Comparing test accuracy: from pairwise to network meta-analysis of tests

Yemisi Takwoingi
Professor of Test Evaluation and Evidence Synthesis
Co-convenor Cochrane Screening & Diagnostic Tests Methods Group
TESTS
Outline

• Introduction/background
  • Comparative accuracy study designs

• Current practice for comparing diagnostic test accuracy (DTA)
  • Test comparison strategy
  • Meta-analysis methods

• Extensions to network meta-analysis of diagnostic test accuracy (DTA-NMA)
Scope of a DTA review

• Multiple objectives are possible
• 3 main types of analyses based on review question and objectives
  
  1) What is the diagnostic accuracy of a test?
  
  2) How does the accuracy of two or more tests compare?

  3) How does test accuracy vary with clinical and methodological characteristics?

(1) & (2) are typically primary objectives of a DTA review
Is the sensitivity and (or) specificity of the new test better than that of existing test(s)?
Index and comparator tests

- **Index test**: “new” test or test strategy we wish to evaluate
- **Comparator test**: existing test or diagnostic management strategy which may be standard practice
- We compare the accuracy of the index and the comparator tests
- The term “comparator test” can be confusing so simply put, we compare the accuracy of index tests
- **Reference standard**: the best available way to verify the presence or absence of the target condition.
  - May be a single test or a combination of tests and clinical information not routinely available in practice.
Study designs for comparing test accuracy

Within-study (controlled) comparison

Within-subject (paired or multiple tests)

Between-subject (unpaired parallel groups)

Takwoingi 2016 [https://etheses.bham.ac.uk/id/eprint/6759/](https://etheses.bham.ac.uk/id/eprint/6759/)
Study designs for comparing test accuracy

Within-study (controlled) comparison

Within-subject (paired or multiple tests)  
Between-subject (unpaired parallel groups)  
Random allocation  
Non-random allocation

Takwoingi 2016 [https://etheses.bham.ac.uk/id/eprint/6759/](https://etheses.bham.ac.uk/id/eprint/6759/)
Study designs for comparing test accuracy

Within-study (controlled) comparison

Within-subject (paired or multiple tests)
Between-subject (unpaired parallel groups)

Random allocation
Non-random allocation

Prospectively controlled
Historically controlled

Takwoingi 2016 https://etheses.bham.ac.uk/id/eprint/6759/
Study designs for comparing test accuracy

Within-study (controlled) comparison

Within-subject (paired or multiple tests)

Between-subject (unpaired parallel groups)

Robust designs

Random allocation

Non-random allocation

Prospectively controlled

Historically controlled

Takwoingi 2016 [https://etheses.bham.ac.uk/id/eprint/6759/]
Robust test comparison designs

Series of patients

R

CT (index test A)

CAG (Reference standard)

Estimate sensitivity and specificity for CT

Compare test accuracy between randomized groups

MRI (index test B)

CAG (Reference standard)

Estimate sensitivity and specificity for MRI

Unpaired (between-subject randomized) design

CAD = coronary artery disease
CAG = coronary angiography
Robust test comparison designs

Unpaired (between-subject randomized) design

Series of patients

CT (index test A)

MRI (index test B)

CAG (Reference standard)

CAG (Reference standard)

Estimate sensitivity and specificity for CT

Estimate sensitivity and specificity for MRI

Compare test accuracy between randomized groups

CAD = coronary artery disease

CAG = coronary angiography

<table>
<thead>
<tr>
<th>CAD</th>
<th>No CAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT+</td>
<td>TP</td>
</tr>
<tr>
<td>CT−</td>
<td>FN</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CAD</th>
<th>No CAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRI+</td>
<td>TP</td>
</tr>
<tr>
<td>MRI−</td>
<td>FN</td>
</tr>
</tbody>
</table>
Robust test comparison designs

Paired (within-subject) design

Series of patients
- CT (index test A)
- MRI (index test B)
- CAG (Reference standard)

