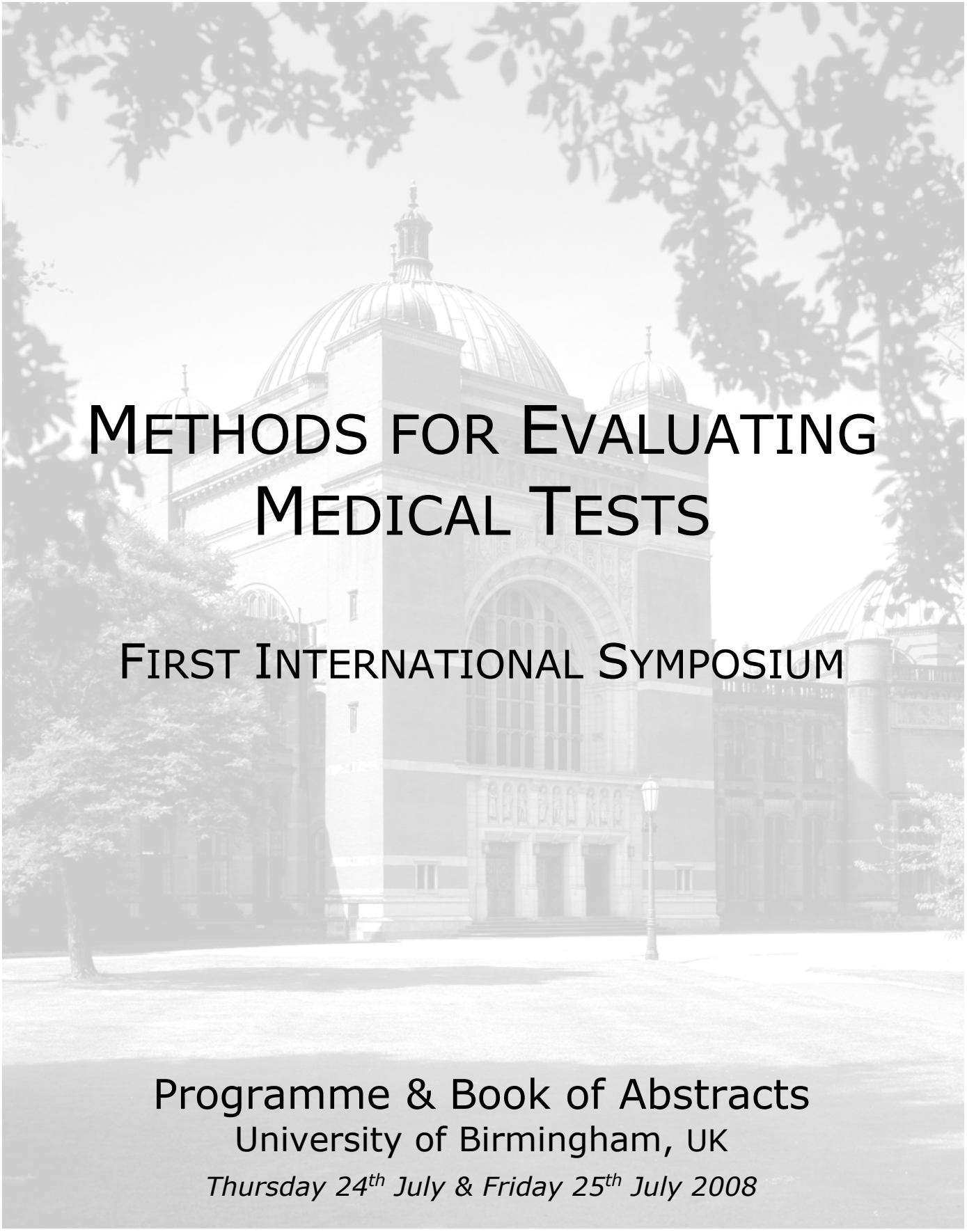
UNIVERSITY OF BIRMINGHAM



METHODS FOR EVALUATING MEDICAL TESTS

FIRST INTERNATIONAL SYMPOSIUM

Programme & Book of Abstracts

University of Birmingham, UK

Thursday 24th July & Friday 25th July 2008

Welcome

The design, execution, analysis, reporting and implementation of evaluations of medical tests present unique methodological challenges, which are currently the subject of research and development.

This multidisciplinary symposium provides a forum for disseminating recent research and stimulating dialogue amongst researchers and healthcare professionals actively involved in evaluating medical tests. Hosted by the Diagnostic Research Group in the Department of Public Health, Epidemiology and Biostatistics, the 2008 event promotes the importance of research into all aspects of medical diagnostics. The two-day conference is organised jointly with the Centre for Evidence Based Medicine, Oxford, and presents an opportunity to debate practice, methodological issues and current/recent research in the field of medical tests.

Themes for this year are:

1. Methodological issues in studies of test accuracy
2. Systematic reviews and meta-analyses of diagnostic tests
3. Monitoring, prognosis and other purposes of tests
4. Evaluating impacts of tests on patients and resources
5. Applying evaluations in practice

We thank you for coming and hope you enjoy the conference.

Jon Deeks
Scientific committee (Chair)

Local Planning Group

Department of Public Health, Epidemiology and Biostatistics, University of Birmingham:

Anne Walker

Lavinia Ferrante di Ruffano

Tess Moore

Chris Hyde

Jon Deeks (Chair)

Scientific Planning Group

Jon Deeks	Department of Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham.
Chris Hyde	Department of Public Health, Epidemiology and Biostatistics, University of Birmingham. Birmingham.
Paul Glasziou	Professor of Evidence-based Medicine, University of Oxford. Oxford.

How to cite this publication

The Abstract book should be cited as:
Methods for Evaluating Medical Tests. Symposium; 2008 Jul 24-25; Department of Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, UK.

Abstracts from this symposium may be cited as:
Author(s). Title [Abstract]. In: Methods for Evaluating Medical Tests. Symposium; 2008 Jul 24-25; Department of Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, UK. Page number(s).

For example

Jon Deeks, Paul Glasziou, Les Irwig. When should a new test replace the gold standard? [Abstract]. In: Methods for Evaluating Medical Tests. Symposium; 2008 Jul 24-25; Department of Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, UK.14.

Abstracts are available at this website

www.medical-test-res.bham.ac.uk/symposium2008

Table of Contents

Programme Overview

Wednesday 23rd July

18:00 – 20:00	Welcome drinks	Barber Institute of Fine Arts

Thursday 24th July

09.30	Registration	Foyer Arts building
10.30	**Session 1** Primary studies of test accuracy	Arts Building Lecture Theatre
12.15	*Lunch*	Public Area 1st floor
13.15	**Session 2** Meta-analysis	Arts Building Lecture Theatre
14.45	**Poster viewing**	Arts Building Lecture room 2 or 3
15.40	**Session 3** Systematic reviews	Arts Building Lecture Theatre
17.00	*Break*	Public Area 1st floor
17.10	**Session 4** Statistical analysis showcase: Meta-analysis using Stata, R and SAS	Arts Building Lecture Theatre
18.10	Close	
19:30 – 23:00	**Conference Dinner**	Birmingham Botanical Gardens

Programme overview

Friday 25th July

09.00	**Session 5** Other applications of tests	Arts Building Lecture Theatre
10.40	*Break*	Public Area 1st floor
11.10	**Session 6** Evaluating patient outcomes	Arts Building Lecture Theatre
12.20	*Lunch*	
13.20	**Session 7** Interpreting and applying findings	Arts Building Lecture Theatre
14.50	*Break*	
15.15	**Session 8** Soap Box	Arts Building Lecture Theatre
16.15	Close	

Programme overview

Full Programme

Wednesday 23rd July

18:00 – 20:00 Welcome drinks Barber Institute of Fine Arts

Thursday 24th July

Registration will take place in the Foyer of the Arts Building

All plenary sessions will be held in the Arts Building Lecture Theatre – 1st floor

Poster session will be held in Lecture Room 2 or 3 – 1st floor

Lunch and coffee will be provided in the public area outside the lecture theatre on the first floor.

Opening

09.30 Registration
Foyer of the Arts Building

10.30 Introduction and welcome
Doug Altman (Keynote Speaker)

Session 1
Primary studies of test accuracy
Arts Building Lecture Theatre 1

10.45 - 11.15	When should a new test replace the gold standard? *Jon Deeks (Keynote Speaker)*	*Page 10*
11.15 - 11.35	Bias in accuracy measures caused by data driven selection of optimal cut-off values for continuous test results: mechanisms, magnitude and solutions. *Mariska Leeflang*	*Page 11*
11.35 - 11.55	How much does prior information sway the diagnostic process? *Robert Newcombe*	*Page 12*
11.55 - 12.15	Diagnostic accuracy when the reference standard is not binary. *Shang-Ying Shiu*	*Page 13*
12.15	**Lunch** ***Public area 1st floor***	

Thursday 24th July continued…

Session 2
Meta-analysis
Arts Building Lecture Theatre 1

13.15 - 13.45	Meta-analysis of diagnostic accuracy: Continuing challenges and future directions *Constantine Gatsonis (Keynote Speaker)*	*Page 14*
13.45 - 14.05	An empirical comparison of three meta-analytic models for the analysis of diagnostic test accuracy studies. *Jac Dinnes*	*Page 15*
14.05 - 14.25	Multivariate meta-analysis of diagnostic test accuracy when the reference standard has four categories. *Mariska Leeflang*	*Page 16*
14.25 - 14.45	Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Richard Riley*	*Page 17*

Poster session
Public area 1st floor

14.45 – 15.40	**Poster viewing**	*Page 37*

Session 3
Systematic reviews
Arts Building Lecture Theatre 1

15.40 – 16.00	How well do published search filters perform in finding diagnostic test accuracy studies? *Julie Glanville*	*Page 18*
16.00 – 16.20	Reasons why sensitivity and specificity do vary with disease prevalence *Mariska Leeflang*	*Page 19*
16.20 – 16.40	Can diagnostic filters offer similar sensitivity and a reduced number needed to read compared to searches based on index test and target condition? *Penny Whiting*	*Page 20*
16.40 – 17.00	Publication bias in studies of diagnostic accuracy in the stroke literature. What is the evidence? *Miriam Brazzelli*	*Page 21*
17.00	**Break** ***Public area 1st floor***	

Full Programme

Thursday 24th July continued…

Session 4
Statistical analysis showcase: Meta-analysis using Stata, R and SAS
Arts Building Lecture Theatre 1

17.10-17.30	Meta-analysis of diagnostic accuracy studies using R *Francesca Chappell*	*Page 22*
17.30–17.50	metandi: Stata software for statistically rigorous meta-analysis of diagnostic accuracy studies *Roger Harbord*	*Page 23*
17.50–18.10	Metadas: A SAS macro for meta-analysis of diagnostic accuracy studies *Yemisi Takwoingi*	*Page 24*
18.10	**Close**	

Full Programme

Friday 25th July

Session 5
Other applications of tests
Arts Building Lecture Theatre 1

9.00 - 9.30	Monitoring in chronic disease *Paul Glasziou (Keynote Speaker)*	*Page 25*
9.30 – 10.00	Lessons from International Surveys on Interpreting Monitoring Tests in General Practice *Andrea R. Horvath (Keynote Speaker)*	*Page 26*
10.00 – 10.20	Methods for Assessing New Biomarkers in Clinical Practice. *Kevin McGeechan*	*Page 27*
10.20-10.40	Outcome and prognostic determinants in patients with traumatic knee injuries in General Practice. *Harry Wagemakers*	*Page 28*
10.40	**Break** ***Public area 1st floor***	

Session 6
Evaluating patient outcomes
Arts Building Lecture Theatre 1

11.10 - 11.40	The role of randomised controlled trials, accuracy studies and other types of comparative evidence for test evaluation *Sally Lord (Keynote Speaker)*	*Page 29*
11.40 - 12.00	Indirect evidence on impact on patients' outcomes. *Jeannine Gailly*	*Page 30*
12.00 – 12.20	A review of the use of randomized trials to assess the impact of diagnostic tests on patient outcomes. *Lavinia Ferrante di Ruffano*	*Page 31*
12.20	**Lunch** ***Public area 1st floor***	

