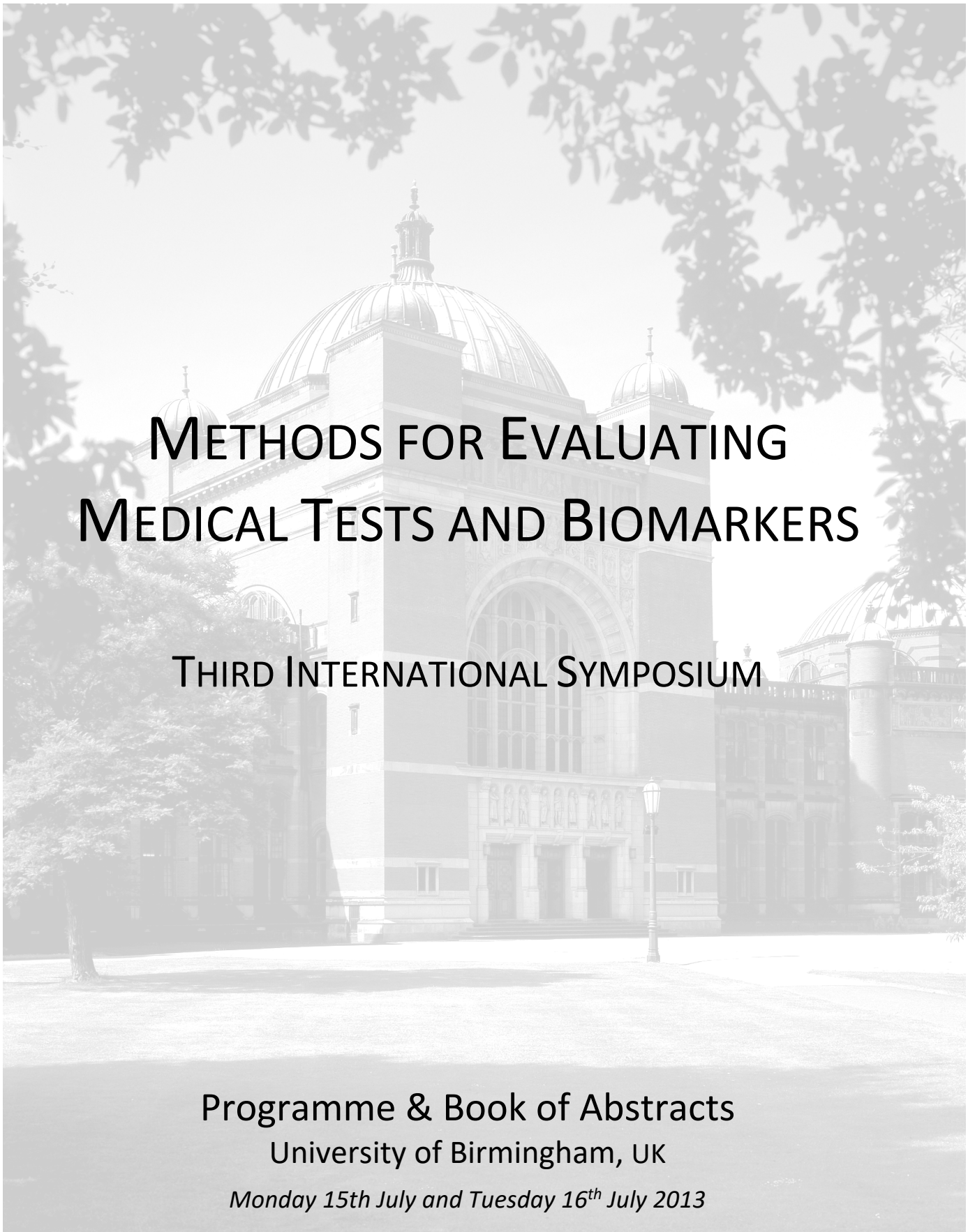
UNIVERSITY OF BIRMINGHAM



METHODS FOR EVALUATING MEDICAL TESTS AND BIOMARKERS

THIRD INTERNATIONAL SYMPOSIUM

Programme & Book of Abstracts

University of Birmingham, UK

Monday 15th July and Tuesday 16th July 2013

Welcome

The importance of rigorous and scientific evaluation of medical tests and biomarkers is increasingly recognised in health care, both through developments in test technology and greater emphasis on identifying cost effective diagnostic and monitoring strategies for use in health care. However, the design, execution, analysis, reporting and implementation of evaluations of medical tests and biomarkers present unique methodological challenges.

This multidisciplinary symposium provides a forum for discussing and disseminating recent research and stimulating dialogue amongst researchers and healthcare professionals actively involved in evaluating medical tests.