Estimate sensitivity and specificity for CT
Estimate sensitivity and specificity for MRI

Compare test accuracy within patients

CAD = coronary artery disease
CAG = coronary angiography
Robust test comparison designs

Series of patients
- CT (index test A)
- MRI (index test B)
  CAG (Reference standard)

Estimate sensitivity and specificity for CT
Compare test accuracy within patients
Estimate sensitivity and specificity for MRI

CAD = coronary artery disease
CAG = coronary angiography

<table>
<thead>
<tr>
<th></th>
<th>CAD</th>
<th>No CAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CT+</td>
<td>CT−</td>
</tr>
<tr>
<td>MRI+</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>MRI−</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Paired (within-subject) design
Joint classification table: an example

Objectives: To compare the diagnostic accuracy and cost-effectiveness of T-SPOT.TB® (Oxford Immunotec, Abingdon, UK) and QuantiFERON® TB GOLD In-Tube (Cellestis, Carnegie, VIC, Australia) for diagnosis of suspected active TB and to estimate the diagnostic accuracy of second-generation IGRA.

Design: Prospective within-patient comparative diagnostic accuracy study.

<table>
<thead>
<tr>
<th>T-SPOT.TB, n</th>
<th>Active TB positive (categories 1 and 2)</th>
<th>Active TB negative (category 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>187</td>
<td>13</td>
</tr>
<tr>
<td>Negative</td>
<td>49</td>
<td>41</td>
</tr>
<tr>
<td>Indeterminate</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>Missing</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

QFT-GIT Total | 253  | 58  | 17  | 17  | 18  | 363   | 51  | 319  | 16  | 37  | 16  | 439   |

Empirical Evidence of the Importance of Comparative Studies of Diagnostic Test Accuracy

Yemisi Takwoingi, DVM; Mariska M.G. Leeflang, PhD; and Jonathan J. Deeks, PhD

Background: Systematic reviews that "compare" the accuracy of 2 or more tests often include different sets of studies for each test.

Purpose: To investigate the availability of direct comparative studies of test accuracy and to assess whether summary estimates of accuracy differ between meta-analyses of noncomparative and comparative studies.

Data Sources: Systematic reviews in any language from the Database of Abstracts of Reviews of Effects and the Cochrane Database of Systematic Reviews from 1994 to October 2012.

Study Selection: 1 of 2 assessors selected reviews that evaluated at least 2 tests and identified meta-analyses that included both noncomparative and comparative studies.

Data Extraction: 1 of 3 assessors extracted data about review and study characteristics and test performance.

Data Synthesis: 248 reviews compared test accuracy; of the 6915 studies, 2113 (31%) were comparative. Thirty-six reviews (with 52 meta-analyses) had adequate studies to compare results of noncomparative and comparative studies by using a hierarchical summary receiver-operating characteristic meta-regression model for each test comparison. In 10 meta-analyses, noncomparative studies ranked tests in the opposite order of comparative studies. A total of 25 meta-analyses showed more than a 2-fold discrepancy in the relative diagnostic odds ratio between noncomparative and comparative studies. Differences in accuracy estimates between noncomparative and comparative studies were greater than expected by chance ($P < 0.001$).

Limitation: A paucity of comparative studies limited exploration of direction in bias.

Conclusion: Evidence derived from noncomparative studies often differs from that derived from comparative studies. Robustly designed studies in which all patients receive all tests or are randomly assigned to receive one or other of the tests should be more routinely undertaken and are preferred for evidence to guide test selection.

Primary Funding Source: National Institute for Health Research (United Kingdom).