Full Programme

Friday 25th July continued…

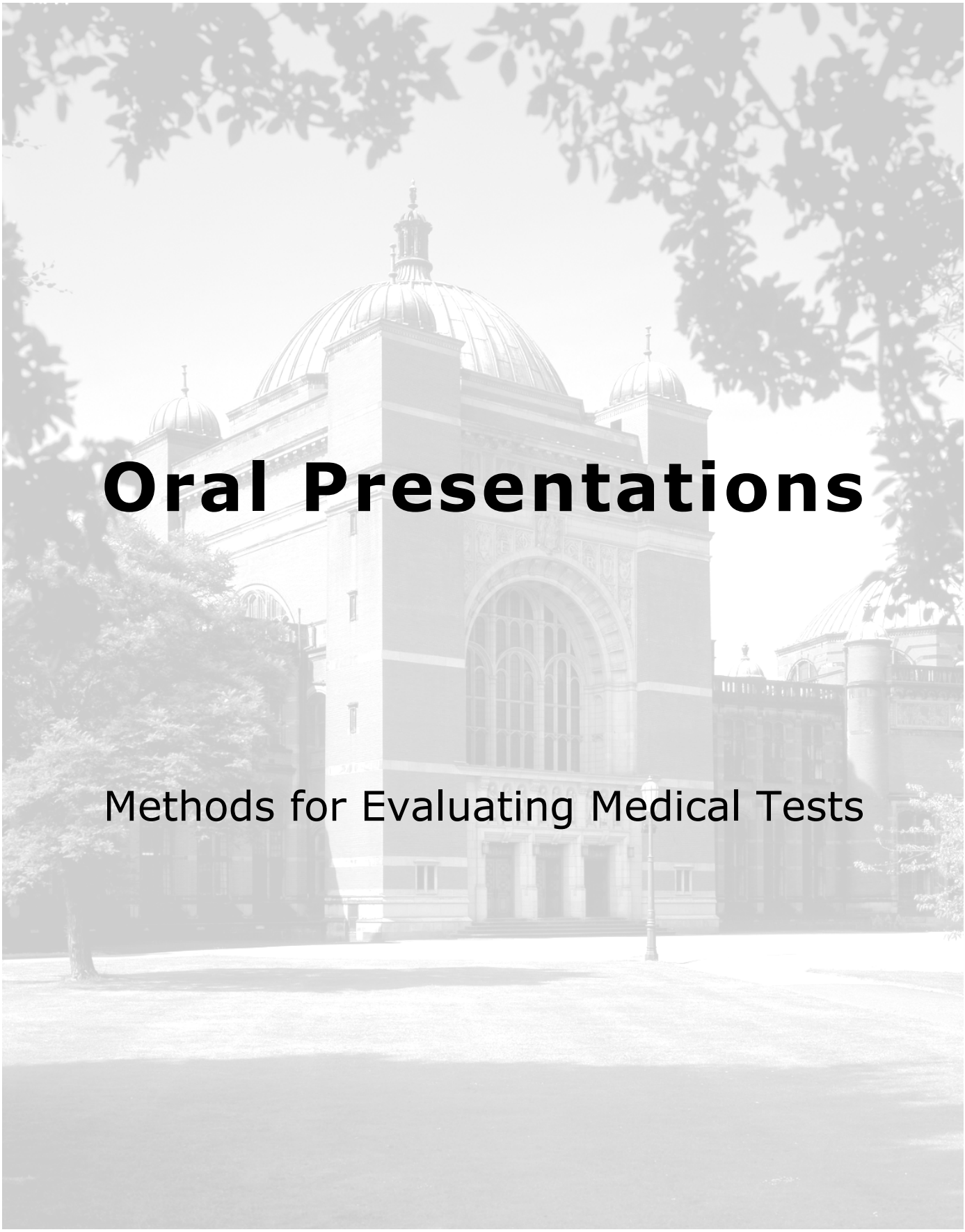
Session 7
Interpreting and applying findings
Arts Building Lecture Theatre 1

13.20 - 13.50	Applying Diagnostic Evidence to Individual Patients *Nick Summerton (Keynote Speaker)*	*Page 32*
13.50 – 14.15	Diagnostic tests for screening: The clinical relevance of positive findings. *Robert Grosselfinger*	*Page 33*
14.15 – 14.35	Nonparametric monotonic regression can illustrate how the likelihood ratio varies with a continuous test result without specifying a test threshold. *Roger Harbord*	*Page 34*
14.35 – 14.55	Grading quality of evidence and strength of recommendations for diagnostic tests and strategies and developing summary of findings tables for diagnostic accuracy studies. *Jan Brozek*	*Page 35*
14.55	**Break** ***Public area 1st floor***	

Session 8
Soap Box – Discussion
Arts Building Lecture Theatre 1

15.15	"Should the government decide to invest a further £20M in test research, how should they best invest it?" *Various Speakers*	*Page 36*

Full Programme



Oral Presentations

Methods for Evaluating Medical Tests

Invited paper

When should a new test replace the gold standard?

Jon Deeks (Presenting), Paul Glasziou, Les Irwig

Professor of Health Statistics, Public Health Epidemiology and Biostatistics, University of Birmingham, UK

New diagnostic tests, in particular new "gold" standard tests, may change whom we classify as having a "disease". For example, PCR tests for infection, new enzyme tests, and new imaging methods such as MRI may identify more abnormality than traditional reference standards and change the spectrum of patients considered diseased. This reclassification usually happens by consensual drift rather than by clear principles. As our diagnostic armamentarium continues to expand and improve, this dilemma will increasingly challenge us. Hence we asked: when is it appropriate to regard the new test as a new reference standard?

Evaluation of new tests appears to be hindered by the lack of a perfect reference standard or 'judge'. However, we may side step this "perfect judge" by instead focusing on the consequences of the decision rather than perfect estimation. Most common is a broadening of the diagnosis by a new test which is apparently more sensitive, but which may also detect earlier or less consequential cases. The assessment of these consequences may be made by an imperfect but fair "umpire" provided that it (a) has some ability to discriminate between disease and non-diseased cases, and (b) is unbiased, that is, its errors must be conditionally independent of the new and old tests.

Using the concepts of consequences and fairness, we set out the following three principles to aid judgments about the new test:

1. The consequences of the new reference test can be understood through the disagreements between the old and new reference tests.

2. Resolving the disagreements between old and new test requires a "fair", but not necessarily perfect, umpire

3. The possible umpires include causal exposures, concurrent testing, prognosis, or the response to treatment.

These principles will be illustrated by considering the comparison of new TIGRA tests for latent tuberculosis infection with the current gold standard tuberculin skin test.

Contact details: j.deeks@bham.ac.uk

Notes

Contributed paper

Bias in accuracy measures caused by data driven selection of optimal cut-off values for continuous test results: mechanisms, magnitude and solutions

Mariska Leeflang, Johannes Reitsma, Karel Moons, Aeilko Zwinderman

Background: Data-driven selection of optimal cut-off values for continuous test results may lead to associated measures of diagnostic accuracy that are overoptimistic.

Aim of study: To determine the magnitude of the bias in diagnostic accuracy measures associated with data-driven selection of optimal cut-off values under a range of conditions using simulated datasets. In addition, we examined potential solutions to reduce this bias.

Methods: Under different scenarios, we compared data-driven estimates of accuracy measures with the true value to determine the magnitude of the bias. We examined the impact of factors like sample size and underlying distribution of test results. Continuous test results for diseased and non-diseased individuals were generated and each simulation war repeated 2000 times in order to determine the median magnitude of bias. Three alternative methods were examined whether they can reduce the magnitude of the bias: leave-one-out procedure, sample characteristics and assuming a specific distribution, and robust fitting of ROC-curves.

Results: The magnitude of bias caused by data-driven optimization of cut-off values was directly and inversely related to sample size. The absolute magnitude of the bias in a study with a total sample of 40 was 5.9% for both sensitivity and specificity. Lowering the prevalence leads to more bias in sensitivity. The absolute value of the true sensitivity and specificity had little impact on the magnitude of the bias, unless they approached 100%. The underlying distribution (normal, log normal, gamma) had no effect on the amount of bias. All three alternative methods led to decrease of bias in Normally Distributed test results. However, when a Normal Distribution was assumed when the true distribution was not, sample characteristics and assuming a specific distribution led to more bias.

Discussion: Data driven choice of optimal cut-off values can lead to overoptimistic estimates of diagnostic accuracy. When underlying assumptions are met, alternative methods may result in more robust estimates.

Contact details: m.m.leeflang@amc.uva.nl

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam, The Netherlands

Notes

Contributed paper

How much does prior information sway the diagnostic process?

Robert Newcombe, Alison Stroud, Mark Wiles

Clinical diagnosis based on the result of a test can utilise prior information on the patient as well as the test result. The present study was prompted by the observation that two pathologists differed diametrically in their views of the influence of prior information. We asked 12 speech and language therapists to assess whether 20 sound recordings of individual swallows were indicative of aspiration, on two occasions. In one round the patient's correct clinical scenario was presented, in the other round an alternative scenario representing a very different degree of prior risk was given. The assessors graded each swallow as normal, abnormal but not aspiration, or aspiration. Overall, moving from the low risk to the high risk scenario resulted in a more severe grade being given in 72 (30%) of instances, unchanged in 150 (63%), and a less severe grade in 17 (7%), a clear preponderance of changes in the predicted direction. The degree to which prior information swayed the diagnosis varied highly significantly between swallows and, more importantly, between the 12 assessors. Such divergence between assessors in the degree to which they are influenced by prior patient information may be an important contributor to the observer variation that is known to occur in routine clinical diagnostic practice.

Contact details: newcombe@cf.ac.uk

Department of Primary Care & Public Health, Cardiff University, UK

Notes

Contributed paper

Diagnostic accuracy when the reference standard is not binary

Shang-Ying Shiu, Constantine Gatsonis

Statistical methods for the evaluation of the accuracy of diagnostic tests typically assume a binary true disease status. However, a binary disease status may often be obtained only after dichotomizing a reference standard which is measured on a continuous or ordinal categorical scale. Such situations can arise in individual studies of the accuracy of diagnostic modalities as well as in systematic reviews of studies. A summary measure of accuracy when the reference standard is continuous was proposed by Obuchowski (2006). In this work, we propose an extension of the common framework for ROC analysis, which allows for a range of threshold values for defining a binary reference standard. We consider the analysis of studies in which both the diagnostic test and the reference standard are reported as continuous measures. Measures of accuracy can then be evaluated for threshold on the reference standard and can be averaged over a range of values for this threshold. We study the geometric properties of the sensitivity, specificity and the ROC curve under the extended framework, and examine the effect of varying reference standard threshold on these quantities. We propose a semi-parametric model for estimating the sensitivity, specificity and the ROC curve in this setting. Under suitable order restrictions on the mean of the test result variable, fitting is done via two alternative approaches: isotonic regression and the monotone smooth splines. This approach applies, with simple modifications, also to the analysis of studies in which the reference standard is defined on an ordinal categorical scale. We apply our method to the analysis of the accuracy of PET in the detection of axillary node involvement in women diagnosed with breast cancer.

Reference

Obuchowski, N. A. (2006). An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. *Statistics in Medicine* 25, 481–493.

Contact Details: shiu@stat.sinica.edu.tw
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Notes

Invited paper

Meta-analysis of diagnostic accuracy: Continuing challenges and future directions

Constantine Gatsonis

Director of the Center for Statistical Sciences, Brown University, USA

Systematic reviews of the accuracy of diagnostic and screening tests have come of age: this much is clear. The Cochrane Library will feature the first such reviews in 2008 and a Handbook will also be released in the same time period. The paradigm is best developed in the setting of studies reporting estimates of sensitivity and specificity, for which hierarchical and mixed models methods are now available. However, significant methodologic challenges persist. In this presentation, we will review the state of the art in statistical methods for meta-analysis of studies reporting pairs of sensitivity and specificity. We will note open methodologic questions, such as how to account for verification bias and errors in the reference standard. We will also survey the methodologic challenges awaiting us in the broader landscape of studies and measures of accuracy, beyond the meta-analysis of sensitivity/specificity data.

Contact details: gatsonis@stat.brown.edu

Notes

Oral Presentations

Contributed paper

An empirical comparison of three meta-analytic models for the analysis of diagnostic test accuracy studies

Jacqueline Dinnes, Paul Roderick, Susan Mallett, Sally Hopewell, Jon Deeks

Aim: To compare meta-analytic methods for the analysis of diagnostic test accuracy studies, their ability to detect spectrum effects and the impact of including interactions with shape in heterogeneity investigations.

Methods: 29 published systematic reviews (reporting 60 investigations of heterogeneity) were identified from the DARE database that presented 2x2 data plus at least one spectrum-related covariate per study. Re-analyses were undertaken using the Moses model (both unweighted and inverse variance weighted) and the hierarchical SROC model. Covariates for differences in accuracy and shape were added to each model for investigations of heterogeneity, with an additional covariate for differences in threshold being added to the HSROC model.

Results: Substantial disagreements were noted between models, both for estimates of diagnostic accuracy and spectrum variables. For the primary analyses, the weighted Moses model on average underestimated the results of the unweighted model by around 30%, whereas the unweighted Moses model and the HSROC showed disagreements in both directions. For the heterogeneity investigations, the Moses models underestimated the size of differences in accuracy between groups compared to the HSROC model; the discrepancies were less with parallel SROC curves. The Moses models found strong evidence of differences in accuracy and shape where the HSROC model did not and vice versa. The within model comparisons showed that including interactions covariates with shape (regardless of their significance) almost always affects a review's conclusions regarding the size and sometimes strength of any differences in accuracy by that covariate. The effect was less for the HSROC model.

Conclusion: The Moses models cannot be relied upon to approximate the results of the 'optimal' HSROC model. The circumstances that lead to bias in estimates SE(lnDOR) in the weighted Moses model are common. The question remains whether an interaction of covariate with curve shape should be routinely modelled.

Table: Comparison of DORs at Q* and at average threshold

	Ratio of DORs at Q*			Ratio of DORs at average threshold		
ROR	Moses (w) vs Moses (eq)	Moses (eq) vs HSROC	Moses (w) vs HSROC	Moses (w) vs Moses (eq)	Moses (eq) vs HSROC	Moses (w) vs HSROC
Maximum	7.78	2.77	1.21	1.27	4.81	5.51
75th pctile	0.87	0.99	0.72	0.91	1.05	0.75
Median	**0.67**	**0.78**	**0.51**	**0.71**	**0.94**	**0.55**
25th pctile	0.50	0.51	0.24	0.54	0.68	0.46
Minimum	0.10	0.07	0.10	0.36	0.05	0.03

Contact details: jac.dinnes@gmail.com

University of Southampton, UK

Notes

Oral Presentations

Contributed paper

Multivariate meta-analysis of diagnostic test accuracy when the reference standard has four categories

Mariska Leeflang, Johannes Reitsma, Aeilko Zwinderman

Background: Current meta-analytic techniques for diagnostic test accuracy use pairs of sensitivity and specificity as the underlying parameters in their models. When the true disease status is categorized into more than two categories, dichotomization may lead to loss of information and will directly affect the resulting two-by-two table.

Methods: We used data from a systematic review about the diagnostic accuracy of galactomannan testing for the diagnosis of invasive aspergillosis (IA). The reference standard defined patients as proven, probable, possible or no IA patients. The data were first analyzed via a bivariate meta-analysis in which we combined the proven and probable IA categories and the possible and no categories, respectively, and subsequently via a multivariate meta-analysis that estimates the proportions of test positives and test negatives in each one of four reference categories within a single model.