Hosted by the Test Evaluation Research Group at the University of Birmingham, the 2013 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:

Evaluating the impact of tests
Primary Studies and Meta-Analysis of Test Accuracy
Prognosis Research
Translating Test Research into Practice – Guidelines, Evidence Reports and Technology Assessments
Industry, Regulation and Test Development
Reference Standards and Comparisons
Monitoring and Longitudinal Studies
Prediction, Classification and Clinical Decision Rules

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:

Paul Good (co-chair)

Susanna Wisniewski (co-chair)

Lavinia Ferrante di Ruffano

Anne Walker

Karen Biddle

Scientific Committee

Professor Jon Deeks – University of Birmingham, UK

Dr. Clare Davenport – University of Birmingham, UK

Professor Patrick Bossuyt – University of Amsterdam, Netherlands

Dr Richard Riley – University of Birmingham, UK

Dr Matthew Thompson – University of Oxford, UK

Dr. Brian Willis – University of Birmingham, UK

How to cite this publication

The Abstract book should be cited as:

Methods for Evaluating Medical Tests and Biomarkers. Symposium; 2013 Jul 15-16; 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; 2013 Jul 15-16; Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, UK. Page number(s).

Table of Contents

Programme Overview

Monday 15th July

09:00	Registration opens	Ground Floor, Wolfson Centre
09:30	**Session 1** **Welcome and Introduction** **Evaluating the impact of tests**	Leonard Deacon Lecture Theatre
10:45	*Morning Coffee/Tea and Poster Viewing*	Ground Floor Wolfson Centre
11:15	**Session 2** **Primary Studies and Meta-Analysis of Test Accuracy**	Leonard Deacon Lecture Theatre
13:00	*Lunch*	Lower Ground Floor Wolfson Centre
13:45	**Session 3** **Prognosis Research**	Leonard Deacon Lecture Theatre
15:15	*Afternoon Coffee/Tea and Poster Viewing*	Ground Floor Wolfson Centre
15:45	**Session 4** **Translating Test Research into Practice – Guidelines, Evidence Reports and Technology Assessments**	Leonard Deacon Lecture Theatre
17:30	*Close*	
19:30	**Conference Dinner**	The Jam House

Tuesday 16th July

09:15	**Session 5** **Industry, Regulation and Test Development**	Leonard Deacon Lecture Theatre
10:15	*Morning Coffee/Tea and Poster Viewing*	Ground Floor Wolfson Centre
10:45	**Session 6** **Reference Standards and Comparisons**	Leonard Deacon Lecture Theatre
12:30	*Lunch*	Lower Ground Floor Wolfson Centre
13:15	**Session 7** **Monitoring and Longitudinal Studies**	Leonard Deacon Lecture Theatre
14:30	*Afternoon Coffee/Tea*	Ground Floor Wolfson Centre
14:45	**Session 8** **Prediction, Classification and Clinical Decision Rules**	Leonard Deacon Lecture Theatre
16:15	*Closing remarks*	

Full Programme

Monday 15th July

09:00 Registration – Ground Floor Wolfson Centre

Session 1 – Evaluating the Impact of Tests - Leonard Deacon Lecture Theatre

Chair – Jon Deeks

09:30 – 09:35 Introduction and welcome
Jon Deeks

09:35 – 10:05 Evaluating the impact of medical tests
Sally Lord (Keynote Speaker)

10:05 – 10:25 Selecting the optimal diagnostic strategy for patients with stable angina: an evidence report for the Dutch Health Care Insurance Board.
Ann van den Bruel

10:25 – 10:45 Cost Benefit Analysis for the evaluation for new genetic tests : a pilot feasibility study
Martin Eden

10:45 – 11:15 *Morning Coffee/Tea and Poster Viewing* Ground Floor Wolfson Centre

Session 2 – Primary Studies and Meta-Analysis of Test Accuracy - Leonard Deacon Lecture Theatre

Chair – Matthew Thompson

11:15 – 11:35 Individual participant data meta-analysis to examine the accuracy of serum mesothelin for diagnosing malignant pleural mesothelioma
Hans Reitsma

11:35 – 11:55 Meta-analysis of diagnostic test accuracy studies with multiple thresholds
Richard Riley

11:55 – 12:15 Remits, roles and working models for Trial Steering Committees and Data Monitoring Committees in studies evaluating diagnostic tests: a survey of current practice
Lavinia Ferrante di Ruffano

12:15 – 12:35 Do we need to improve the study design framework to evaluations of test accuracy?
Chris Hyde

12:35 – 12:55 Diagnostic accuracy studies: How to report and analyse inconclusive test results
Bethany Shinkins

