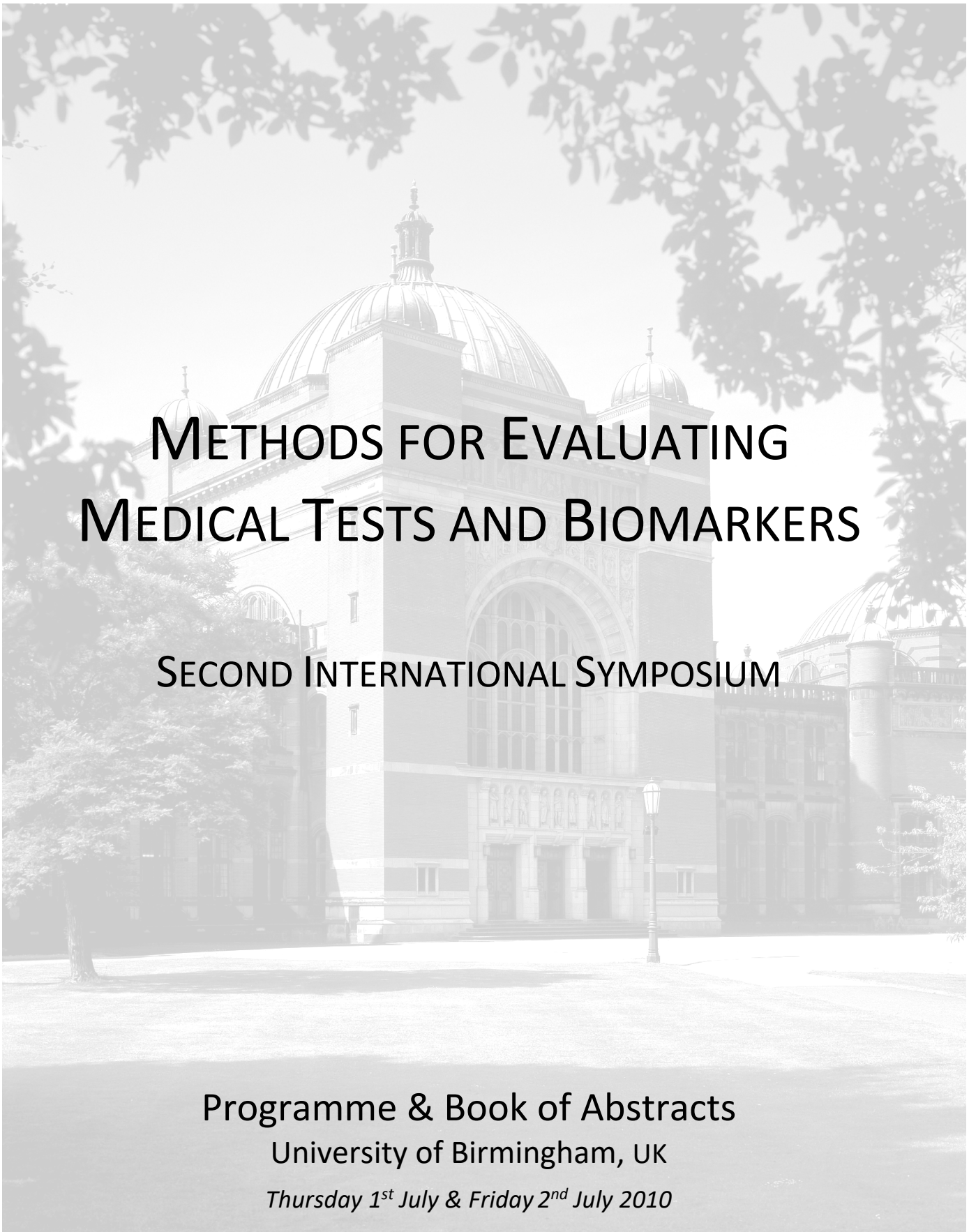
UNIVERSITY OF BIRMINGHAM



# METHODS FOR EVALUATING MEDICAL TESTS AND BIOMARKERS

## SECOND INTERNATIONAL SYMPOSIUM

Programme & Book of Abstracts

University of Birmingham, UK

*Thursday 1st July & Friday 2nd July 2010*

# Welcome

The design, execution, analysis, reporting and implementation of evaluations of medical tests and biomarkers 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 2010 event promotes the importance of research into all aspects of medical diagnostics, and presents an opportunity to debate practice, methodological issues and current/recent research in the field of medical tests.

**Themes for this year are:**

- Clinical prediction models
- Developments in meta-analysis
- Models and health economic assessments
- Impact of tests
- Guidelines and policy
- Industry, regulation and cutting-edge issues
- Issues in verification and reference standards
- Monitoring

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

Jon Deeks

Scientific committee (Chair)

# Local Planning Committee

Public Health, Epidemiology and Biostatistics, University of Birmingham:

Lavinia Ferrante di Ruffano (co-chair)

Susanna Wisniewski (co-chair)

Anne Walker

Alice Sitch

Yemisi Takwoingi

Boliang Guo

# Scientific Committee

Professor Jon Deeks – University of Birmingham, UK

Professor Lucinda Billingham – University of Birmingham, UK

Professor Patrick Bossuyt – University of Amsterdam, Netherlands

Professor Chris Hyde – Peninsula College of Medicine and Dentistry, Exeter, UK

Dr Rafael Perera – University of Oxford, UK

Dr Richard Riley – University of Birmingham, UK

Dr Matthew Thompson – University of Oxford, UK

# How to cite this publication

The Abstract book should be cited as:

Methods for Evaluating Medical Tests and Biomarkers. Symposium; 2010 Jul 1-2; 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 and Biomarkers. Symposium; 2010 Jul 1-2; Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, UK. Page number(s).

**Abstracts are available at this website**

www.medical-test-res.bham.ac.uk/symposium2010

# Table of Contents

# Programme Overview

## Thursday 1st July

| | | |
|---|---|---|
| 08:00 | Registration opens | Ground Floor, Wolfson Centre |
| 09:30 | **Session 1**<br>Clinical prediction models | Leonard Deacon Lecture Theatre |
| 11:15 | *Morning Coffee/Tea and Poster Viewing* | Ground Floor Wolfson Centre |
| 11:45 | **Session 2**<br>Developments in meta-analysis | Leonard Deacon Lecture Theatre |
| 13:05 | *Lunch* | Lower Ground Floor Wolfson Centre |
| 14:00 | **Session 3**<br>Models and health economic assessments | Leonard Deacon Lecture Theatre |
| 15:40 | *Afternoon Coffee/Tea and Poster Viewing* | Ground Floor Wolfson Centre |
| 16:30 | **Session 4**<br>Impact of tests | Leonard Deacon Lecture Theatre |
| 18:10 | *Close* | |
| 19:30 | **Conference Dinner** | Birmingham Council House |

## Friday 2nd July

| | | |
|---|---|---|
| 08:30 | **Session 5**<br>Guidelines and policy | Leonard Deacon Lecture Theatre |
| 10:00 | *Morning Coffee/Tea and Poster Viewing* | Ground Floor Wolfson Centre |
| 10:30 | **Session 6**<br>Industry regulation and cutting-edge issues | Leonard Deacon Lecture Theatre |
| 12:00 | *Lunch* | Lower Ground Floor Wolfson Centre |
| 12:45 | **Session 7**<br>Issues in verification and reference standards | Leonard Deacon Lecture Theatre |
| 14:05 | *Afternoon Coffee/Tea and Poster Viewing* | Ground Floor Wolfson Centre |
| 14:30 | **Session 8**<br>Monitoring | Leonard Deacon Lecture Theatre |
| 16:00 | *Close* | |

# Full Programme

## Thursday 1st July

08:00 Registration – Ground Floor the Wolfson Building

**Session 1 - Clinical prediction models - Leonard Deacon Lecture Theatre**

09:30 – 09:35 Introduction and welcome

*Jon Deeks*

09:35 – 10:05 Evaluation of diagnostic and prognostic (bio)markers and prediction models: from development to implementation

*Karel Moons (Keynote Speaker)*

10:05 – 10:35 Implementation of clinical prediction rules in primary care evidence: evidence, challenges and possible solutions

*Tom Fahey (Keynote Speaker)*

10:35 – 10:55 Biomarkers of tolerance in renal transplantation: Pipeline from discovery to translation in the clinic

*Irene Rebollo Mesa*

10:55 – 11:15 Validation and refinement of a model to predict the risk of endometrial cancer in patients with postmenopausal bleeding

*Merel Breijer*

11:15 – 11:45 *Morning Coffee/Tea and Poster Viewing* ***Ground Floor Wolfson Centre***

**Session 2 - Developments in meta-analysis - Leonard Deacon Lecture Theatre**

11:45 – 12:05 Bivariate meta-analysis of predictive values

*Mariska Leeflang*

12:05 – 12:25 Cervical length measurement for the prediction of pre-term birth in multiple pregnancies: a systematic review using bivariate meta-analysis on stratified bootstrap samples to deal with studies reporting multiple accuracy estimates

*Brent Opmeer*

12:25 – 12:45 An empirical assessment of the validity of uncontrolled comparisons of the accuracy of diagnostic tests

*Yemisi Takwoingi*

12:45 – 13:05 Meta-analysis of the accuracy of sequences of diagnostic tests by allowing for between tests correlation

*Nicola Novielli*

13:05 – 14:00 *Lunch Break* ***Lower* Ground Floor Wolfson Centre**

**Session 3 – Models and health economic assessments - Leonard Deacon Lecture Theatre**

14:00 – 14:30 Health economic assessment of diagnostic tests

*Matt Stevenson (Keynote Speaker)*

14:30 – 15:00 Assessing the cost effectiveness of using prognostic biomarkers: a case study in prioritising patients waiting for coronary artery surgery

*Steve Palmer (Keynote Speaker)*

15:00 – 15:20 Using a natural history model in combination with screening data to estimate the test characteristics of the FOB test in the English bowel screening programme

*Sophie Whyte*

15:20 – 15:40 Cost-effectiveness analysis of diagnostic tests of biomarkers: an example study of transoesophageal echocardiography to diagnose the presence of ascending aortic atherosclerosis in cardiac surgery patients

*Hendrik Koffijberg*

15:40 – 16:30 *Afternoon Coffee/Tea and Poster Viewing* ***Ground Floor Wolfson Centre***

**Session 4 - Impact of tests - Leonard Deacon Lecture Theatre**

16:30 – 16:50 Sensitivity and specificity rarely vary with prevalence

*Mariska Leeflang*

16:50 – 17:10 Empirical evidence that disease prevalence does affect diagnostic test performance – the effects of prevalence on the interpretation of x-rays by junior doctors

*Brian Willis*

17:10 – 17:30 What do 'test-treat' trials measure?

*Lavinia Ferrante di Ruffano*

17:30 – 17:50 Identification of additional effects of medical testing: towards more comprehensive test evaluation

*Jolande Vis*

17:50 – 18:10 Statistical design and preliminary analysis of eye-tracking studies to investigate diagnostic performance in CT colonography

*Susan Mallett*

## Friday 2nd July

**Session 5 - Guidelines and policy - Leonard Deacon Lecture Theatre**

| | | |
|---|---|---|
| 08:30 – 09:00 | Diagnosis in NICE clinical guidelines<br>*Phil Alderson (Keynote Speaker)* | |
| 09:00 – 09:20 | Commissioning reviews of diagnostic test accuracy: Lessons from a systematic review of positron emission spectroscopy and positron emission spectroscopy/ computed tomography<br>*Mary Pennant* | |
| 09:20 – 09:40 | Defining the role of a diagnostic test does not (yet) allow to limit the level of eligible evidence<br>*Stefan Sauerland* | |
| 09:40 – 10:00 | Developing evidence-based recommendations for tests and markers<br>*Patrick Bossuyt* | |
| 10:00 – 10:30 | *Morning Coffee/Tea and Poster Viewing* | ***Ground Floor Wolfson Centre*** |

**Session 6 - Industry, regulation and cutting-edge issues - Leonard Deacon Lecture Theatre**

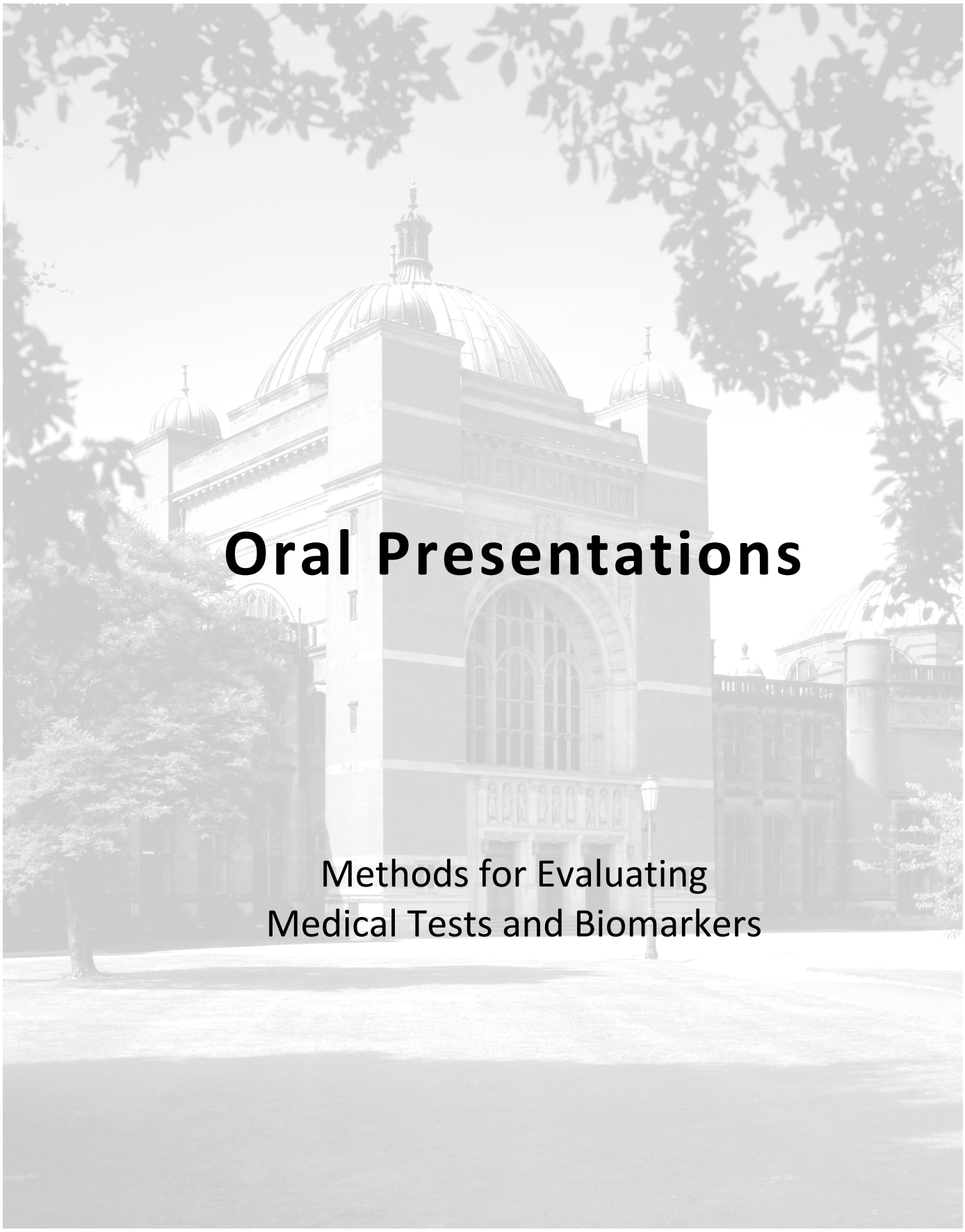
| | | |
|---|---|---|
| 10:30 – 11:00 | The issues of implementing IVD technology in healthcare settings outside the hospital<br>*Malcolm Luker (Keynote Speaker)* | |
| 11:00 – 11:20 | Exploration of methods to analyse MRMC diagnostic studies using CT colonography: why ROC AUC is not the answer<br>*Susan Mallett* | |
| 11:20 – 11:40 | Evaluation of longitudinal biomarkers<br>*Ruwanthi Kolamunnage-Dona* | |
| 11:40 – 12:00 | Assessing the additional value of diagnostic markers: a comparison of traditional and novel measures<br>*Ewout Steyerberg* | |
| 12:00 – 12:45 | *Lunch* | ***Lower Ground Floor Wolfson Centre*** |

**Session 7 - Issues in verification and reference standards - Leonard Deacon Lecture Theatre**

| | |
|---|---|
| 12:45 – 13:05 | Correcting for partial verification bias: a comparison of methods<br>Joris de Groot |
| 13:05 – 13:25 | PET/CT in cancer: Moderate sample sizes may suffice to justify replacement of a regional Gold Standard<br>*Oke Gerke* |
| 13:25 – 13:45 | Using patient management as a proxy for patient outcomes in test evaluation<br>*Lukas Staub* |
| 13:45 – 14:05 | Adjusting for differential verification bias in diagnostic accuracy studies: a Bayesian approach<br>*Joris de Groot* |
| 14:05 – 14:30 | *Afternoon Coffee/Tea and Poster Viewing* **Ground Floor *Wolfson Centre*** |

**Session 8 – Monitoring - Leonard Deacon Lecture Theatre**

| | |
|---|---|
| 14:30 – 15:00 | How to choose the best test for clinical monitoring<br>*Les Irwig (Keynote Speaker)* |
| 15:00 – 15:20 | Monitoring intraocular pressure as a marker of glaucoma<br>*María Vásquez Montes* |
| 15:20 – 15:40 | Can we assess adherence to medication by measuring change in blood pressure and cholesterol?<br>*Andrew Hayen* |
| 15:40 – 16:00 | How well is monitoring of risk factors reported in clinical guidelines?<br>*Ivan Moschetti* |
| 16:00 – 16:10 | Closing remarks<br>*Patrick Bossuyt* |



# Oral Presentations

Methods for Evaluating
Medical Tests and Biomarkers

*Keynote speaker*

## Evaluation of diagnostic and prognostic (bio)markers and prediction models: from development to implementation

Karel Moons, Prof. Clinical Epidemiology, Julius Centre for Health Sciences and Primary Care, Netherlands.

Reports of novel tests or (bio)markers, either prognostic or diagnostic, are abundant in the medical literature. They range from simple blood or urine tests/markers to those obtained from e.g. genomics, proteomics and imaging techniques. They greatly vary in accuracy, invasiveness of their measurement, and costs. The challenge is to optimally exploit existing and new tests/markers/models. I discuss why and how such markers need to be evaluated in their clinical context and whether they truly add new information on patients. I notably discuss the design and analysis of quantifying the added value of (new) markers/tests using multivariable modelling, how to validate such models in new patients, and how to implement the results in clinical practice. I aim to provide an overview of the process of these scientific evaluations, to further guide researchers, practicing physicians and laboratory workers involved in the study or application or diagnostic and prognostic tests or markers.

**Contact details:** k.g.m.moons@umcutrecht.nl

**Notes**

*Keynote speaker*

## Implementing clinical prediction rules in primary care: evidence, challenges and possible solutions

Tom Fahey, HRB Centre for Primary Care Research, RCSI Medical School, Division of Population Health Sciences, Dublin, Ireland.

The potential of improving patient care by using clinical prediction rules (CPRs) is not yet realised. Implementing CPRs in clinical practice means that several barriers have to be addressed and overcome including finding; understanding and applying CPRs at the time of clinical contact with the patient. The HRB Centre for Primary Care Research (www.hrbcentreprimarycare.ie) is a nationally funded research programme that aims to develop a register of CPRs that will be available via the Cochrane Primary Health Care Field. CPRs will be classified according to level of evidence (derivation, validation, impact) as well as clinical domain. Searching will be facilitated by means of recognised primary care taxonomy (International Classification of Primary Care) and the register will be updated on a regular basis. In terms of further evidence synthesis and implementation, Computer Based Clinical Decision Support Systems (CDSSs) are being developed that will integrate into the Electronic Health Record and enable use of CPRs in real-time clinical practice. Critical discussion of the likely barriers and potential solutions to enhancing the use of CPRs in this way will be discussed using clinical examples as the focus for the presentation.

**Contact details:** tomfahey@rcsi.ie

**Notes**

*Contributed paper*

## Biomarkers of tolerance in renal transplantation: pipeline from discovery to translation into the clinic

Irene Rebollo Mesa, David A. Stephens, Robert I. Lechler, Maria Hernandez Fuentes

Background: Identifying patients in whom immunological tolerance is established or is developing would allow an individually tailored approach to post-transplant management of kidney allograft recipients.

Aim: To this end we wanted to define a set of biomarkers of transplantation tolerance.
We undertook a multi-centre study aimed at identifying a "signature" of clinical transplant tolerance combining several biomarkers and bioassays.

Measures and Participants: The discovery study included a European training set (TrS), and an American test set (TS) of kidney transplant recipients. The TrS included 11 tolerant patients (Tl), and 60 more classified as either stable (St) or chronic rejectors (CR), as well as 19 healthy controls (HC). The TS comprised 89 patients of which 24 where Tl, and 31 HC. For the validation study, new patients are being recruited in both cohorts (20 for each group Tl, St and CR) , and the former patients are still followed up.
A series of biomarkers where measured in peripheral blood, including gene expression, immunophenotyping and anti-donor responses.