Results: Twenty-eight studies, containing 4501 participants were included. The bivariate meta-analysis resulted in a mean sensitivity of 0.70 (95% CI 0.59 to 0.79) and a mean specificity of 0.91 (95% CI 0.87 to 0.94. The results of the multivariate analysis showed a proportion of test positives in the proven IA patients of 0.74 (95% CI 0.53 to 0.87) and in the probable IA patients of 0.63 (0.50 to 0.74). The proportion of test negatives in the possible IA patients was 0.55 (95% CI 0.40 to 0.69) and for the patients without IA 0.92 (95% 0.80 to 0.97).

Conclusions: The multivariate model is a useful extension of the bivariate model to meta-analyze diagnostic accuracy data where the reference standard has more than two categories. The multivariate method uses all available data, which leads to additional insight into the performance of a test. Furthermore, it avoids the decision which reference standard categories have to be combined, a decision which is frequently made in an arbitrary way and a potential source of variation in results between studies applying different combinations.

Contact details: m.m.leeflang@amc.uva.nl

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam, The Netherlands

Notes

Oral Presentations

Contributed paper

Meta-analysis of diagnostic test studies using individual patient data and aggregate data

Richard Riley, Susanna Dodd, Jean Craig, Paula Williamson

Background

A meta-analysis of diagnostic test studies provides evidence-based results regarding the accuracy of a particular test, and usually involves synthesising aggregate data (AD) from each study, such as the two by two tables of diagnostic accuracy. A bivariate random-effects meta-analysis (BRMA) can appropriately synthesise these tables [1], and leads to clinical results such as the mean sensitivity and mean specificity across studies. However, translating such results into practice may be limited by between-study heterogeneity and that they relate to some 'average' patient across studies.

Methods

This talk will describe how the meta-analysis of individual patient data (IPD) from diagnostic studies can lead to more clinically meaningful results tailored to the individual patient. IPD models will be introduced that extend the BRMA framework to include study-level covariates, which help explain the between-study heterogeneity, and also patient-level covariates, which allow the interaction between test accuracy and patient characteristics to be assessed. It will be shown that the inclusion of patient-level covariates requires careful separation of within-study and across-study accuracy-covariate interactions, as the latter are particularly prone to confounding. The models will be assessed through simulation, and are extended to allow IPD studies to be combined with AD studies, as IPD are not always available for all studies.

Application

Application is shown to 23 studies assessing the accuracy of ear temperature for diagnosing fever in children, with 16 IPD studies and 7 AD studies. The models reveal that between-study heterogeneity is partly explained by the use of different measurement devices, and importantly there is no evidence that individual age modifies diagnostic accuracy.

Conclusion

Meta-analysis of IPD from diagnostic test studies can be performed in a bivariate meta-analysis framework. It also allows one to assess how patient-level covariates modify diagnostic accuracy, and it thus can produce more clinically meaningful results than the traditional AD approach.

[1] Chu H, Cole SR: Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol 2006, 59:1331-1332.

Contact details: richard.riley@liv.ac.uk

Centre for Medical Statistics and Health Evaluation, University of Liverpool, UK

Notes

Contributed paper

How well do published search filters perform in finding diagnostic test accuracy studies?

Julie Glanville, Gill Ritchie, Carol Lefebvre

Objectives

Systematic reviews of diagnostic test accuracy studies need to be able to identify relevant research efficiently. A large number of search filters to identify diagnostic test studies in the major databases have already been published. This research was undertaken to assess the performance of search filters in finding diagnostic test accuracy studies for systematic reviews, where sensitivity is important.

Methods

Diagnostic test accuracy search filters were identified by searching MEDLINE, our own files and by contacting colleagues. We applied the filters to a case study review of diagnostic test accuracy studies for urinary tract infections (UTI) in young children. The review was informed by a wide ranging and sensitive search. We used a relative recall approach to create a gold standard. The studies included in the review with records in MEDLINE formed the gold standard. The performance of the filters in finding those gold standard records was assessed.

Results

We identified twenty-three diagnostic test accuracy search filters for use with MEDLINE. The case study systematic review of UTI included 179 studies of diagnostic test accuracy, of which 138 were available in MEDLINE. The filters showed a wide range of sensitivities (range: 20.6% to 86.9%). Precision was inconsistent and did not compensate for the poor sensitivity (range: 1% to 9.4 %.).

Conclusions

Our results broadly support those reported in two other studies. The search filters tested do not offer an adequate trade-off between sensitivity and precision to be used to identify studies for systematic reviews. However, there are methods available to explore whether better performing search filters are viable based on an objective statistical analysis of the text and indexing used in records.

Contact details: jmg1@york.ac.uk

York Health Economics Consortium Ltd, University of York, UK

Notes

Contributed paper

Reasons why sensitivity and specificity do vary with disease prevalence

Mariska Leeflang, Patrick Bossuyt, Les Irwig

Background: The sensitivity and specificity of a diagnostic test are often assumed to be independent of disease prevalence. Yet several studies and systematic reviews have reported differences in sensitivity and specificity related to prevalence. We explored the mechanisms that may be responsible for diagnostic accuracy varying with prevalence.

Methods: Conceptual exploration of real and artefactual reasons why diagnostic accuracy may vary with disease prevalence, illustrated by examples from the literature.

Results: Factors responsible for differences in prevalence between studies or study subgroups can also be responsible for differences in sensitivity and specificity.

Real variability is usually associated with spectrum effects:

- Patient spectrum itself: higher prevalence often results in a more severely diseased population in which the test performs better;
- Referral filter: refers patients with certain characteristics to the study, these characteristics influence both prevalence and diagnostic accuracy;
- Reader expectation: diagnostic accuracy of readers can be influenced by the (supposed) prevalence in the study group.

Artefactual differences can result from study design related effects:

- Additional exclusion criteria: when patients that are more difficult to classify are excluded, prevalence may change and the test will seem to have a higher diagnostic accuracy;
- Verification bias: partial or differential verification leads to differences in prevalence as well as differences in diagnostic accuracy;
- reference standard misclassification: when the reference standard is no 'gold standard', sensitivity will be less and specificity will be more underestimated as prevalence increases.

Conclusion: Sensitivity and specificity may vary with prevalence via several mechanisms. Differences in prevalence between studies are a first indicator for differences in study population or flaws in study design. We encourage authors of systematic reviews to explore associations between prevalence and diagnostic accuracy. We hope that more future systematic reviews will analyze and report such associations, and provide helpful explanations for these patterns, if they find them.

Contact details: m.m.leeflang@amc.uva.nl

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam, The Netherlands

Notes

Contributed paper

Can diagnostic filters offer similar sensitivity and a reduced number needed to read compared to searches based on index test and target condition?

Penny Whiting, Marie Westwood, Margaret Burke, Jonathan Sterne, Roger Harbord, Julie Glanville

Background

Literature searches involve searching electronic databases, which requires a sensitive search strategy to capture as many relevant records as possible. Filters to identify test accuracy studies can lead to the omission of a considerable number of relevant studies. We have previously shown that even searches designed to be very sensitive based on index test and target condition miss relevant studies indexed on Medline.

Objectives

To compare the performance of search strategies that do not incorporate a diagnostic filter (based on the index test and target condition) with searches that add a filter to these strategies.

Methods

We included seven reviews that had each carried out extensive sensitive searches of multiple databases. We identified studies included in these reviews that were indexed on MEDLINE and these were used as our "reference" set (523 studies). We ran searches of MEDLINE for each review based on the index test + target condition alone. We then combined these searches with each of 22 published diagnostic filters that we had translated to run on Ovid MEDLINE. We evaluated how many of the "reference" studies for each review were identified by each of these 23 searches. We also looked at the number of records produced by each of the searches and their ability to reduce the number needed to read (NNR).

Results

Searches designed to be very sensitive based on index test and target conditions miss an average of around 9% of studies indexed on MEDLINE (range 0-13% across reviews). Searches that included a filter missed between 14 and 58% of studies indexed on Medline (ranged 3-88%). The NNR was reduced from 55 for index test + target condition searches to between 7 and 51 (median 27) for searches that incorporated a filter. None of the filters offer acceptable sensitivity for a reasonable decrease in the number needed to read. For moderate sensitivity (>80%) the NNR was not substantially reduced (range 29-51).

Conclusions

Currently available diagnostic filters should not be used to identify studies for inclusion in test accuracy reviews because they are unable to offer both acceptable sensitivity and a reduced NNR.

Contact details: penny.whiting@bristol.ac.uk
Department of Social Medicine, University of Bristol, UK

Notes

Contributed paper

Publication bias in studies of diagnostic accuracy in the stroke literature. What is the evidence?

Miriam Brazzelli, Peter Sandercock, Steff Lewis, Jonathan Deeks

Background

Whilst there is substantial literature on publication bias in systematic reviews of randomised controlled trials[1] there is little evidence on publication bias in systematic reviews of diagnostic studies.

We evaluated the proportion of diagnostic studies presented at international stroke meetings that were later published in full and assessed which study characteristics influenced full publication.

Methods

We reviewed all diagnostic abstracts presented at the International Stroke Conference and the European Stroke Conference between 1995 and 2004 and subsequently published in special issues of Stroke and Cerebrovascular Diseases. Abstracts were selected if they reported findings of diagnostic studies of accuracy. Full-text publications of diagnostic abstracts were identified through MEDLINE and EMBASE searches. We assessed the features and findings of all abstracts. Determinants of publication were assessed by a series of univariate Cox regression analyses.

Results

Of the 160 identified abstracts, 121 (76%) were subsequently published in full. 62% of them were published in full within 24 months of presentation. Median time to publication was 16 months. Only inter-observer agreement between test readers predicted full publication (p = 0.02). The clinical utility of results did not affect publication, neither did the type of study design, the type of test, the country of origin of the corresponding author, the multi-centre status, or the Youden's Index.

Conclusions

We did not find clear evidence of bias in the process of publication that occurs after abstract acceptance. We were unable to assess bias in abstract submission or acceptance. Amongst 121 published diagnostic studies on stroke 'inter-observer agreement' was the only factor statistically associated with full publication. Clinical utility of results and other study characteristics did not seem to predict publication of diagnostic research submitted to international stroke meetings. Overall diagnostic abstracts fail to report many relevant methodological aspects.

References

1. Scher RW, Dickersin K, Langenberg P. *Full publication of results initially presented in abstracts: a meta-analysis.* JAMA 1994; 272: 158-162

Acknowledgements: This project is funded by the Scottish Executive Health Department Chief Scientist Office http://www.show.scot.nhs.uk/CSO

Contact details: m.brazzelli@ed.ac.uk

Division of Clinical Neurosciences, University of Edinburgh, UK

Notes

Contributed paper

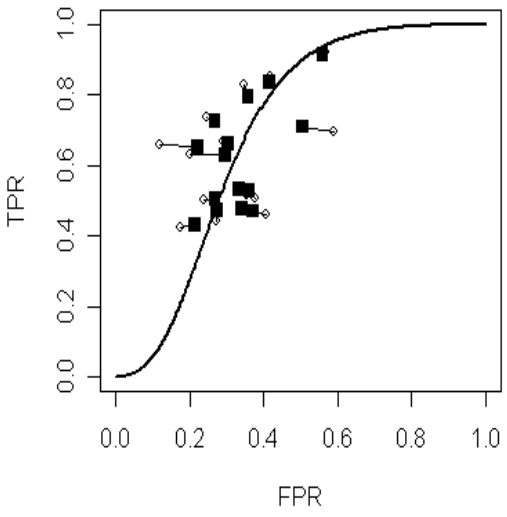
Meta-analysis of diagnostic accuracy studies using R

Francesca Chappell, Gillian Raab, Joanna Wardlaw

Introduction

Recent research suggests that meta-analysis of diagnostic accuracy studies requires a nonlinear bivariate random effects model, which produces estimates of the mean sensitivity and specificity and a summary ROC curve (1). Nonlinear models are complex, and fitting them can lead to computational problems. Computer programs using SAS PROC NLMIXED and WinBUGS are available in the literature. However, systematic reviewers may not be able to use these programs, either because of the expense of the software (in the case of SAS) or because of a lack of programming skill. An alternative to SAS or WinBUGS is R (http://cran.r-project.org/), which has the advantages of being free and extensive graphics capabilities. Although R is also a complex package it can be programmed so that the end-user needs very little programming skill.

Development of R functions Meta-analysis of diagnostic accuracy studies should include assessments of whether the model is appropriate for the data. The R functions plotfor and bivarROC produce forest and ROC plots (see figure) to help assess heterogeneity, correlation, and suitability of data for meta-analysis. The bivarROC function generates both maximum likelihood and Bayesian estimates with intervals. Further graphical output includes posterior densities from the Bayesian analysis. Non-graphical output includes assessment of the adequacy of fit and appropriateness of the bivariate model. Another function, two_uni, conducts separate meta-analyses for sensitivity and 1-specificity. We also present an algorithm to guide the meta-analyst through appropriate statistical procedures for meta-analysing data from diagnostic accuracy studies.



Testing of R functions Analyses of data from real systematic reviews are presented to demonstrate the use of the R functions and the algorithm, highlighting different key issues regarding meta-analysis of diagnostic accuracy studies.

Conclusion These R functions are designed to be easier to use than the SAS or WinBUGS programs and have the advantage of being free. They also provide very useful plots to aid the meta-analyst in the assessment and interpretation of the results. They are freely available at *http://www2.napier.ac.uk/depts/fhls/DiagMeta/* along with instructions for their use.

Reference

1. Harbord RM et al. A unification of models for metaanalysis of diagnostic accuracy studies. Biostatistics 2007;8(2):239-251.