Ann Intern Med. 2013;158:544-554. For author affiliations, see end of text.
What is the test comparison strategy in the comparative DTA review?
**Test comparison strategy**

**Direct (head-to-head) comparison**
- **Index test A**
- **Index test B**
- **Reference standard**
- Estimate sensitivity and specificity for test A
- Estimate sensitivity and specificity for test B
- Compare test accuracy using same set of studies

**Indirect (between-study) comparison**
- **Index test A**
- **Index test B**
- **Reference standard**
- Estimate sensitivity and specificity for test A
- Estimate sensitivity and specificity for test B
- Compare test accuracy using different sets of studies
Cochrane DTA Review examples

Xpert MTB/RIF and Xpert MTB/RIF Ultra assays for active tuberculosis and rifampicin resistance in children

Alexander W Kay1, Lucia González Fernández2, Yemisi Takwoingi3, Michael Eisenhut4, Anne K Detjen5, Karen R Steingart66, Anna M Mandalakas1b

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

Katharine Abba1, Jonathan J Deeks2, Piero L Olliaro3, Cho-Min Naing4, Sally M Jackson1, Yemisi Takwoingi2, Sarah Donegan1, Paul Garner1

First trimester serum tests for Down's syndrome screening

S Kate Alldred1, Yemisi Takwoingi2, Boliang Guo3, Mary Pennant4, Jonathan J Deeks2, James P Neilson1, Zarko Alfirevic1
Example 1: Xpert MTB/RIF and Xpert MTB/RIF Ultra assays for active tuberculosis and rifampicin resistance in children

Direct comparison of Xpert MTB/RIF and Xpert Ultra (3 studies)

Kay AW et al. 2020. CD013359
Example 2: Rapid diagnostic tests for *P. falciparum* malaria

- HRP-2 antibody-based tests
- RDTs
- pLDH antibody-based tests

Abba K et al. 2011. CD008122
Example 2: Rapid diagnostic tests for *P. falciparum* malaria

Abba K et al. 2011. CD008122
Example 2: Rapid diagnostic tests for *P. falciparum* malaria

- 6 RDT types within 2 groups of antibody-based tests
- Type 1: 10 brands
- Type 2: 2 brands
- Type 3: 3 brands
- Type 4: 4 brands
- Type 5: 2 brands
- Type 6: none
- Total of 21 RDT brands included in 74 studies

Abba K et al. 2011. CD008122
Example 3: First trimester serum test strategies for Down’s syndrome screening

Takwoingi 2016 [https://etheses.bham.ac.uk/id/eprint/6759/]

Alldred SK et al. 2015. CD011975
Methods and reporting of systematic reviews of comparative accuracy were deficient: a methodological survey and proposed guidance

Yemisi Takwoingi\textsuperscript{a,b,*}, Christopher Partlett\textsuperscript{c}, Richard D. Riley\textsuperscript{d}, Chris Hyde\textsuperscript{e}, Jonathan J. Deeks\textsuperscript{a,b}

\textsuperscript{a}Test Evaluation Research Group, University of Birmingham, Birmingham, UK
\textsuperscript{b}NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK
\textsuperscript{c}Nottingham Clinical Trials Unit, Faculty of Medicine and Health Sciences, University of Nottingham, Nottingham, UK
\textsuperscript{d}Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University, Staffordshire, UK
\textsuperscript{e}Exeter Test Group, College of Medicine and Health, University of Exeter, Exeter, UK

Accepted 11 December 2019; Published online 14 December 2019

Abstract

Objective: The objective of this study was to examine methodological and reporting characteristics of systematic reviews and meta-analyses which compare diagnostic test accuracy (DTA) of multiple index tests, identify good practice, and develop guidance for better reporting.
How can I statistically combine the studies to compare test accuracy?
Key challenges for DTA meta-analysis

• Two summary statistics for each study
  • sensitivity and specificity and each have different implications

• Threshold effects induce correlations between sensitivity and specificity and often seem to be present
  • thresholds can vary between studies
  • same threshold can imply different sensitivities and specificities in different groups

• Heterogeneity is the norm
  • substantial variation in sensitivity and specificity are observed in most reviews
Additional key challenges for comparative DTA meta-analysis

