta_{Alteplase}$, $\beta_{Reteplease,}$ β_{PTCA} are estimates of the Log Odds Ratio (LOR) of Alteplase, Reteplase and PTCA compared to a reference comparator (e.g. Streptokinase).
2. $LOR_{Alteplase\ vs\ Streptokinase} = \beta_{Alteplase}$
3. $LOR_{Reteplase\ vs\ Streptokinase} = \beta_{Reteplase}$
4. $LOR_{PTCA\ vs\ Streptokinase} = \beta_{PTCA}$
5. $LOR_{Alteplase\ vs\ PTCA} = \beta_{Alteplase} - \beta_{PTCA}$ (consistency assumption)

Estimating Uncertainty

$$OR_{AB} = OR_{AC} \Big/ OR_{BC}$$

$$\log OR_{AB} = \log OR_{AC} - \log OR_{BC}$$



$$\operatorname{var}\left(\log OR_{AB}\right) = \operatorname{var}\left(\log OR_{AC}\right) + \operatorname{var}\left(\log OR_{BC}\right)$$

Estimated uncertainty in indirect estimates

- 95% confidence (credible) intervals are estimated by adding the variance for the contributing indirect comparisons
- Only represents uncertainty arising from the sampling error in the contributing trials
- Does not represent uncertainty in the fundamental assumptions
- Absolute ‘Best Case’ estimate of uncertainty

"Indirect comparisons are not randomized comparisons, and cannot be interpreted as such. They are essentially observational findings across trials, and may suffer the biases of observational studies."

THE COCHRANE COLLABORATION®

Cochrane Handbook for Systematic Reviews of Interventions

Version 5.1.10 (updated March 2011)

Cochrane Reviews using network meta-analysis

FREE

Oral antiviral therapy for prevention of genital herpes outbreaks in immunocompetent and nonpregnant patients

Genital herpes is caused by herpes simplex virus 1 (HSV-1) or 2 (HSV-2). Some infected people experience outbreaks of genital herpes, typically characterised by vesicular and erosive localised painful genital lesions. This review compares the effectiveness and safety of three oral antiviral drugs (aciclovir, famciclovir, and valaciclovir) prescribed to suppress genital herpes outbreaks in non-pregnant patients.

FREE

Methods to decrease blood loss and transfusion requirements for liver transplantation

Excessive blood loss and increased blood transfusion requirements may have significant impact on the short-term and long-term outcomes after liver transplantation. This review compares the potential benefits and harms of different methods of decreasing blood loss and blood transfusion requirements during liver transplantation.

FREE

Immunomodulators and immunosuppressants for multiple sclerosis: a network meta-analysis

There are several therapeutic strategies for the treatment of multiple sclerosis, including immunosuppressants, immunomodulators, and monoclonal antibodies. Their relative effectiveness in the prevention of relapse or disability progression is unclear due to the limited number of direct comparison trials. A summary of the results, including both direct and indirect comparisons of treatment effects, may help to clarify the above uncertainty. This review estimates the relative efficacy and acceptability of interferon ß-1b, interferon ß-1a, glatiramer acetate, natalizumab, mitoxantrone, methotrexate, cyclophosphamide, azathioprine, intravenous immunoglobulins, and long-term corticosteroids versus placebo or another active agent in participants with multiple sclerosis. It ranks the treatments according to their effectiveness and risk-benefit balance.

FREE

Methods to decrease blood loss during liver resection: a network meta-analysis

Liver resection is a major surgery with significant mortality and morbidity. Various methods have been attempted to decrease blood loss and morbidity during elective liver resection. These methods include different methods of vascular occlusion, parenchymal transection, and management of the cut surface of the liver. A surgeon typically uses only one of the methods from each of these three categories, but the optimal treatment strategy for liver resection is unknown. This review compares the benefits and harms of different treatment strategies to decrease blood loss during elective liver resection.

Discussion points

- Can we add value to existing reviews using network meta-analysis?
- Can this be readily incorporated in your current reviews?
- Is network meta-analysis within the remit of Cochrane reviews?

Methodological Challenges of Diagnostic Test Accuracy Reviews

Alex Sutton
Nicola Cooper
Rhiannon Owen
Keith Abrams

Biostatistics Research Group, University of Leicester
NIHR Complex Reviews Support Unit (CRSU) Diagnostic Lead Centre www.nihrcrsu.org