Methods: 1) Discovery Study: Significantly altered expression detected by microarray was statistically determined using four-class analysis and the Kruskal-Wallis test with Benjamini-Hochberg adjustment for False Discovery Rate (FDR) at 1%, using the TrS. A similar procedure was used to rank the biomarkers. We then evaluated the predictive power of the top 10-most differentially expressed genes, and the top 4 biomarkers, using receiver operating characteristic (ROC) curves. In order to combine their predictive power, we included them as predictors in a binary regression model to perform classification within sample. The binary regression procedure was used to compute probabilities p[1],….,p[n] of being a Tol-DF patient for each subject. The ROC curve was produced by varying a probability threshold between zero and one; for each value of the threshold t, a 2x2 classification table of Actual class versus Predicted class for subject i set equal to "Tol-DF" if p[i] > t. In order to select the most predictive subset of biomarker-predictors, and ensure the stability of the solution in other samples, we used 5-fold cross validation. We then validated the best subset by estimating the probability of tolerance in the TrS, using the parameters obtained in the TS. We used a ROC curve, and evaluated the optimal sensitivity and specificity provided by the optimal cut-off estimated in the TrS. 2) Validation Study for Translation into the Clinic: The primary objective is to observe the frequency of patients that display the tolerance signature as defined in the discovery study. We expect 80% of the new Tl patients to be classified as such; 90% of the previous cohort of Tl to maintain the signature; and 20% of patients with stable function, on standard immunosuppression 5 years after transplant to express the biomarkers of tolerance.

Results: The diagnostic capabilities of the combined results of several of the above mentioned biomarkers and bioassays result in a high specificity and sensitivity both in the training set and the test set. We are currently analysing the newly collected data from the validation study, to explore the possibility of using the discovered tolerance signature to inform drug weaning protocols in kidney transplant recipients.

Contact: irene.rebollo_mesa@kcl.ac.uk

**Notes**

*Contributed paper*

## Validation and refinement of a model to predict risk of endometrial cancer in patients with postmenopausal bleeding

Merel Breijer, Anne Timmermans, Helena van Doorn, Lil Valentin, Ben Willem Mol, Brent Opmeer

Objective: We previously showed that a multivariable prediction model based on clinical characteristics accurately estimated the probability of endometrial carcinoma in women presenting with postmenopausal bleeding[1]. The current study aims to externally validate this model and to derive and validate a simplified risk score to help clinicians in the management of women with postmenopausal bleeding.

Methods: The dataset used for validation was obtained for an individual patient data (IPD) meta-analysis to determine the diagnostic accuracy of endometrial thickness measurement in the detection of endometrial cancer among women with postmenopausal bleeding. We included only studies for which the following clinical characteristics were available: age, diabetes mellitus, use of anticoagulants, and body mass index (BMI). Since the original model was developed in patients not using hormone replacement therapy (HRT) those patients were excluded. We used multiple imputation to deal with missing data within studies. Prognostic accuracy was quantified with the area under the ROC-curve (AUC). Negative predictive value (NPV) for different cut-off values of estimated probability was calculated. A new risk score based on the original multivariable model was derived and internally and externally validated. Points (pts) were assigned to each patient for the presence of diabetes mellitus (6 pts), obesity (BMI>26 kg/m2, 5 pts), for age: every 4 years over 55 (1 pt) or abstract points for every year under 55 (-2 pts), for the use of anticoagulants points were abstracted (-4 pts).

Results: For the original IPD meta-analysis, 79 primary investigators were contacted of which 14 could provide data. Besides the dataset used for derivation of the model (540 patients), three out of these 14 authors provided information on patient characteristics of 1194 patients. After excluding patients with HRT data on 870 patients was available for validation purposes. The prognostic value of the model in the three validation cohorts (AUC 0.74, 0.80 and 0.81, respectively) was comparable to the accuracy found in the derivation cohort (AUC 0.72). Negative predictive value was 100% for a cut-off value of estimated cancer risk of 1% and 2% and 99% for a cut-off value of estimated cancer risk of 3%, 4% and 5%. The accuracy of the risk score was similar to the accuracy of the original model in the derivation dataset (AUC 0.73 vs. 0.72) and in the three validation cohorts (AUC 0.75, 0.80 and 0.81 respectively).

Conclusion: The existing model for the probability of endometrial cancer risk in patients with postmenopausal bleeding was found to maintain its diagnostic accuracy in multiple independent cohorts. Transforming the model into a simplified risk-score usable in clinical practice did not alter its accuracy. This risk score could be used on its own, e.g. to triage women that can be directly reassured or referred to a gynecologist, or in combination with endometrial thickness measurement, where the optimal positivity thresholds could depend on women's individual risk score.

1. Opmeer B, van Doorn H, Heintz A, Burger C, Bossuyt P, Mol B. Improving the existing diagnostic strategy by accounting for characteristics of the women in the diagnostic work up for postmenopausal bleeding. BJOG 2007;114:51-58.

Contact: M.C.Breijer@amc.uva.nl

**Notes**

*Contributed paper*

**Bivariate meta-analysis of predictive values**

Mariska Leeflang, Lotty Hooft, Johannes Reitsma, Jon Deeks, Patrick Bossuyt

Background: It is common in meta-analysis of test accuracy studies to obtain summary estimates of sensitivity and specificity, or to produce a summary ROC curve. Usually, meta-analyzing estimates of positive predictive values directly is discouraged, because they tend to vary more with changes in prevalence. However, as there is anecdotal evidence that sensitivity and specificity can also vary with changes in prevalence, it is unclear whether meta-analyzing predictive values directly will produce different results than first obtaining summary estimates of sensitivity and specificity and calculating predictive values by Bayes theorem.

Objective: To compare the conventional bivariate logitnormal model for the meta-analysis of test accuracy studies, which results in summary estimates of sensitivity and specificity, with an alternative bivariate logitnormal model of the positive and negative predictive value.

Methods: From a set of 30 meta-analyses, containing 487 diagnostic accuracy studies, we selected meta-analyses included a consecutive series of eligible patients. We obtained summary estimates of sensitivity and specificity for each review, using the bivariate logitnormal method for meta-analysis. The mean prevalence was used to calculate positive and negative predictive values. The same model was also used to directly obtain summary estimates for positive and predictive values. The final estimates for predictive values were compared, as well as the -2 log likelihood to compare how the models fitted the data.

Results: Sixteen reviews fulfilled our criteria and allowed both the conventional bivariate logitnormal model and the bivariate model for predictive values to fit the data. Of these 16 reviews, 10 showed a lower -2LL for the predictive values model, and 6 showed a lower -2LL for the conventional model. The estimated predictive values did not differ significantly.

Discussion: Our results do not show a significant preference for either the conventional model or the bivariate model for predictive values. Although modelling predictive values may result in outcomes that are preferred by clinicians, the question remains whether for example effects of covariates on predictive values can be interpreted the same way as the effects of these covariates on sensitivity and specificity.

Further research: Recently, a trivariate model of sensitivity, specificity and prevalence was published. We will compare the models above with this model as well and present those results during the symposium.

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

**Notes**

*Contributed paper*

## Cervical length measurement for the prediction of preterm birth in multiple pregnancies: a systematic review using bivariate meta-analysis on stratified bootstrap samples to deal with studies reporting multiple accuracy estimates

Brent Opmeer, Arianne Lim, Maud Hegeman, M.A. Huis- in 't Veld, Hein Bruinse, Ben-Willem Mol

Objective: Midpregnancy cervical length (CL) is thought to be associated with the risk of preterm birth, and its measurement with transvaginal sonography could be used to predict preterm birth (PTB). We aimed to assess the predictive accuracy of cervical length in twin pregnancies in a systematic review of the literature.

Methods: We identified studies that reported on sonographically measured CL as a predictor of preterm birth in twin and higher multiple pregnancies. We searched studies on women with a multiple pregnancy, in which CL was measured during pregnancy, and in whom gestational age at birth (GAB) was known.
Two-by-two tables cross-classifying CL and PTB before 34 weeks were extracted. Relevant study characteristics were documented, including quality and design, gestational age used to define PTB (reference test), and gestational age (GA) at which CL was measured (time).

The bivariate model was used to estimate sensitivity, specificity and sROCs using Proc NLMixed in SAS 9.1. As the basic model does not accommodate correlated data from studies reporting results for multiple thresholds for the index test (CL), for the reference test (GAB), and/or for time, we used average model estimates based on 50 stratified samples, each with a single bivariate datapoint (sensitivity/specificity) per study. We estimated overall accuracy for different CL thresholds, irrespective of PTB definition or timing, as well as accuracy for different definitions of PTB (<30 weeks; 30-34 weeks; 34-37 weeks) and different timings (<20 weeks; 20-24 weeks; > 24 weeks), using the same procedure on subsets of the sample.

Main Results: We identified 12 studies reporting the predictive accuracy of CL for PTB, for a total of 1642 pregnant women. Overall accuracy was moderate, with sensitivity increasing and specificity decreasing with increasing C-L thresholds (mean; 95%CI): 20mm (0.30; 0.13 to 0.55; 0.94; 0.77 to 0.99), 25mm (0.36; 0.10 to 0.75; 0.94; 0.86 to 0.97), 30mm (0.41; 0.22 to 0.63; 0.87; 0.85 to 0.89) and 35mm (0.78; 0.42 to 0.94; 0.66; 0.44 to 0.83), respectively. Accuracy was slightly better for more extreme criteria for PTB (< 30 weeks) and for CL measured later in GA (> 24 weeks). sROC curves reflecting these differences will be presented.

Discussion: The basic model to simultaneously estimate sensitivity and specificity with the bivariate model cannot accommodate multiple observations per study. In order to assess the impact of study characteristics, we used repeated stratified sampling to adjust for these within study dependencies. Until researchers can accommodate the correlational structure of such more complex data structures from diagnostic test accuracy reviews, the proposed approach could be a feasible and acceptable alternative.

Contact: B.C.Opmeer@amc.uva.nl

**Notes**

*Contributed paper*

## An empirical assessment of the validity of uncontrolled comparisons of the accuracy of diagnostic tests

Yemisi Takwoingi, Mariska Leeflang, Mary Pennant, Jac Dinnes, Jon Deeks

Background: Diagnostic test accuracy reviews can provide evidence to support the selection of diagnostic tests by comparing the performance of tests or test combinations. Studies that directly compare tests within patients or between randomized groups are preferable but are uncommon. Consequently, between-study uncontrolled (indirect) comparisons of tests may provide the only evidence of note. Such comparisons are likely to be more prone to bias like indirect comparisons between healthcare interventions (Glenny et al., 2005), and maybe more severely due to considerable heterogeneity between studies and the lack of a common comparator test.

Objectives: To estimate bias and reliability of meta-analyses of uncontrolled comparisons of diagnostic accuracy studies compared to those of comparative studies.

Methods: Meta-analyses that included test comparisons with both comparative studies and uncontrolled studies were identified from a cohort of higher quality diagnostic reviews (Dinnes et al., 2005) indexed in the Database of Reviews of Effects up to December 2002 supplemented by more recent searches. The hierarchical summary ROC model was used for meta-analysis to estimate and compare accuracy measures. The degree of bias and variability of relative sensitivities, specificities and diagnostic odds ratios (DOR) between comparative and uncontrolled comparisons was assessed

Results: Twenty-two comparative reviews with sufficient data to conduct both direct and uncontrolled test comparisons were identified. Large discrepancies in test performance between direct and uncontrolled comparisons were found with a halving or doubling of the relative DOR in about 50% of reviews. Direction of the bias was unpredictable. A trade-off between sensitivity and specificity was also observed.

Conclusions: Test selection is critical to health technology assessment. In the absence of comparative studies, selection has often relied on comparisons of meta-analyses of uncontrolled studies. Limitations of such comparisons should be considered when making inferences on the relative accuracy of competing tests, and in encouraging funders to ensure future test accuracy studies address important comparative questions.

Contact: y.takwoingi@bham.ac.uk

**Notes**

*Contributed paper*

## Meta-analysis of the accuracy of sequences of diagnostic tests by allowing for between tests correlation

Nicola Novielli, Nicola Cooper, Alex Sutton, Keith Abrams

Objective: To develop a Bayesian modelling approach to the meta-analysis of the accuracy of sequences of diagnostic tests by relaxing the assumption of independence between tests.

Background: The inclusion of the accuracy of diagnostic tests into economic decision models is crucial to the correct evaluation of the cost-effectiveness of diagnosis-to-treatment pathways. However, there is debate on which information should be used to inform such decision models; whether single studies or systematic reviews need to be used. From our perspective, meta-analyses allow all the available information to be included and, within the Bayesian framework, full account of uncertainty into parameter estimation. Diagnostic strategies are rarely composed by one single test. Often, a sequence of tests is used instead (Fig 1).

Example dataset: We performed a systematic review of the accuracy of DDimer and Wells Score in combination for the accuracy of DVT (Deep Vein Thrombosis). Six different types of data have been extracted from 21 studies, comprising complete and partial count data for DDimer, proportion for DDimer, count data for Wells score.

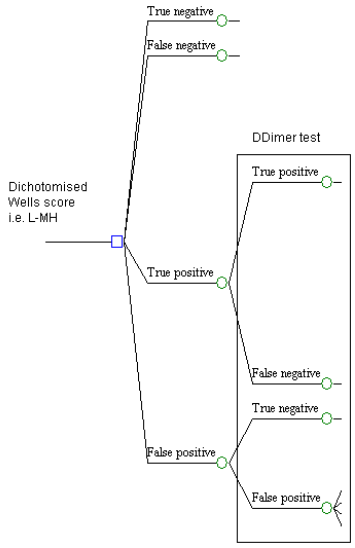



Figure 1. Example of a sequence of Dichotomised diagnostic tests

Results: Our novel random effect modelling approach is based on the bivariate random effect model first adapted to diagnostic data by Reitsma (2005) and 1) incorporates all the available evidence into the same model; 2) deals with the issue of multiple thresholds (Wells score test); 3) fully accounts for missing data; 4) incorporates the heterogeneity due to explicit or implicit threshold variability (DDimer); 5) includes an estimate of the between study heterogeneity.

Conclusions: A proper evaluation of the accuracy of diagnostic sequences potentially produces less biased economic evaluations, leading to efficacious decision-making. However, a direct estimation of the between test correlation does not exist yet. Our research is moving in this direction.

Reitsma, J. B., A. S. Glas, A. W. S. Rutjes, R. J. P. M. Scholten, P. M. Bossuyt, and A. H. Zwinderman. 2005. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology 58 (10):982-990.

Contact: nn40@leicester.ac.uk

**Notes**

*Keynote speaker*

**Health economic assessment of diagnostic tests**

Matt Stevenson, Senior Research Fellow, School of Health and Related Research, University of Sheffield, UK

The National Institute for Health and Clinical Excellence have produced a clear reference case for undertaking assessments of pharmaceuticals. A small number of diagnostic technologies have been appraised through this route, however, a bespoke process has been deemed preferable, leading to the formation of the Diagnostic Advisory Committee, a methods working group to provide a diagnostic reference case and a pilot evaluation. This talk will focus on the key differences between the appraisal of diagnostic tests and the appraisal of pharmaceuticals and highlight potential areas for methodological development with references to the pilot evaluation and other diagnostic evaluations undertaken.

Contact: m.d.stevenson@sheffield.ac.uk

**Notes**

*Keynote speaker*

## Assessing the cost effectiveness of using prognostic biomarkers: a case study in prioritising patients waiting for coronary artery surgery

Steve Palmer, Senior Research Fellow, Centre for Health Economics, York University, UK.

Objective: To determine the effectiveness and cost effectiveness of a range of strategies based on conventional clinical information and novel circulating biomarkers for prioritising patients with stable angina awaiting coronary artery bypass grafting (CABG).

Methods: We carried out systematic reviews and meta-analyses of literature-based estimates of the prognostic effects of circulating biomarkers in stable coronary disease. The cost-effectiveness of prioritising patients on the waiting list for CABG using circulating biomarkers was compared against a range of alternative formal approaches to prioritisation as well as no formal prioritisation. A decision-analytic model was developed to synthesise data on a range of effectiveness, resource use and value parameters necessary to determine cost-effectiveness.

Results: We included 390 reports of biomarker effects in our review. The quality of individual study reports was variable, with evidence of small study (publication) bias and incomplete adjustment for simple clinical information such as age, sex, smoking, diabetes and obesity. The risk of cardiovascular events while on the waiting list for CABG was 3 per 10,000 patients per day within the first 90 days (184 events in 9935 patients with a mean of 59 days at risk). Risk factors associated with an increased risk, and included in the basic risk equation, were age, diabetes, heart failure, previous myocardial infarction and involvement of the left main coronary artery or three-vessel disease. The optimal strategy in terms of cost-effectiveness considerations was a prioritisation strategy employing biomarker information. Evaluating shorter maximum waiting times did not alter the conclusion that a prioritisation strategy with a risk score using estimated glomerular filtration rate (eGFR) was cost-effective.

Conclusion: Formally evaluating the clinical and cost effectiveness of prognostic biomarkers is important even when effects at an individual level are small. Routinely employing more information in the prioritisation of patients awaiting CABG appears to be a cost-effective approach and may result in improved health outcomes. The most cost-effective strategy employed a risk score using conventional clinical information together with a single biomarker (eGFR). The additional prognostic information conferred by collecting the more costly novel circulating biomarker CRP, singly or in combination with other biomarkers, in terms of waiting list prioritisation is unlikely to be cost-effective.

Contact: sjp21@york.ac.uk

**Notes**

*Contributed paper*

## Using a natural history model in combination with screening data to estimate the test characteristics of the FOB test in the English bowel cancer screening programme

Sophie Whyte, Cathal Walsh, Jim Chilcott

Background: Due to the invasive nature of colonoscopy large scale trials of the faecal occult blood test (FOBt) in which all persons receive colonoscopy are unfeasible. Hence, methods used to estimate FOBt characteristics are subject to limitations and the range of estimates is large. We used a natural history model in combination with screening data to estimate the FOBt characteristics for the English bowel cancer screening programme. This approach is preferable as the model implicitly includes asymptomatic cancers, colonoscopy miss rates and embeds the problem in the framework of Bayesian inference.

Methods: A published Excel Markov state transition model for the natural history of colorectal cancer was modified and updated with recent English data. The Metropolis Hastings algorithm was used to estimate the natural history parameters and FOBt test characteristics by generating multiple sets of parameters. These parameter sets form the posterior distribution which is compatible with the observed data, so accurately represent parameter uncertainty. Observed data included incidence categorised by age and stage, autopsy data on polyp prevalence, and FOBt positivity rates and cancer and polyp detection rates from the first round of screening. Due to the methodology used, correlated parameter sets are produced which are used for probabilistic sensitivity analyses.

Results: The estimates for sensitivity of FOBt to polyps and cancer were 0.08 (95% credible interval 0.06, 0.13) and 0.35 (0.33, 0.39) respectively. The estimate for the specificity of FOBt was 0.9945 (0.9940, 0.9952). The values for FOBt sensitivity obtained were comparable to those reported elsewhere. The value for specificity was higher than that reported elsewhere which may be due to the referral threshold which is used within the England bowel cancer screening programme.

Conclusions: Our analysis demonstrates that estimates for FOBt characteristics can be produced using a natural history model and screening data. These estimates are not subject to the limitations associated with trials and are directly relevant to the English bowel cancer screening programme.

Contact: sophie.whyte@sheffield.ac.uk

**Notes**

*Contributed paper*

## Cost-effectiveness analysis of diagnostic tests or biomarkers: an example study of transesophageal echocardiography to diagnose the presence of ascending aortic atherosclerosis in cardiac surgery patients

Hendrik Koffijberg, Bas van Zaane, Arno Nierich, Karel Moons

Background: Most diagnostic tests or biomarkers are (only) evaluated on their discriminative or predictive accuracy. However, such (commonly cross-sectional) accuracy studies do not provide direct evidence on the tests ability to change patient outcomes, let alone on their cost-effectiveness. Recent guidelines suggest that a proper cost-effectiveness analysis of (novel) tests or biomarkers should be done before implementation in practice, using randomised diagnostic strategy studies. We present an example study how the cost-effectiveness of diagnostic tests/markers can be done without such randomised studies but making use of cross sectional accuracy study data and therapeutic intervention studies on patient outcome. The example study concerns post-operative stroke in on-pump cardiac surgery. This is often caused by emboli merging from the atherosclerotic ascending aorta (AA) after manipulation. A new diagnostic transesophageal echocardiography showed to accurately determine the presence (and extent) of atherosclerosis prior to sternotomy, such that it allows the surgeon to change surgical strategy to reduce/avoid emboli production. We here assessed the tool's cost-effectiveness when it would be applied in practice, as compared to care as usual (manual palpation for detection of AA atherosclerosis).

Methods: A Markov decision-analytical model, using input from the cross sectional accuracy study and longitudinal patient outcome studies, was developed used to assess differences in costs and health effects between the two strategies. The incremental cost-effectiveness ratio was calculated for various subgroups. Conservative as well as observed prevalence rates of AA atherosclerosis were defined per subgroup. Probabilistic sensitivity analysis was used to determine the robustness of the model results.

Results: Using the new tool consistently resulted in more adapted procedures and, consequently, in a lower risk of stroke and a (slightly) higher number of life-years. The incremental costs decreased whereas incremental effects increased with patient age. The incremental costs-effectiveness ratio (ICER) ranged from €4937/QALY for 55-year-old men to €-6191/QALY for 75-year-old women.