• Many DTA studies are not comparative
• Different study designs
  • Correlated data
  • Availability of fully cross-classified data
## Meta-analysis methods for comparing test accuracy

*(up to July 2014)*

<table>
<thead>
<tr>
<th>Reference</th>
<th>Method</th>
<th>Test accuracy measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Moses et al 1993; Littenberg and Moses 1993</td>
<td>Comparison of Q*</td>
<td>Q*</td>
</tr>
<tr>
<td>2 Hasselblad and Hedges 1995</td>
<td>Standardized distance between the means of two populations</td>
<td>Effectiveness measure (d) proportional to log DOR</td>
</tr>
<tr>
<td>3 <strong>Rutter and Gatsonis 2001</strong></td>
<td>HSROC meta-regression</td>
<td>Diagnostic odds ratio (DOR)</td>
</tr>
<tr>
<td>4 Kowalski et al 2001</td>
<td>Generalized estimating equation</td>
<td>Sensitivity and specificity</td>
</tr>
<tr>
<td>5 Lijmer et al 2002</td>
<td>Moses SROC meta-regression</td>
<td>DOR</td>
</tr>
<tr>
<td>6 Worster et al 2002</td>
<td>General linear mixed model</td>
<td>Likelihood ratios</td>
</tr>
<tr>
<td>7 Suzuki et al 2004</td>
<td>Conditional relative odds ratio</td>
<td>DOR</td>
</tr>
<tr>
<td>8 Siadaty and Shu 2004</td>
<td>Proportional odds ratio</td>
<td>DOR</td>
</tr>
<tr>
<td>9 Siadaty et al 2004</td>
<td>Repeated measures modelling</td>
<td>DOR</td>
</tr>
<tr>
<td>10 <strong>Reitsma et al 2005; Hamza et al 2009</strong></td>
<td>Bivariate meta-regression</td>
<td>Sensitivity and specificity</td>
</tr>
<tr>
<td>11 Cheng et al 2013‡</td>
<td>Network meta-analysis</td>
<td>Sensitivity and specificity</td>
</tr>
<tr>
<td>12 Verde 2013‡</td>
<td>Bivariate meta-analysis of paired data</td>
<td>Sensitivity and specificity</td>
</tr>
<tr>
<td>13 Trikalinos et al 2014</td>
<td>Bivariate meta-analysis of paired data</td>
<td>Sensitivity and specificity</td>
</tr>
</tbody>
</table>

‡Conference presentation

Takwoingi 2016 [https://etheses.bham.ac.uk/id/eprint/6759/](https://etheses.bham.ac.uk/id/eprint/6759/)
Hierarchical meta-regression

• Hierarchical models can incorporate a study-level covariate to compare test accuracy

• Different questions can be addressed
  • Bivariate model
    • differences in summary points of sensitivity and/or specificity
  • HSROC model
    • differences in overall accuracy
    • differences in threshold
    • differences in shape of SROC curve
Comparing test accuracy

CT vs MRI for CAD

CT: 89 studies
MRI: 19 studies

BNP vs NT-proBNP for heart failure

BNP: 20 studies, 16 cut-offs
NT-proBNP: 16 studies, 13 cut-offs
Hierarchical meta-regression models

**Bivariate model**

Primary objective: to compare summary points

**HSROC model**

Primary objective: to compare summary curves
Let’s get technical...
Bivariate model specification

Models the proportion in each study \((i)\) that have correct test results in diseased and non-diseased groups

\[
\begin{pmatrix}
\mu_{Ai} \\
\mu_{Bi}
\end{pmatrix} \sim N\left(\begin{pmatrix}
\mu_A \\
\mu_B
\end{pmatrix}, \sum \right)
\]

with \(\sum = \begin{pmatrix}
\sigma_A^2 & \sigma_{AB} \\
\sigma_{AB} & \sigma_B^2
\end{pmatrix}\)