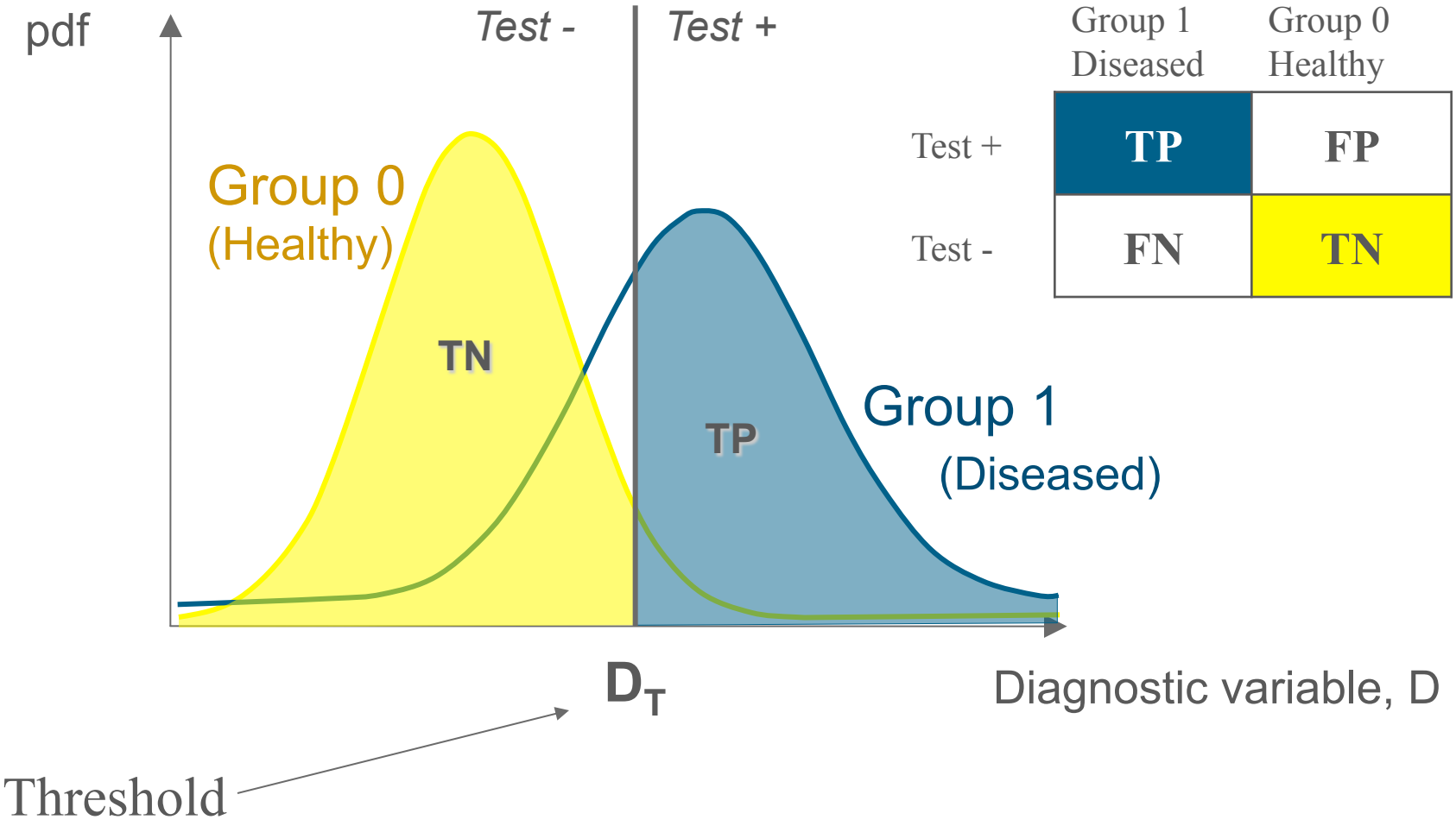
Outline

- Background
- Challenges
- Possible solutions
- Moving forward

Background: Evaluation of a diagnostic test

- Consider a population to be made up of 2 groups:
  - Those with a disease
  - Those without the disease
- A test aims to identify people as belonging to one of these two groups
- Often a 'Gold Standard' test can perfectly distinguish groups, but cannot be used in routine practice (eg pathology)
- Other imperfect tests (often quicker and cheaper) are available, yielding continuous diagnostic markers
  - Scale may be explicit (e.g. chemical level)
  - Or implicit (e.g. interpretation of an image)