Conclusions: The new tool reduces costs and increases health benefits in patients older than 65 years. In some subgroups the additional costs will likely be small compared with the additional health benefits. More general: cost-effectiveness estimations of diagnostic tests/biomarkers can be inferred without specifically designed randomised diagnostic studies. Moreover, such estimations may guide more efficient designs of randomised diagnostic studies.

Key reference:

van Zaane B, Nierich AP, Buhre WF, Brandon Bravo Bruinsma GJ, Moons KGM: Resolving the blind spot of transoesophageal echocardiography: a new diagnostic device for visualizing the ascending aorta in cardiac surgery. Br. J. Anaesth. 2007; 98: 434-41.

Contact: h.koffijberg@umcutrecht.nl

**Notes**

*Contributed paper*

## Sensitivity and specificity rarely vary with prevalence

Mariska Leeflang, Lotty Hooft, Johannes Reitsma, Patrick Bossuyt

Background: Anecdotal evidence shows that sensitivity and specificity tend to vary with prevalence, but empirical studies systematically examining this relationship are not available. Therefore, our objective was to investigate the impact of prevalence on sensitivity and specificity in a large set of meta-analyses of diagnostic accuracy studies.

Methods: From a set of 30 meta-analyses, containing 487 diagnostic accuracy studies, we selected meta-analyses with at least 10 studies that had included a series of equally suspected patients. We obtained summary estimates of sensitivity and specificity for each review, using the bivariate logitnormal method for meta-analysis, including prevalence as a continuous covariate (linear, quadratic relationships) to investigate its effect on sensitivity and specificity.

Results: Twenty-three reviews fulfilled our criteria. They contained between 10 and 39 studies, varying in prevalence from 0.1% to 98% (IQR 15.7% to 61.8%). Overall, prevalence had a significant effect on both sensitivity and specificity. However, when analyzing the reviews separately, prevalence had a significant effect on both sensitivity and specificity in only two reviews. In six reviews, prevalence had a significant effect on specificity only. In none of the reviews prevalence had a significant effect on sensitivity only.

Conclusions: Diagnostic accuracy may change with varying prevalence, but it is not common to find significant effects on accuracy statistics in individual systematic reviews. This may in part be due to the limited power of meta-regression to explain heterogeneity with study level covariates, aggravated by the generally low prevalence (median 37%; 25th percentile 16%).

Further investigations: Because the main underlying mechanism how prevalence may alter diagnostic accuracy are differences in disease spectrum and case-mix, we will further investigate these relationships as well and present those data during the symposium.

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

**Notes**

*Contributed paper*

## Empirical evidence that disease prevalence does affect diagnostic test performance – the effects of prevalence on the interpretation of x-rays by junior doctors

Brian Willis

Background: Recently it has been recognised that the sensitivity and specificity may be affected by the prevalence of disease[1].

Objective: To evaluate the effects of the prevalence of abnormality on junior doctors' performance in interpreting x-rays.

Method: A systematic sample of 2593 patients' records was collected from an attending cohort at a UK emergency department. In the sample 1053 x-rays were interpreted by junior doctors following their triage into high and low probability of abnormality populations by radiographers. Following exclusion of 86 x-rays due to incomplete data, 967 were analysed. The main outcomes were sensitivity, specificity, likelihood ratios, diagnostic odds ratios and ROC curve.

Results: For the high probability (77%) and low probability (13%) x-rays, the overall sensitivities were 94.1 (89.5 – 96.8) and 75.0 (65.5 - 82.6) and the specificities were 56.0 (42.3 - 68.8) and 92.3 (90.0 - 94.1) respectively. These were statistically significant as were the differences in the positive likelihood ratio between the two populations: 2.14 (1.56 - 2.93) and 9.78 (7.32 - 13.07) respectively. To assess potential differences in patient spectrum between the high and low probability populations the distributions of x-ray type were compared. Although in some individual categories there were differences in the proportions of x-rays examined, logistic regression showed that overall, x-ray type did not significantly affect test performance. Differences in the diagnostic odds ratios were also found not to be significant which is consistent with a common ROC curve and with doctors changing their implicit threshold between the two populations.

Conclusions: The results support the hypothesis that the sensitivity and specificity of a diagnostic test may be affected by the disease prevalence. More generally this has implications for clinicians when applying multiple tests, as each test leads to a revision of the probability of disease, and this may influence the performance of subsequent tests. Ultimately this may affect the transferability of study results to practice.

[1] Leeflang MM, Bossuyt PM, Irwig L., Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis, J Clin Epidemiol. 2009 62(1):5-12

Contact: brian.willis@manchester.ac.uk

**Notes**

*Contributed paper*

## What do 'test-treat' trials measure?

Lavinia Ferrante di Ruffano, Chris Hyde, Jonathan Deeks

Background: Will introducing a new diagnostic test benefit patient health? This question underpins the evaluation of diagnostic tests. Trials that randomise patients to diagnostic strategies (RCTs) and evaluate patient outcomes after the implementation of subsequent treatment, are the methodological gold-standard. However RCTs are only as informative as the outcomes they measure. This choice is particularly tricky in trials comparing test-treatment interventions, since outcomes must evaluate the effects of a whole care pathway: diagnosis, consequent decision-making and treatment. Outcomes that fail to capture the effects of these complex interventions are likely to produce misrepresentative evaluations of performance.

Aim: To comprehensively map out how a test can impact patient health.

Methods: An explanatory framework was developed in two stages: firstly by reference to existing diagnostic test research frameworks in order to create a hypothetical model of the test-treat process, and secondly through a systematic analysis of randomised test-treatment comparisons. Published test-treatment RCTs were identified through the Cochrane Central Register of Controlled Trials (Issue 2 2009, years 2004-7). Included studies evaluated test-treat strategies of any diagnostic modality and assessed patient outcomes after treatment.

Results: The authors present a comprehensive framework that sets out 13 mechanisms through which a diagnostic test can change patient health. Mechanisms are grouped according to the manner in which they exert their effect: direct effects of tests, altered decisions and actions, changed timeframes and altered perceptions; these routes of effect are illustrated using real clinical questions addressed by the 78 randomised trials identified from the world-wide literature. A summary of which of these routes test-treatment trials have measured is also presented.

Conclusion: Diagnostic tests can influence health in ways more numerous and complex than often recognised. Our framework emphasises the risks of focussing on evaluations of accuracy and decision-making, since in doing so we risk missing other important potential benefits such as when a test alters the way in which treatment is delivered. The framework is presented as a tool for identifying all likely processes of change in any test-treat comparison, so aiding the selection of important outcomes for measurement. It is recommended that future test evaluations incorporate direct measurement of each potentially relevant mechanism to assist in the interpretation of results, particularly to differentiate truly ineffective diagnostic interventions from underpowered studies and poorly implemented test-treatment strategies. An accompanying checklist has been designed for guiding the evaluation of test-treatment strategies: for assessments of existing evidence, identifying the need for new primary studies, and for informing trial design.

Contact: l.ferrantediruffano@bham.ac.uk

**Notes**

*Contributed paper*

## Identification of additional effects of medical testing: towards more comprehensive test evaluation

Jolande Vis, Myra van Zwieten, Patrick Bossuyt, Karel Moons, Marcel Dijkgraaf, Kirsten McCaffery, Ben-Willem Mol, Brent Opmeer

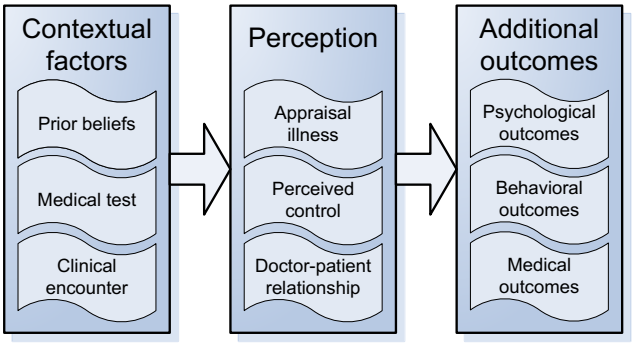
Background: At present, diagnostic and prognostic tests are mainly evaluated for their capacity to support clinical decision making. Yet, tests may also affect patients' outcomes regardless of clinical decisions. For instance, tests may reassure patients or motivate health behaviour change. Ignoring such additional effects could lead to incomplete or biased test evaluations and thus result in suboptimal use of tests in clinical practice.

Objective: To evaluate whether medical testing can influence patients' clinical outcomes other than through their effect on clinical decision making.

Design: Qualitative study using focus groups with medical professionals. Four focus groups consisted of gynaecological residents, three groups of gynaecologists and two groups of gynaecological M.D. researchers (n=43). The focus groups were audio recorded and transcribed verbatim. Focus group transcriptions were coded inductively by three independent researchers. Recurrent themes were identified and analysed. The data were used to construct a framework of additional effects of medical testing.

Results: All participants contributed various clinical examples in which medical testing led to additional effects on patients' outcomes. Recurrent themes of additional effects were found throughout the data. The clinical examples illustrated that test results themselves or in interaction with the context of the test situation may produce additional effects. Our data showed that testing may influence patients' illness perception, their perceived control, or affect the doctor-patient relationship. This may lead to changes in the emotional state of patients, their behaviour and/or medical outcomes. We found that additional effects may favorably or unfavorably affect patients, regardless of potential effects on clinical decision making. The figure shows a conceptual framework of additional effects of testing based on the data from the focus groups.

Conclusions: Besides supporting clinical decision making, medical testing may have additional effects on health in both positive or negative ways. The conceptual framework may trigger researchers to include additional effects of medical testing in future comprehensive test evaluations.




**Contact:** j.y.vis@amc.nl

**Notes**

*Contributed paper*

## Statistical design and preliminary analysis of eye tracking studies to investigate diagnostic performance in CT colonography

Sue Mallett, Darren Boone, Peter Phillips, Stuart A. Taylor, Michael Steward, Douglas G. Altman, David J. Manning, Steve Halligan

Introduction: Understanding the nature of missed diagnoses in radiological studies will provide insight into interventions to improve diagnostic performance. In CT colonography, readers examine a computer generated 3D flythrough video of colon for the presence or absence of lesions. Identification of a lesion depends on two steps - firstly the perception of a potential abnormality and secondly the classification of the abnormal area as a lesion or not. Advances in visual perception have enabled eye tracking of radiologists during the image reading process, and more recently whilst reading moving 3D images such as used in CT colonography. In a collaborative programme with research radiologists and visual perception experts, we are providing statistical design and analysis for a series of clinical studies to understand the nature of errors in radiologists' reading of 3D videos of the colon. Analysis of full patient video clips is planned to examine issues in reader fatigue.

Study design: A series of pilot experiments have been designed to record radiologists' visual search characteristics and their recognition of polyps recorded using a Tobii X50 eye-tracker using both short (30 sec) and long (15 min) clips of 3D flythrough CT colon videos. Patients with a mix of previously reported true and false-positive lesions were selected from previous studies to enrich for videos likely to include both types of errors for missed diagnoses. Characteristics and parameters of visual search patterns are being analysed.

Results: Preliminary examination of pilot experiments has proven proof of concept. In three radiologists reading the same video, we were able to identify a radiologist who did not record identification of a lesion, but where the eye tracking showed that eye dwell was co-incidental to the lesion. This suggests that the source of reader error in this instance was not in the visual perception, but in the reader classification of the lesion. We will discuss issues in the design of these investigations and present results from a series of readers.

Conclusions: Based on this proof of concept, it is technically feasible to collect data on visual search patterns during interpretation of 3D videos of CT colonography. Further studies are planned in a more representative selection of patient videos, to obtain insight into the nature of errors in diagnostic accuracy and to examine reader fatigue.

Contact: susan.mallett@csm.ox.ac.uk

**Notes**

*Keynote speaker*

**Diagnosis in NICE clinical guidelines**

Phil Alderson, Associate Director, Centre for Clinical Practice, National Institute for Clinical Excellence, UK

Diagnostic questions feature in the majority of NICE clinical guidelines, sometimes being the dominant area of interest in a guideline. Diagnostic topics have been dealt with in a variety of ways, from literature review of test accuracy through to complex modelling of pathways. Some examples will be presented, and the practical and technical problems that have been encountered will be summarised, together with some thoughts about the future.

**Contact:** philip.alderson@nice.org.uk

**Notes**

*Contributed paper*

## Commissioning reviews of diagnostic test accuracy: Lessons from a systematic review of positron emission spectroscopy and positron emission spectroscopy/computed tomography

Mary Pennant, Clare Davenport, Chris Hyde

Background: When reviewing test accuracy studies to inform practice recommendations, possible strategies include 1) reviewing single test accuracy studies, 2) using indirect comparisons to compare tests with those conventionally used or 3) using direct comparisons with conventional tests, where analysis is restricted to studies with within-study comparisons. The HTA commissioned a review of the test accuracy of positron emission spectroscopy (PET) and positron emission spectroscopy/computed tomography (PET/CT) for breast cancer (BC) recurrence and we attempted to address the question in the most informative way.

Methods: A systematic review of studies of PET or PET/CT for the detection of BC recurrence was conducted. Data for conventional imaging tests (CITs) was only included if it came from studies that also investigated the accuracy of PET or PET/CT.

Results: 1) Single tests: Estimates of sensitivity and specificity were 91% (CI 86-94%) and 86% (CI 79-91%) for PET (n=25) and 96% (CI 89-99%) and 89% (CI 75-95%) for PET/CT (n=5) respectively.
2) Indirect comparisons: A systematic review of the accuracy of all CITs in addition to PET and PET/CT was not feasible in the time available. For comparisons of PET and PET/CT with CITs in studies assessing PET or PET/CT, estimates for relative sensitivity and specificity were similar to those for direct comparisons.
3) Direct comparisons: Relative sensitivity and specificity were 1.12 (CI 1.04-1.21, p=0.005) and 1.12 (CI 1.01-1.24, p=0.036) for PET (n=10) and 1.19 (CI 1.03-1.37, p=0.015) and 1.15 (0.95-1.41, p=0.157) for PET/CT (n=4) compared to CITs respectively. Estimates for sensitivity and specificity from these studies were 89% (CI 83-93%) and 93% (CI 83-97%) for PET and 95% (CI 88-98%) and 89% (CI 69-97%) for PET/CT respectively and were similar to single test accuracy findings.

Conclusions: In this review, direct comparisons were considered to be the most useful measure for informing on the value of PET and PET/CT for current practice. True indirect comparisons were unfeasible due to the scale of research in this area and the proxy indirect comparisons made added little to the review. Single test accuracy measures alone were not useful to address the research question. However, these estimates were useful in assessing whether the smaller number of direct PET versus PET/CT comparisons gave representative estimates of the accuracy of PET and PET/CT.

Contact: m.pennant@bham.ac.uk

**Notes**

*Contributed paper*

## Defining the role of a diagnostic test does not (yet) allow to limit the level of eligible evidence

Stefan Sauerland, Fülöp Scheibler, Inger Janssen, Robert Grosselfinger, Milly Schröer-Günther, Stefan Lange

Background: The German Institute for Quality and Efficiency in Health Care (IQWiG) evaluates various types of medical interventions in order to inform national decisions on reimbursement. As many new and costly diagnostic tests have recently been developed (e.g. in oncology or radiology), much work at IQWiG has been devoted to methods that permit us to determine the benefits of diagnostic tests. The aim of this presentation is to summarise and analyse IQWiG's current concept of diagnostic test evaluation.

Methods: Using all completed and ongoing health technology assessments (HTAs) conducted by IQWiG, we examined 1) which role the diagnostic tests under consideration had and 2) what type of evidence was available. 1) follows the definitions by Bossuyt et al., who described three roles of a new diagnostic test (replacement, triage or add-on). 2) refers to the study types used in the assessment of diagnostic tests: various types of randomised controlled trials (RCTs), as well as non-randomised controlled clinical trials (CCTs), are suited to assess add-on or triage tests; diagnostic accuracy studies (DASs) are only suited to assess replacement tests.

Results: A total of 18 reports dealt with diagnosis and screening. Their proportion among all reports (drug and non-drug) increased over time from 13% (6 of 47 completed reports) to 52% (12 of 23 ongoing projects), mainly because of the rising interest in positron emission tomography (n= 10 new projects). The 18 reports addressed 19 key questions in which the role of the diagnostic test was either add-on (n= 11), triage (n= 7) or replacement (1). In 17 projects RCTs and CCTs were or will be searched for, mostly with additional consideration of DASs. (One project had just begun and was thus excluded.) Literature searches were limited to DASs only in the single question on test replacement. RCTs (n= 3) and CCTs (n= 8) were found for only 3 of the 7 research questions for which preliminary or final results were available. However, DASs were found for 6 of the 7 questions and their number was usually large (n> 10 in 4 questions).

Conclusions: It is crucial within the context of HTA to define the exact role of a new diagnostic test as add-on, replacement or triage, as this allows tailoring literature searches accordingly. Currently, however, most HTAs have to include DASs regardless of the role of the test. This problem is due to the scarcity of RCTs assessing diagnostic tests, but may also be caused by the initially unclear role of a new diagnostic test. As replacement tests are seldomly assessed by HTA, the relevance of DASs appears to be generally limited.

Reference: Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006; 332: 1089-92.

Contact: stefan.sauerland@iqwig.de

**Notes**

*Contributed paper*

## Developing evidence-based recommendations for tests and markers

Patrick Bossuyt

Background: As the availability of medical tests and markers increases and some of these carry hefty price tags, the need for developing evidence-based recommendations about their use grows. Several groups and health care systems have developed manuals for developing recommendations about the use and reimbursement of tests and markers, but it is unclear to what extent they agree in their methods.

Methods: We collected, analyzed and compared guidance documents for developing recommendations about tests and markers from the German Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, the NHS National Institute for Health and Clinical Excellence, the U.S. Preventive Services Task Force and the Agency for Healthcare Research and Quality, the Australian Medical Services Advisory Committee, the Dutch platform for Evidence-based Guideline Development, and the GRADE working group.

Results: All systems focus on the collection and synthesis of diagnostic accuracy data. Some systems do not move beyond diagnostic accuracy, while others systems more explicitly emphasize the need for patient outcome data, without making randomized clinical trial data necessary. Several systems allow the use of indirect data in analytic frameworks, although these systems are often more causal than comparative in nature.

Conclusion: Systems for developing evidence-based recommendations about testing vary. Diagnostic accuracy features in all of them, even though many tests are not used for diagnostic purposes. The use of analytic frameworks, which incorporate indirect evidence about patient benefit, deserves further development and possible harmonization.

Contact: p.m.bossuyt@amc.uva.nl

**Notes**

*Keynote speaker*

## The issues of implementing IVD technology in healthcare settings outside the Hospital

Malcolm Luker, CEO Philips Healthcare Incubator

In vitro diagnostics (IVD's) are widely used throughout professional healthcare settings to assist in the diagnosis and monitoring of a patient's condition. However the use of IVD's in non-professional healthcare settings is far lower, with only blood glucose monitoring having any widespread adoption. There have been a variety of proposals on how IVD's could be used effectively in home and other primary care settings to bring about an improvement in healthcare outcomes and avoid costly secondary care intervention, but the implementation of these initiatives has been negligible.

The talk will focus on two key issues that impact the adoption of remote patient monitoring (RPM):

- Evidence needs, including:
    - Current situation, including costs and accepted "issues"
    - What clinical data will be needed to secure adoption and payment of the system
    - What endpoints will be needed to demonstrate the value to potential users and payers
    - What health economic data must be captured in studies

- Reimbursement factors & Payment models
    - What are the drivers for reimbursement
    - What will the costs of RPM do on rate of adoption
    - Are new or existing payment frameworks in place to assist reimbursement
    - Will choice of model influence adoption

Examples will be drawn from the commercial world where RPM programmes have been implemented, and the talk will also discuss the Regulatory challenges facing RPM, even before evidence and reimbursement has been addressed.

Contact: malcolm.luker@philips.com

**Notes**

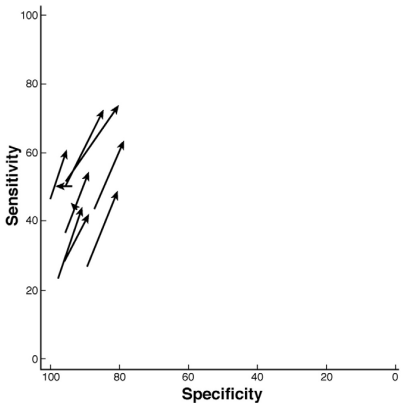
*Contributed paper*

## Exploration of methods to analyse MRMC diagnostic studies using CT colonography: Why ROC AUC is not the answer

Sue Mallett, Gary Collins, Steve Halligan, Justine McQuillan, Douglas Altman

Background: We investigated methods for analysing multi-reader multi-case studies to compare the diagnostic accuracy of two tests, interpretation with and without Computer-Aided Design (CAD), in CT colonography. Our preferred method of analysis was using the average differences in readers' sensitivity and specificity. Due to a preference of regulatory authorities for using the area under the ROC curve (ROC AUC) for analysis of MRMC radiological studies, we investigated the suitability of this method for our CT colonography study.