\(\mu_A\) is the mean logit sensitivity
\(\mu_B\) is the mean logit specificity
\(\sigma_A^2\) is the variance of the logit sensitivity
\(\sigma_B^2\) is the variance of the logit specificity
\(\sigma_{AB}\) is the covariance of logit sensitivity and logit specificity
Bivariate model with a covariate

Assuming a test type covariate $t$ that may affect both sensitivity and specificity, the model can be extended as follows:

$$
\begin{pmatrix}
\mu_{Aik} \\
\mu_{Bik}
\end{pmatrix} \sim N
\begin{pmatrix}
\begin{pmatrix}
\mu_A \\
\mu_B
\end{pmatrix}
+ \begin{pmatrix}
\nu_A t_k \\
\nu_B t_k
\end{pmatrix},
\begin{pmatrix}
\sigma_A^2 & \sigma_{AB} \\
\sigma_{AB} & \sigma_B^2
\end{pmatrix}
\end{pmatrix}
$$

Effect of test type on variance parameters can also be investigated.
CT versus MRI for CAD example: direct and indirect comparisons

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CT</strong></td>
<td>0.97 (0.96, 0.98)</td>
<td>0.87 (0.84, 0.90)</td>
</tr>
<tr>
<td><strong>MRI</strong></td>
<td>0.88 (0.84, 0.91)</td>
<td>0.70 (0.59, 0.79)</td>
</tr>
</tbody>
</table>

5 CT versus MRI studies

103 studies: 84 CT only, 14 MRI only, 5 CT vs MRI studies
Meta-regression not limited to pairwise comparisons

Sensitivity at a 5% false positive rate for 9 first trimester serum test strategies for Down’s syndrome screening

Each circle represents the summary sensitivity for a test strategy and the size of each circle is proportional to the number of Down's cases.

The test strategies are ordered according to decreasing sensitivity. The number of studies, cases and women included for each test strategy are shown on the horizontal axis.

A=Age, PI GF, PAPP-A and free βhCG; B=Age, PAPP-A, free βhCG and AFP; C=Age, ADAM 12, PAPP-A and free βhCG; D=Age, PAPP-A and free βhCG; E=Age, PAPP-A; F=PAPP-A; G=Age, free βhCG and AFP; H=Age, free βhCG; I=Free βhCG

Alldred SK et al. 2015. CD011975
HSROC model specification

The model takes the form

$$\text{logit}(\pi_{ij}) = (\theta_i + \alpha_i \text{dis}_{ij}) \exp(-\beta \text{dis}_{ij})$$

- **threshold**
  - i.e. proportion test positive
    - (random effect)

- **accuracy**
  - (random effect)

- **dependence of accuracy on threshold**
  - i.e. shape of the summary curve
    - (fixed effect)
HSROC model with a covariate

• Assuming a test type covariate $Z$ that may affect accuracy, threshold and shape, the model can be extended as:

$$\text{logit}(\pi_{ij}) = \left((\theta_i + \gamma Z_i) + (\alpha_i + \lambda Z_i)dis_{ij}\right) \exp\left(-(\beta + \delta Z_i) \cdot dis_{ij}\right)$$

• Shape parameter is estimated as $\beta$ for one test and $\beta+\delta$ for the other test.

• If $\delta = 0$ is assumed and covariate terms are removed for shape, SROC curves for the tests will have the same shape ($\beta$)

$$\text{logit}(\pi_{ij}) = \left((\theta_i + \gamma Z_i) + (\alpha_i + \lambda Z_i)dis_{ij}\right) \exp\left(-\beta \cdot dis_{ij}\right)$$

• Relative diagnostic accuracy of the two curves can be summarized using the relative DOR = $\exp(\lambda)$
A ‘non-technical’ summary of the methods

Meta-analysis of diagnostic accuracy studies in mental health

Yemisi Takwoingi,¹ Richard D Riley,² Jonathan J Deeks¹

¹Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, UK; ²Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, UK

Correspondence to Dr Yemisi Takwoingi, y.takwoingi@bham.ac.uk

ABSTRACT

Objectives To explain methods for data synthesis of evidence from diagnostic test accuracy (DTA) studies, and to illustrate different types of analyses that may be performed in a DTA systematic review.