Sensitivity vs. Specificity

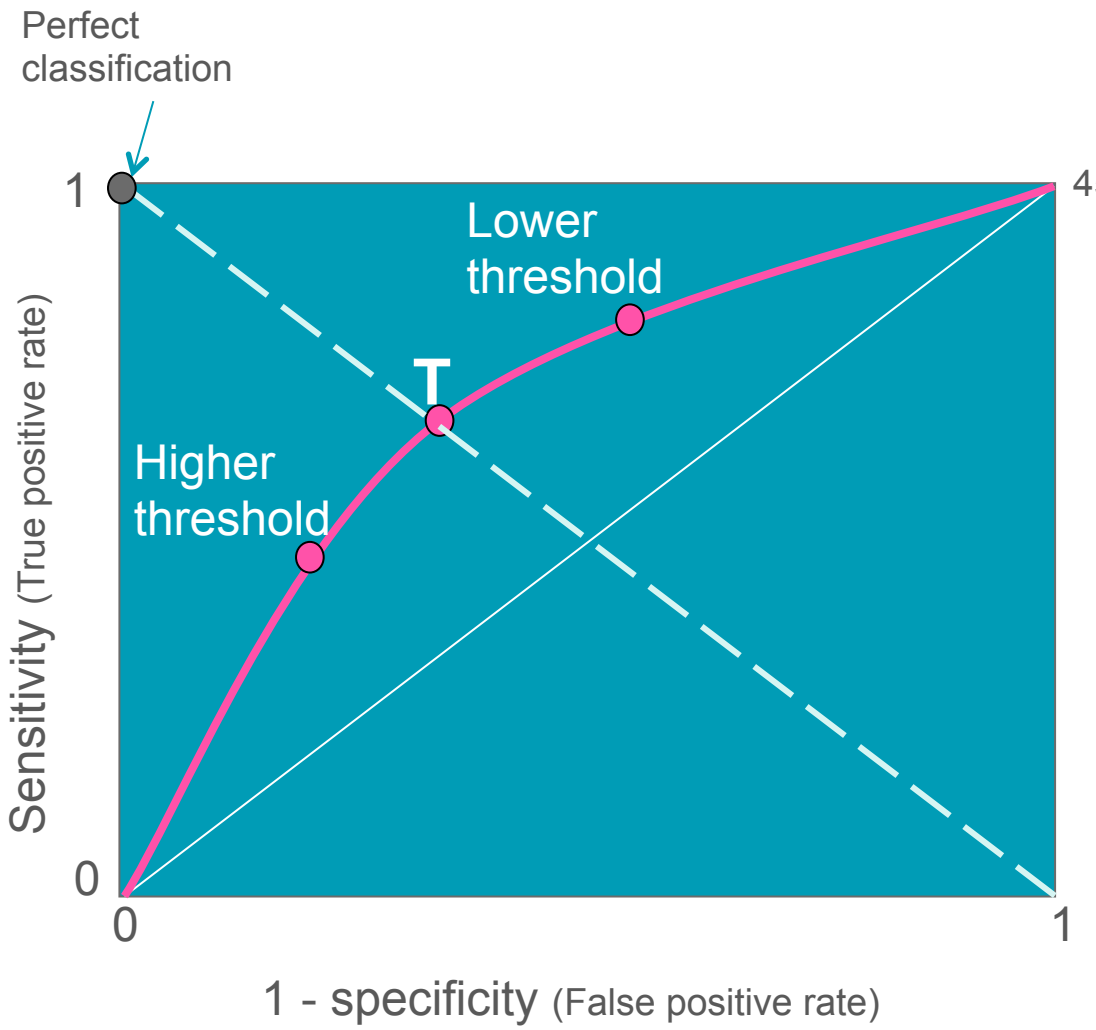


	Group 1 Diseased	Group 0 Healthy
Test +	TP	FP
Test -	FN	TN

Sensitivity = number of true positives/total with disease

Specificity = number of true negatives/total without disease

Receiver Operating Characteristic (ROC) Curve: Selecting the Threshold



Point T gives Max. accuracy threshold BUT
ignores relative opportunity costs of FP and FN results

	Group 1 Diseased	Group 0 Healthy
Test +	TP	FP
Test -	FN	TN

Aim 1

Raise awareness of the challenges of conducting diagnostic test accuracy (DTA) reviews and offer potential (simple to more complex) solutions to some but not all of the challenges

Challenges of meta-analysing diagnostic test accuracy data

More complex than for effectiveness data due to:

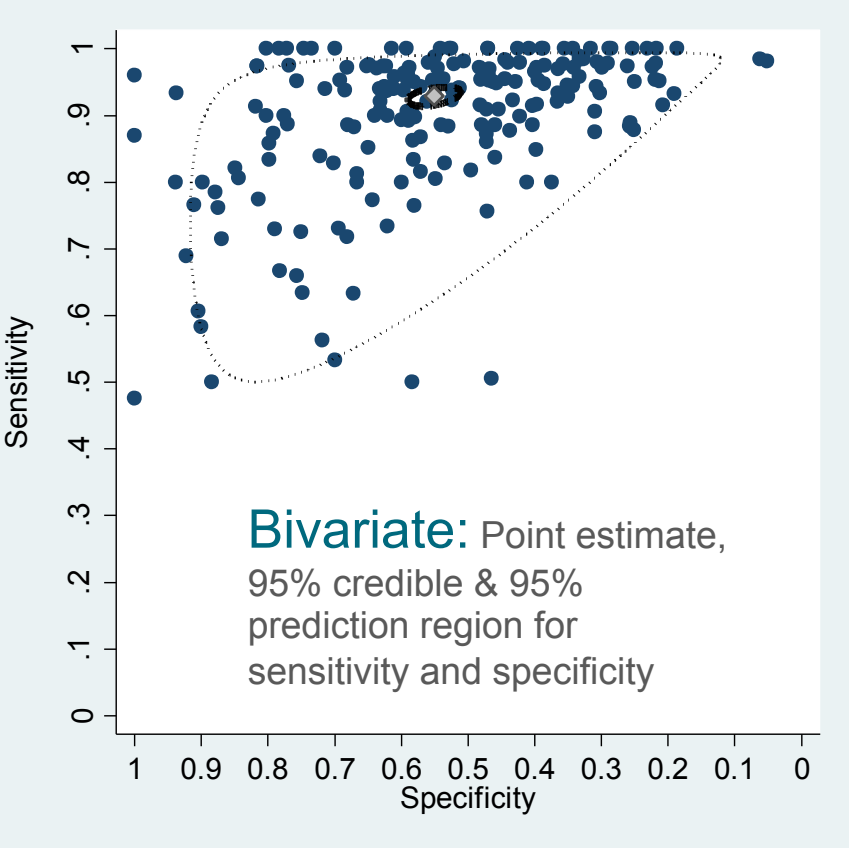
- Two dependent outcomes – **sensitivity** and **specificity**
- Variable test threshold levels (either explicit or implicit)
- Different reference tests (imperfect gold standard)