13:00 – 13:45 *Lunch Break* Lower Ground Floor Wolfson Centre

Session 3 – Prognosis Research - Leonard Deacon Lecture Theatre

Chair – Richard Riley

13:45 – 14:15 Prognosis research: new opportunities in linked electronic health records
Harry Hemingway (Keynote Speaker)

14:15 – 14:35 The interplay of sensitivity and specificity when comparing risk models and classification rules
Ben van Calster

14:35 – 14:55 Net Reclassification Risk: a graph to clarify the potential prognostic utility of new markers
Ewout W. Steyerberg

14:55 – 15:15 An overview of systematic reviews of the prognostic utility of platelet function tests for predicting vascular event rates in patients on aspirin therapy: implications of discrepancies in study identification and selection
Janine Dretzke

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

Session 4 – Translating test research into practice – Guidelines, evidence reports and technology assessments - Leonard Deacon Lecture Theatre

Chair – Clare Davenport

15:45 – 16:15 *The Role of Modelling in Test Evaluation: Experience in NICE Diagnostic Appraisals.*
Chris Hyde (Keynote Speaker)

16:15 – 16:35 *Bringing risk prediction models into the 21st century*
Patrick Bossuyt (Keynote Speaker)

16:35– 16:55 A Systematic Review of Grading Systems for Medical Tests for Guideline Developers
Gowri Deurenberg Gopalakrishna

16:55 – 17:55 Interactive Session: The identification and prioritisation of diagnostic research: the experiences of three European Agencies
Nick Hicks

Tuesday 16th July

Session 5 - Industry, regulation and test development - Leonard Deacon Lecture Theatre

Chair – Patrick Bossuyt

09:15 – 09:45 *Ajit Lalvani (Keynote Speaker)*

09:45 – 10:15 *Matthew Thompson (Keynote Speaker)*

10:15 – 10:45 Morning Coffee/Tea and Poster Viewing **Ground Floor Wolfson Centre**

Session 6 – Reference Standard & Comparisons - Leonard Deacon Lecture Theatre

Chair – Karel Moons

10:45 – 11:05 Methods for diagnostic meta-analysis in the absence of a gold-standard reference
Nandini Dendukuri

11:05 – 11:25 Evaluating Diagnostic Accuracy in the Face of Multiple Reference Standards
Christiana Naaktgeboren

11:25 – 11:45 Shortcomings in Reporting and Methodology of Latent class models (LCMs) in Diagnostic Research – A Systematic Review
Maarten van Smeden

11:45 – 12:05 Reporting and methods in systematic reviews of comparative accuracy
Yemisi Takwoingi

12:05 – 12:25 Direct versus Indirect Comparisons in Systematic Reviews of Test Accuracy Studies: An IPD Case Study in Ovarian Reserve Testing
Junfeng Wang

12:30 – 13:15 Lunch Lower Ground Floor Wolfson Centre

Session 7 – Monitoring & Longitudinal Studies - Leonard Deacon Lecture Theatre

Chair – Brian Willis

13:15 – 13:35 How to model longitudinal data for the prediction of hemoglobin level in whole blood donors
Yvonne Vergouwe

13:35 – 13:55 Comparing strategies for interpreting longitudinal CEA measures when screening for colorectal cancer recurrence
Bethany Shinkins

13:55 – 14:15 Evaluation of time dependent diagnostic accuracy using repeated measurements of a biomarker
Ruwanthi Kolamunnage-Dola

14:15 – 14:35 Has the randomised controlled trial design been successfully used to evaluate new monitoring strategies? An assessment of experience to date
Jac Dinnes

14:35 – 14:45 Afternoon Coffee/Tea Ground Floor Wolfson Centre

Session 8 –Prediction, classification and clinical decision rules - Leonard Deacon Lecture Theatre