Methods We described properties of meta-analytic methods for quantitative synthesis of evidence. We used a DTA review comparing the accuracy of three screening questionnaires for bipolar disorder to illustrate application of the methods for each type of analysis.

Results The discriminatory ability of a test is commonly expressed in terms of sensitivity (proportion of those with the condition who test positive) and specificity (proportion of those without the condition who test negative). There is a trade-off between sensitivity and specificity, as an increasing threshold for defining test positivity will decrease sensitivity and increase specificity. Methods recommended for meta-analysis of DTA studies —such as the bivariate or hierarchical summary receiver operating characteristic (HSROC) model —jointly summarise sensitivity and specificity while taking into account this threshold effect, as well as allowing for between-study differences in test performance beyond what would be expected by chance. The bivariate model focuses on estimation of a summary sensitivity and specificity at a common threshold while the HSROC model focuses on the estimation of a summary curve from studies that have used different thresholds.

Conclusions Meta-analyses of diagnostic accuracy studies can provide answers to important clinical questions. We hope this article will provide clinicians with sufficient understanding of the terminology and methods to aid interpretation of systematic reviews and facilitate better patient care.
Software for meta-analysis of DTA studies

Resources for authors

- DTA Handbook
- Software for meta-analysis

Researchers have prepared macros or modules for statistical models for meta-analysis of data from diagnostic test accuracy studies for several statistical analysis software programs. As these become available we will add them to this page. Currently, there is a macro available for SAS and a package for STATA.

SAS

MetaDAS: A SAS macro for meta-analysis of diagnostic accuracy studies, contains both the bivariate and the HSROC model. Please find the required documents hereunder:

- **User guide** version 1.3 (2012). (PDF 2.7MB, opens in new window)
- **Quick reference** and worked example (2012). (PDF 2.6MB, opens in new window)
- The SAS macro itself: **METADAS v1.3**. This is provided as a text file and opens in a new window.

R

There are several user-written packages for conducting meta-analysis of diagnostic test accuracy (DTA) studies in R. This tutorial summarises and illustrates some of the packages. Step-by-step instructions are also provided for carrying out the bivariate binomial method by fitting a generalized linear mixed model (GLMM) using the glmer function in the R package lme4. A .R file, "Bivariate binomial meta-analysis of diagnostic test accuracy studies.R" and example dataset based on a review by Schuetz et al. 2010, are included with the **tutorial** in the zipped folder.

STATA

METANDI: A Stata user-written package for meta-analysis of diagnostic accuracy studies (Harbord and Whiting 2008;...
Can I do network meta-analysis of diagnostic test accuracy?
What are the DTA-NMA methods and which one should I use?
Network meta-analysis for DTA (DTA-NMA)
Identification of NMA-DTA methods
10 methodological studies and 40 empirical studies

From inception to end of July 2019
# DTA-NMA methods

## 10 methodological studies of 9 different DTA-NMA methods

<table>
<thead>
<tr>
<th>Model</th>
<th>Arm-based</th>
<th>Bayesian setting</th>
<th>Imperfect reference standard</th>
<th>Multiple thresholds</th>
<th>Joint classification tables</th>
<th>2x2 tables/index test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trikalinos et al. 2014</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Ma 2015</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Menten &amp; Lesaffre 2015 (Model A)</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Menten &amp; Lesaffre 2015 (Model B)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Menten &amp; Lesaffre 2015 (Model C)</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Dimou et al. 2016</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Cheng 2016 (Model A)</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Cheng 2016 (Model B)</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Cheng 2016 (Model C)</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nyaga et al. 2018a</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Nyaga et al. 2018b</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Owen et al. 2018</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Lian et al. 2019</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Evaluating multiple diagnostic tests: An application to cervical cancer