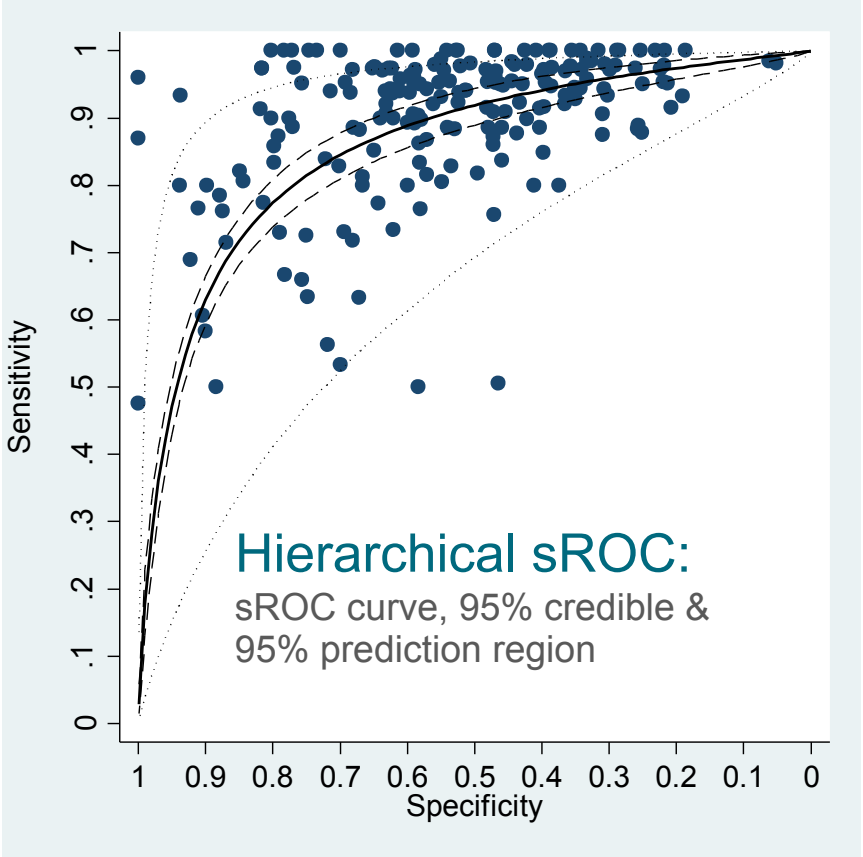
Other issues include:

- Different populations / study conduct (leading to between-study heterogeneity)
- Data quality / risk of bias

Two dependent outcomes
- Sensitivity and Specificity

- Requires a meta-analysis model that models sensitivity, specificity and their correlation simultaneously





- Statistical models are equivalent although presentation of results are different

Challenges of meta-analysing diagnostic test accuracy data

More complex than for effectiveness data due to:

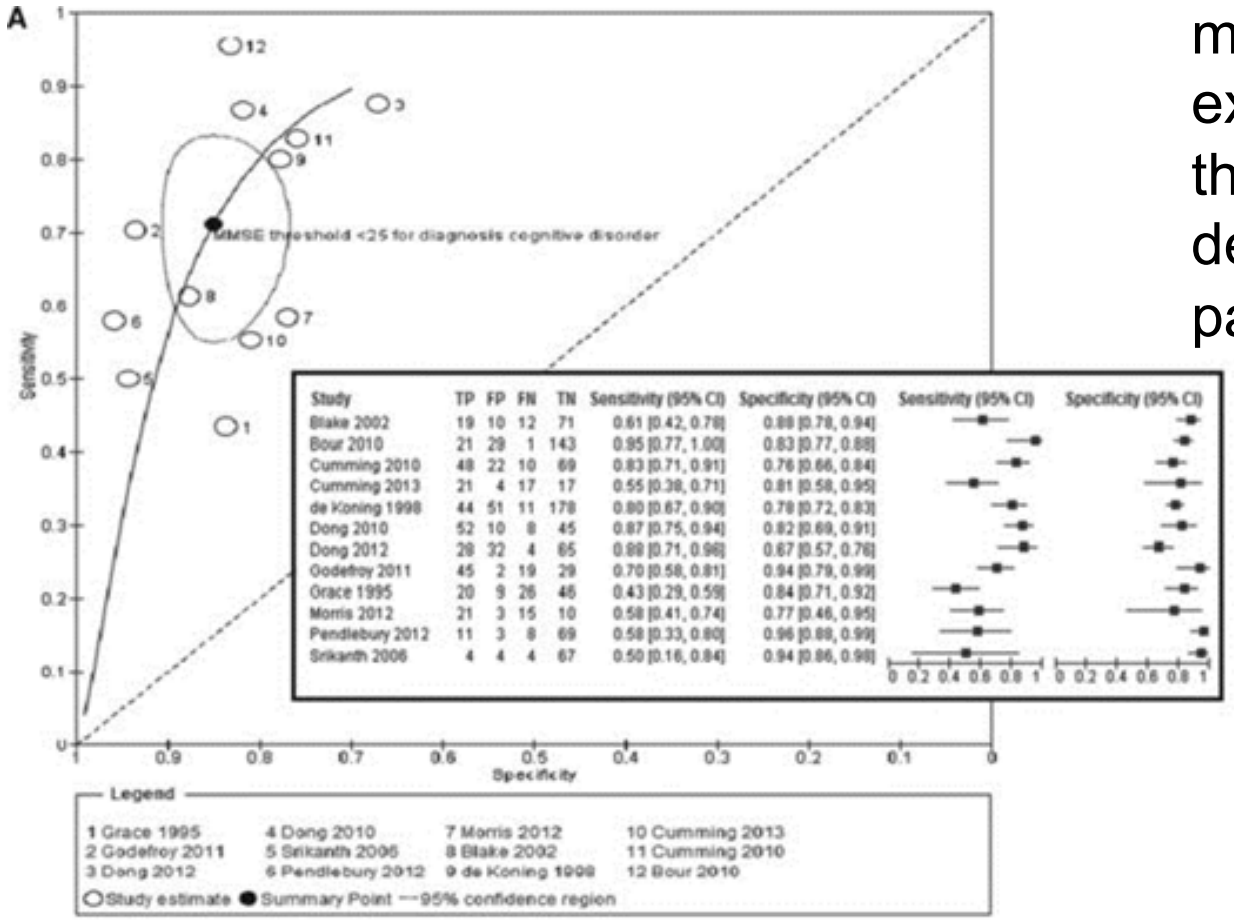
- ✓Two dependent outcomes – **sensitivity** and **specificity**
- ✓ Variable test threshold levels (either explicit or implicit)

  - BUT data on test threshold in primary studies (if known) ignored
- Different reference tests

Other issues include:

- Different populations / study conduct (leading to between-study heterogeneity)
- Data quality / risk of bias