Study design and analysis methods: A multi-reader multi-case study was designed using 107 patient CT colonography videos. Ten radiologist readers read all patient videos for each of the two test methods. Radiologists reported their interpretation of each patient video as either normal (no polyps present) or abnormal (polyps present). Radiologists also reported a confidence score according to a scale of 1 to 100 for the presence or absence of polyps. Analysis of the diagnostic measures of sensitivity and specificity used a per patient analysis based on radiologists' reporting of patients as with or without polyps, as clinical management decisions on referral of patients for optical colonoscopy are made on such reports. Assessment of radiologist performance with and without CAD was based on a difference in sensitivity and difference in specificity between the two tests. Confidence intervals were calculated using bootstrap methods. Analysis using ROC AUC was attempted using the Dorfman Berbaum Metz method using LabMRMC software (LabMRMC 1.0Bv beta3 & Proproc). Assumptions of this ROC AUC method were examined.




Results: Figure 1 shows the change in per patient sensitivity and specificity for each of 10 readers when assisted by CAD, for all polyp sizes. The arrows point in the direction of each reader's change. Analysis of ROC AUC using the Dorfman Metz method experienced problems due to issues with using confidence scores in our study and the resulting difficulties in fitting ROC curves. Other characteristics of our study led to additional practical and theoretical issues in the calculation and interpretation of ROC AUC measures.

Conclusions: We found ROC AUC methods were unsuitable for analysis of diagnostic accuracy for CT colonography tests using a study design to detect presence or absence of polyps based on patient videos. We detail where problems arose due to assumptions of the method and issues with data fitting, against a background of the current literature. Using a difference in sensitivity and difference in specificity, we were able to make a meaningful comparison between radiologist performance in CT colonography in two tests, with and without CAD.

Contact: susan.mallett@csm.ox.ac.uk

**Notes**

*Contributed paper*

## Evaluation of Longitudinal Biomarkers

Ruwanthi Kolamunnage-Dona, Paula Williamson, Cheng-Hock Toh, Ingeborg Welters, Colin Downey

It is common in most diagnostic studies that repeated measurements of biomarkers are available alongside event history data, in which times to terminating events are recorded. A well-known example is in HIV research in which a biomarker such as CD4 count is determined intermittently and its relationship with time to AIDS onset is of interest. Another common situation is where some patients withdraw from the study before completing the biomarker measurement schedule but the dropout may be informative. In such cases, the longitudinal profiles alone may not characterise the accuracy of a biomarker, it may be an artefact caused by selective dropout. Assuming study dropout is independent of biomarker accuracy induces bias in the ROC analysis. Statistical methods for the proper treatment of data of this form, merging information from the two sources, are currently both under-developed and under-used in biomedical research.

One approach to evaluate the strength of association between a longitudinal biomarker process and an event time is through time-dependent covariate Cox proportional hazards model, which estimates the risk of the event as a function of time-varying biomarker process. However, this method assumes that longitudinal biomarkers are measured without error. Our approach models the longitudinal biomarker measurements incorporating the impact of informative dropout on time to disease onset. In this methodology, we jointly model the repeated measurements of biomarkers and event history data to estimate the ROC curve, and thereby evaluate the biomarker accuracy. We assume that the biomarker data or some transformation thereof are normally distributed and utilise random-effects to capture the association between longitudinal and event history processes.

We have applied this methodology to clarify whether the aPTT (activated partial thromboplastin time) waveform can be a time-dependent biomarker of sepsis. Sepsis is the leading cause of death worldwide. Although treating sepsis in the early stages will reduce mortality, a fundamental challenge is that prompt diagnosis is difficult. Microbiological culture results are considered the gold standard for diagnosing sepsis but may take up to 48-96 hours to process. aPTT is one of the newer biomarkers of sepsis, and when the analyser is available, aPTT waveform analysis is an inexpensive, rapid and readily available tool.

Contact: Ruwanthi.Kolamunnage-Dona@liverpool.ac.uk

**Notes**

*Contributed paper*

## Assessing the additional value of diagnostic markers: a comparison of traditional and novel measures

Ewout Steyerberg

Background: New markers and diagnostic tests have to be assessed for their value in addition to simple, readily available diagnostic characteristics. There is currently much scientific confusion on the measures to quantify such additional value. Traditionally, we tended to focus on the improvement in the area under the receiver operating characteristic (ROC) curve (AUC). A new but already quite popular measure is the net reclassification improvement (NRI), which is calculated as the sum of the improvements in sensitivity and specificity[1]. Moreover, decision–analytic measures have been proposed, including decision curves[2]. These quantify the improvement in net benefit (fraction true positive classifications penalized for false positive classifications) over a range of decision thresholds.

Aim: We aimed to define the role of these two relatively novel approaches (NRI and net benefit calculations) in the evaluation of diagnostic markers.
Methods: For illustration we present a case study of diagnosing the presence of residual tumor versus benign tissue in 544 patients with testicular cancer. We consider 3 tumor markers (AFP, HCG, and LDH) in addition to post chemotherapy size (the current clinical diagnostic indicator).

Results: We find inconsistent results in ranking of the importance of the 3 tumor markers. The AUC increase was largest for LDH (+0.021), and lowest for HCG (+0.013). For a clinically defendable decision threshold (20% risk of residual tumor), AFP and HCG ranked best (both net benefit +.6), and LDH worst (net benefit +.1). The NRI showed the same ranking as the net benefit analyses, but the magnitude of improvement was similar for all 3 markers (AFP: 9.6%; HCG: 9.2%; LDH: 7.7%).

Conclusions: Our judgment of the additional value of a diagnostic marker depends on the measure chosen. A decision-analytic perspective is not compatible with an overall judgment as obtained from the AUC in ROC analysis nor with NRI calculations. The current practice of reporting AUC and NRI needs to be replaced by routinely reporting net benefit analyses.

1. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27(2):157-72; discussion 207-12.
2. Vickers AJ. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. Am Stat 2008;62(4):314-320.

Contact: e.steyerberg@erasmusmc.nl

**Notes**

*Contributed paper*

**Correcting for partial verification bias: a comparison of methods**

Joris de Groot, Kristel Janssen, Koos Zwinderman, Patrick Bossuyt, Johannes Reitsma, Karel Moons

Background: a common problem in diagnostic research is that the reference standard has not been performed in all patients. This partial verification may lead to biased accuracy measures of the test under study. Several solutions have been proposed to alleviate this bias. The authors studied the performance of multiple imputation and the conventional correction method proposed by Begg and Greenes under a range of different situations of partial verification, to examine under which circumstances they produce similar results and when their results differ.

Methods: in a series of simulations, using a previously published Deep Venous Thrombosis dataset (N=1292), the authors deliberately set the outcome of the reference standard to missing based on various underlying mechanisms and by varying the total number of missing values. They then compared the performance of different correction methods (ie Multiple Imputation and the Begg and Greenes correction method) in each of these patterns of verification, in particular their ability to reduce the bias in estimates of accuracy by comparing it with the true value in the complete dataset.

Results: the results of the study show that when the mechanism of missing reference data is known, accuracy measures can easily be correctly adjusted using either the Begg and Greenes method, or multiple imputation. In situations where the mechanism of missing reference data is complex or unknown, multiple imputation is more flexible and straight forward than the Begg and Greenes correction method.

Conclusion: partial verification by design can be a very efficient data collection strategy. In that case the pattern of missing reference data will be known and accuracy measures can easily be correctly adjusted using either Begg and Greenes method, or (Multiple) Imputation. If not defined by design, partial verification should be avoided, as it can seriously bias the results. There are however situations where the mechanism of missing reference data is not known and partial verification can not be avoided. In these situations we strongly recommend to use Multiple Imputation methods to correct. These methods are more flexible and straight forward than the Begg and Greenes correction method and give reliable estimates of the missing reference data.

Contact: j.degroot-17@umcutrecht.nl

**Notes**

*Contributed paper*

## PET/CT in Cancer: Moderate Sample Sizes may Suffice to Justify Replacement of a Regional Gold Standard

Oke Gerke, Mads Poulsen, Kirsten Bouchelouche, Poul-Flemming Høilund-Carlsen, Werner Vach

Background & Objectives: For certain cancer indications, the current patient evaluation strategy comprises a perfect, but locally restricted gold standard procedure. If positron emission tomography/computed tomography (PET/CT) can be shown to be reliable within the gold standard region and if it can be argued that PET/CT also performs well in adjacent areas, then sample sizes in accuracy studies can be reduced.

Methods: Traditional standard power calculations for demonstrating sensitivities of both 80% and 90% are shown for indications with a low prevalence of only 20%. The argument is then described in general terms and demonstrated by an ongoing study of metastasized prostate cancer.
Results: An added value in accuracy of PET/CT in adjacent areas can outweigh a downsized target level of accuracy in the gold standard region, justifying smaller sample sizes.

Conclusions: If PET/CT provides an accuracy benefit in adjacent regions, then sample sizes can be reduced and the conduct of trials accelerated, leading to earlier decisions on the use of PET/CT while exposing fewer patients and reducing overall costs.

Reference : Gerke O, Poulsen MH, Bouchelouche K, Høilund-Carlsen PF, Vach W. PET/CT in Cancer: Moderate Sample Sizes may Suffice to Justify Replacement of a Regional Gold Standard. Molecular Imaging and Biology 2009; 11(6):381-385.

Contact: oke@stat.sdu.dk

**Notes**

*Contributed paper*

## Using patient management as a proxy for patient outcomes in test evaluation

Lukas Staub, Sarah Lord, R. John Simes, Suzanne Dyer, Nehmat Houssami, Robert Chen, Les Irwig

Background: The clinical value of a new test is ideally assessed by randomised controlled trials (RCTs) that measure its impact on patient health outcomes. Given the different practical challenges of these trials, this evidence is rarely available and test accuracy is often used as a surrogate. However, improved accuracy does not necessarily lead to improved patient health. Therefore, the consequences of testing on patient management are often investigated as an intermediate step in the pathway. Due to the lack of guidance on the interpretation of this evidence, patient management studies often neglect a discussion of the limitations of measuring patient management in test evaluation.

Methods: We discuss the rationale for measuring patient management, describe the common study designs and provide guidance to the reader about how this evidence can be interpreted.

Results: Two purposes of patient management studies are, first, to confirm that additional findings from the new test translate to a defined change in management, and second, to explore the link between test results and patient management where this is not yet well defined. Three designs are commonly used:

1. RCTs that measure patient management as a primary endpoint. This design directly measures the consequences of both alternative test strategies and does not rely on assumptions of planned management.

2. Diagnostic before-after studies that compare planned patient management before and after testing, thus providing an overview of management in broad patient groups where multiple differential diagnoses are considered.

3. Accuracy studies that are extended to report on the actual treatment or further tests received following a positive and negative test result. This evidence can support assumptions about the impact of test results on the use of treatments and further tests when it is uncertain whether all additional findings from the new test lead to a change in management.

We provide examples to discuss how the validity of this evidence depends on both the susceptibility of study designs to biases and the level of certainty surrounding assumptions linking changes in patient management to improved outcomes.

Conclusions: Patient management studies can support recommendations about the use and funding of a test when accuracy evidence is available but assumptions about changes in management associated with the test are uncertain and pivotal to conclusions. However, if it is unclear to what extent patient management is a good proxy for patient outcomes, further evaluation of the test remains ethical and essential.

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

**Notes**

*Contributed paper*

## Adjusting for differential verification bias in diagnostic accuracy studies: a Bayesian approach

Joris de Groot, Nandini Dendukuri, Kristel Janssen, Johannes Reitsma, Patrick Bossuyt, Karel Moons

Background: in studies of diagnostic accuracy, the performance of the test under study (index test) is determined by verifying its results against the results of a reference standard applied to the same patients. If verification of index test results by the preferred reference standard can not be performed in all study subjects, an alternative reference test could be given to subjects in whom the result of the preferred reference test is not available. The difficulty when using this so called differential verification is that the two reference standards often are of different quality, or define the target condition differently. Incorrectly treating results of the alternative reference standard as if they were from the preferred reference standard will lead to differential verification bias. In this study the authors propose a Bayesian model to simultaneously correct for the problem of differential verification bias and adjust for the imperfect nature of the reference standards.

Methods: a Bayesian method to correct for differential verification bias is presented, using two simulated examples of differential verification. Furthermore, the techniques are applied to estimate the accuracy of the Elbow Extension Test for diagnosing elbow fractures, using data of a recently published study.

Results: the Bayesian model provides reliable estimates for all accuracy measures compared to the true values used for the simulations.

Conclusion: we would like to reiterate the important message to verify as many patients as possible with the preferred reference standard in diagnostic studies in order to avoid verification bias. However, in clinical practice complete verification is commonly impossible for various reasons such as patient burden and costs. In some situations where verification by the preferred reference standard is impossible or unethical in specific groups, verification by a different reference standard can be considered. The discussed model helps the researcher to make unbiased inferences about a variety of specifics of the index test under study.

The method presented here is useful in drawing the best possible inferences from diagnostic tests in the presence of differential verification.

Key reference: Appelboam A et al. Elbow extension test to rule out elbow fracture: multicentre, prospective validation and observational study of diagnostic accuracy in adults and children. BMJ 2008;337:a2428.

Contact: j.degroot-17@umcutrecht.nl

**Notes**

*Keynote speaker*

**How to choose the best test for clinical monitoring**

Les Irwig, Prof. Screening and Test Evaluation Program, School of Public Health, University of Sydney, Australia.

Clinical monitoring is a common medical activity to assess response to treatment [Initial Response Monitoring] or whether changes occur subsequently that may need additional treatment [Long Term Monitoring]. In both instances, the difficulty is separating signal [real change] from noise [the background variability of monitoring measurements, usually due mostly to short-tem biological fluctuations]

In this talk, I will discuss criteria for choosing the best test for monitoring, how criteria differ between initial response and long-term monitoring, and why different tests may be appropriate for these phases of monitoring.

Contact: lesi@health.usyd.edu.au

**Notes**

*Contributed paper*

## Monitoring Intraocular Pressure as a Marker of Glaucoma

María Vázquez-Montes, Rafael Perera, Ryo Asaoka, Paul Glasziou, David Crabb, David Garway-Heath

Glaucoma is a chronic progressive optic neuropathy leading to impaired vision and sometimes blindness if untreated. Open angle glaucoma (OAG) is the most common form of glaucoma, affecting about 2% of the population aged over 40, and is second to age related macular degeneration as a main cause of blindness in the UK. There are no acute attacks in OAG with the only signs being a gradual and progressive loss in the visual fields (VF) and optic nerve changes. Although VF are the preferred measures for identifying OAG, little agreement on which to chose among the several possible summary measures available has limited their use for monitoring patients at risk.

Among the risk factors associated with OAG, Intra-ocular pressure (IOP) is the measure of choice for monitoring as part of a strategy to prevent OAG. It is the only treatable risk factor while changes in IOP are assumed to detect early changes in OAG. However it is unclear if the random variation of IOP makes it a precise enough measure to establish risk levels based on a single measurement.

We use data from two RCTs treating patients with high IOP to obtain estimates of IOP random variation and the group natural progression/deterioration (placebo and treatment groups analyzed separately). These estimates should help to define an optimal monitoring criterion for IOP to identify individuals requiring treatment adjustments.

As part of this talk we will show relevant methodological challenges (and some solutions) in the analysis of monitoring measurements such as: the use of measurements on one or both eyes, data transformation to fit model assumptions, partial missing data, and model validation.

Contact: maria.vazquezmontes@dphpc.ox.ac.uk

**Notes**

*Contributed paper*

## Can we assess adherence to medication by measuring change in blood pressure and cholesterol?

Andrew Hayen, Katy Bell, Paul Glasziou, Adrienne Kirby, Bruce Neal

Context: Guidelines recommend clinicians monitor patients' blood pressure and cholesterol levels to assess adherence after starting therapy. However the capacity of monitoring to detect non-adherence is unknown.
Objective: To estimate the accuracy of blood pressure/cholesterol monitoring for detecting non-adherence.

Design: Secondary analysis of blood pressure and cholesterol data in the PROGRESS (the perindopril protection against recurrent stroke study) and LIPID (Long-Term Intervention with Pravastatin in Ischemic Disease) trials. To assess the ability of monitoring to detect complete non-adherence, we compared change in blood pressure after 4 months or cholesterol after 12 months of treatment in those who continued (adherent) or discontinued treatment (non-adherent), and in those in active (adherent) or placebo (non-adherent) groups. For the LIPID analysis, we also assessed the ability of monitoring to detect partial non-adherence by comparing change in those who took at least 80% of their pills with those who did not.

Participants: 3433 patients with previous stroke or TIA (PROGRSS trial) and 9014 patients with prior coronary heart disease (LIPID).

Main Outcome Measures: Sensitivity, specificity, area under ROC curve.

Results: The absence of a fall in systolic blood pressure at 3 months had a sensitivity of 41% and a specificity of 80% for detecting complete non-adherence. Discriminatory power was modest over the range of cutoffs (area under the ROC curve 0.67). Increasing blood pressure measurements enhances the accuracy: using 2 sets (10 sets) of readings before and after treatment, the area under the ROC rises to 0.73 (0.92). The absence of a fall in LDL cholesterol at one year had a sensitivity of 50% and a specificity of 94% for detecting complete non-adherence. Discriminatory power was reasonable over the range of cut-offs (area under the ROC curve 0.89). Accuracy was more modest for detecting partial non-adherence: a change in LDL cholesterol at 12 months >-19.3mg/dL had a sensitivity of 37% and specificity of 87% (area under the ROC curve 0.65).

Conclusions: Standard monitoring of blood pressure is poor at detecting non-adherence to therapy, but does better with more measurements. Monitoring LDL cholesterol can detect patients who are completely non-adherent with therapy, but is less useful in detecting partial non-adherence

Contact: andrew.hayen@sydney.edu.au

**Notes**

*Contributed paper*

**How well is monitoring of CVD risk factors reported in clinical guidelines?**

Ivan Moschetti, Daniel Brandt, Rafael Perera, Carl Heneghan

Objective: One of the most common actions in clinical practice is cardiovascular disease (CVD) management and subsequent chronic disease monitoring. Clinical guidelines aim to raise the overall quality of care by standardizing decisions regarding diagnosis, management and treatment. To date, there has been no systematic examination on the reporting of monitoring recommendation from clinical practice guidelines and a better understanding of the monitoring process could impact substantially on patient's outcomes, clinical decision and overall costs to the health care system. Apart from age and sex, three modifiable risk factors: smoking, blood pressure and cholesterol- make a substantial contribution to CVD.

Practically clinicians need to know: what to monitor, how frequently to monitor, and what to respond to if the parameter index is out of range. To better understand the problem we undertook a systematic analysis of reported monitoring effects in current CVD prevention guidelines.

Design: systematic analysis of reported monitoring effects in current CVD prevention guidelines for three major CVD risk factors: cholesterol, smoking and hypertension.

Setting: primary and secondary care. We included guidelines published in English without any limitation on country or region of publication.

Participant: Participants receiving CVD risk factors evaluation for primary and secondary prevention.

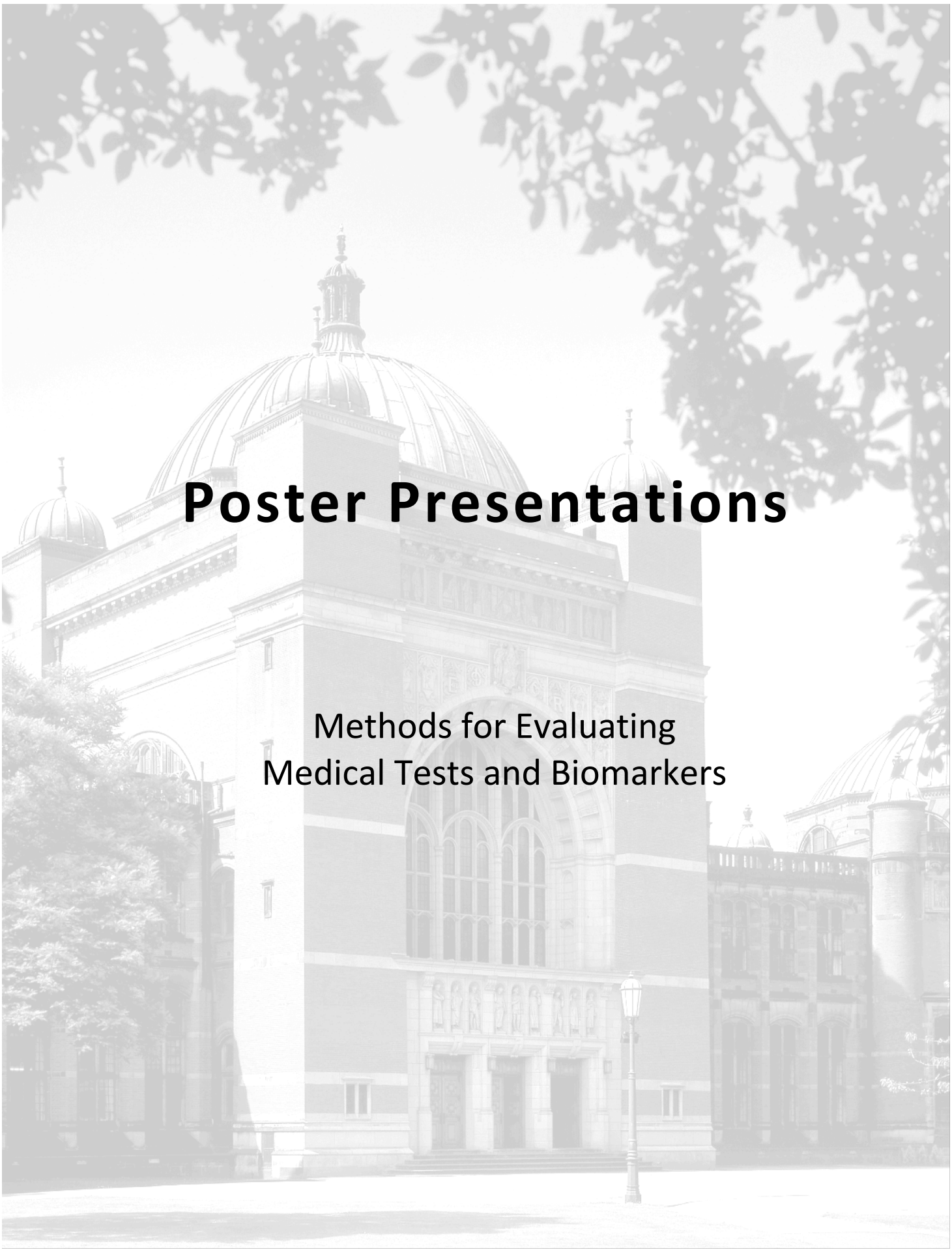
Main outcome measures: The primary outcome was the extent to which monitoring was addressed within the guidelines. Secondary to this is the completeness of monitoring recommendations, defined by the presence of three components: 1) a specific target or parameter to monitor, 2) frequency with which the specific target should be monitored, and 3) changes to consider if the monitored targets or parameters are not met.