Contact details: francesca.chappell@ed.ac.uk
Division of Clinical Neurosciences, University of Edinburgh, UK

Notes

Contributed paper

metandi: Stata software for statistically rigorous meta-analysis of diagnostic accuracy studies

Roger M Harbord, Penny Whiting, Jonathan AC Sterne

Background

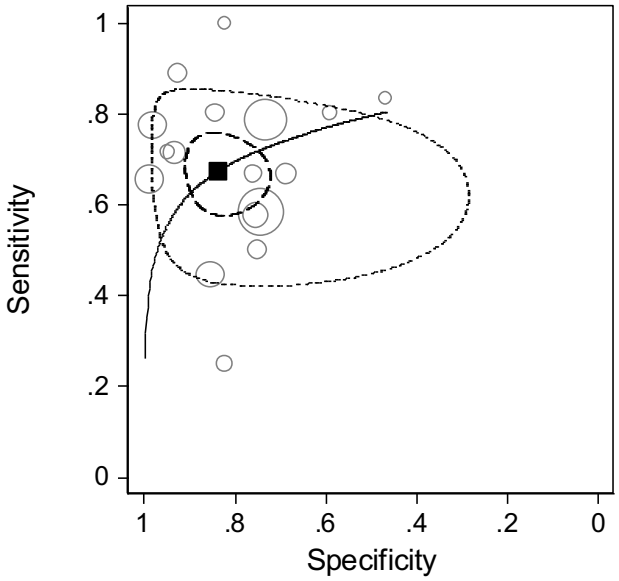
Many approaches to the meta-analysis of diagnostic accuracy are currently in use. Consensus is building that statistically rigorous methods involving hierarchical models are necessary to ensure valid results. Such methods correctly handle the correlation between sensitivity and specificity and the binomial distribution of the data within each study. There are two such models, the hierarchical summary ROC (HSROC) model and the bivariate random-effects model, which have been shown to be equivalent when no covariates are fitted, as well as in certain cases with covariates.

Aim

To develop a user-written module metandi (Harbord 2008) for the statistical software package Stata that performs meta-analysis of diagnostic accuracy studies without covariates and displays the results in both bivariate and HSROC parameterisations, as well as on a graph.

Results

The user-written command gllamm and the (faster) official command xtmelogit introduced in Stata 10 can both be used to fit the bivariate model: the corresponding HSROC parameter estimates can also be produced after some extra work. The metandi module provides a straightforward interface in which a single command fits the model and (optionally) graphs the results, e.g.:




```
. metandi tp fp fn tn, plot
```

Conclusion

Increasing accessibility of statistically rigorous methods will increase their use and facilitate appropriate analyses.

Reference: Harbord, RM. (2008). "METANDI: Stata module to perform meta-analysis of diagnostic accuracy."

Statistical Software Components S456932, Boston College Department of Economics. http://ideas.repec.org/c/boc/bocode/s456932.html.

This module may be installed from within Stata by typing "`ssc install metandi`".

Contact details: roger.harbord@bristol.ac.uk
Department of Social Medicine, University of Bristol, UK

Notes

Oral Presentations

Contributed paper

Metadas: A SAS macro for meta-analysis of diagnostic accuracy studies

Yemisi Takwoingi, Boliang Guo, Jonathan Deeks

Introduction:
Hierarchical or multilevel methods are advocated for the meta-analysis of diagnostic accuracy studies (Gatsonis and Paliwal, 2006). The Cochrane Collaboration has begun registering diagnostic test accuracy reviews but RevMan 5, the review authoring tool, only enables summary ROC regression and external analyses is required. We present metadas, a SAS macro, developed as a wrapper for Proc NLMIXED and compare it with the Stata user-written programs, metandi and midas.

Results:
Metadas reduces the problem of selecting starting values for model parameters in Proc NLMIXED. The macro can run any number of tests consecutively and has several options which include model choice (hierarchical summary receiver operating characteristic or bivariate model), predictions based on the empirical Bayes estimates, covariate inclusion, likelihood ratio tests, and model checking. The output of the analysis is summarised in a Word document with all parameter estimates in a format suitable for input into RevMan 5 in order to produce SROC plots. In addition, estimates of summary measures of test accuracy such as the expected sensitivity, specificity, likelihood ratios and diagnostic odds ratio are produced, as well as relative measures when there is a covariate in the model.

Conclusions:
Metadas is a versatile program that renders meta-analysis of diagnostic accuracy studies in SAS more accessible. It is easy to use and although it has no graphical capability in terms of SROC plots, it provides more flexibility in model fitting and result output than either metandi or midas.

Reference:
Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: Methodologic primer. Am. J. Roentgenol. 2006;187:271-281

Contact details: y.takwoingi@bham.ac.uk

Department of Public Health, Epidemiology and Biostatistics, University of Birmingham, UK

Notes

Invited paper

Monitoring in chronic disease

Paul Glasziou

Professor of Evidence-based Medicine, University of Oxford, UK

Managing long term illness is an important and costly element of health care, and accounts for 80% of GP consultations. Monitoring forms a major part of this management, however it has been neglected as an area for research, despite the substantial costs it entails. For example, despite a lack of evidence of effectiveness of self-monitoring in Type 2 diabetes, the costs of monitoring strips alone in 2002 in the UK was £118 Million per year, which was larger than the expenditure on oral medications for diabetes. Despite the investment in monitoring, in many patients chronic disease is poorly controlled. For example, in a UK study before the new GP contract, only 14% of 21,024 newly diagnosed patients with hypertension had met target blood pressure after 12 months, and among treated patients about 40% of INRs are outside target ranges, compared with the ideal of 5%.

We suggest that optimal monitoring should be considered in five phases:
(1) pre-treatment, (2) initial titration, (3) maintenance, (4) re-establish control, and (5) cessation. Though randomised trials trials of monitoring showed a mixed pattern of effects, but few appear to have optimized monitoring protocols prior to the trial. There appears to be a great potential for both clinical improvement and reduced costs by better monitoring of chronic conditions in primary care.

Contact details: paul.glasziou@dphpc.ox.ac.uk

Centre for Evidence-Based Medicine, University of Oxford, UK.

Notes

Oral Presentations

Invited paper

Lessons from International Surveys on Interpreting Monitoring Tests in General Practice

Andrea R. Horvath

Professor of Clinical Chemistry, University of Szeged, Hungary.

Andrea R. Horvath[1], Sverre Sandberg[2,3] on behalf of the authors of the studies

[1] Department of Clinical Chemistry, University of Szeged, Hungary; [2] Laboratory of Clinical Biochemistry, Haukeland University Hospital and [3] Norwegian Quality Improvement of Primary Care Laboratories (NOKLUS), Department of Public Health and Primary Health Care, University of Bergen, Norway

Introduction: Monitoring is repeated testing aimed at guiding and adjusting the management of conditions. Information on clinically significant changes between successive measurements may be important in assessing progression of disease or effects of therapeutic interventions. Interpretation of serial data in monitoring can be approached by statistical process control and control charts or by the use of the 'reference change value' (RCV) or critical difference (CD). CD is defined as the minimum difference needed between two consecutive test results to be certain (with a given degree of confidence) that the two results are truly different and not simply a result of analytical and intra-individual biological variation. Despite the availability of these interpretative approaches and that testing for monitoring purposes accounts for the majority of the workload of laboratories, many patients are poorly controlled and doctors fail to interpret monitoring test results correctly. Therefore we investigated the clinical judgment of practising physicians in different countries when monitoring patients treated for diabetes mellitus or with oral anticoagulants.

Methods: Interpretive skills of general practitioners (GPs) were investigated in 3 separate post-analytical surveys using case history-based questionnaires in 10 European countries and Australia. The first two surveys organized by NOKLUS and the IFCC Global Campaign of Diabetes Mellitus focused on interpretation of glucose, HbA_{1c} and microalbuminuria test results in a type 2 diabetic patient. The third survey, carried out in Norway, tested interpretation of prothrombin time and INR results in a warfarin treated patient. Unanimous, coded replies were registered on a web-based application in each country. All participants received a feedback report after the survey, stating the GP's own answers, the pooled results of other participants in that country, and a discussion on the concept of CD, and the clinical implications and conclusions of the case, together with relevant recommendations of available practice guidelines.

Results: A total of 6390 GPs returned the questionnaires at a response rate of 7-83%. When interpreting the CD between two glucose values 80% of the GPs of all countries responded with changes lower than the 'true' CD value of 14%. For interpretation of HbA_{1c} in monitoring diabetes, a CD of 12% was quoted as reasonable in the feedback to participants. Interpretations at the 95% confidence level were unrealistically strict for up to 50% of the GPs for decreases in HbA_{1c} and for almost all of them for increases in HbA_{1c}. The range of CDs stated for increasing or decreasing the warfarin dose from an INR result of 3.3 was substantial and CDs were significantly smaller when an increase in the INR result was considered. In other words, GPs act at a lower threshold when HbA_{1c} or INR increases than when it decreases, probably because increase in these test results indicate higher risk for diabetes or bleeding complications, respectively. Similarly to the INR survey, estimates of CDs for microalbuminuria had a large range among GPs, i.e. the action limits of GPs were highly variable. However, the pattern of estimating CDs were surprisingly similar in all countries regardless of the differences in social, cultural, and organizational aspects of their healthcare systems.

Conclusions: Our studies indicate that GPs in different countries tend to misjudge the critical difference indicating that a true change has occurred between successive test results. This might lead to inappropriate medical decisions on patient management. Our findings highlight the need for better communication of test results with the users of laboratory services to increase the awareness of clinicians of how analytical and biological variation can influence test interpretation in monitoring. Further research is needed to investigate how better reporting and similar post-analytical surveys with direct feedback to doctors on their interpretive skills could reduce the misinterpretation of laboratory results and thus improve patient safety and outcomes.

References:

1. Skeie S. *et al.* Postanalytical external quality assessment of blood glucose and hemoglobin A1c: an international survey. Clin Chem. 2005;51:1145-1153.
2. Kristoffersen AH, *et al.* Postanalytical external quality assessment of warfarin monitoring in primary healthcare. Clin Chem. 2006;52:1871-1878.
3. Aakre KM *et al.* Postanalytical external quality assessment of urine albumin in primary health care: an international survey. Clin Chem 2008, in press.

Contact details: ahorvath@clab.szote.u-szeged.hu

Contributed paper

Methods for Assessing New Biomarkers in Clinical Practice

Kevin McGeechan, Petra Macaskill, Les Irwig, Gerald Liew, Tien Wong

The estimation of an individual's risk of cardiovascular disease is the foundation of CVD prevention around the world. In the UK the NHS has recently proposed that everyone over 40 undergo a vascular check up and have their risk of cardiovascular disease estimated. The risk of cardiovascular disease will be estimated using measurements of the traditional risk factors (eg age, family history, smoking, blood pressure, cholesterol and glucose levels). However, newer biomarkers (eg C-reactive protein and coronary artery calcium score) are regularly proposed which aim to improve cardiovascular risk prediction. The clinician must decide whether to measure these new risk factors and how this additional information should be utilized.

The potential gain in using a biomarker can be assessed in terms of discrimination, calibration and the number of individuals who are reclassified into a different treatment group. Criticism that evaluation of new biomarkers has relied too heavily on measures of discrimination has led to a greater emphasis on the summary measures associated with calibration and reclassification. However, these also have limitations. For example, the Hosmer-Lemeshow test, a measure of calibration, is overly sensitive when sample sizes are large. Also, the newly proposed Net Reclassification Information is affected by the choice of categories used. Graphical displays have also been proposed that may be more useful to the clinician than these summary measures in deciding whether to measure an additional biomarker.

The strengths and limitations of the existing approaches outlined above will be discussed and illustrated using data from the Atherosclerosis Risk in Communities Study (ARIC) an ongoing community cohort study in the USA. We will then outline an alternative graphical approach that provides greater clarity for the clinician, and patient, to determine at what level of predicted risk additional testing may be worthwhile and what is the likelihood that the patient being tested would have their risk changed by a meaningful amount.

Contact details: kevinm@health.usyd.edu.au

School of Public Health, University of Sydney

Notes

Contributed paper

Outcome and prognostic determinants in patients with traumatic knee injuries in General Practice

Harry Wagemakers, Pim Luijsterburg, Bart Heintjes, Marjolein Berger, Bart Koes, Sita Bierma-Zeinstra

Introduction

A wait and see policy is advocated in patients with traumatic knee injuries. Outcome and prognostic factors in patients with traumatic knee injuries in primary care setting is yet unknown as is management by the GP.

Objective of this study

To gain insight in the outcome and prognostic determinants of traumatic knee injuries after one year in patients consulting the GP.

Methods

This study was part of a large prospective cohort study on knee complaints in general practice[1]. Fourty GP's participating in a research network included patients with new (traumatic) knee complaints. MR imaging was used to determine the nature and severity of lesions. History taking was performed at baseline and after 3, 6 and 12 months of follow-up. Physical examination and MRI were performed at baseline and after 12 months. Prognostic determinants for perceived recovery and presence of a lesion are determined.

Results

Of the 134 included patients 122 reported on their perceived recovery; 101 patients (84%) reported clinically relevant recovery. There is no significant difference in perceived recovery between patients with and without lesions as detected with MRI. The pain severity score decreased the most during the first 3 months after injury. The Lysholm score increased the most during these 3 months. Medical consumption (re-consultation with the GP and referral to physical therapy and secondary care) also shows no difference between patients with and without lesions.

With regard to perceived recovery high workload, effusion, crepitation, pain at passive flexion and the Apley grinding test showed association. In relation the presence/absence of lesions age over 40, rotational trauma, continuation activity impossible, genu flexum and pain palpation MCL showed association.

Conclusions

The vast majority of patients is clinically recovered after a knee injury. Medical consumption during 12 months of follow-up is considerable. History taking and physical examination show some prognostic value where MRI does not show any value regarding prognosis.