Different reference tests



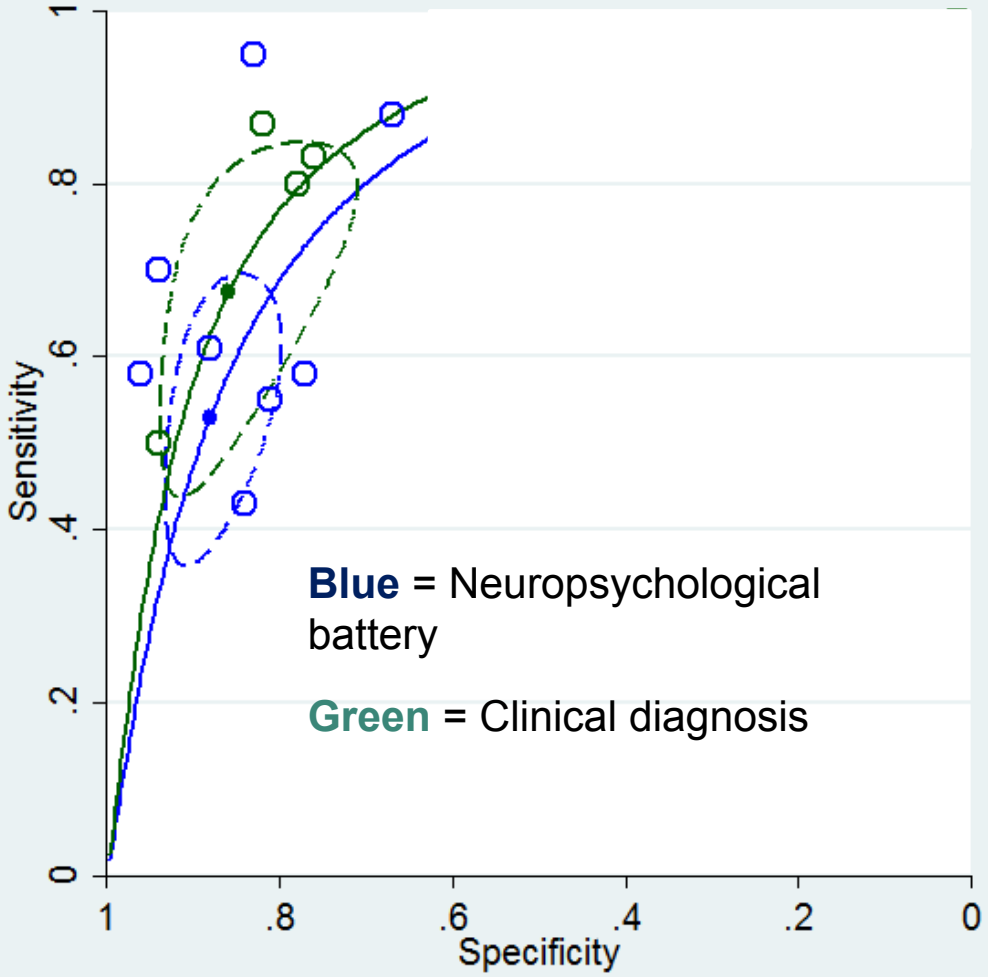
Lees et al Stroke 2014

Evaluating the Folstein's mini-mental state examination < 25/30 for the diagnosis of dementia in stroke patients

Studies included in the meta-analysis used two different reference standards:

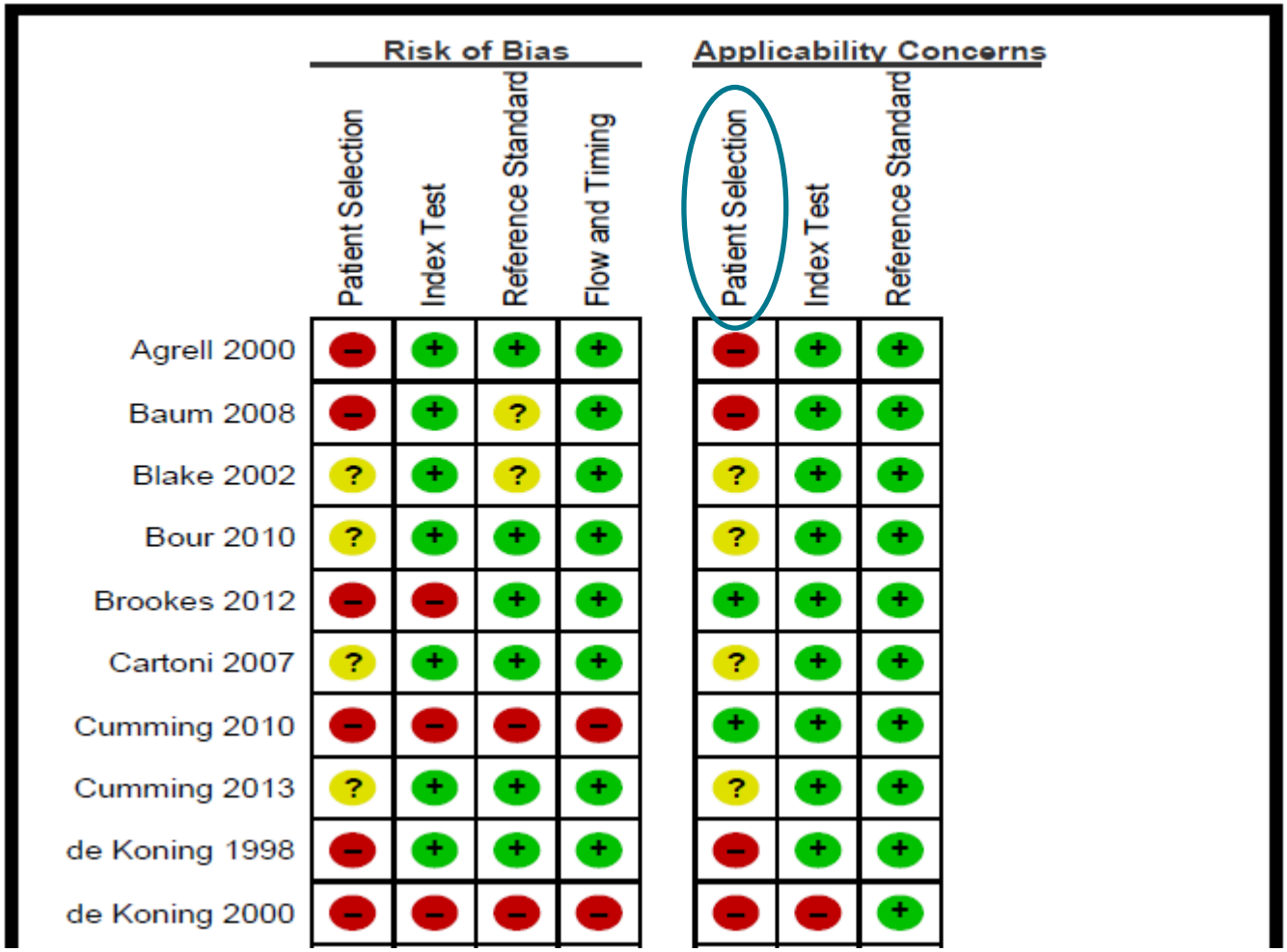
- Neuropsychological battery (NPB)
- Clinical diagnosis