Results: We assessed 117 guidelines published or updated between 2002 and 2009. More than half of all guidelines in our sample did not address the monitoring of one or more main CVD risk factors: 84/ 117 (72%) guidelines contained a section on Lipid monitoring, and of these 63% (53/84) mentioned monitoring. Specific information was reported in 47% for what to monitor, 49% for when to monitor and only 36% for what to do if target is out of range. 79/117 (68%) contained a section on Hypertension monitoring, and approximately a half, 51% (40/79) mentioned monitoring. Specific information was reported in 37% for what to monitor, 35% for when to monitor and only 30% for what to do if target is out of range. 65 of 117 (55%) contained a section on Smoking, and 57% (37/65) mentioned monitoring. Specific information was reported in 46% for what to monitor, 31% for when to monitor and 35% for what to do if target is out of range.

Conclusion: Monitoring is currently poorly reported in CVD guidelines, specifically for what and when to monitor important modifiable risk factors that require substantial monitoring in routine clinical practice. In addition what actions to take in response to monitoring targets that are out of range is not well defined.

Contact: ivan.moschetti@dphpc.ox.ac.uk

**Notes**



# Poster Presentations

Methods for Evaluating
Medical Tests and Biomarkers

# Poster Summary

| *No.* | *Title* | *Author* |
|---|---|---|
| P1 | Rapid diagnosis of *Clostridium difficile* using a device responsive to faecal volatile organic compounds | Katy Garner |
| P2 | Development of a panel of biomarkers for the diagnosis of prostate cancer | Colin Wheeler |
| P3 | Urine biomarkers for detecting bladder cancer: systematic review and meta-analysis | Graham Mowatt |
| P4 | PCA3 gene test in diagnosing prostate cancer; MUMM review no. 25 | Maija Saijonkari |
| P5 | Surveying the role of CK7/20 in differentiation of Barrett's oesophagus from Gastric Intestinal Metaplasia | Fariborz Mansour Ghanaei |
| P6 | Serum pepsinogen levels as a marker of gastritis in children | Susumu Iwasaki |
| P7 | Using clinical criteria to improve the use of genetic testing in diabetes | Beverley Shields |
| P8 | Clinical criteria do not precisely classify which insulin treated patients have Type 1 diabetes | Beverley Shields |
| P9 | Postal urinary C-peptide creatinine ratio can discriminate Type 1 from Type 2 diabetes and identify Type 1 patients with persistent endogenous insulin production | Tim McDonald |
| P10 | Are 'test-treat' trials as rare as suspected? A capture-recapture estimate of numbers published. | Lavinia Ferrante di Ruffano |
| P11 | Sample size calculations for 'test-treat' trials | Alice Sitch |
| P12 | Measuring the clinical validity of imaging findings: using MRI to detect lumbar spinal stenosis in patients with low back pain | Lukas Staub |
| P13 | Prostate specific antigen following primary treatment for prostate cancer: a review of clinical guidelines' recommendations on its use as a monitoring test. | Jac Dinnes |
| P14 | Practical problems in guideline development – diagnostic accuracy and Guideline Development Groups | Elizabeth Shaw |
| P15 | The assessment of diagnostic technologies: A NICE approach | Hanan Bell |
| P16 | Towards an economic evaluation of high throughput genetic testing for hereditary breast cancer: a systematic review of the current economic evidence | William Sullivan |
| P17 | Performance of methods for meta-analysis of diagnostic test accuracy studies when there are few studies | Yemisi Takwoingi |

| | | |
|---|---|---|
| P18 | Meta-analysis methods of the continuous relationship between age and diagnostic sign: heart rate | Richard Stevens |
| P19 | Investigation of heterogeneity in meta-analyses of diagnostic tests – a systematic review of methods used in research | Brian Willis |
| P20 | Investigating and improving the understanding of Cochrane Diagnostic Test Accuracy Reviews (DTARs) | Chris Hyde |
| P21 | Challenges of diagnostic systematic reviews in primary care settings: an example of rectal bleeding in cancer diagnosis | Margaret Astin |
| P22 | Meta-analysis challenges in pharmacogenetics: A case study of CYP2D6 genetic variant in breast cancer | Nigel Fleeman |
| P23 | Systematic review of test performance of mammography in detecting ipsilateral breast tumour recurrence and metachronous contralateral breast cancer | Clare Robertson |
| P24 | Focused boosting (reweighting) to improve diagnostic classifiers | Philip Gichuru |
| P25 | Pre-selection for screening by prediction models | Inge Stegeman |
| P26 | Systematic review of test performance when test positives and test negatives have a different reference standard | Clare Robertson |
| P27 | Verification problems in diagnostic accuracy studies: consequences and solutions | Karel Moons |
| P28 | Presenting clinically relevant information of diagnostic performance in systematic reviews: the use of predictive values (PPV and 1-NPV) | Henrika de Vet |
| P29 | Pin the threshold on the ROC curve: How to indentify relevant decision thresholds. | Colin Everett |
| P30 | Joint evaluation of sensitivity and specificity and its impact on the sample size | Werner Vach |
| P31 | Non-parametric estimation of ROC curves based on Bayesian models when the true disease state is unknown | Philip Gichuru |
| P32 | Interim analyses in diagnostic studies differ from interim analyses in treatment studies | Oke Gerke |
| P33 | Reporting of primary diagnostic studies – an example from a national guideline | Elizabeth Shaw |
| P34 | Evaluation of the methodological quality of diagnostic studies: experience with QUADAS and suggestions for amendments | Heike Ratz |
| P35 | Sales and self-tests on the world wide web | Geraldine van der Meer |

*Poster 1*

## Rapid diagnosis of *Clostridium difficile* using a device responsive to faecal volatile organic compounds

Christopher Probert, Katy Garner, Steve Smith, Ben DeLacy Costello, Rick Ewen, Norman Ratcliffe, Rosemary Greenwood

Introduction: *C. difficile* is reported in 50,000 people in England and Wales each year; of these patients 9000 die. *C. difficile* costs the EU 4B Euro pa. The HPA CDI Working Group reported toxin assay kits misdiagnose 1 in 5 to 1 in 10 cases of *C. difficile*. We have reported that the volatile organic compounds emitted from faeces change in disease (Garner, FASEBJ 2007). We have developed a prototype device that rapidly responds to volatile organic compounds VOCs with an on-screen diagnosis in 30 minutes.

Aims and methods: We aimed to assess the ability of a fast sensor device to diagnose *C. difficile*.

247 microbiologically-characterised faecal samples were sourced from HPA: 53 had *C. difficile*, 54 had *Campylobacer jejuni*, 30 had hospital acquired diarrhoea without *C. difficile*, the rest were an assortment of conditions and controls. Gas was collected from the faecal sample by heating and injected into the device which is based on a long GC column with a heated metal oxide sensor. Sensor outputs for the first 132 samples were analysed by logistic regression, the data was split 80:20, arbitrarily by the software, to create test and validation sets and the process run 10 times using different selections. From the 120 items, 18 were chosen that appeared discriminating and the models from these items assessed four times.

The output of the sensor also provides input to an Artificial Neural Network (ANN) program: decision *C. diff* or not *C. diff*. Samples were divided into training (n=148) and validation (n=99) sets.

Results**:** Statistic analysis

1. The Kappa statistic, to rate the classification from the discriminant analysis, of the first four validation data sets was 0.92, 0.79, 1 and 0.8. A kappa of 0.61-0.8 shows substantial agreement between two tests, a value >0.8 is 'almost perfect'; each model shows substantial/almost perfect agreement, or the correct diagnosis.

2. The model correctly differentiated normal and non-*C. difficile* samples from *C. difficile* samples in all cases on another four models. The diagnosis of *C. difficile* was made correctly in 86%, 80%, 100% and 100% of samples in these four validation data sets.

3. The ROC to show the ability to separate *C. difficile* from all other samples in a logistic model. The AUC is 0.92. This also implies significant agreement between the model and the laboratory diagnosis.

ANN analysis

The ANN correctly diagnosed 94% of samples in both sets.

Conclusion: Our device has the capability to correctly diagnose *C. difficile* infection. The device is easy to operate and gives the diagnosis in 30 minutes. Furthermore, early experiments (n=64) with a superfast device promise to be equally accurate, but in just 10 minutes. These devices have potential for point of care applications.

Contact: katy.garner@bristol.ac.uk

**Notes**

*Poster 2*

## Development of a panel of biomarkers for the diagnosis of prostate cancer

Colin Wheeler, Michael McAndrew, Jens Koopmann, Nick Workman, Rachel Fallon

The issues surrounding the use of PSA in the diagnosis of prostate cancer (PC) are well documented. In community practice the PSA cut-off point of 4ng/ml is considered sensitive but relatively non-specific (sensitivity of 86% and specificity of 33%)[1] and there is a need for a diagnostic test with greater discriminatory power. The development of autoantibodies associated with prostate cancer has been described[2]. In general, the appearance of such antibodies can precede disease symptoms by many years, making them attractive as potential biomarkers for early diagnosis.

We have developed a unique "functional protein" array platform which utilises 925 correctly folded proteins chosen for their roles in disease and used this to detect autoantibodies in PC serum samples. We have taken advantage of the multiplex nature of array assays and applied complex data analysis strategies to identify panels of biomarkers which may have clinical utility in the diagnosis of prostate cancer.

Serum samples (case; n = 73 and control; n = 60) were analysed using protein arrays. Data analysis was performed using a multi-scaling and outlier removal step for quadruplicate signals from all 133 protein arrays. Data were arbitrarily split into test and training sets and the data from the training set was then used with both support vector machines and genetic programming to identify classifiers which would successfully distinguish case from control samples. Classifiers were assessed for performance by referring to the combined sensitivity and specificity (S+S score) using the test set. Data were repeatedly split into test and training sets and analysis cycles repeated until a stable set of classifiers was identified.

A set of biomarkers was identified which can distinguish PC from control samples with both sensitivity and specificity above 90%. Further development and validation of this panel is on-going with a larger cohort.

[1] Prostate-specific antigen testing accuracy in community practice. BMC Family Practice (2002) 3:19 Hoffman et al

[2] Autoantibody Signatures in Prostate Cancer. N Engl J Med. (2005) 353:1224 Wang et al

Contact: colin.wheeler@senseproteomic.com

**Notes**

*Poster 3*

**Urine biomarkers for detecting bladder cancer: systematic review and meta-analysis**

Graham Mowatt, Shihua Zhu, Mary Kilonzo, Charles Boachie, Cynthis Fraser, TR Leyshon Griffiths, James N'Dow, Ghulam Nabi, Jonathan Cook, Luke Vale

Background: Bladder cancer is the fifth most common cancer in the European Union. Since the mid 1990s many urine biomarker tests for detecting bladder cancer have been developed, driven by the fact that there is a lack of reliable non-invasive methods for diagnosis and disease surveillance.

Objectives: To assess the test performance of three biomarkers generally regarded as being amongst the most clinically relevant at present - fluorescence in situ hybridisation (FISH), ImmunoCyt and nuclear matrix protein (NMP22) - and cytology for the detection of bladder cancer.

Methods: Major electronic databases including MEDLINE, MEDLINE In-Process, EMBASE, BIOSIS, Science Citation Index and Health Management Information Consortium were searched until April 2008. Types of studies considered were randomised controlled trials, non-randomised comparative studies and diagnostic cross-sectional studies that reported the absolute numbers of true and false positives and negatives. Participants had symptoms suspicious for bladder cancer or were previously diagnosed with non-muscle-invasive disease. Biomarker tests considered were FISH, ImmunoCyt, NMP22, or cytology, with a reference standard of histopathological examination of biopsied tissue. The results of the individual studies were tabulated. Meta-analysis models were fitted using hierarchical SROC curves. Summary sensitivity, specificity, positive and negative likelihood ratios and diagnostic odds ratios for each model were reported as point estimate and 95% confidence interval.

Results: A total of 71 studies were included. In the pooled estimates, sensitivity was highest for ImmunoCyt (84%, 95% CI 77 to 91%) and lowest for cytology (44%, 95% CI 38 to 51%). FISH (76%, 95% CI 65 to 84%), ImmunoCyt and NMP22 (68%, 95% CI 62 to 74%) all had higher sensitivity than cytology. This situation was reversed for specificity, which was highest for cytology (96%, 95% CI 94 to 98) and lowest for ImmunoCyt (75%, 68 to 83%). Cytology had higher specificity than FISH (85%, 95% CI 78 to 92%), ImmunoCyt or NMP22 (79%, 95% CI 74 to 84%).

Conclusions: This is the first systematic review specifically comparing the test performance of FISH, ImmunoCyt, NMP22 and cytology for bladder cancer. All three contemporary urinary biomarkers had higher sensitivity, but lower specificity than urine cytology. The lower specificity (more false positives) of currently available urinary biomarkers could lead to further unnecessary investigations.

Contact: g.mowatt@abdn.ac.uk

**Notes**

*Poster 4*

## PCA3 gene test in diagnosing prostate cancer: MUMM review no. 25

Kimmo Taari, Kristina Hotakainen, Maija Saijonkari, Riitta Grahn, Jaana Leipala

Managed Uptake of Medical Methods (MUMM) - a tool for decision making in Finland.

Managed Uptake of Medical Methods (MUMM) is a joint programme of Finnish specialised care providers (represented by the hospital districts) and Finnish health technology assessment agency Finohta launched in 2005.

The main objective of the program is to assure sufficient safety and effectiveness of new emerging health technologies as well as information about their costs. Each evaluation process is carried out by a review group consisting of two clinicians familiar with the technology and two HTA-methodology specialists and an information specialist from Finohta. Based on the MUMM review, a recommendation for the use of the technology in Finland is given by MUMM Advisory Committee along with the hospital districts debutes.
PCA3 gene test in diagnosing prostate cancer

Background: Serum PSA is widely used in the diagnostics and surveillance of prostate cancer, although its non-specifity to prostate cancer limits its usefulness. The prostate cancer gene 3 (PCA3) urine test has been proposed a novel marker of prostate cancer.

Aim: The aim of this systematic literature review was to assess the sensitivity and specificity of the PCA3 test commercially available in Finland in indicating prostate cancer as compared to pathological examination of biopsy specimens. In addition, we assessed the correlation of the PCA3 test to the characteristics of the prostate tumours and whether the number of prostate biopsies could be reduced by using the test.

Methods: The Medline, Cochrane and Journals@Ovid databases were searched for literature.
Results: 144 research articles and reviews were found. Six original research articles published in 2006–2008 fulfilled the inclusion criteria. They reported results of the PCA3 urine test taken before a prostate biopsy or radical prostatectomy. The AUC values of the PCA3 test for predicting a positive biopsy finding were 0,66−0,78. PCA3 correlated with the size and extracapsular growth of the tumour. There were no studies reporting patient follow-up after a biopsy or prostatectomy.

Conclusions: There is so far scarce evidence on the benefits of the PCA3 test. PCA3 as such is probably not an adequately sensitive indicator of prostate cancer. When combined to other diagnostics methods the PCA3 test may, however, provide some additional information to the evaluation of the probability and aggressiveness of prostate cancer.

Contact: maija.saijonkari@thl.fi

**Notes**

*Poster 5*

## Surveying the role of CK7/20 in differentiation of Barrett's esophagus from Gastric Intestinal Metaplasia

Fairborz Masour Ghanaei, Hadi Hajizadeh Fallah, Ali Ghanbari Motlagh, Alee Koloukani

Introduction: Replacement of the Normal squamous mucus of distal esophagus with metaplastic columnar epithelium including gablet cells names Intestinal metaplastia. This phenomenon resulted from gastric-esophageal reflux as its main cause, called Barrett's esophagus, increases the risk of esophageal adenocarcinoma of esophagus up to 30-40 times in comparison to normal population. Chronic gastritis which is important pathological cause is chronic infection of H.Pylori by inducing chronic inflammatory changes which finally result into changes like mucosal atrophy at epithelial metaplasia in case of involving of gastro-esophageal junction (GEJ) produce metaplasia the same as the one produce by Barrett's esophagus [1]. But, it doesn't increase the rate of esophageal adenocarcinoma [10, 11]. If we do endoscopy and biopsy procedure for diagnosis Barrett's esophagus, it may be imagined that the specimen has been taken from distal esophagus however unwillingly taken from proximal parts of gastrum. This reminds an important point which either this specimen which included chronic metaplasia is from gastric origin [5,6] or it is due to Barrett esophagus originally[3,7]. One of the new diagnostic way for differentiating these two metaphlasia from each other is using of Immunohistochemical staining specifically by using of Cytokeratin 7 , 20 (CK7/20). The staining pattern in this technique as described by Morsby et al is staining of superficial metaplastic epithelium with CK20 and both deep and superficial metaplastic epithelium with CK7[13].

Methodology: In this study we used endoscopic specimens of 30 patients with H.pylori infection and 31 ones with Barrett's esophagus which exist in our laboratories from past. All Barrett's specimens had been taken from a place higher than GEJ and all gastric specimens where from anterum area. We stained with CK7/20 with IHC staining procedure and then studied them by optical microscope. Two professional pathologists from two different research centers which collaborated in this study studied them separately and then were given each other's results afterward. SPSS software was used for analysis data and also we determined sensivity and specificity of the test.

Results and conclusion: According to achieved results 61.3% of Barrett's specimens fallowed the same pattern as Morsby defined (18 out of 30). 56% of anterum specimens had gastric pattern respectively. The rest of patterns in both two groups were completely from each other. Sensivity and specificity of the test were 61.3% and 100% respectively. Therefore diagnosis of Barrett's esophagus should not be made on basis of one single procedure.

Contribution: Our findings, which have been performed for the first time in Iran, confirmed the previous results gained by the other researchers around the world.

Contact: sahar_allie@yahoo.co.in

**Notes**

*Poster 6*

## Serum pepsinogen levels as a marker of gastritis in children

Susumu Iwasaki, Yoshihisa Urita, Kazumasa Miki, Monotobu Sugimoto, Yoshinori Igarashi, Yasukiyo Sumino

Background**:** Serum PGs have been used as biomarkers of gastric mucosal status, including atrophic change, inflammation, and tumor markers in some malignant diseases, before the discovery of *Helicobacter pylori* (*H. pylori*). Although serum PG1 and PG2 levels are known to increase in the presence of *H. pylori*-related non-atrophic chronic gastritis in adults, there are few reports of the relationship between serum PG levels and *H. pylori* infection in children who are considered to keep the infection for shorter periods. The aim of this study is to evaluate the difference in serum PG levels between *H. pylori*-infected and non-infected children.

Materials and Methods: *H. pylori* IgG antibody and serum PG levels were measured in 522 consecutive children (age range, 1-18 years) and 2111 adults (age range, 20-88 years) after an overnight fast at the first visit regardless of their symptoms. Serum PG concentrations were assayed using PG1 and PG2 Riabead Kits. Serum samples were also examined for *H. pylori* antibody by an enzyme-linked immunosorbent assay (ELISA) using the EPI HM-CAP IgG (Enteric Products, Inc., N.Y.) assays. In general, the prevalence of *H. pylori* infection increased with age, especially from the 16-year age group and peaked at the 60-69 year age group.

Results: *H. pylori* IgG antibody was detected in 85 children (16.3%) and 1478 adults (70.0%). In *H. pylori*-negative subjects, PG1 and PG2 levels gradually increased with age in subjects under 19 of age, and rose to the same levels as adults in their middle twenties. Particularly, children aged younger than 10 years had lower levels in both PG1 and PG2 compared to those older than 10 years. PG1/2 ratios are almost unchanged from the 8-9-year age group to the 40-49-year age group, and gradually decreased with advancing age. In *H. pylori*-infected subjects, serum PG1 was increased with age and peaked at 20-29 year of age, whereas PG2 was peaked at 17-18 year of age and plateau later. Both PG1 and PG2 levels were increased remarkably in infected children aged more than 10 years and significantly higher than in those without *H. pylori* infection. In age group of <10, there was little difference in serum PG1 levels between *H. pylori*-positive and -negative subjects. PG2 levels rose earlier than PG1 and were higher in *H. pylori*-positive subjects than in *H. pylori*-negative subjects in all age groups, resulting in that PG1/2 ratios increased gradually but did not peaked clearly.