Reference:

1.Heintjes EM, Berger MY, Koes BW, Bierma-Zeinstra SM: Knee disorders in primary care: design and patient selection of the HONEUR knee cohort. BMC Musculoskelet Disord 2005; 6: 45.

Contact details: hpa.wagemakers@dordrecht.nl

Erasmus Medical Center, The Netherlands

Notes

Invited paper

The role of randomised controlled trials, accuracy studies and other types of comparative evidence for test evaluation

Sally J Lord

Epidemiologist, NHMRC Clinical Trials Centre, University of Sydney, Australia.

The goal of test evaluation is to provide evidence that the new test improves patient outcomes or produces other benefits without adversely affecting patient outcomes. Tests may improve patient outcomes if they improve the selection of treatment by providing more accurate diagnostic, prognostic or predictive information than existing tests; are safer; or offer other attributes such as improved patient acceptability.

Randomized controlled trials (RCTs) comparing the new test strategy and subsequent treatment with current best practice will provide the best evidence about its impact on patient outcomes. However these RCTs are not always available.

This presentation describes an approach for deciding when evidence of test accuracy and safety can be linked to evidence from existing treatment trials to infer patient outcomes and when new RCTs are required. This approach involves specifying the potential benefits of the new test and whether it will be used as an add-on, triage or replacement to existing tests to identify the critical questions for evaluation.

If the new test is proposed as a more sensitive add-on or replacement test, the critical question is the efficacy of treatment for the new cases detected. New RCTs will be required if these patients represent a different spectrum of disease to those included in existing treatment trials.

If the new test is proposed to provide other benefits with no change in treatment, the type of evidence needed depends on the proposed outcomes. For example, if a new triage test is proposed to be safer by avoiding invasive testing in some patients, comparative evidence of test strategy accuracy and safety may suffice. However, if a new replacement test is proposed to improve patient acceptability, short-term RCTs assessing this outcome will also be required.

The assumptions made when linking evidence of test accuracy, safety and treatment efficacy to infer patient outcomes must be explicitly stated and should be tested in RCTs if uncertainties exist.

Contact details: sally.lord@ctc.usyd.edu.au

Notes

Contributed paper

Indirect evidence on impact on patients' outcomes

Jeannine Gailly, Anne Van den Bruel

Background: A systematic review on fluoresceine angiography (FA) and indocyanine green angiography (ICGA) as diagnostic tests for exudative age related macular degeneration (AMD) identified a lack of direct evidence on patient outcomes. Alternatively, indirect evidence on the impact on patient management was searched.

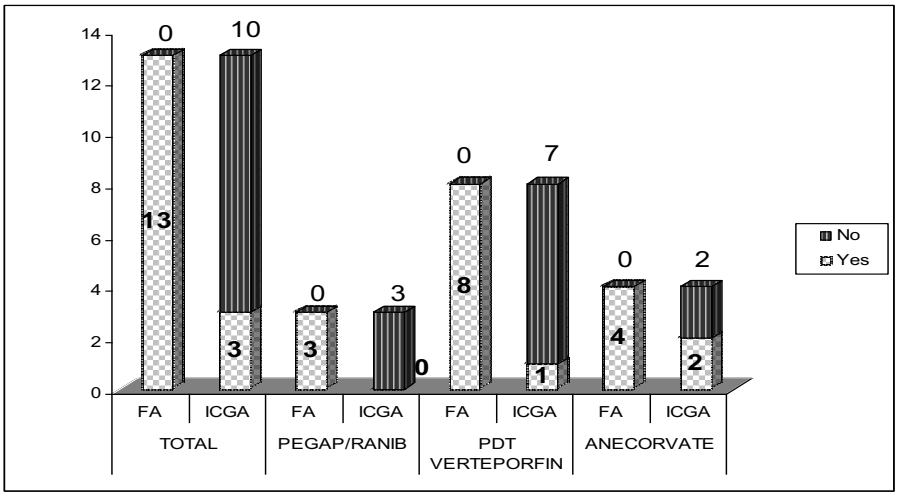
Method: The use of these tests as selection criteria was evaluated for treatments with pegaptanib, ranibizumab, photodynamic therapy with verteporfin, and anecorvate acetate. For these, a systematic search was performed in Medline and Embase for RCT published in the last five years.

Results: Thirteen RCT were found. FA was used to select patients for treatment in all 13 RCT; ICGA is mentioned in 3, it is never used for pegaptanib or ranibizumab treatment, and used in 1of the 8 RCT with verteporfin and in 2 of 4 RCT with anecorvate acetate. Stratified by type of AMD, ICGA is used in 1 of 7 RCT including predominantly classic choroidal □haracterization□n (CNV), in 1 of 3 RCT with minimally classic or occult CNV, in none of the 2 RCT including classic, occult or mixed CNV and in 1 RCT including CNV with retinal angiomatous proliferation (RAP). A meta-analysis with similar selection groups was not possible because studies' heterogeneity.

Conclusion: Indirect evidence was searched by studying the inclusion criteria in RCT on effective treatments in patients with exudative AMD. FA was used as inclusion criterion in all the 13 RCT, whichever treatment was studied or type of patients was included. Despite a lack of direct evidence, FA may be considered as having an impact on disease management. ICGA was only used in 3 RCT, regardless of the type of treatment or the type of AMD included, and was always used in combination with FA. The estimation of the independent impact of ICGA can not be determined from this data.

Reference:
Van den Bruel A, Gailly J, Vrijens F, Devriese S. Guidance for the use of five ophthalmic tests in clinical practice. Brussels: Belgian Health Care Knowledge Centre (KCE); 2008. KCE Reports C (D/2008/10.273/06) www.kce.fgov.be

Figure: Use of FA and ICGA in RCT according to type of treatment



Contact details: jeannine.gailly@kce.fgov.be

Belgian Health care knowledge centre, Brussels, Belgium

Notes

Contributed paper

A review of the use of randomized trials to assess the impact of diagnostic tests on patient outcomes

Lavinia Ferrante di Ruffano, Jacqueline Dinnes, Chris Hyde, Jon Deeks

Background and Objectives: The value of a diagnostic test to the healthcare system ultimately lies in its ability to be beneficial to patients in terms of improving patient outcomes. It is commonly accepted that the randomised controlled trial provides the most methodologically sound vehicle for establishing clinical effectiveness of tests and their contingent treatments – so called test-plus-treatment interventions. However, the logistical challenges in undertaking such trials and the numbers of participants required make such studies challenging. In this review we describe the characteristics of two cohorts of trials of tests and subsequent management, and their findings.

Methods: Two cohorts of test-treatment intervention trials were identified for analysis: The Cochrane Central Register of Controlled Trials, searched for years 2004-7, and all published and ongoing trials funded by the NHS Health Technology Programme, including unpublished archival material such as original study protocols and monitoring reports. Studies were included if they evaluated a management strategy involving randomisation of subjects to one or more diagnostic tests, and assessed their impact on patient outcomes after subsequent treatments. Tests used for screening or monitoring were excluded. A subset of studies was used to develop an abstraction technique allowing the characterisation of studies which underpin evaluations of test-treat packages. This was achieved through the classification of aspects of test-interventions and study design. Data were abstracted independently and in duplicate.

Results: Of 8,975 potentially relevant trials identified in CENTRAL 2004-7, 70 (0.8%) fulfilled the inclusion criteria; 19 of 45 trials identified in the HTA cohort were eligible. In addition to demonstrating the scarcity of consideration given to how a test has impacted on patient outcomes, we present an analysis of the key characteristics of the trials.

Conclusions: Trials of test and treatment interventions are currently rare in the medical literature, and used is a restricted set of situations. We will provide a synthesis of the key methodological challenges that they face, and comment on the circumstances in which they have yielded useful evidence.

Contact details: l.ferrantediruffano@bham.ac.uk

Department of Public Health Epidemiology and Biostatistics, University of Birmingham, UK.

Notes

Invited paper

Applying Diagnostic Evidence to Individual Patients

Nick Summerton

General Practitioner & Clinical Lead BMI Health Screening, Yorkshire, UK.

The diagnostic tools available to the clinician include the history, the examination, specific questionnaires, physiological testing, imaging, pathology (including genetics), endoscopy and the simple passage of time. There may be published evidence available on the analytical validity, the clinical validity and the clinical utility of such technologies.

In using any potential diagnostic tool it is very important to be clear about the precise purpose for which the information obtained is being used. Furthermore, in addition to the clinician having an appreciation for the validity and the reliability of such information in his/her hands (and in relation to his/her patients), there is also a requirement for careful clinical scrutiny. For example data from the medical history always needs to be interpreted in relation to the individual's physical, psychological and social circumstances.

The medical setting (e.g. community-based versus hospital-based) is not only important in relation to the performance characteristics of the diagnostic data but also the precision required for the diagnostic outputs. To maximise diagnostic efficiency and effectiveness it is further suggested that all diagnostic evidence should also be considered in the context of a diagnostic processing pathway or a diagnostic processing web

In the future more careful consideration needs to be given to how best to deliver diagnostic evidence to clinicians in a useful and useable fashion.

Contact details: n.summerton@hull.ac.uk

Notes

Contributed paper

Diagnostic tests for screening: The clinical relevance of positive findings

Robert Grosselfinger, Julia Hommerich, Julia Kreis, Fueloep Scheibler, Stefan Lange

Background: Several researchers have suggested a stepwise evaluation of diagnostic procedures [1]. One important step is the measurement of test accuracy, including properties both independent (sensitivity, specificity) and dependent (predictive values) of prevalence. Furthermore, an improvement in clinical outcomes associated with the test result should be shown (corresponding to level 5 by Fryback et al [1]). To evaluate the effects of a diagnostic test on clinical outcomes, a common approach is to compare different strategies as a whole (combination of test and intervention), based on the diagnostic test under investigation. For example, in the case of therapy failure, it is impossible to determine whether this is caused by ineffective treatment or the inadequacy of the diagnostic test to identify potentially responsive patients. This imprecision in validating diagnostic tests may result in over-treatment, as not all diagnoses are clinically relevant and there may, for example, be spontaneous regression to normal health status, especially in early disease phases. This may lead to the (paradoxical) result that the efficacy of two tests (in terms of clinical outcomes) may vary, even if the tests have the same sensitivity and specificity.

Objective: We describe the clinical relevance of positive test results as an additional characteristic in the evaluation of the appropriateness of screening tests.

Methods: On the basis of our experience in the preparation of systematic reviews on screening strategies, we suggest a possible extension of validation studies in order to consider the clinical relevance of positive test results.

Results: On the one hand, the consideration of the clinical relevance of a positive test result may result in the choice of an alternative diagnostic test with equivalent or even lower sensitivity (assuming the same specificity). On the other hand, the supplementary information gained by the consideration of clinical relevance could lead to an improvement in clinical outcomes associated with the screening test. Intervention-related adverse effects could be reduced, especially in the case of potentially harmful interventions.

Conclusions: The appropriateness of a screening test may also depend on the clinical relevance of a positive test result.

Reference:

1. Fryback, DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11(2): 88-94.

Contact details: robert.grosselfinger@iqwig.de

Institute for Quality and Efficiency in Health Care (IQWiG), Cologne, Germany

Notes

Oral Presentations

Contributed paper

Nonparametric monotonic regression can illustrate how the likelihood ratio varies with a continuous test result without specifying a test threshold

Roger M. Harbord

Background

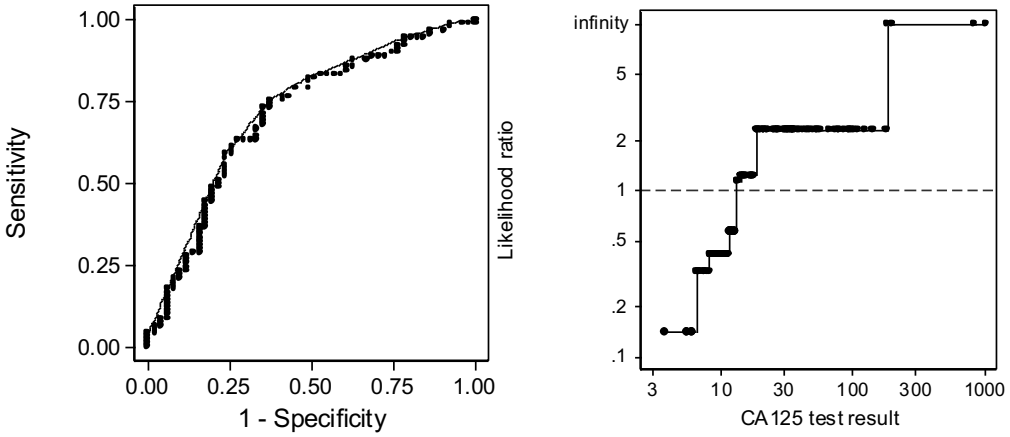
Interpretation of a ROC plot assumes that the continuous test result will ultimately be dichotomised at a single threshold. Likelihood ratios provide an alternative means of interpreting test results that do not require dichotomisation. However, traditional methods require that either categorisation of the test result or specification of a parametric model. Nonparametric monotonic (isotonic) regression provides an alternative which assumes only that higher tests results imply greater likelihood of disease.

Objectives

1. To illustrate the use of nonparametric monotonic regression for analysing continuous test results.
2. To show how a simple extension allows estimation of likelihood ratios for continuous tests without the need to choose cut-points or parametric forms.

Methods and Results

We apply the methods of Lloyd (2002) to standard data sets from the literature to illustrate that nonparametric isotonic regression corresponds to drawing a series of straight line segments of decreasing slope (a "convex hull") around the data on a ROC plot. We extend his work by constructing a plot of the likelihood ratio against the test result as a series of steps, and discuss the interpretation of such a plot and procedures for adding confidence bands to illustrate the degree of uncertainty.