Different reference tests

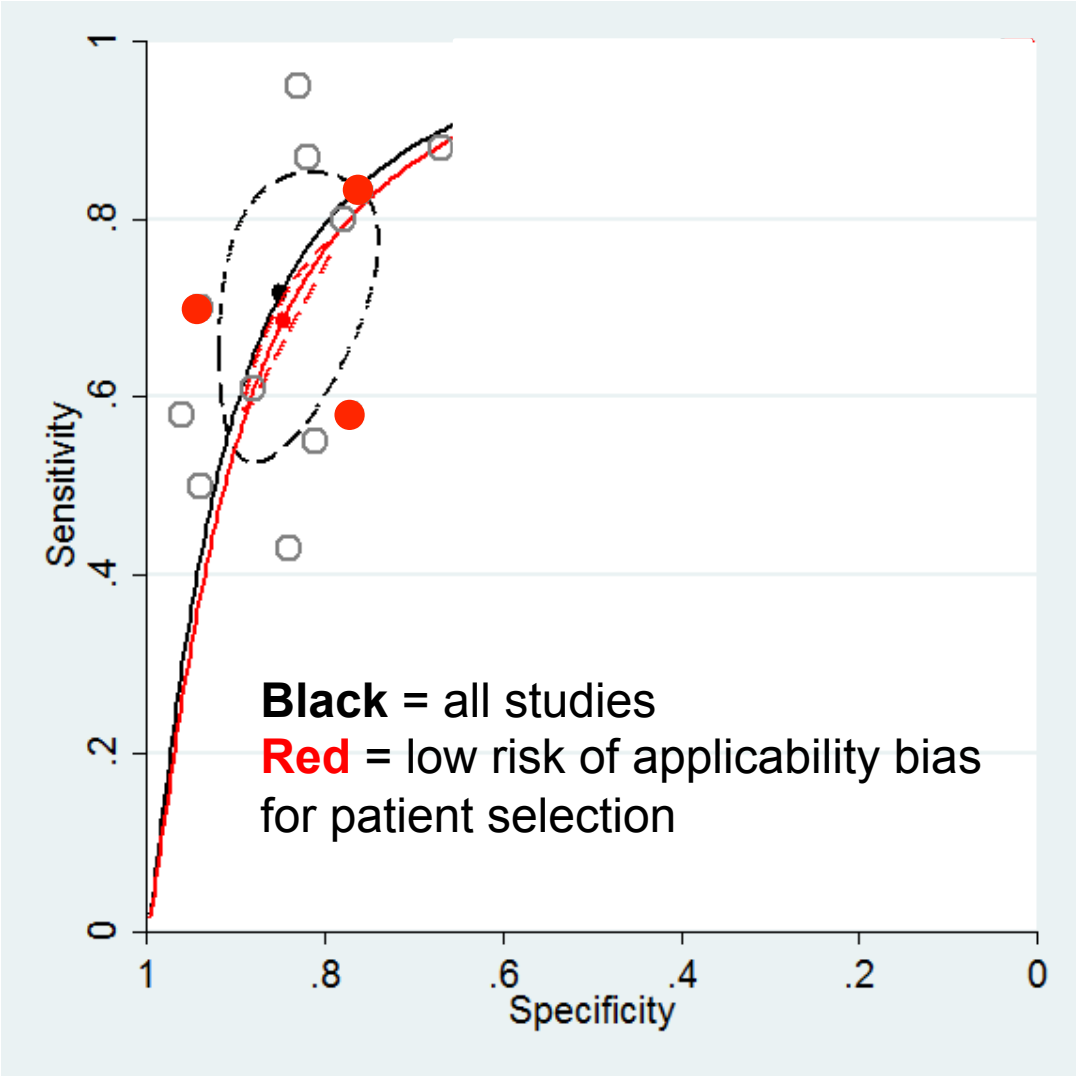


Exploring whether results vary by reference test using meta-regression

Assessing study quality / risk of bias



Assessing study quality / risk of bias



- Exploring impact of risk of applicability bias due to approach to patient selection using meta-regression
- 95% credible region reduced

UNIVERSITY OF LEICESTER

Challenges of meta-analysing diagnostic test accuracy data

More complex than for effectiveness data due to:

- ✓Two dependent outcomes – **sensitivity** and **specificity**
- ✓ Variable test threshold levels (either explicit or implicit)
  - BUT data on test threshold in primary studies (if known) ignored
- ✓ Different reference tests (imperfect gold standard)

Other issues include:

- ✓ Different populations / study conduct (leading to between-study heterogeneity)
  - Limited by the data available and number of studies
- ✓ Data quality / risk of bias
  - Limited by the data available and number of studies

Software

- All analyses presented so far are possible to fit using Stata macros
- Similar functionality available in R
- Bespoke macro for SAS developed specifically for Cochrane use
- WinBUGS can fit all of the above and beyond (but not graphics!)