Areti Angeliki Veroniki¹,²,³, Sofia Tsokani¹, Evangelos Paraskevaidis⁴, Dimitris Mavridis¹,⁵
**Hierarchical meta-regression and DTA-NMA methods**

<table>
<thead>
<tr>
<th>Bivariate meta-regression model</th>
<th>Normal-binomial model</th>
<th>Hierarchical latent class model</th>
</tr>
</thead>
<tbody>
<tr>
<td>- A covariate for test type is used to explore sensitivity and specificity between tests</td>
<td>- Hierarchical model using the logit transformation of sensitivity and specificity</td>
<td>- Based on differences (contrasts) between the different tests in the network</td>
</tr>
<tr>
<td>- Assumes that participants undergoing different tests are independent subgroups within each study</td>
<td>- Allows for correlation between tests</td>
<td>- Allows for different reference standards</td>
</tr>
<tr>
<td>- Does not account for the within-study correlation between tests</td>
<td><strong>Beta-binomial model</strong></td>
<td>- Correlations between tests from the same study are ignored</td>
</tr>
<tr>
<td><strong>Nyaga et al. (2018b)</strong></td>
<td><strong>Normal-binomial model</strong></td>
<td></td>
</tr>
<tr>
<td>- Sensitivity &amp; specificity are directly modelled using a beta-binomial defined in $[0,1]$</td>
<td><strong>Variance component model</strong></td>
<td></td>
</tr>
</tbody>
</table>
Network plot of cytology, HPV DNA, and mRNA tests for CIN2+

Closed triangle solid line: triple-test studies (n = 4)
Black circle: cytology only (n = 1)
Dotted-dashed line: HPV DNA vs cytology (n = 32)

Summary of application to cervical cancer

• Different DTA-NMA methods may lead to different results
  • Differences in point estimates and their uncertainty

• Differences in results across models may be due to differences in how the models deal with
  • Heterogeneity
  • Sensitivity and specificity (logits or proportions)

• Choice of a DTA-NMA method depends on the available data
Are we there yet with DTA-NMA?

FOR THE LAST TIME... NO! AND STOP ASKING!!!

ARE WE THERE YET?
Limitations of DTA-NMA

- Comprehensive evaluation is needed to assess the performance of the models
- Complexity: as number of tests increase, number of additional parameters to estimate increase, and so does risk of convergence issues
- Data availability
- Lack of easy to use programs in popular statistical software
“Meta-analytic models that account for pairing of test results within an individual within each study have been developed as an extension of the bivariate model. The method proposed by Trikalinos (2014) ... The approach of Dimou (2016) ... These methods require further evaluation before they are recommended for routine use. However, as suggested by Trikalinos (2014) they may be useful as a sensitivity analysis.

Network meta-analysis models have also been developed that utilise data from both direct and indirect comparisons of multiple tests... However, further evaluation of these methods for dealing with complex correlational structures is required before they are implemented in Cochrane reviews."
Take home message

• Be clear about the test comparison strategy and strength of the evidence
  • All studies (comparative and non-comparative studies)
  • Restricted to comparative studies that have directly compared the tests
  • Analyses using relevant comparative studies are desirable but may not be feasible

• Hierarchical meta-regression models for comparison of points (bivariate model) or curves (HSROC model) are the norm.

• More complex methods are being published but evaluations are required before they can be adopted in Cochrane DTA reviews.

• A rapidly developing field so watch this space.
Acknowledgements

• Thanks to Argie Veroniki and Sofia Tsokani for slides on the scoping review and the case study.

• The views expressed are mine and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.
References


References