Conclusion

Nonparametric monotonic regression should be used to present likelihood ratios in test accuracy studies of continuous tests.

Reference

Lloyd, CJ. Estimation of a convex ROC curve. *Statistics & Probability Letters* 2002;**59**:99-111

Contact Details roger.harbord@bris.ac.uk
Department of Social Medicine, University of Bristol, UK

Notes

Contributed paper

Grading quality of evidence and strength of recommendations for diagnostic tests and strategies and developing summary of findings tables for diagnostic accuracy studies

Holger Schünemann, Andrew D Oxman, Jan Brozek, Paul Glasziou, Roman Jaeschke, Gunn E Vist, John W Williams Jr., Regina Kunz, Jonathan Craig, Victor M Montori, Patrick Bossuyt, Gordon H Guyatt

Many organizations apply the GRADE approach to grading the quality of evidence and strength of recommendations for interventions. Cochrane review authors use the GRADE approach to grade the quality of evidence for interventions studies in Cochrane summary of findings (SoF) tables – a presentation of the most important information and findings of a review in a table format. The GRADE working group has suggested a separate approach to grading the quality of evidence for questions of diagnostic accuracy (Schünemann et al. BMJ, 2008, in press). Using this approach, cross sectional or cohort studies can provide high quality evidence of test accuracy if they are linked to direct information about patient-important outcomes. Test accuracy in itself, however, is a surrogate for patient-important outcomes, so that these studies often provide low quality evidence for recommendations about diagnostic tests, even when the studies do not have serious limitations. This assessment is due to the recognition that inference from data on accuracy of a diagnostic test or strategy requires information whether applying the test improves patient-important outcomes (i.e. because of availability of effective treatment, reduction of test related adverse effects or anxiety, or improvement of patients' wellbeing from prognostic information). Therefore, studies that provide high quality information about accuracy may provide only low quality evidence of impact on patient-important outcomes, and thus low quality evidence for recommendations about diagnostic test use. Judgments are needed to assess the directness of test results in relation to consequences on patient-important outcomes. Furthermore, systematic reviews of diagnostic test accuracy require adequate, transparent and easy-to-understand information for users of the reviews. Summary of findings tables are one approach. The GRADE working group has developed approaches to presenting this information. This presentation will describe the GRADE approach to grading the quality of evidence for diagnostic test accuracy and suggest presentation formats for summary of findings tables for systematic reviews of diagnostic test accuracy.

Contact details: hjs@buffalo.edu, mmbrozek@cyf-kr.edu.pl

Italian National Cancer Institute "Regina Elena", Rome, Italy

Notes

Oral Presentations

Soap Box

"Should the government decide to invest a further £20M in test research, how should they best invest it?"

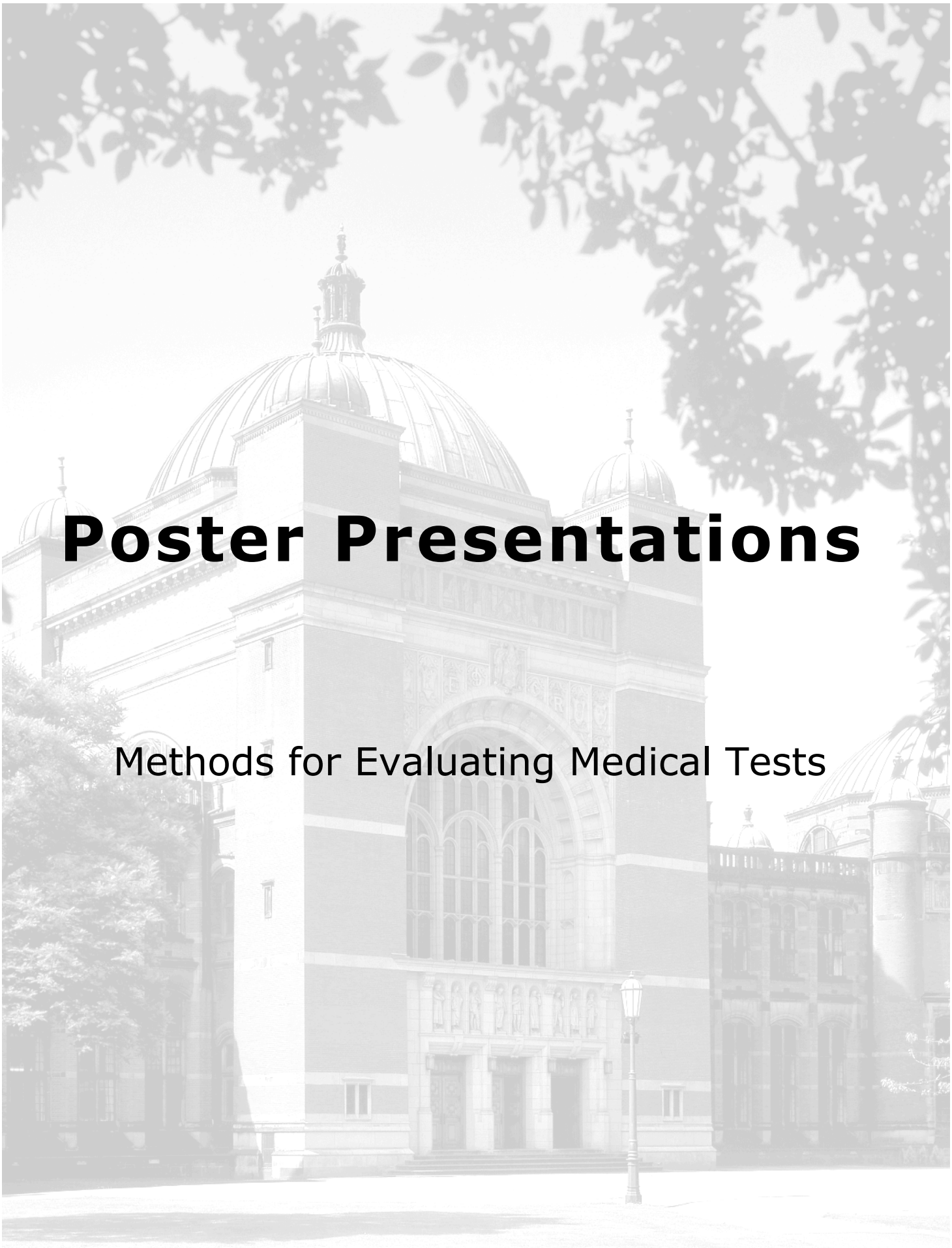
A series of speakers from varied backgrounds will address this question to the house.
Each speaker will be asked to give a 5 minute response taking a particular point of view. The aim is to persuade you, the audience (who will be asked to vote), that theirs would be the best way of spending the hypothetical fund of £20million.

There will be no restrictions on the persuasive tactics (humour, entertainment, emotional blackmail, etc) that the speakers can use!

Proposed means of spending £20million:

1. In primary studies of test accuracy
2. In RCTs of tests evaluating impact on patient outcomes
3. In developing multivariable diagnostic prediction models
4. In systematic reviews of test accuracy
5. In health economic analyses
6. In analyses of routine data
7. In employing more methodologists

Soap Box



Poster Presentations

Methods for Evaluating Medical Tests

Poster Summary

Poster 1

The Epidemiology of Reviews of Test Performance: an analysis of the content of 5 specialist review databases

Clare Davenport, Sue Bayliss

Background

Systematic reviews are an important resource for summarizing existing knowledge about test performance and for undertaking methodological research. Given the substantial growth in test performance reviews (TPRs) observed over the last decade, it would be useful for potential users of reviews to know which resources are the most appropriate for their research purpose.

Objectives

To examine the epidemiology of TPRs located in five specialist review databases :York CRD's DARE and HTA databases, Medion (University of Maastricht), C-EBLM (International Federation of Clinical Chemistry) and the ARIF in-house database (University of Birmingham) with respect to number of reviews of test performance, disease topic area, purpose of review, test application and clinical setting.

Results

A large degree of overlap existed between databases. Medion contained the largest number (n=672) and the largest number of unique (n=328) TPRs. The HTA database contained the smallest number of TPRs (n=333) and DARE the smallest number of unique TPRs (n=93). A combination of three databases identified only 69% of TPRs. Obstetrics and Gynaecology (18%; range 8-18%, median 13%); Gastrointestinal disease (15%, range 8-14%, median 10%) and Cardiology (15%, range 6-14%, median 9%) were the most prominent disease areas across databases. Most reviews were evaluating test accuracy (85%; range 14-81%, median 66%) with only 19% of reviews concerned with test effectiveness or cost-effectiveness. Diagnosis in secondary care was the most common application of tests being evaluated (61%, range 12-55%, median 44%) followed by screening in any setting (48%, range 26-75%, median 28%). The HTA database had a relatively large proportion of cost-effectiveness and screening reviews and C-EBLM a large proportion of reviews concerned with infectious disease, haematology and early test development.

Conclusions

Specialist review databases offer an addition to general bibliographic databases where application of diagnostic method filters can compromise search sensitivity. Important differences between databases in terms of coverage and content should be considered when choosing a resource. Our findings raise the question whether the current balance of test evaluation (test accuracy dominating test effectiveness) and research setting (the predominance of TPRs in secondary care) matches the needs of decision makers.

Contact Details: c.f.davenport@bham.ac.uk
Department of Public Health Epidemiology and Biostatistics, University of Birmingham, UK

Notes

Poster 2

PET/CT in cancer: Analysis strategies in comparative diagnostic studies of accuracy with paired binary data

Oke Gerke, Werner Vach, Poul Flemming Høilund-Carlsen

Background & Aim

An ongoing debate takes place whether diagnostic studies should be planned as randomized controlled trials to compare two groups like therapeutic studies of active drugs usually do. Where the latter is disabled for testing of different drugs in the same patients (except for cross-over trials), diagnostic studies comparing different diagnostic modalities offer the possibility that new non-invasive imaging techniques may, due to their comparably safe application, accompany standard diagnostic imaging in the same sample of patients. This possibility of evaluating PET/CT in cancer is exploited here.

Design and Methods

Patients participating in this kind of studies benefit from both the new imaging technique under consideration and the current standard imaging techniques. Following Freedman (1987), such studies are called diagnostic phase II studies. In these, a working diagnosis with both a standard procedure and PET/CT is compared to a final diagnosis with respect to a binary outcome, e.g. “cancer = yes/no”, “staging=N0/N1+”. Some important aspects in which the study design might vary (aim of the study, availability and quality of a gold standard) are discussed, and a statistical analysis strategy is presented. The latter comprises diagnostic measures like “change in diagnostic accuracy” (sensitivity, specificity) and “change in clinical value” (positive/negative predictive values) as well as sample size considerations.

Results

Formulas for approximate 95% confidence intervals for the differences in sensitivity, specificity, positive and negative predictive values between PET/CT and the standard procedures are given, respectively. Implications of incomplete gold standard procedures (e.g., no true state confirmation in patients diagnosed as disease-free by both standard imaging and PET/CT) on the statistical analysis strategy are analyzed.

Conclusions

Diagnostic studies to assess the merit of PET/CT in the diagnostic work-up of cancer patients can and should start with phase II studies allowing each patient to benefit from both standard diagnostic imaging and PET/CT. Primary focus in the analysis strategy should be 95% confidence intervals for differences in diagnostic measures. Even if the gold standard procedure is incomplete, the statistical analysis strategy given here may still be applicable.

Reference:

Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. The British Journal of Radiology, 1987; 60: 1071-1081.

Contact details: oke@stat.sdu.dk

Department of Nuclear Medicine, Odense University Hospital & Department of Statistics, University of Southern Denmark, Denmark

Notes

Poster 3

Experience of producing recommendations for diagnosis for a NICE guideline on Glaucoma

Kate Homer, Jennifer Hill

Background

The National Collaborating Centre for Acute Care (NCC-AC) has been commissioned by the National Institute for Health and Clinical Excellence (NICE) to develop a clinical guideline on the diagnosis, treatment and service provision of glaucoma for the NHS in England and Wales. The development of the guideline involves the recruitment of a guideline development group (GDG) made up of clinicians, nurses and patient representatives who consider evidence resulting from systematic reviews based around clinical questions for each area of the guideline.

Purpose

To discuss the particular issues around guideline development in the diagnosis of glaucoma and to highlight any similarities or differences to the same process performed for clinical questions on treatment.

Method

Clinical questions were formulated for diagnosis by the GDG and inclusion/exclusion criteria devised for sifting the literature search results. The NCC-AC team of systematic reviewers and health economists selected studies of a suitable design to answer the clinical questions and reviewed their methodological quality using a checklist for diagnostic studies. Data was then extracted and presented to the GDG for consideration.

Results

The methodology for developing clinical questions on diagnosis and the subsequent interpretation of the results were less familiar to the GDG members compared to the more frequent process of designing specific 'PICO' questions and analysis of randomised controlled trials for treatment. Sifting the literature and quality assessment of the diagnostic studies presented particular hurdles, particularly due to the poor quality of studies and inadequate reporting. Describing the data from multiple diagnostic studies in an intuitive and easily interpretable manner to the GDG was also challenging.

Discussion

Making recommendations for clinical practice can be particularly challenging in diagnostic areas such as glaucoma. There are a number of ways the guideline development process could be made easier including improved study design and reporting of diagnostic studies.

Contact Details: khomer@rcseng.ac.uk
National Collaborating Centre for Acute Care (NICE), UK

Notes

Poster Presentations

Poster 4

Systematic reviews of diagnostic research; promises and pitfalls

Petra Jellema, Daniëlle van der Windt, Riekie de Vet, Henriëtte van der Horst

The Cochrane Diagnostic Reviewers' Handbook was developed to help reviewers to be systematic and explicit about the questions they pose and how they derive answers to those questions. Though not available for circulation yet, we were in a position to use the Handbook while carrying out several diagnostic systematic reviews. In this presentation we would like to compare some of the Handbook's recommendations with our experiences.