Aim 2

Provoke discussion regarding how to ensure reviews of diagnostic tests answer clinically-relevant questions

How do we compare performance of different tests?

- Paucity of direct comparative studies of test accuracy
- Systematic reviews of comparative accuracy often undertake separate meta-analyses for each test and then compare their results implicitly:
  - Does not ensure like-with-like comparisons (i.e. test accuracy may be confounded by patient group, study methods, etc.)
  - Often no common control/reference test
  - Diagnosis often requires the use of multiple tests in combination

Beyond “simple” pairwise meta-analysis

- Methods have been generalised allowing synthesis of studies including multiple index tests on the same patients *(ARHQ 2013)*
- Similarly, methods generalised to include multiple threshold points for the same test from each study
- Several groups working on network meta-analysis in a diagnostic test context
- Individual patient data potentially offers the ability to perform more powerful analyses

Evaluating sequences of tests to optimise diagnosis

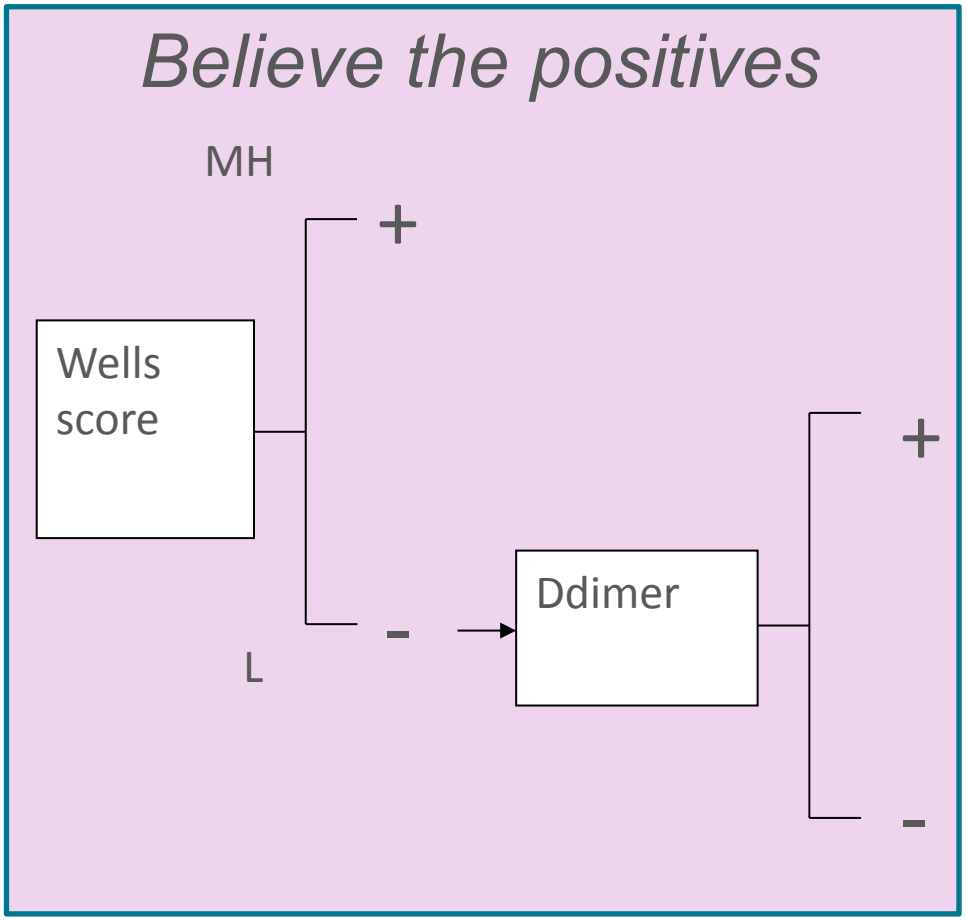
Often single tests evaluated in studies but multiple tests used in combination for diagnosis