Conclusions: In non-infected subjects, PG1 and PG2 levels gradually increased with age in subjects younger than 19 years. Since PG1 and PG2 levels were higher in *H. pylori*-infected subjects aged older than 10 years than in non-infected subjects, these biomarkers are considered a useful screening test for *H. Pylor i* - associated gastritis in the adolescents like adults. However, these biomarkers were unable to distinguish younger children with *H. pylori*-associated gastritis from those without *H. pylori* infection, suggesting that serum pepsinogens are not useful as a marker for gastritis in younger children before entrance to an elementary school. In contrast, PG1/PG2 ratio should be used as an index of atrophic gastritis in adults aged older than 40 years.

Contact: foo@eb.mbn.or.jp

**Notes**

*Poster 7*

## Using clinical criteria to improve the use of genetic testing in diabetes

Beverley Shields, Timothy McDonald, Sian Ellard, Andrew Hattersley

Background: Maturity-onset diabetes of the young (MODY) is a genetic form of familial, young-onset diabetes. MODY is rare (~1% of diabetes) and is often misdiagnosed as the more common Type 1 diabetes (T1D) or Type 2 diabetes (T2D). A correct genetic diagnosis of MODY is important, as these patients needs different treatment from other types of diabetes.

Aims/Methods: We aimed to determine clinical criteria that could be used in young onset diabetes (diagnosed <35y) to discriminate MODY (n=522) from T1D (n=348) and T2D (n=253), and to produce a probability model for predicting a patient's likelihood of MODY. Data were available on BMI standard deviation score, age at diagnosis, duration of diabetes (y), treatment, family history of diabetes, and diabetes control (HbA1c). Logistic regression analysis was used to determine clinical characteristics predictive of MODY and regression equations were derived to enable calculation of individual probabilities for MODY. ROC curves were used to determine optimal cut-offs for sensitivity and specificity.

Results: The key discriminator of T1D and T2D was insulin treatment from diagnosis (99% v 0.4%, respectively). Therefore, T1D v MODY comparisons were carried out in those insulin treated at diagnosis, and T2D v MODY comparisons were tested in the rest. The optimal model for T1D v MODY (Model 1) produced a logistic regression equation for a log odds ratio (logOR) of 6.5-(1.2*HbA1c)+(3*if parent has diabetes)+(0.06* age diagnosis)-(0.06*duration diabetes). Model 1 showed good discrimination (area under the ROC curve= 0.967). The best T2 v MODY model (Model 2) provided the regression equation for logOR of 19.7-(0.35*age diagnosis)-(0.88*BMI SDS)-(1.1*HbA1c) (area under the ROC curve = 0.975). In patients treated with insulin from diagnosis, a cutoff of >50% probability from Model 1 gave a likelihood ratio (LR) for MODY of 25. For patients not on insulin at diagnosis, a cutoff of >80% probability from Model 2 gave a LR for MODY of 18. Given a pre-test probability of 1%, these models would give post-test probabilities for MODY of 20% and 15%, respectively.

Conclusion: Clinical criteria can discriminate well between MODY and T1 or T2 diabetes, and could be used in clinical practice to improve selection of patients for genetic testing. With the considerable benefit of making a correct diagnosis of MODY, the post-test probabilities of 15% and 20% would represent appropriate levels at which to request expensive molecular genetic testing.

Contact: beverley.shields@pms.ac.uk

**Notes**

*Poster 8*

## Clinical criteria do not precisely classify which insulin treated patients have Type 1 diabetes

Beverley Shields, Timothy McDonald, Neelima Bhupathi-Raju, Maggie Shepherd, Andrew Hattersley

Background: There are no agreed clinical criteria for classifying Type 1 diabetes (T1D). Quality Outcomes Framework (QOF) advised using insulin treatment <1y of diagnosis or age at diagnosis <30y. T1D is defined by the World Health Organisation as beta-cell destruction leading to absolute insulin deficiency. Therefore, absent endogenous insulin secretion in long duration diabetes robustly defines T1D.

Aim: To determine which clinical criteria can identify T1D, as defined by absolute insulin deficiency, in insulin treated patients.

Methods: We studied 72 insulin-treated adult diabetic patients (>5y duration, insulin within 2y diagnosis). Endogenous insulin secretion was measured using post prandial urinary C-peptide creatinine ratio (UCPCR). Absolute insulin deficiency was defined as <0.2nmol/mmol.

Results: 40/72 patients had T1D as defined by insulin deficiency. QOF criteria were sensitive (93%), but not specific (22%) resulting in an inflated T1D prevalence: 40% QOF "T1D" had measurable C-peptide. Cutoffs derived from Receiver Operating Characteristic curves, had better specificity: age at diagnosis <39 (68% sensitivity, 97% specificity); <=1.5mths to insulin treatment (80% sensitivity, 56% specificity); BMI<29 (78% sensitivity, 56% specificity). Age at diagnosis performed best with cutoff<39y misclassifying fewer patients than QOF criteria (14/72(19%) v 28/72(39%), p=0.01). Combined criteria obtained through regression tree analysis improved the classification modestly (13/72(18%) misclassified).

Conclusion: Compared with the gold standard of chronic insulin deficiency, QOF criteria markedly overestimated the number of insulin treated patients with T1D. Despite being better than QOF, individual or combinations of clinical criteria could not accurately classify all patients. The limitations of clinical criteria have important implications for service planning as well as management of patients.

Contact: beverley.shields@pms.ac.uk

**Notes**

*Poster 9*

## Postal Urinary C-peptide creatinine ratio can discriminate Type 1 from Type 2 diabetes and identify Type 1 patients with persistent endogenous insulin production

Timothy McDonald, Beverley Shields, Rachel Besser, Maggie Shepherd, Bridget Knight, Andrew Hattersley

Introduction: Serum C-peptide is a measure of persistent endogenous insulin production useful in identifying diabetes subtypes. Urinary C-peptide creatinine ratio (UCPCR) is stable over 3 days at room temperature so offers a potential practical outpatient alternative.

Aims/Objectives: To determine if a patient posted postprandial UCPCR can:(1) discriminate Type 1 (T1D) and Type 2 (T2D) diabetes;(2) detect T1D patients with persistent endogenous insulin secretion.

Methods: Postprandial UCPCR was measured in adult Caucasian patients (diabetes duration ?5y, eGFR ?60mL/min/1.73m2) with T1D (n=61; diagnosed<30y, insulin since diagnosis) or T2D (n=54; diagnosed ≥35y, not insulin treated in year following diagnosis, treated with diet (6), oral agents (30), or insulin (18)). T1D patients with significant UCPCR and matched negative T1D controls, undertook a mixed meal tolerance test (MMTT).

Results: UCPCR was lower in T1D than T2D:(median(IQR) <0.02(<0.02-<0.02) v 2.47(1.36-4.13)nmol/mmol, p=3.7x10-21). Receiver Operating Characteristic Curves identified a cut-off of UCPCR ≤0.2nmol/mmol to be 95% sensitive and 98% specific for discriminating T1D from T2D (area under curve 0.99 perfect test 1.0). All 3 T1D patients with UCPCR>0.2 remained positive on repeat testing. In a MMTT their stimulated serum C-peptide was higher than in the controls (mean 579 v 24pmol/l) with 2/3 exceeding the C peptide negative cut off (200pmol/L) despite being GAD positive and 25 and 31 years post diagnosis.

Conclusions: Posted postprandial UCPCR can discriminate Type 1 from Type 2 diabetes and detect patients with long duration Type 1 diabetes with significant insulin production. Practical and pragmatic classification using UCPCR may be more relevant to patient care than an aetiological approach in long duration diabetes.

Contact: tim.mcdonald@rdeft.nhs.uk

Notes

*Poster 10*

## Are test–treat trials as rare as suspected? A capture-recapture estimate of numbers published

Lavinia Ferrante di Ruffano, Clare Davenport, Anne Eisinga, Sue Bayliss, Anne Fry-Smith, Chris Hyde

Background: The ultimate aim of diagnostic test evaluation is to determine which tests have the most favourable impact on patient health. Randomised controlled trials are the methodological gold-standard for evaluating these questions, however it is a common anecdote that such publications are 'rare'. But how rare are they? To date none has attempted to enumerate the true extent of this important evidence-base.

Objectives: To estimate the number of test-treat randomised trials published within a defined time-period, as indexed in the Cochrane Central Register of Controlled Trials (CENTRAL).

Methods: A capture-mark-recapture technique was used to estimate the total number of test-treat trials published between 2004 and 2007. CENTRAL (Issue 2, 2010) was searched using two separate ascertainment strategies. The first (S1) used terms related to diagnostic research, and was not restricted by test modality. The second (S2) was a test-specific search that targeted the names of five mainstream imaging modalities. Studies were included if they randomised patients with suspicious signs/symptoms to receive a diagnostic test followed by the administration of treatment. Trials that did not assess patient outcomes were excluded, as were those that evaluated serial testing or population-based screening. Each search was screened independently by two researchers.

Results: Search strategies yielded 13,576 (S1) and 12,041 (S2) titles and are currently under review by the second screener. Upon completion, findings will include a 2x2 contingency cross-tabulation of the number of relevant studies retrieved by both or only one of the searches from which an analysis of ascertainment overlap will produce the estimate of total number published. Agreement between screeners will be presented as kappa values.

Conclusions: The estimate as a proportion of all trials indexed on CENTRAL will be discussed, and challenges encountered in constructing effective strategies for identifying test-treat trials commented on.

Contact: l.ferrante@bham.ac.uk

**Notes**

*Poster 11*

**Sample size calculations for 'test-treat' trials**

Alice Sitch, Lavinia Ferrante di Ruffano, Jon Deeks

Background: In a 'test-treat' trial patients are randomised to receive different test strategies. According to the result of the test the appropriate course of treatment is administered. The outcomes for each individual are measured after they have received both the test and the subsequent treatment and from analysing these outcomes conclusions can be made regarding the best treatment pathway. The way in which sample size calculations are made depends on the outcome assessed, particularly whether it assesses clinical processes (such as diagnoses made, use of further tests, and treatments given) or patient outcomes subsequent to treatment. The aim of this work is to understand how sample sizes for trials of this type are being calculated, how well they are reported, and whether they are appropriate.

Methods: A cohort of 121 articles, published between January 2004 and May 2007, reporting on 101 test-treat trials was examined. The aim of the review being to understand how well the calculations are reported and if they have been done in a way that provides the study with adequate power to answer the primary hypothesis. This also allowed us to compare the magnitude of the hypothesised difference with what was observed.

Results: Analysis of the cohort is ongoing. We will report on the frequency of the sample size calculations, choice of outcome measure (process or patient) and the actual power of the studies.

Contact: a.sitch@bham.ac.uk

**Notes**

*Poster 12*

## Measuring the clinical validity of imaging findings: using MRI to detect lumbar spinal stenosis in patients with low back pain

Lukas Staub, Sarah Lord, Thomas Barz, Markus Melloh, Patrick Bossuyt

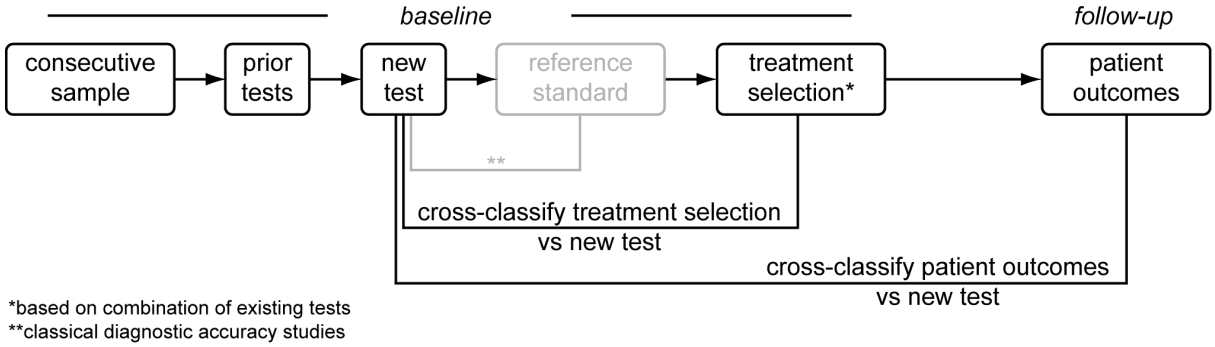
Background: Lumbar spinal stenosis can be effectively treated with decompressive surgery. However, the radiological findings in the diagnostic work-up of these patients do not always correlate well with clinical symptoms, and guidance about when to proceed to surgery is inconsistent. A recently described sign in MRI is proposed to help detect clinically relevant stenosis. However, due to the lack of an accepted reference standard the sign's diagnostic accuracy cannot be easily determined.

Objective: To describe a practical alternative to the classical accuracy study which assesses the clinical validity of the new MRI sign to detect lumbar spinal stenosis in patients with low back pain.

Methods: We defined the intended benefits of the MRI sign to guide treatment selection and improve patient outcomes. We then designed a study to measure its clinical validity to predict these consequences.

Results: In the absence of an established reference standard, the MRI sign findings will be cross-classified with decisions for surgery based on the existing tests and patient outcomes in 24 month follow-up examinations (figure). The results will be used to estimate: i) how well the MRI sign can distinguish between patients that do or do not benefit from surgery, and ii) the concordance between the MRI sign and existing tests to explore its possible value as a triage test. The observed proportion of discordant test results will help inform the design of future randomised trials of the MRI sign.

Conclusions: Measuring associations between test results and their downstream consequences may be the best way to judge the clinical validity of a new imaging test. This approach takes into account the diagnostic pathway in which the test will be used, requiring a common understanding between clinicians and epidemiologists of how to apply the principles of test evaluation to address clinical questions about tests.




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

**Notes**

*Poster 13*

## Prostate specific antigen following primary treatment for prostate cancer: a review of clinical guidelines' recommendations on its use as a monitoring test

Jac Dinnes, Jenny Hewison, Doug Altman, Jon Deeks

Introduction: Monitoring involves repeated testing on an individual over time to make decisions about the management of a disease or condition. It is a central activity in the management of patients, taking up a considerable part of the clinical workload.[1] In contrast, literature on the use of tests for monitoring purposes is relatively sparse. Ongoing research largely focuses on monitoring to optimise treatment in chronic disease, e.g. to establish and maintain control of the disease. Our interest is in monitoring for the detection of decline, progression or recurrence of a condition, where at some predefined threshold a change in patient management is indicated. An example is repeated measurement of prostate specific antigen (PSA) following primary treatment for prostate cancer (PCa). There is an abundance of literature on use of PSA in this context and it is in widespread use. We undertook a review of clinical guidelines on the management of PCa to try to identify why PSA has been such a 'success story' as a monitoring test and to determine whether there is sufficient underlying evidence to justify its use in this context. In particular, we were looking for evidence indicating a) how often the test should be applied, and b) at what point a change in patient management is indicated.

Methods: MEDLINE searches (1990-2009) were undertaken to identify guidelines related to management of PCa. The National Library of Guidelines and the Trip database were also accessed and reference lists of retrieved papers checked for further relevant studies. Recommendations or statements relating to the use of PSA following primary treatment were extracted by one author; references to any supporting evidence were also noted. A narrative synthesis was undertaken to determine the consistency of recommendations between guidelines and the type/volume of evidence cited. We also considered whether any of the guidelines made cautionary remarks regarding: PSA variability, between-study differences affecting threshold for intervention, and uncertainty in natural history of PSA and PCa recurrence.

Results: Identified 10 guidelines (n=7) or best practice statements (n=3). Six guidelines recommended two follow-up regimes (4 supported with primary literature); two PSA thresholds following prostatectomy and two following radiotherapy treatment. Mainly non-systematic reviews and up to 5 primary studies cited in support. The only identified systematic review found insufficient evidence for any particular threshold. In the main, guideline recommendations were not tempered by reference to uncertainties in the evidence base.

Conclusions: There is considerable inconsistency in guideline recommendations and substantive recommendations are made where the evidence is inconclusive. Guidelines have not paid due attention to issues important for monitoring, such as technical and biological variability in test results and interpretation of serial measurements. This is due to inadequacies in the evidence base and to inappropriate use of available evidence. Given the ever increasing number of tests that may be used for monitoring purposes, it seems timely to consider the lessons that can be learned from the failures in the PSA development process.

[1] Glasziou PP and Aronson JK. An introduction to monitoring therapeutic interventions in clinical practice. In: Glasziou PP, Irwig L, Aronson JK (eds), Evidence-Based Medical Monitoring: From Principles to Practice. Oxford: BMJ Books, 2008, Chapter 1.

Contact: j.dinnes@bham.ac.uk

**Notes**

*Poster 14*

## Practical problems in guideline development – diagnostic accuracy and Guideline Development Groups

Elizabeth Shaw, Phil Alderson, Roberta Richey

Background: Developing NICE guidelines involves many stages, including a review, or reviews, of the evidence, and drafting recommendations based on this evidence. These two stages involve both a technical team, responsible for the evidence synthesis, and a Guideline Development Group (GDG), responsible for drafting recommendations based on the evidence reviews. However, although it is accepted that the methods for evidence synthesis of diagnostic accuracy studies are less well developed than for intervention reviews, less is known about how GDGs interpret and use reviews of diagnostic accuracy studies to draft recommendations.

Aims: To describe our experience across several guidelines involving reviews of diagnostic accuracy, and to report how GDGs worked with the presented data. To draft recommendations on how diagnostic accuracy reviews can be presented in guideline development.

Methods: Several guidelines have been developed including diagnostic accuracy studies. We report anecdotal evidence of how the GDGs worked with the diagnostic accuracy data, and reflect views of the technical teams involved on how the process worked.
Practical recommendations were drafted, based on the experience of the technical team.

Results: Reviews of diagnostic accuracy studies present many challenges to a GDG. These differ by topic, by type of review, and by the type of evidence. From observation, GDGs react differently to each challenge. We will present examples of how GDGs work with diagnostic accuracy reviews to illustrate each of these and propose practical recommendations on who these could be tackled.

Discussion: Guideline developers should consider how the type of diagnostic accuracy review will be understood and interpreted by a GDG. The lack of methodological guidance on the synthesis of evidence for guideline development poses significant challenges, and this can impact on how the presented evidence is interpreted.

Contact: beth.shaw@nice.org.uk

**Notes**

*Poster 15*

## The Assessment of Diagnostic Technologies: A NICE Approach

Hanan Bell, Mirella Marlow, Nick Crabb

In 2009, NICE began a new programme to assess diagnostic technologies. This programme is separate from any evaluation of diagnostics under the guidelines activities of NICE. The programme is in the process of testing potential assessment methods with a pilot assessment due to be completed in the fourth quarter of 2010. An interim methods statement has been developed to inform the pilot processes and is available at the NICE website[1].

NICE diagnostic assessments share many features with the methods of other NICE programmes, including being based on cost-effectiveness modelling (reference case), being openly developed using external assessment teams, having concern for equality issues, and using systematic reviews and meta-analysis to estimate costs and outcomes.

However, diagnostics differ from the bulk of previously developed NICE technology appraisals which are primarily evaluations of pharmaceuticals. Many diagnostic assessments will be more complex than typical pharmaceutical assessments because diagnostic tests can be used in multiple ways, either due to different disease entities, or in different portions of the care pathway. Often, multiple tests or sequence may be alternatives to the technology being evaluated. Moreover, some tests require investment in costly capital equipment making the total usage of the equipment important in determining the cost of the technology. The relevant outcomes for the cost-effectiveness analysis usually do not occur directly from the diagnostic test but as a result of the care pathway which follows. Moreover, there may not always be an agreed care pathway.

Diagnostics also often differ from pharmaceuticals in the nature of the evidence available. The licensing process for pharmaceuticals requires controlled trials assessing the relevant outcomes. Such trials are often not available for diagnostics. Often data are only available for test characteristics such as sensitivity and specificity, which can be difficult to meta-analyze.

The NICE methods statement outlines alternatives for extending the methods for pharmaceutical assessment to deal with these differences. These include methods for dealing with extended scoping of the assessments, options for outcomes estimation, and options for costing the technologies. These methods are currently under test as part of the pilot assessment. A final methods guide will be issued for public comment in 2011.

[1] The interim methods statement is available at:
http://www.nice.org.uk/aboutnice/whatwedo/aboutdiagnosticsassessment/diagnosticsassessmentprogramme.jsp

Contact: hanan@hsbell.com

**Notes**

*Poster 16*

## Towards an economic evaluation of high throughput genetic testing for hereditary breast cancer: A systematic review of the current economic evidence

William Sullivan, Katherine Payne

Introduction: Genetic testing for mutations in the hereditary breast cancer (HBC) genes BRCA1/2 has been available since 1996 using Sanger sequencing methods. In the UK, women are only eligible for testing if they reach the threshold of having at least a 20% risk of carrying a BRCA1/2 mutation. High throughput sequencing (HTS) techniques are currently being developed that can increase the potential number of patients to be offered testing and possibly lower the testing threshold. The incremental costs and benefits of HTS compared to current methods are not known. This study is motivated by the need to understand current care pathways and the economic evidence presently available to support genetic testing for BRCA1/2 mutations to inform the structure of an economic model to evaluate HTS in BRCA1/2 testing.

Study aim: To identify and critically appraise published economic evaluations of genetic tests for HBC.