We carried out systematic reviews on the diagnosis of the Irritable Bowel Syndrome, Colorectal Cancer, Inflammatory Bowel Disease, Lactose Intolerance and Celiac Disease, while using version 0.3 (updated July 2005) of the Cochrane Diagnostic Reviewers' Handbook. Although the handbook was very helpful in many steps, we would like to discuss several issues, such as:

- Use of a methodological filter: this is strongly discouraged but what are the consequences of not using such a filter?
- Use of additional search strategies: according to the Handbook searching in electronic databases is only the start of a search strategy, while further strategies should consist of searching for other reviews, reference checking, hand searching and trying to identify unpublished studies. However, does the yield outweigh the costs of these time consuming strategies?
- Reporting *versus* quality assessment: do QUADAS items 10 ('were uninterpretable/ intermediate test results reported?') and 11 ('were withdrawals from the study explained?') actually measure bias?
- Use of a summary quality score: the Handbook strongly discourages the use of a summary quality score or levels of evidence approach. Instead, reviewers are recommended to incorporate quality differences into their reviews by considering individual quality items as potential sources of heterogeneity. But what are the consequences of this recommendation?
- Use of the SROC analysis: this approach characterizes the relationship between sensitivity and 1-specificity across studies and takes into account variation in the threshold for test positivity between studies. However, what is the interpretation and clinical implication of SROC analysis when studies show heterogeneity of sensitivity and specificity but use the same threshold?

Contact details: p.jellema@vumc.nl
Department of General Practice, University Medical Center, Amsterdam. The Netherlands

Notes

Poster 5

Renal ultrasonography to predict vesico-ureteral reflux after urinary tract infection in childhood: systematic review and meta-analysis

Sandrine Leroy, Jeremy Friedman, Nadjette Mourdi, Isabelle Colombet, Gérard Bréart, Martin Chalumeau

Introduction: Vesicoureteral reflux (VUR) is found in 20-40% of children with febrile urinary tract infection (UTI). Renal ultrasonography (US) is recommended at the time of the first UTI to detect renal and urinary tract abnormalities. Many authors have proposed using US findings to predict VUR, however there are discrepancies in the results reported, particularly because different US criteria were used in the various studies.

Aim: To perform a systematic review and meta-analysis of studies which evaluated the diagnostic accuracy of renal US for all grade and high grade VUR in children with UTI.

Methods: Studies were identified by a systematic electronic search in MedLine, Embase, Cochrane Library and Google Scholar databases from 1985 to 2006. We evaluated the quality of studies according to methodological standards of diagnostic studies and pooled data of studies which evaluated the same US criteria using a random effect model of the diagnostic odds ratio (DOR). We explored the threshold effect by SROC analysis and the heterogeneity by univariate meta-regression.

Results: From the 1456 potentially relevant articles, 33 observational cohort studies were included (5054 patients, 27% with VUR) to study the diagnostic accuracy of renal US for all grade VUR. Pelvi-calyceal, ureteral, urinary tract dilation, and abnormal renal length had a significant pooled DOR respectively of: 3.3 [95% CI: 1.5-7.2], 1.8 [1.0-3.0], 2.3 [1.6-3.5], 4.0 [1.6-9.6]. There was evidence for a strong heterogeneity for all these criteria (I^2 >50%) due to a threshold effect, except for ureteral dilation (I^2=30%). The pooled sensitivity and specificity of the ureteral dilation were 13% [10-17] and 92% [90-93] respectively. The diagnostic accuracy of renal US for high-grade VUR was analysed in 10 studies (1857 patients, 10% with high-grade VUR). Only the ureteral dilation and the abnormal renal length had a significant pooled DOR without evidence of heterogeneity (I^2 <50%): 5.5 [1.3-22] and 3.9 [1.7-8.6] respectively. The pooled sensitivity and specificity of the ureteral dilation were 17% [9-29] and 96% [94-98] respectively.

Conclusion: Ureteral dilation seemed the best renal US criteria for the prediction of both all grade and high-grade VUR in children with UTI, with a low sensitivity but a very high specificity. Its combination with other sensitive predictors, such as procalcitonin, in a clinical decision rule could be useful.

Contact Details: sandrine.leroy@csm.ox.ac.uk
INSERM U149, Saint-Vincent-de-Paul hospital, Paris, France.

Notes

Poster 6

Dynamic tests of ovarian reserve - A systematic Review of diagnostic accuracy

Abha Maheshwari, Ahmed Gibreel, Siladitya Bhattacharya, Neil Johnson

Objective- To determine the diagnostic accuracy of dynamic tests of ovarian reserve including Clomiphene citrate challenge test (CCCT), GnRH agonist stimulation test (GAST) and exogenous FSH ovarian reserve test (EFORT) or any other dynamic endocrine test of ovarian reserve in prediction of fertility outcomes.

Design- Systematic review

Methods- Studies were identified without language restrictions from MEDLINE, EMBASE, PASCAL, Biosis, Cochrane Library, National Research Register, SCISEARCH, conference papers, and manual searching of the bibliographies of known primary and review articles. Studies were selected if accuracy of dynamic tests were evaluated for predicting fertility outcome using one or more of the outcomes measures as the reference standard.

Main outcome measures- live birth rate, pregnancy rate, number of oocytes retrieved and cycle cancellation rate.

Results- The positive and negative likelihood ratios of CCCT in the prediction of non-pregnancy were 1.77 (1.01-3.11) and 0.84 (0.74-0.99) at FSH >10 (day 3 or 10). The diagnostic odds of abnormal CCCT for non-pregnancy were 2.11 (95% CI, 1.04-4.29). We could not determine the diagnostic accuracy of GAST and EFORT, due to inconsistencies in the way these tests were conducted.

Conclusions- This systematic review and meta-analysis of dynamic tests of ovarian reserve was limited by heterogeneity in terms of the population sampled and the index and reference tests. There is an urgent need for consensus on how to perform these tests and the definition of normality. Evidence was particularly insufficient, owing to a lack of diagnostic test accuracy studies amenable to meta-analysis, for GAST and EFORT. With the present level of evidence, none of these tests should be used to predict non–pregnancy.

Key Reference

Broekmans FJ, Kwee J, Hendriks DJ, Mol BW, Lambalk CB. A systematic review of tests predicting ovarian reserve and IVF outcome. *Hum.Reprod.Update* 2006; 12 : 685-718.

Contact Details: abha.maheshwari@abdn.ac.uk
Obstetrics and Gynecology, University of Aberdeen, Aberdeen, UK

Notes

Poster 7

Meta-analysis of Diagnostic Test accuracy data and Bayesian model-choice criteria: Deep-venous thrombosis example

Nicola Novielli, Nicola J. Cooper, Alex J. Sutton, Keith R. Abrams

Background: Several meta-analysis models for combining diagnostic test data have been described in the literature. These models vary in the assumptions they make regarding i) the variability in test thresholds between studies and ii) incorporation of variability beyond that expected by chance (between-study heterogeneity). In any particular situation it is unclear which model is the most appropriate for the data. Traditionally, goodness of fit criteria are used to choose between different statistical models. Where complex non-nested models with random effects are considered (as in this situation) the Deviance Information Criteria (DIC) provides a criterion for choosing between models.

Objectives: To explore the use of DIC to choose between different meta-analysis models applied to 198 studies evaluating DDimer for deep-vein thrombosis - DVT (Goodacre 2005).

Methods: To meta-analyse the DVT diagnostic test data the following Random/Fixed effect models are fitted: 1. Independent estimates of Sensitivities/Specificities; 2. Symmetric summary-ROC curves estimation; 3. Asymmetric summary-ROC curve estimation; 4. Bivariate estimate of Sensitivities/Specificities.

Table 1. Model and DICs

Model	Effects	DIC	pD
1. Independent	Fixed	5560	1.9
	Random	2148	320
2. Symmetric summary	Fixed	2434	199
	Random	2142	318
3. Asymmetric summary	Fixed	2411	192
4. Bivariate model	Random	2133	301
	Random- mean age	2120	290

In addition, these models are extended to include covariates. The fit of the different models is assessed using the DIC.

Results: As can be observed in Table 1, for this example the Bivariate model fits best (DIC: 2133) and the fit is improved by including a covariate for mean age of patients (DIC: 2120). The mean age of the study population in the bivariate model affects either

Sensitivity [+0.039 increase per year, 95%CI (-0.008 to 0.087)] or Specificity [-0.040 increase per year, 95%CI (-0.074 to -0.006)].

Conclusions: With numerous alternative approaches available for meta-analysis of diagnostic test accuracy data, each making different assumptions, a way of choosing between models is required. The use of DIC seems to be well suited.

Goodacre, S., F. C. Sampson, et al. (2005). "Variation in the diagnostic performance of D-dimer for suspected deep vein thrombosis." Qjm **98**(7): 513-27.

Contact Details: nn40@le.ac.uk
University of Leicester, UK

Notes

Poster Presentations

Poster 8

The Eunet HTA Core Model for Diagnostic technologies: How to assess effectiveness

Iris Pasternack; Tuija Ikonen, Sigurdur Helgason, Heikki Ukkonen, Sami Kajander

Background: EUnetHTA was established in 2006 to connect national health technology assessment (HTA) agencies to enable knowledge exchange and support. One goal was to create a generic methodological HTA framework, the Core Model, based on methodological evidence and current best practices. The Core model for medical and surgical interventions was divided into nine domains: current use, description of technology, safety, effectiveness, costs, organisational, social, ethical and legal issues.
Objective: To present the work done in the effectiveness domain of the EUnetHTA´s Core Model for diagnostic technologies.
Methods: The work is done in multidisciplinary teams with participants from 25 EU countries.
Results: Effectiveness domain was split into two domains: accuracy and effectiveness. Effectiveness relevant topics, divided into several generic issues, were identified (see examples below).
Conclusion: The work will be completed by November 2008. All comments are valuable.

Topic	Issue	Importance *3=critical 2=important 1=optional*	Transfera-bility *3=complete 2=partially 1=not*
Comparative accuracy of a replacement technology	Is there evidence that the replacing technology is more specific or safer than the old one?	2	2
Safety	What is the mortality related to the diagnostic technology?	3	3
Change-in management	Does the use of the technology lead to a change in the physicians' management decisions?	2	2
Change-in management	How does the technology modify the need for other tests and use of resources?	2	2
Change-in management	How does the technology modify the need for hospitalization?	2	2
Health outcomes	Is there an effective treatment for the condition the technology is detecting?	3	2
Health outcomes	What is the effect of the test-treatment intervention on mortality?	3	2
Health outcomes	How does the technology modify the effectiveness of subsequent interventions?	2	2
Health outcomes	What is the effect of the technology on health-related quality of life?	3	2
Health outcomes	What are the overall benefits and harms in health outcomes considering the amount of false positive and false negative.	3	2

Contact Details: iris.pasternack@stakes.fi
Finnish Office for Health Technology Assessment (Finohta), Finland

Notes

Poster 9

Assembling the evidence for new cancer staging tests: A systematic review of positron emission tomography (PET) in patients with colorectal liver metastases

Lukas Staub, Suzanne Dyer, Sarah Lord

Background: PET has been proposed as an additional test for staging patients with potentially resectable colorectal liver metastases (CLM). The primary potential benefit is the detection of additional metastases and avoidance of surgery in these patients. Whether this provides a benefit depends on a series of factors: the sensitivity of PET incremental to conventional imaging; the proportion of patients with a true positive result who avoid surgery; and the relative benefits of surgery versus no surgery on patient mortality/morbidity/quality-of-life. Any benefit needs to be weighed against the harms of false positive findings.

Objectives: To conduct a systematic review of the effectiveness of PET in addition to computed tomography (CT) for detecting metastases in patients with potentially resectable CLM and to explicate the assumptions needed when relying on indirect evidence to assess new cancer staging tests.

Methods: Studies from the most recent high-quality PET HTA were supplemented with an updated search of MEDLINE & EMBASE to December 2006. Included were studies reporting (i) incremental accuracy of PET over CT; (ii) change in patient management and outcomes following PET. Negative PET results have no impact on patient management for this indication, thus accuracy studies were included if they verified all PET results or all PET-positive results.

Results: Four accuracy (n=259) and three patient management (n=162) studies, but no studies comparing patient outcomes were identified. PET was positive in 18-40% of patients, detecting additional sites of true metastases in 11-39% of patients with false-positive findings in 0-7%. The consequences of false-positive findings were poorly reported. Seventeen percent of all patients (86% of PET-positives; two studies) avoided planned surgery.

Conclusions: PET detects additional true metastases in up to 39% of patients and avoids surgery in most of these patients. These findings suggest PET may improve outcomes by identifying patients who are unlikely to benefit from surgery. These conclusions are based on two critical assumptions that: (i) the harms of surgery outweigh the benefits in patients with PET-detected additional true metastases; and (ii) these benefits outweigh the harms in patients with a false positive finding. Randomised trials are needed if these assumptions are unacceptable.

Contact details: lukas.staub@ctc.usyd.edu.au

NHMRC Clinical Trials Centre, Systematic Reviews and Health Care Assessment, The University of Sydney, Australia

Notes

Poster 10

Diagnostic value of history-taking and physical examination in patients with a knee injury in General Practice

Harry Wagemakers, Simone Boks, Bart Heintjes, Marjolein Berger, Bart Koes, Sita Bierma-Zeinstra

Introduction

History-taking and physical examination should help the GP establish a clinical diagnosis. However, the diagnostic accuracy is often questioned.

Objective of this study

To determine the prevalence of meniscal tears and anterior cruciate ligament (ACL) lesions and to determine the diagnostic value of isolated determinants as well as composite examination.