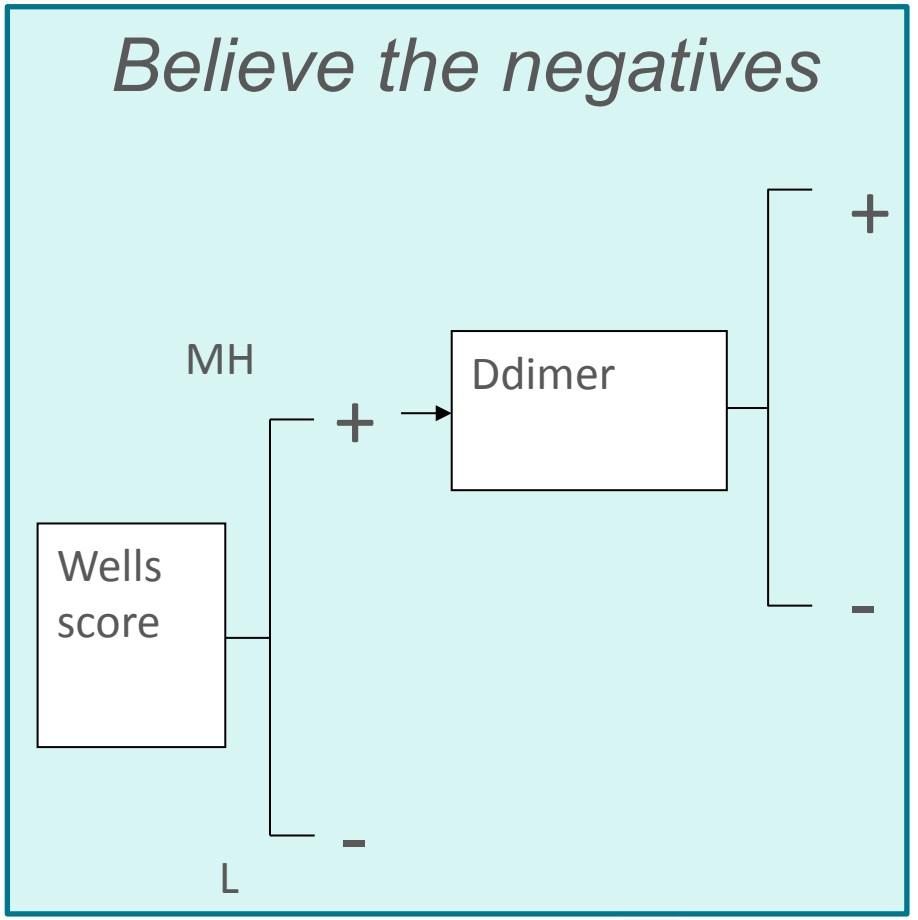
Performance of tests may differ depending where they are in the diagnostic pathway (i.e. test performance is not independent)

Ideally want to estimate (meta-analytically) the accuracy of combinations of diagnostic tests (reflecting clinical practice), acknowledging the **likely** *non-independence* of the tests

Evaluating sequences of tests

Example: Evaluation of *Ddimer (a blood test)* and *Wells Score (checklist of symptoms & clinical history)* tests for diagnosing Deep Vein Thrombosis





Deep Vein Thrombosis example

Examples of data types	Wells Score		Ddimer Accuracy data			
	WS group	Diseased/ Total	True +ve	False +ve	False -ve	True –ve
TYPE A: complete data (n=11)	high	10/17	8	2	2	5
	moderate	6/44	6	18	0	20
	low	1/41	1	8	0	32
TYPE B: (n=4)	high	26/29	25	2	1	1
	moderate	4/15	-	-	-	-
	low	2/32	-	-	-	-
TYPE C (n=4)	high	-	-	-	-	-
	moderate	-	-	-	-	-
	low	2/149	2	76	0	71
TYPE D (n=20)	high	26/29	-	-	-	-
	moderate	4/15	-	-	-	-
	low	2/32	-	-	-	-
TYPE E (n=94)	N/A	-	2	76	0	71

Novielli et al Value in Health 2013

Is the data fit for purpose?

- Many DTA studies small and poor quality focusing on a single index test
- Are exhaustive all-inclusive systematic reviews of these studies the optimal way to answer relevant clinical/policy questions?
  - i.e. Although this approach proven to be successful for RCTs of interventions, do we need to innovate rather than simply translate methodology?
- Alternative approaches:
  - Review only large, good-quality studies
  - Conduct new primary studies of multiple index tests evaluating the whole diagnosis/treatment pathways
    - More reliable and efficient than trying to combine heterogeneous, often poor quality studies, on different parts of the “puzzle”??

Discussion points

What are the most clinical/policy relevant questions to answer when evaluating diagnostic test performance?

- What is the accuracy of test X (sensitivity & specificity)?
- What factors affect test Xs accuracy?
- At what threshold should test X be used at?
- Which test, X or Y, is the most accurate?
- Where in the diagnostic pathway should test X be used, and at what threshold? Should other tests (Y, Z, etc.) be included in the diagnostic pathway?
- What is the most cost-effective diagnostic strategy for a given disease? (i.e. test sequence and thresholds)

  - Requires modelling of full clinical pathway including subsequent treatments and beyond

How can Cochrane reviews help to answer these questions?

- How well are we doing currently?
- What could be improved?

Complex Reviews Support Unit (CRSU)

Expertise within CRSU

Key Areas of Support

- Diagnostic test accuracy (DTA) reviews
- Network meta-analysis (NMA)
- Individual participant data (IPD)/clinical study report meta-analysis

Other Areas of Support

- Economic evaluation
- Realist synthesis
- Qualitative reviews
- Use of routine data
- Non-randomised studies
- Prognostic reviews
- Prevalence reviews
- Causal pathway analysis

Conclusions

- We don’t have all the answers
  - We do have some
  - Perhaps we are (at least) starting to ask the right questions?
- Important to work closely with clinicians
  - As analyses get more complex the results obtained are relevant to clinical practice (i.e. answer clinically meaningful questions)
- CRSU funded by the NIHR to offer support all complex reviews
- Please let us know how the CRSU can offer assistance

NIHR CRSU - Complex Reviews Support Unit

- About us
- FAQs
- Open Business
- Apply for CRSU Support
- Links to Guidance
- Courses and Conferences
- Contact us

Contact us

Moira Aitken
CRSU
1 Lilybank Gardens
Glasgow
G12 8RZ

Telephone: 0141 330 6826
Email: moira.aitken@glasgow.ac.uk

Website: www.nihrcrsu.org
Twitter: @NIHRCRSU

University of Glasgow
UNIVERSITY OF LEICESTER
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE
NHS National Institute for Health Research