Methods: Centre for Reviews and Dissemination (CRD) systematic review methods were used to systematically identify relevant studies. The search strategy combined terms relevant to economic evaluations and interventions relevant to HBC. MEDLINE, EMBASE, PsycINFO and NHSEED (NHS Economic Evaluation) databases were searched to identify English Language studies between January 1996 and January 2010. Bibliographies of included studies were also checked. Identified studies were critically appraised according to NHSEED Database guidelines. Titles and abstracts were assessed and critically appraised for inclusion by 2 independent reviewers. Results: The search identified 359 studies, of which 24 abstracts potentially met the inclusion criteria and of these 15 studies were identified as full economic evaluations relevant to HBC. Five of the identified economic evaluations were directly relevant to genetic testing for HBC. Nine of the studies evaluated preventive treatments (n=5) or screening programs (n=4) for BRCA1/2 carriers. One study evaluated testing for BRCA1/2 mutations to prevent ovarian cancer. The five economic evaluations directly relevant to genetic testing for HBC were of varying quality. One of these four studies was appraised as a good quality study. All were retrospective modeling studies. Four of the studies compared a genetic testing program with no genetic testing; one compared the cost-effectiveness of different laboratory testing techniques. None of the studies reported using systematic review methods to identify existing evidence to inform model parameters. Outcome measures differed across the studies and included, for example, total number of mutations detected or QALY gains. All studies made some attempt at evaluating uncertainty using sensitivity analysis. Two studies reported probabilistic sensitivity analyses. The results of all five studies supported the use of genetic testing for risk of carrying the HBC gene for a population of women. None of the studies were relevant to UK practice.

Conclusions: Current UK guidelines for genetic testing to identify the BRCA1/2 mutations in HBC are not informed by robust economic data. Existing economic evaluations of genetic testing for HBC are of variable quality. The variation in approaches to measure outcomes in these studies may reflect the current challenge of how best to quantify benefit for diagnostic tests but also the lack of robust evidence to populate economic models. Funding: This study is part of the project TECHGENE funded by EC Framework 7 (no 223143).

Key reference: CRD (2007) 'NHS EED Handbook', http://www.york.ac.uk/inst/crd/pdf/nhseed-handb07.pdf

Contact: william.sullivan@manchester.ac.uk

**Notes**

*Poster 17*

## Performance of methods for meta-analysis of diagnostic test accuracy studies when there are few studies

Yemisi Takwoingi, Boliang Guo, Richard Riley, Jon Deeks

Background: Hierarchical methods recommended (Leeflang et al., 2008) for meta-analyses of diagnostic test accuracy studies are complex, relying on iterative procedures for the estimation of multiple model parameters. In certain circumstances, for instance when there are few studies in a meta-analysis, such models may not converge, fail to compute standard errors or give unstable parameter estimates.
Objectives: To evaluate the performance of meta-analytic approaches for test accuracy studies when few studies are available, and to develop recommendations for proceeding with meta-analysis when the suggested hierarchical methods fail.

Methods: Ten-thousand meta-analysis datasets were simulated for each of a range of scenarios typical of test accuracy studies varying according to important factors including: the number of studies, number of patients within studies, disease prevalence, and heterogeneity in test threshold and accuracy across studies. A variety of meta-analysis models were fitted and performance was assessed according to the bias, mean-square error and coverage of parameter estimates.

Results: Estimation of hierarchical model parameters and their standard errors often fail in the absence of heterogeneity in threshold and accuracy, and when prevalence is low or the diagnostic odds ratio (DOR) is high. For example, in one scenario with a prevalence of 5% and DOR of 38, the proportion of convergence failures for the hierarchical summary ROC (HSROC) model were 57%, 47%, and 42% for meta-analyses with 5, 10 and 20 studies respectively. Simpler models converge more easily but are clearly biased in a number of scenarios. Univariate random effects logistic regression models to pool sensitivity and specificity separately and the symmetric HSROC model were the 2 approaches that gave the least biased diagnostic odds ratio estimates.

Conclusions: Hierarchical methods can be problematic with convergence sensitive to the number of studies in a meta-analysis and sparse data. Where hierarchical methods fail, researchers may be forced to use simpler but less statistically rigorous methods, such as two separate random effects logistic regression models for sensitivity and specificity. Even with this approach, quantitative synthesis may sometimes be impossible and a narrative synthesis becomes the only option.

Contact: y.takwoingi@bham.ac.uk

**Notes**

*Poster 18*

## Meta-analysis of the continuous relationship between age and a diagnostic sign: heart rate

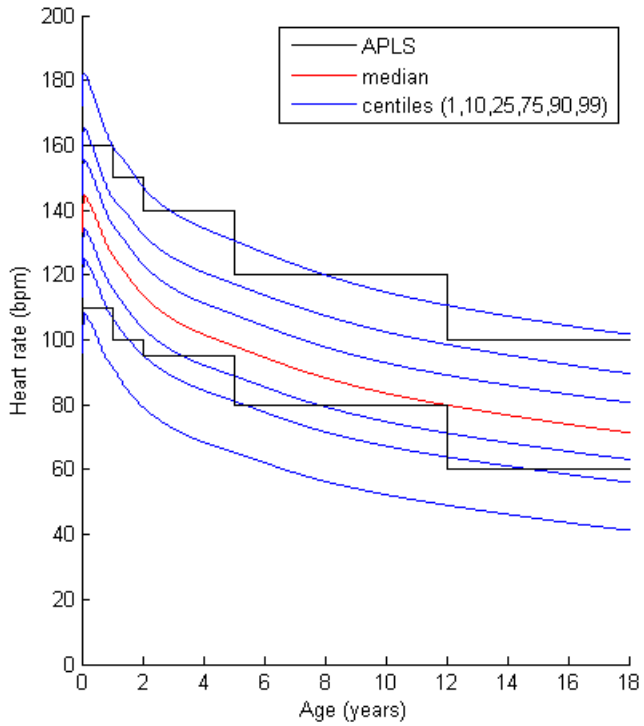
Richard Stevens, Susannah Fleming, Matthew Thompson, Rafael Perera, Carl Heneghan

Background: Heart rate is a key vital sign widely used in the diagnosis of children who may be seriously ill. It is therefore important to know the normal range of heart rate in a child of any given age, but the reference ranges in current use have no clear evidence base. Following a systematic review by our collaborators, we sought a method to combine 59 studies that reported heart rate by age into new centile charts.

Methods: We adapted 'kernel regression' to perform non-parametric regression on the median heart rate of children in different age groups as follows. Each study (or each subgroup of a study) contributed a kernel centered on the mean age with height proportional to the median heart rate and kernel width proportional to the age range in the study. All kernels were combined across studies with a weighted average in which weights were proportional to the number of children in the study, to give a smooth estimate of the relationship between median heart rate and age. The method was repeated for other centiles of heart rate.

Results: The Figure shows our new charts for the 1st, 10th, 25th, 50th, 75th, 90th and 99th centiles and, for comparison, the centile charts in the existing UK guidelines (APLS).

Conclusion: Non-parametric regression has allowed the creation of centile charts that improve on existing charts in realism, detail and evidence.




Contact: richard.stevens@dphpc.ox.ac.uk

**Notes**

*Poster 19*

## Investigation of heterogeneity in meta-analyses of diagnostic tests – a systematic review of methods used in research

Brian Willis, Muireann Quigley

Background: In the last decade there have been a number of developments in both the meta-analysis of diagnostic tests and the investigation of heterogeneity including the emergence of new models (bivariate random effects, hierarchical summary ROC curves) and their extension using meta-regression.

Objective: To review the methods used for identifying and investigating heterogeneity in meta-analyses of diagnostic tests and their implementation by investigators in applied research.

Methods: A systematic review of the diagnostic test literature was carried out to identify meta-analyses applied to diagnostic tests. Eight databases (MEDLINE, EMBASE, Cochrane, CINHAL, PsychInfo, HMIC, Global Health and AMED) were searched, and studies were included if they satisfied all of the following: Evaluated a diagnostic test; measured test performance; searched two or more databases; stated the search terms and inclusion criteria; used a statistical method to summarise performance. The included studies were critically appraised for the methods used to identify and explore the sources of heterogeneity.

Results: From over 6000 citations, 237 studies were included for appraisal. Nearly 10% did not test for heterogeneity and over half failed to investigate its causes. Although the most commonly used tests were Cochran's Q test and the I2 test, this belies the diverse approaches taken in their implementation; decisions on heterogeneity were based on critical values ranging from 0.01-0.25, absolute values of Q, and I2 ranges of 0-75% with some using a nonzero gradient coefficient in the SROC model as evidence of heterogeneity. Meta-regression was used to analyse the sources; the most popular model being to add covariates to the SROC curve. Correlation between the sensitivity and specificity and its effects on estimates of heterogeneity was rarely considered – only 6 studies extended the new models.

Conclusion: This review suggests a lack of consensus and clarity in testing for heterogeneity in diagnostic test research and limited uptake of the newer methods.

Contact: brian.willis@manchester.ac.uk

**Notes**

*Poster 20*

## Investigating and improving the understanding of Cochrane Diagnostic Test Accuracy Reviews (DTARs)

Chris Hyde, Clare Davenport, Ruth Garside, Mariska Leeflang, Patrick Bossuyt, Jon Deeks

Background: There is very little empirical research on barriers to understanding evaluations of tests and reviews of them, or indeed means to overcome these barriers. Over the past 5 years, The Cochrane Collaboration has been developing its approach to the synthesis of test evaluations culminating in the publication of the first Cochrane DTAR in 2008. During the process of providing support it has become clear that the complexity of many parts of the DTARs will be challenging even for those who are familiar with the general Cochrane review format. There is a need to further explore understanding of completed Cochrane DTARs in order to refine their presentation and improve impact.

Objective: To explore understanding of the three currently available Cochrane DTARs

Method: We have been funded to examine the perceptions of two groups: policy-makers and their advisers and clinical groups to whom the three reviews should be relevant. The initial step will be to get opinions on what are the key features of each of the completed reviews from authors. The second will involve sending a brief questionnaire to about 150 individuals to identify potential candidates for in-depth interviews. They will be asked about their overall understanding of one of the three completed DTARs. The third step will comprise approximately 40 individual, face to face interviews drawing on cognitive interview techniques with purposive sampling of respondents. For those in whom the initial questionnaire indicates little experience in interpreting test accuracy reviews the main focus will be on whether they can locate, understand and interpret the key elements of the DTAR. For those in whom the initial questionnaire indicates an advanced level of understanding of test accuracy reviews, the focus will be on eliciting suggestions for improvement and exploring the reason for any disagreements about the DTAR's interpretation. We believe the findings will be generalisable to systematic reviews of test accuracy published outside the Cochrane Library.

Results: The one year project will start in mid-2010 so results will not be available at the symposium. The poster will expand on the rationale for the project and its methods.

Contact: christopher.hyde@pcmd.ac.uk

**Notes**

*Poster 21*

## Challenges of diagnostic systematic reviews in primary care settings: an example of rectal bleeding in cancer diagnosis

Margaret Astin, Tom Griffin, Richard Neal, Peter Rose, William Hamilton

Introduction: General Practitioners face diagnostic decisions every day. Much of the consultation between doctor and patient is taken up with clinical assessment of signs and presenting symptoms. Furthermore many symptoms occurring in primary care patients are common to a range of conditions. Risk estimates underpin national guidance, such as the NICE Referral Guidelines for Suspected Cancer (2000, 2005). We conducted a systematic review of symptoms associated with colorectal cancer to assess the risk of cancer.

Methods: We searched MEDLINE, EMBASE, MEDLINE in process, the Cochrane Library and CINAHL in April 2009, updated in February 2010, for studies of any design in symptomatic adult patients in primary care. We excluded studies of asymptomatic patients, screening, referred populations, patients with colorectal cancer recurrences or fewer than 100 participants. The target condition was carcinoma of colon or rectum. We extracted data for 2x2 tables to estimate performance characteristics for each symptom. Data were pooled in a meta-analysis. Quality of studies was assessed with the QUADAS tool.

Results: The searches identified 1896 papers; 50 were appraised, and 23 met inclusion criteria. We grouped all data on rectal bleeding together. Thirteen papers of 18,634 patients with rectal bleeding provided positive predictive values (PPVs) ranging from 2.2% to 15.8% with a pooled estimate of 6.3% (95% confidence interval 4.0% to 10.0%). A subgroup analysis of five studies with 887 patients aged over 50 found a pooled estimate of 8.1% (95% CI 6.0% to 10.8%). For rectal bleeding accompanied by anaemia, weight loss, or change in bowel habit positive likelihood ratios (PLRs) were 7.88, 1.88 and 1.81 respectively. Abdominal pain, diarrhoea or constipation accompanied by rectal bleeding yielded PLR values between 0.3 and 1.03. Findings from large case control or cohort studies using electronic databases differed from smaller prospective studies. Heterogeneity was variable between studies and performance characteristics. The QUADAS items that scored the lowest were differential and partial verification bias.

Conclusions: This review illustrates some challenges in primary care research for evidence synthesis of diagnostic studies. When patients present with particular symptoms or conditions the lack of true and false negatives limits the analyses that may be undertaken. Large retrospective studies using GP databases may yield very different findings to prospective designs. Primary care is at the interface of secondary care where most investigations are undertaken, yet complete validation of tests by reference standards that may be invasive may not be practical or ethical.

Reference: NICE Referral Guidelines for Suspected Cancer. London 2005.

Contact: m.p.astin@bristol.ac.uk

**Notes**

*Poster 22*

## Meta-analysis challenges in pharmacogenetics: A case study of CYP2D6 genetic variant in breast cancer

Nigel Fleeman, James Oyee, Rumona Dickson

Background: In the UK, NICE recommends tamoxifen or aromatase inhibitors for breast cancer patients undergoing hormonal therapy. Recently, the potential effect of CYP2D6 genetic variants (polymorphisms) on clinical response to tamoxifen treatment in breast cancer patients has gained much interest to the scientific community and there is growing expectation that testing for CYP2D6 status may be used to guide future decision making regarding an individual patient's hormonal therapy. In this paper, we highlight obstacles in attempting to evaluate the evidence for differences in overall survival (OS) by predicted clinical response.

Methods: A systematic review was undertaken to identify all studies on CYP2D6 genotyping related to tamoxifen treatment. An attempt was made to classify patients based on predicted metaboliser status as follows: poor metaboliser (PM), intermediate metaboliser (IM), extensive metaboliser (EM) and ultra-rapid metaboliser (UM).

Results: Nine studies from 8 different cohorts reported OS by CYP2D6 status. Not all the studies measured CYP2D6 status in the same way. For example, some studies simply examined one allele (e.g. *4) and presented data by genotypes, other studies used phenotypes (e.g. PM although these were not always defined in the same way), while others considered enzymatic function (e.g. decreased activity). Aside from the fact that some heterogeneity across studies was also evident, it was thus impossible to pool data from studies and therefore no meta-analysis was performed.

Conclusions: Despite recent development in genotyping platforms which enable concurrent investigation of several alleles, our review has identified a lack of agreement about which common alleles to test for and how to classify these even when the same alleles are tested across studies (e.g. should a patient with the *4/wt genotype be considered an EM, IM or PM?). Agreement about which alleles to test for would help interpret findings to describe any association of polymorphisms and clinical outcomes in pharmacogenetic studies. Moreover, to date, studies have only retrospectively examined associations between outcomes and genotype/phenotype/enzyme function, i.e. "clinical validity". Decision makers ideally require evidence from prospective studies in which patients are tested for CYP2D6, divided into groups that are exposed or not exposed to the intervention(s) of interest before the outcomes have occurred, i.e. "clinical utility". However, until agreement about which alleles to test for is reached, such studies are unlikely to be either possible or desirable.

Contact: nigel.fleeman@liverpool.ac.uk

**Notes**

*Poster 23*

## Systematic review of test performance of mammography in detecting ipsilateral breast tumour recurrence and metachronous contralateral breast cancer

Clare Robertson, Senthil Kumar Arcot Ragupathy, Charles Boachie, Cynthia Fraser, Steve Heys, Graeme MacLennan, Graham Mowatt, Ruth Thomas, Fiona Gilbert and the Mammographic Surveillance Health Technology Assessment

Introduction: X-ray mammography (XRM) is the reference standard imaging technique for detecting or confirming the absence of ipsilateral tumour recurrence or ipsilateral second primary cancer (IBTR) and metachronous contralateral breast cancers (MCBC) in women previously treated for primary breast cancer. IBTR and MCBC are usually detected by either clinical examination and/or XRM. Approximately 50% of local disease recurrences in the breast will be detected by XRM, with the remainder detected clinically. Approximately 10% of palpable tumours are not demonstrated on XRM and require additional imaging and investigation. Diagnostic tests for IBTR and MCBC can, therefore, be used in two ways; as a diagnostic test forming part of a routine surveillance regimen or as part of non-routine surveillance to evaluate a suspicious result on a prior diagnostic test. This also raises the question of whether XRM should be replaced with a more sensitive diagnostic test, although this carries the possibility of increased false positive results, causing further unnecessary investigations (invasive and non-invasive). We conducted a systematic review to determine the diagnostic accuracy of surveillance XRM for detecting IBTR and MCBC in women previously treated for primary breast cancer.

Methods: Studies were identified by a systematic electronic search from 1990 onwards. The index test for the review was XRM. Comparator tests included ultrasound, MRI, specialist led clinical examination and unstructured primary care follow-up. The reference standard was histopathological assessment for test positives and a period of follow up for test negatives. We evaluated the quality of studies using an adapted version of the QUADAS tool. We calculated sensitivity, specificity, positive and negative likelihood ratios and diagnostic odds ratio for each included study. We planned to combine results in a meta-analysis using the hierarchical summary receiver operating characteristic (HSROC) framework.

Results: From 246 potentially relevant titles and abstracts, 7 direct head-to-head cohort and 2 single cohort studies were included. Variations in study comparisons precluded meta-analysis. None of our considered tests were used for the same purpose (i.e. routine or non-routine surveillance, IBTR or MCBC detection). Three studies evaluated the performance of surveillance XRM, and MRI, for detecting IBTR in routine surveillance patients but most tests, and combinations of tests, were reported only by single studies. None of the studies evaluated test performance for detecting MCBC in non-routine surveillance patients. Findings suggest that although XRM is associated with a high sensitivity and specificity, MRI is the most accurate test for detecting IBTR and MCBC.

Conclusion: MRI is the most accurate test for detecting IBTR and MCBC in women previously treated for primary breast cancer. However these results should be interpreted with caution as they are based on only a few studies. This review also demonstrates some of the methodological complexities in conducting systematic reviews of diagnostic test performance where tests can be used for multiple purposes.

Contact: c.robertson@abdn.ac.uk

**Notes**

Poster 24

## Focused boosting (reweighting) to improve diagnostic classifiers

Philip Gichuru, Marc Aerts

Background: Misclassifications of diagnosis either false positives or false negatives do not occur symmetrically. Boosting is a general technique to combine several weak classifiers to produce a powerful committee that is able to correctly classify the misclassified subjects. The boosting algorithm focuses on the misclassified subjects, assigns increasing weights to these hard to correctly classify cases iteratively, and averages these analysis. In this way it can achieve substantial improvements in terms of error rates. The boosting (reweighting) algorithm however treats these two types of errors with equal importance not differentiating whether an error is false positive or false negative[1]. The primary objective was to modify the boosting algorithm and propose a method that focuses or attaches "more importance" to a specific type of misclassification.

Methods: Classification trees were used as the diagnostic instrument and misclassification error rate was used to compare and contrast the results with conventional boosting. Two approaches were set up to focus the boosting of errors. The first involved just altering the boosting algorithm to subjectively focus on given misclassifications and leaving the construction of the tree unaltered. The second approach used loss matrices to alter the construction of the tree but leave the normal boosting algorithm unchanged. In every scenario several bootstrap samples were used and averaged for each algorithm. The effect of increasing the number of iterations in the boosting algorithm was also checked.

Results: Results show that misclassifications do not occur in equal proportions. The proposed focused boosting algorithm succeeded in attaching more importance to a particular type of misclassification of interest. However in most of the cases, the lowest error was where false positives or false negatives got equal importance. Increasing the number of boosting iterations helped to some extent but this depended on the structure of misclassification and the importance attached at that time.

Conclusion: For good results thorough knowledge of the structure of misclassifications is required in order to choose most appropriate weights that attach subjective importance to the "more important" type of misclassification. Focused boosting will be much more advantageous than conventional boosting as one has more flexibility of directing the weighting of hard to classify observations.

References: [1]. Freund and Schapire, (1997). Adaboost.M1.

Contact: phillip.gichuru@csm.ox.ac.uk

**Notes**

*Poster 25*

## Pre-selection for screening by prediction models

Inge Stegeman, Roderik Kraaijenhagen, Patrick Bossuyt

Introduction: The invitation to population screening in the Netherlands is based on age criteria. Screening for breast, cervix or colorectal cancer is done in women between 50 and 75 years of age, women between 30 and 60 years, and in men and women between 55 and 75 years, respectively. Screening is not offered to younger or older participants, because the benefit-risk and/or the benefit-cost ratios are regarded to be too low. One could argue that, as compared to age alone, these ratios can be more accurately assessed if all known determinants of the benefits and harms are taken into account. This could be guided by a more complete evaluation of cancer risk.

Objectives: To evaluate the effectiveness of using cancer risk models in inviting people to population-based cancer screening programs.

Methods: We systematically searched the literature for risk factor models for the three types of cancer for which the Netherlands has or plans population screening: breast, cervix and colorectal cancer.