Methods

This study was part of a large prospective cohort study on knee complaints in general practice[1]. Fourty GP's included patients with traumatic knee complaints. MR imaging was used to determine the nature and severity of lesions. History taking was performed by means of a self-report questionnaire. MRI and physical examination were performed based on a standard protocol. Diagnostic value was expressed in terms of sensitivity (Se), specificity (Sp), predictive values (PVP,PVN) and likelihood ratios (LR+,LR-).

Results

134 patients were included; the prevalence of meniscal tears in this study was 0.35 while the prevalence of ACL lesions was 0.21.

	TP*	Se$^{\Phi}$	Sp$^{\Phi}$	PVP$^{\Phi}$	PVN$^{\Phi}$	LR+$^{\Phi}$	LR-$^{\Phi}$
Meniscal tear							
Age over 40 years	33	0.70 (0.57-0.83)	0.64 (0.54-0.74)	0.52 (0.39-0.64)	0.80 (0.71-0.89)	2.0 (1.4-2.8)	0.5 (0.3-0.7)
History = 3 + phys. ex.	7	0.15 (0.05-0.25)	0.97 (0.94-1.00)	0.78 (0.51-1.00)	0.66 (0.57-0.74)	5.8 (1.3-26.8)	0.9 (0.8-1.0)
ACL lesion							
"Popping" sensation	44	0.63 (0.45-0.81)	0.73 (0.64-0.82)	0.39 (0.24-0.53)	0.88 (0.81-0.95)	2.3 (1.5-3.6)	0.5 (0.3-0.8)
Anterior drawer test (ADT)	64	0.83 (0.68-0.98)	0.57 (0.48-0.67)	0.31 (0.20-0.43)	0.94 (0.88-1.00)	2.0 (1.5-2.6)	0.3 (0.1-0.7)
History 3-3 + ADT	5	0.16 (0.02-0.30)	0.99 (0.98-1.00)	0.80 (0.60-1.00)	0.82 (0.75-0.89)	15.4 (1.8-131)	0.8 0.7-1.0)

Conclusions

The diagnostic value of isolated determinants from history-taking and physical examination in detecting meniscal tears is small. Composite examination does alter the diagnostic value only by a small degree.
There is some diagnostic value from history-taking and physical examination in detecting ACL lesions and composite examination is meaningful.

1.Heintjes EM, Berger MY, Koes BW, Bierma-Zeinstra SM: Knee disorders in primary care: design and patient selection of the HONEUR knee cohort. BMC Musculoskelet Disord 2005; 6: 45.

Contact details: hpa.wagemakers@dordrecht.nl
Erasmus Medical Center, The Netherlands

Notes

Poster 11

Is Independent Monitoring Needed For Test Accuracy Studies?

Jane Daniels, Khalid S Khan, Shakila Thangaratinam, Richard Gray

Background

It is standard practice for randomised controlled trials (RCTs) to include some form of independent monitoring. Although not legally required, UK funding agencies mandate both an independent trial steering committee (TSC) to oversee RCTs on their behalf and an independent data monitoring committee (DMC) who review confidential interim data to protect the safety of the participants and to ensure that randomising patients between the trial treatment arms remains ethical. These committees are composed of a small number of individuals, who have pertinent expertise and are independent of the study. There have been numerous recommendations regarding DMCs for RCTs, most notably the DAMOCLES project. However, none of these provide guidance as to how and when independent monitoring should be included in primary test accuracy studies.

Experience

We conducted a systematic review of 40 test accuracy studies reported in 7 major journals over 2 years and found no reference to independent monitoring. In the course of conducting a large study, we developed a model for independent oversight and some of the issues we considered are discussed here.

Issues

The sample size for test accuracy studies should be sufficient to ensure the lower limit of the 95% confidence interval of the sensitivity and specificity of a test do not fall below an acceptable lower limit. TSCs are useful to provide objective monitoring of recruitment rates, data completeness and adverse events associated with test and the assumptions regarding prevalence of the underlying condition, as the precision of estimates of sensitivity depends critically on the proportion of cases detected. No formal stopping rules for test accuracy studies have been developed so, for independent monitors, the risks of missing a true case (false negative) and the impact of unnecessary treatment of false positives should be considered in deciding whether to recommend halting the study. In observational test accuracy studies, all participants are subject to both the index and reference test and outcomes are often not altered by participation. If test results will not influence management of the patient a single monitoring body may be sufficient. However, if knowledge of interim results may adversely impact on study recruitment, the traditional TSC/ DMC division of responsibilities, and consequential confidentiality of interim data, is appropriate. These facets of monitoring need further discussion and formal guidelines would be helpful.

Contact details: j.p.daniels@bham.ac.uk
Birmingham Clinical Trials Unit, University of Birmingham, UK

Notes

Poster 12

TIGRAs and challenges on the diagnosis of latent TB infection

Kerry A Millington, Louisa Gnatiuc, Suranjith Seneviratne, Onn Min Kon, Melissa Wickramasinge, Loong-Yuan Han

The recent introduction of T-cell based interferon-gamma release assays (TIGRAs): QuantiFERON™-TB Gold in-tube (QFT-IT, Cellestis, Carnegie, Australia) and T-SPOT™.*TB* (Oxford Immunotec, Abingdon, U.K.) into clinical practice could enhance the accuracy of latent tuberculosis infection (LTBI) diagnosis, due to better specificity in BCG-vaccinated persons and probable increased sensitivity than the currently used 100 year old tuberculin skin test (TST).

The only gold standard for LTBI is subsequent development of tuberculosis (TB) but generation of such data requires large longitudinal clinical outcome studies to identify and treat persons at risk of progression. In the absence of this data, the amount of exposure to an infectious TB source case, identified epidemiologically as a risk factor for progression to TB, has been extensively used to validate currently available TIGRAs. Now TIGRAs are being used clinically, the opportunity arises to correlate discordant results with the following well-established clinical and radiological factors associated with the risk of progression to TB:

- epidemiological diagnosis
- suggestive radiological results
- TST conversion in recent contact
- very large TST induration

In some cases known limitations of the TST may explain discordant results; false-negative TST results occur in immunosuppressed or very young patients and false-positive TST can occur in BCG-vaccinated persons due to cross-reactivity between purified protein derivative (PPD) used in TST and BCG. Depending on the size of the risk of progression to TB, the positive result of one test can overrule the discordant result of the second test. For example, a strongly positive (>25mm), ulcerating TST will over-ride a negative TIGRA whilst a negative TIGRA result will over-ride a borderline, weak or moderate positive TST in BCG vaccinated persons. Radiographic evidence of calcified nodules of old untreated TB, within a suggestive epidemiological context (i.e. born/resident of a country with high prevalence of TB or close contact of a confirmed pulmonary case) will override a negative result of either test.

In conclusion, the new TIGRAs are a guide to health professionals in diagnosing LTBI, but until there is an evidence-base on interpretation of discordant TST and TIGRA results, a holistic approach should be used on a case-to-case basis.

Contact Details: k.millington@imperial.ac.uk
Imperial College Healthcare NHS Trust, London, UK

Notes

Poster 13

Quality assessment of diagnostic before-after studies.

Catherine Meads, Clare Davenport, Esther Alborn

Background:
Quality assessment tools for primary studies of test accuracy are relatively well developed, although only one is validated (QUADAS), but very little work has been done to develop tools to quality assess studies evaluating the impact of diagnostic testing on management of patients (diagnostic or therapeutic yield). The recent draft NICE Guide to the Methods of Technology Appraisal (2007) suggests QUADAS "as a useful starting point for appraising studies that evaluate the sensitivity and specificity of a test" but does not mention how to quality assess diagnostic or therapeutic yield studies, in particular diagnostic before-after studies. In the context of undertaking a rapid systematic review of structural neuroimaging in psychosis for NICE, we describe the modifications that we made to QUADAS, our experience of this in practice and in relation to published theory on diagnostic or therapeutic yield studies.[1,2]

Methods:
The QUADAS tool was assessed for use in the review by two systematic reviewers with in-depth knowledge of the clinical area being reviewed and the types of studies being found in the searches that could answer the clinical question. Modifications were made following discussion as considered appropriate.

Results:
Two QUADAS questions were removed altogether and wording of remaining questions modified to make them more readily understandable and precise in this context. Other quality aspects not captured by QUADAS were also felt to be important so four additional questions were developed. However, the developed checklist only partially helped to discern implications of the study designs on the results given.

Discussion:
With more time, further work could have been done to create a better quality assessment tool, for example by incorporating some of the issues mentioned in the paper by Guyatt.[1] This paper is a discussion around quality assessment rather than a checklist for practical use but it does have much valuable insight into the types of issues that should be assessed. The division between topic specific and more generic quality items of relevance to diagnostic before-after studies is important. Further work should be done to validate, on range of topic areas, a quality assessment tool that incorporates items from QUADAS and published theory.[1,2]

References:
1. Guyatt GH, Tugwell PX, Feeny DH, Drummond MF, Haynes RB. The role of before-after studies of therapeutic impact in the evaluation of diagnostic technologies.
2. Knottnerus JA, Dinant G-J, van Schayck OP. The diagnostic before-after study to assess clinical impact. In Knottnerus JA (ed) The evidence base of clinical diagnosis. BMJ Books, London 2002.

Contact Details: c.a.meads@bham.ac.uk
Department of Public Health Epidemiology and Biostatistics, University of Birmingham, UK

Notes

Poster 14

Procalcitonin to predict vesico-ureteral reflux in children with DMSA scintigraphy confirmed acute pyelonephritis: a multicenter European Study

Sandrine Leroy, Annick Galetto-Lacour, Anna Fernandez-Lopez,David Tuerlinckx, Vladislav Smolkin, Dominique Gendrel, Gérard Bréart, Martin Chalumeau

Background:
Febrile urinary tract infection (FUTI) reveals vesicoureteral reflux (VUR) in 20-40% of children. Voiding cystourethrogram (VCUG) is then recommended systematically, but is irradiating, painful, expensive and a posteriori normal in 60-80% of cases. Then, selective approaches for VCUG are needed. Procalcitonin (PCT), a new inflammatory marker, has been identified (in a single-center study) and validated (in a multicenter study) to be a strong and sensitive predictor of VUR in patients with a first FUTI diagnosed by positive urine culture alone. However, early DMSA scan is the gold standard examination for acute pyelonephritis (FUTI with confirmed renal involvement).

Objective:
To study the relationship between VUR and PCT in children with a first acute pyelonephritis confirmed by an early DMSA scan.

Methods:
This secondary analysis of prospective published series included children aged 1 month to 4 years with a first FUTI and a positive early DMSA scan.

Results:
203 patients (62 boys, mean age of 13.3 months, VUR in 29%) were included in 5 European centres. The median value of PCT increased significantly with the grade of VUR (p=0.005), but was not significantly higher in children with *vs* without VUR: 2.3 *vs* 1.5 ng/mL (p=0.2). After dichotomisation around the previously defined 0.5 ng/mL threshold, there was a significant association between high-grade VUR and high PCT [OR=14.6, 95% CI 1.6-247, p=0.004]. However, this relationship between all-grade VUR and high PCT did not remain statistically significant (p=0.8). The sensitivity of high PCT was 78% (95% CI: 66-87) for all-grade VUR and 100% (95% CI: 88-100) for high-grade VUR, both with 21% specificity (95% CI: 15-28).

Conclusions:
Among patients with a first FUTI confirmed by early DMSA scan, a high serum PCT concentration is a significant and sensible predictor of high-grade VUR.

Contact details: sandrine.leroy@csm.ox.ac.uk
INSERM U149, Saint-Vincent-de-Paul hospital, Paris, France

Notes

Poster 15

Horizon scanning for in vitro diagnostic tests: The development of prioritisation criteria for emerging IVDs.

Luan Linden, Sara Trevitt, Claire Packer.

Background: The National Horizon Scanning Centre (NHSC) is part of the National Institute for Health Research (NIHR) and provides advanced notice to NICE and national policy makers in England of key emerging health technologies prior to their availability in the NHS. The NHSC's remit includes all heath technology types, but the identification and prioritisation of in-vitro diagnostic tests (IVDs) has required the development of new methods. Whilst there are large numbers of IVDs reaching the UK market each year, the great majority are incremental in nature and have little evidence of clinical utility. The ability to identify and prioritise potentially disruptive innovative IVDs is required to ensure that evaluation agencies can develop timely and authoritative advice on their adoption and use.

Aim: To develop and evaluate filtration and prioritisation criteria for the selection of emerging IVDs for evaluation.

Methods: We consulted regulatory and professional bodies, and IVD manufacturers to discuss regulatory requirements and the potential for timely information exchange; and undertook an analysis of existing prioritisation criteria used by HTA agencies. We developed and applied pilot criteria on a selection of emerging IVDs.

Outcome: The prioritisation criteria we developed include:

1. Degree of innovativeness
   - **Completely new:** A new test in an area where diagnosis is not currently done or can't be done.
   - **Novel approach where alternatives already exist:** Testing is already available but this is a new way of performing it.
   - **Incremental development**: Improved or extended version of current tests that is likely to have significant advantages.
2. Time to launch (within 12-18 months before launch)
3. Potential to impact upon the patient pathway: evidence of clinical utility or plans to generate evidence within the specified timeframe.
4. Plus one or more other significance factors e.g. patient group size, mortality and morbidity, and cost impact or savings.

The NHSC is using these criteria on all newly identified IVDs. The NHSC has put in place a programme of contacts with major IVD developers for the identification of emerging IVDs. Evaluation of the effectiveness of these criteria is not possible until our policy customers have completed their evaluations.

Contact details: l.p.linden@bham.ac.uk
National Horizon Scanning Centre, Department of Public Health Epidemiology and Biostatistics, University of Birmingham, UK

Notes

Notes:

Notes:

Notes:

Notes:

This page is intentionally blank