Results:
Breast Cancer
From multiple prediction models useful for pre-selection, the Gail model is the most widely known. Several studies have shown that the Gail model is more accurate in calculating risk than age alone, and that risk profiling can be helpful in decision making for mammographic screening, especially in the age range of 40-49.

Colorectal Cancer
Freedman et al. developed a risk algorithm, validated by Park et al; the model is regarded to be useful as a pre-screening tool for colorectal cancer.

Cervix Cancer
To date, no risk prediction model for cervical cancer has been validated. However, compared to human papillomavirus (HPV) age is only a weak predictor for cervical cancer. Beside HPV several other, weaker risk factors, have been identified.

Conclusion: Preliminary results show a potential to improve the effectiveness of screening by taking into account other established risk factors, not just age. This hypothesis is to be tested in clinical studies, as well as practicalities and issues of equity and reliability.

Contact: istegeman@niped.nl

**Notes**

*Poster 26*

## Systematic review of test performance when test positives and test negatives have a different reference standard

Clare Robertson, Senthil Kumar Arcot Ragupathy, Charles Boachie, Cynthia Fraser, Steve Heys, Graeme MacLennan, Graham Mowatt, Ruth Thomas, Fiona Gilbert and the Mammographic Surveillance Health Technology Assessment

Background: The Cochrane Handbook for Diagnostic Test Accuracy Reviews advocates the use of a single reference standard for all included studies. This is not appropriate for all health conditions, however. For example, in women undergoing surveillance following treatment for primary breast cancer, X-ray mammography (XRM) is the standard imaging technique for detecting or confirming the absence of ipsilateral breast tumour recurrence and second primary cancers in the same breast (IBTR) and metachronous contralateral breast cancers (MCBC). Histopathological examination is the reference standard for confirming presence or absence of cancer if suspicion of malignancy is raised by XRM and other investigations. In contrast, there is no recognised reference standard for ascertaining the true negative and false negative measures of a surveillance test. This is usually ascertained by a negative or a positive test result at subsequent testing after a period of follow-up has elapsed. This time interval is variable as highly sensitive tests are capable of detecting smaller tumours and are usually associated with a longer mean sojourn time. Consequently, longer follow-up time intervals are accepted in practice, e.g. one year for XRM or two to three years for magnetic resonance imaging (MRI).

Objective: To describe our experience of conducting a systematic review to determine the diagnostic accuracy of surveillance XRM for detecting IBTR and MCBC in women previously treated for primary breast cancer, following the principles recommended in the Cochrane Handbook.

Methods: Studies were identified by a systematic electronic search of the literature from 1990 onwards. The index test for the review was XRM. Comparator tests included ultrasound, MRI, specialist led clinical examination and unstructured primary care follow-up. The reference standard was histopathological assessment for test positives along with a follow-up period of up to 3 years for test negatives. We adapted the QUADAS tool to assess disease progression bias for test positive and test negative results separately.
Results: From 246 potentially relevant titles and abstracts, 7 direct head-to-head cohort and 2 single cohort studies were included. All studies used histopathology as the reference standard for positive test results and follow up for negative test results. All studies were within the 3 year limit set for follow up of test negatives, although across studies the individual time period varied for each of the considered tests. For the majority of studies the time period between positive testing and reference standard confirmation was unclear. Variations in study comparisons precluded meta-analysis.

Conclusion: The Cochrane Handbook was a very useful resource in conducting this review. However, adopting a single reference standard for test positives and negatives may not always be appropriate. Further guidance as to how Handbook recommendations might be adapted to suit individual clinical conditions would be beneficial to other researchers.

Contact: c.robertson@abdn.ac.uk

**Notes**

*Poster 27*

**Verification problems in diagnostic accuracy studies: consequences and solutions**

Karel Moons, Patrick Bossuyt, Johannes Reitsma, Nandini Dendukuri, Kristel Janssen, Joris de Groot

Background: in studies of diagnostic accuracy, results of the diagnostic test or model under study (index test or model) are verified by comparing them with results of a reference standard applied to the same patients. Accuracy measures, such as test sensitivity, specificity, likelihood ratios, predictive values, ROC area, or diagnostic odds ratio, express how well the results of the index test/model agree with the outcome of the reference standard. Ideally, every patient undergoes the index tests and the same reference standard. Moreover, the reference standard provides error-free disease verification (classification). Unfortunately, all three situations are often not met in diagnostic studies. We describe the most important types of disease verification problems using clinical examples, and proposes solutions to alleviate the associated bias.

Partial verification bias occurs when part of the patients does not undergo the reference standard (such that there is missing outcome data). Usually this is related to other, previous index test results.

Differential verification bias occurs when an alternative or second best reference test is applied to verify disease presence/absence in subjects where the result of the first (preferred) reference test can not be obtained for some reason (e.g. ethical or clinical). Bias arises when the results of the alternative reference standard are treated as if they came from the preferred reference standard, which is often done. The reason is that the two reference tests are of different quality, or define the target condition differently.

Imperfect Reference standard. For many diseases a reference standard is not 24 carat, but also yields misclassification on the outcome, i.e. 'true' presence/absence of the target condition. Bias in the estimated accuracy of the index tests occurs when this imperfectness is simply neglected.

Conclusions and Solutions: if the preferred reference standard has not been applied in all patients, mathematical correction methods (i.e. 'Begg and Greenes' method or Multiple Outcome Imputation) can be used to correct (as much as possible) for the partial verification bias. If in those patients perhaps a second best reference test (with likely a higher threshold of detecting disease presence) is conducted, will produce important additional information for the imputation model. When differential verification methods are used, overall accuracy estimates of the index test/model are very difficult to interpret. Results should be reported separately for each reference standard to provide informative and unbiased measures of the index test/model accuracy. When an imperfect reference test is used, Bayesian correction methods have been introduced to correct for (as much as possible) for this imperfection in the estimation of the accuracy of the index test/model.

Contact: k.g.m.moons@umcutrecht.nl

**Notes**

*Poster 28*

## Presenting clinically relevant information on diagnostic performance in systematic reviews: the use of predictive values (PPV and 1-NPV)

Petra Jellema, Danielle van der Windt, Christian Mallen, Stijn van Weijenberg, David Bruinvels, Henrica de Vet

Objective: We recently conducted a systematic review summarising available evidence on diagnostic tests that may assist the primary care physician in the identification of patients with an increased risk for colorectal cancer (CRC) among those consulting for non-acute lower abdominal symptoms. We aimed to present the results in a way that is informative for primary care physicians.

Data Sources: PubMed, EMBASE and reference screening.

Study eligibility criteria: Studies were selected if the design was a diagnostic study; the patients were adults consulting because of non-acute lower abdominal symptoms; tests included signs, symptoms, blood tests, or faecal tests.
Study appraisal and synthesis methods Quality assessment, using a modified version of the QUADAS tool, and data extraction was performed by two reviewers independently. We presented pooled estimates of sensitivity and specificity using bivariate analysis. We also calculated pooled estimates for the positive predictive value (PPV) and 1 minus the negative predictive value (1-NPV), representing the risk of CRC in patients with a positive test result and in those with a negative test result. We refrained from pooling when there was considerable clinical or statistical heterogeneity.

Results: Evidence for the performance of tests in diagnosing CRC in symptomatic patients presenting in primary care was scarce and varied widely. In contrast to our expectations, however, PPV and 1-NPV more often showed statistical homogeneity than sensitivity and specificity. This enabled statistical pooling of predictive values for age, gender, rectal bleeding and constipation, where this was not possible for sensitivity and specificity. Predictive values were informative when sensitivity or specificity showed wide heterogeneity and results could not be easily interpreted. For example, sensitivity of constipation was very poor and specificity ranged between 0.53 and 0.90. Pooled estimates of predictive values clearly showed that constipation was not associated with an increased risk of CRC, with a pooled estimate for PPV of 0.06 (0.0-0.18) and 0.10 (0.07-0.14) for 1-NPV, which helps to inform and reassure patients presenting with constipation. When pooled estimates could not be presented ranges of predictive values were still of interest. For example, the risk of CRC in patients with a positive iFOBT (PPV) ranged from 0.07 to 0.59, while this was 0.001 to 0.07 for a negative iFOBT result (1-NPV). This information may be more clinically useful than values of sensitivity and specificity ranging from 0.70 to 1.00.

Conclusions: Estimates of PPV and 1-NPV seem to provide information on diagnostic performance that is clinically useful. However, caution in their use is needed: the prior probability of disease may be an important source of heterogeneity determining the results of predictive values. Furthermore, additional methodological development may be needed to specify methods for the assessment of statistical heterogeneity, and explore other prerequisites for meta-analysis of predictive values.

Contact: hcw.devet@vumc.nl

**Notes**

*Poster 29*

## Pin the threshold on the ROC Curve: How to identify relevant decision thresholds?

Colin Everett, Julia Brown, Jane Nixon, Joanne Copeland

Background: Coronary Heart Disease (CHD) is typically diagnosed using X-ray Angiography: an invasive procedure, which - in addition to the risks involved with ionising radiation - carries a small but significant risk of morbidity. To reduce the numbers of unnecessary angiograms, patients are risk-stratified by undergoing a series of non-invasive tests such as Exercise Treadmill Testing (ETT) and Single Photon Emission Computed Tomography. (SPECT) This strategy of several non-invasive tests being performed is due to limitations in the test performance as well as their limited diagnostic accuracy.

Another recently-developed alternative is Cardiac Magnetic Resonance imaging (CMR) which can assess multiple aspects of CHD in a single protocol, and offers greatly-improved image spatial resolution compared to SPECT. To assess CMR's ability to diagnose CHD, CEMARC (Clinical Evaluation of Magnetic Resonance imaging in Coronary heart disease) - a prospective cohort study of 752 patients – will estimate the diagnostic accuracy of a CMR protocol in diagnosing CHD – using X-ray Angiography as the reference standard - and compare it to the accuracies of SPECT and ETT.

Issues: CMR and SPECT allow investigators to score individual segments of the three main coronary arteries according to any defects found and their severity. Summing these provide overall "Stress Perfusion Scores" for the patient. ETT provides us with measures of ST Segment Deviation, Heart Rate recovery in 1 minute and the Duke Treadmill Score. To compare the diagnostic accuracy of these scores against the reference standard, Receiver Operating Curve (ROC) Analysis is undertaken, presenting the sensitivity and false-positive rate of each cut-off value along with the estimated Area Under the Curve. However, with so little information on the use of some of these measures as prognostic indicators, how should appropriate decision thresholds be identified from these?

Various statistical approaches have been suggested. A discussion of the some of the approaches and methods available will be presented, along with the consequential decision thresholds they produce for a simulated dataset.

Contact: c.c.everett@leeds.ac.uk

**Notes**

*Poster 30*

**Joint evaluation of sensitivity and specificity and its impact on the sample size**

Wener Vach, Oke Gerke, Poul Flemming Hoilund-Carlsen

Background: It is common knowledge today that sensitivity and specificity need to be evaluated together in the analysis of the accuracy of a diagnostic procedure. However, these target parameters are usually evaluated separately in the statistical analysis.

Methods: At hand of the simple setup of a diagnostic study comparing a single binary test with a gold standard, we discuss weighted averages of and two-dimensional confidence regions for sensitivity and specificity as alternative concepts. These concepts enable a joint evaluation and, hence, also a trade-off between sensitivity and specificity.

Results: We illustrate the impact of using the different concepts on the power of a study and on sample size calculations. With respect to the latter we point out that different concepts may lead to differences in sample sizes up to a factor of 6. In particular quarter elliptic confidence regions present a powerful approach. Weighted averages are a competitor, but require to specify a plausible range of weights.

Conclusions: We conclude that it is time to take the joint evaluation of sensitivity and specificity more serious in analyzing the accuracy of binary diagnostics tests.

Contact: wv@imbi.uni-freiburg.de

**Notes**

*Poster 31*

## Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown

Philip Gichuru, Niel Hens

Background: Evaluation of diagnostic tests for disease is difficult due to lack of a gold standard test. We begin by reviewing the work of Young-Ku Choi et al (2006). We propose a Bayesian method to overcome the lack of a gold standard test in ROC curve estimation. A Bayesian approach assures the natural monotonicity property of the resulting ROC curve estimate that caters for correlation between tests of a subject. Little work has been done to incorporate covariates in the analysis without a gold standard. In this paper, we propose a method for estimating ROC curves based on Bayesian models while adjusting for covariate effects when the true disease status of tested subjects is unknown. The covariates may be correlated with the disease process, with the diagnostic testing procedure, or both. However we only highlight the former case.

Methods: A linear model was used to fit the covariates to the distribution of test scores. Markov chain Monte Carlo (MCMC) methods are employed to get posterior estimates of the sensitivities and specificities that facilitate inference about the diagnostic accuracy. This involves plotting sensitivity versus (1-specificity) over possible cutoffs making up the ROC plot. Our method can be applied to tests with and without gold standard and to multiple correlated tests on the same patients. A simulation study was set up to check the discrimination ability of the Gold Standard (GS) versus Non Gold Standard (NGS) on the one hand and the same when adjusted for covariates on the other hand. The area under the ROC curve quantified the diagnostic accuracy of the tests and the difference between Area Under Curve (AUC) to assess the tests' accuracy.

Results: The NGS method performs relatively well in comparison to the GS method more so when covariate adjustment is incorporated. Our findings motivated a different approach to measuring delta, the closeness of distribution of diseased and non diseased test values. The methodology was applied to diabetes and pancreas data sets. Findings were in accordance with our simulation study.

Conclusion: Covariates adjustment if correlated with disease process proved beneficial towards correcting diagnostic test accuracy as far as AUC is concerned. Catering for correlation between subjects' tests gave more accurate estimates. It affected the widths of posterior probability intervals.

References: Y.K Choi, W.O. Johnson, M.T. Collins, and I.A. Gardner. Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. JABES, 11:210-229, 2006.

Contact: phillip.gichuru@csm.ox.ac.uk

**Notes**

*Poster 32*

## Interim analyses in diagnostic studies differ from interim analyses in treatment studies

Werner Vach, Oke Gerke, Poul-Flemming Hoilund-Carlsen

Interim analyses are today an integral part of most late phase II and phase III treatment trials. Similarly, if a new diagnostic procedure turns out to be inferior to the current standard diagnostic tool, both the patients in the study as well as the scientific community can benefit from this information. To some extent, interim analyses in diagnostic studies impose similar issues as therapeutic studies, but we think that there are four substantial differences:

1) Basing the analysis of diagnostic studies on confidence intervals as means to estimate differences in sensitivity and specificity with a certain precision, the decision on early stopping will be no longer based on p-values, but on the length of the confidence intervals which do not affect the validity of the confidence interval itself, therefore circumventing the need for adjustment for multiple inferences.

2) In a paired design setting, continued application of both the inferior and the superior procedure in all patients after interim analysis may contribute to the subsequent treatment planning for the patient, especially if the inferior procedure does no or only minimal harm to the patient, but may contribute to the subsequent treatment planning for the patient with additional information.

3) In a paired design setting, the sample size depends heavily on the degree of agreement between the two diagnostic procedures for which a reliable estimate is hard to guess at planning stage. Starting with a probably wrong working assumption, the degree of agreement can be estimated at interim (even if the gold standard results are unknown) and the sample size may be re-adjusted properly. Due to the difficulties in obtaining a reliable initial estimate, it may be argued that such adjustments should be allowed for any diagnostic study using a paired design anyway. It is very likely to regularly observe the need for such an adjustment.

4) In diagnostic studies, blinding at the individual level requires only that both diagnostic procedures are performed without knowledge of the results of the other. Once both procedures are carried out, optimal treatment decision can be based on the results recorded earlier. Moreover, it cannot be avoided that some doctors, nurses, or other people involved in the study build up some vague knowledge about emerging trends which is likely to differ between different people. Therefore, interim analyses with subsequent sharing of knowledge about emerging trends can contribute to a unification of the common knowledge.

Interim analyses in diagnostic studies cannot imply the logic of interim analyses from treatment studies. In the diagnostic setting, some issues disappear, whereas other challenges arise. The possibility for long-term blinding in diagnostic studies is much more limited (or even impossible) which may impose a larger need for interim analyses to control the flow of information on emerging trends.

Contact: wv@imbi.uni-freiburg.de

**Notes**

*Poster 33*

**Reporting of primary diagnostic studies – an example from a national guideline**

Roberta Richey, Elizabeth Shaw, Judith Thornton

Background: Systematic reviewing methods rely heavily on the level and accuracy of reporting in published papers. Although ideal practice is to confirm study methods with study authors, this is not always possible. As statements of reporting of primary studies (such as CONSORT for randomised controlled trials, and STARD diagnostic tests) have been developed, reporting should be improved allowing easier data extraction.

Aims: To describe the adherence to STARD in reported studies reviewed for a national clinical guideline on diagnosis, and to describe the issues encountered when extracting data for the 2x2 tables.

Methods: We re-examined a sub-set of 19 included studies to assess and describe

- the reporting standards against STARD
- quality against QUADAS
- issues and problems in extracting information for the 2x2 table.

Results: Studies varied in both quality and levels of reporting. However, it was not clear to what extent quality assessment would be improved if study authors were contacted for clarification of methods. The significant challenge was extraction of data for the 2x2 table. Often, this was not reported and had to be 'back-calculated' from other results reported.

Discussion: Initiatives such as STARD and QUADAS have made assessment and extraction from published diagnostic accuracy papers easier, but there are still significant challenges, particularly when reviewing papers published pre-STARD. This may have particular relevance to guidelines where a grading system, such as GRADE is being used, and we will discuss this further.

Contact: beth.shaw@nice.org.uk

**Notes**

*Poster 34*

## Evaluation of the methodological quality of diagnostic studies: experiences with QUADAS and suggestions for amendments

Heike Raatz, Katja Suter, Inger Janssen, Regina Kunz

Background: Systematic reviews on the diagnostic accuracy of a test require a thorough methodological assessment of the included studies. The QUADAS instrument[1], developed for this purpose, detects risk of bias inherent in patient selection, the execution of the index test, and the reference standard.

Objectives: We explored whether QUADAS captured all relevant sources of bias when the index test was compared to a concurrent routine test and when the reference standard is follow-up.

Methods: We applied the QUADAS tool in a systematic review on positron-emission-tomography (PET) compared to conventional tests for assessing the diagnostic / prognostic value of interim-PET in patients with lymphoma. A comprehensive literature search with 1144 references yielded 7 included studies. Some but not all compared PET to conventional tests. All used follow-up as reference standard.

Results: We found several limitations:

- QUADAS requests a short interval between index test and reference standard. With follow-up as reference standard, studies need to demonstrate sufficiently long follow-up to distinguish recurrence and healing.
- Reviewers need to assess the possibility of confounding during follow-up.

Limitations when reviewing studies with 2 or more comparison tests:

- QUADAS examines the performance of the index, but not the comparator test.
- QUADAS does not inquire about mutual blinding of readers reviewing 2 tests with subjective reading (e.g. PET vs. CT).
- QUADAS does not explore whether the statistical method takes into account the lack of independence of the results of index and routine test when derived from the same patients.

Conclusions: Currently, QUADAS lacks certain aspects in assessing risk of bias when comparing a new index test to a concurrent routine test and when the reference standard is follow-up. A QUADAS update should consider these additional criteria.

(1) Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol 2006; 6:9.

Contact: hraatz@uhbs.ch

**Notes**

*Poster 35*

## Sales and self-tests on the world wide web

Geraldine van der Meer, Marieke den Breejen

Background: In-vitro diagnostics are medical devices that are used for the examination of body material to provide information for prevention, diagnosis, prognosis and therapy. This study focused primarily on in-vitro diagnostic devices for self-testing that are offered on the internet. Self-tests can be used by any individual with access to the internet to diagnose conditions without the involvement of a health professional. There are increasing signs that counterfeit self-tests are being offered for sale on the internet.

Objective: This study was intended to make an inventory of the self-tests that are offered on the internet and asses if self-tests sold on the internet meet the essential requirements stated in the European and Dutch legislation.

Methods: Design A multi-method approach was used containing qualitative observational and quantitative research. It was executed as a mystery shop study conducted on the internet during March 2010. Setting Data sources were websites written in English and Dutch. In- and exclusion criteria Self-tests that were included are tests laymen could execute themselves and interpret the results without involving a health professional. Self-tests were excluded when the website did not have them in stock or did not sell the tests. Selection criteria First criteria 'two high risk in-vitro diagnostics and two low risk in-vitro diagnostics'. The second criteria 'a large demand from the public for specific self-tests'. HIV-, hepatitis-, cholesterol- and diabetes tests met the selection criteria. Measurements Data were collected from the online offer of self-tests and the self-tests that were purchased. In addition the Dutch National Institute for Public Health and the Environment analyzed the quality of information given with the received self-tests. Analyses The information in the Instructions For Use (IFU) and on the label of the self-tests were analyzed with a quality criteria checklist. The data were coded and analyzed in Microsoft access.

Results: After internet research 16 websites were identified. In total 22 of the selected self-tests were available and ordered on the internet, 64% of them have been delivered. Of the 7 purchased high risk diagnostics, 4 were delivered. According to the Dutch Decree on in-vitro diagnostics high risk diagnostics should only be provided with the involvement of a physician of a pharmacist. It is notable that a urine HIV self-test has been delivered, when HIV can only be diagnosed with a blood sample.

Conclusion: The results will be available halfway June 2010

Contact: geraldinevdmeer@hotmail.com

**Notes**

*This page is intentionally blank*

*This page is intentionally blank*

*This page is intentionally blank*

*This page is intentionally blank*