Chair – Jon Deeks

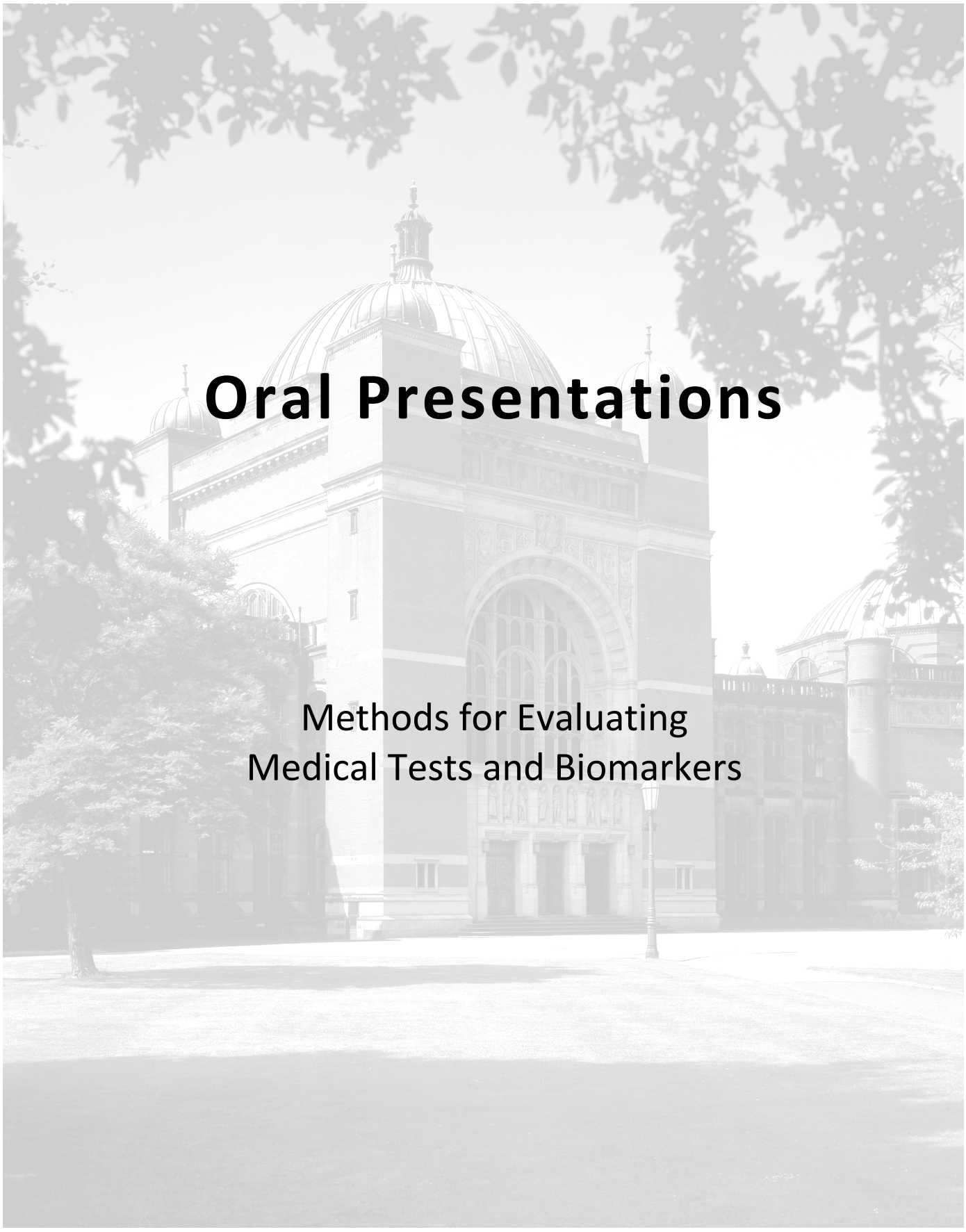
14:45 – 15:15 Prediction, Classification and Clinical Decision Rules: Methodological Conduct and Reporting
Gary Collins (Keynote Speaker)

15:15 – 15:35 A framework for developing and implementing clinical prediction models across multiple studies
Thomas Debray

15:35 – 15:55 The Leicester Diabetes Risk Assessment Tools
Laura Gray

15:55 – 16:15 Prediction study risk of bias assessment tool (PROBAST)
Karel Moons

16:15 – 16:30 Closing remarks
Jon Deeks



Oral Presentations

Methods for Evaluating
Medical Tests and Biomarkers

Keynote speaker

Evaluating the impact of medical tests

Sally Lord, NHMRC Clinical Trials Centre, The University of Sydney, Australia

Medical tests and biomarkers can improve our understanding of disease and provide insights for developing new treatments. Ideally these advances also lead to better tools for diagnosis and disease classification to improve decisions about clinical care.

The potential benefits of using new tests to improve clinical care are substantial, but also demand considerable research resources to validate clinical performance and verify claims of better outcomes. While there is broad agreement that the clinical benefits of a new medical test must be evident before recommending its use in practice, there is less agreement on what the minimum evidence requirements are. There are also unique practical challenges for demonstrating improved health outcomes and evaluating other impacts.

This session will present a framework for evaluating the impact of medical tests, and discuss the principles to address these issues. Using this framework, test evaluation begins by defining the clinical claim for the test and how it will alter the current clinical pathway for the proposed indication. The clinical pathway is central to defining the research questions and determining what type of comparative evidence is needed to evaluate clinical performance and impact on downstream outcomes. Examples will be presented to illustrate the importance of studies beyond traditional measures of test accuracy, to assess the clinical significance of test results relevant to decisions about clinical care; and judgements about the need for new trials to evaluate the impact of subsequent changes in clinical care.

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

Notes

Contributed paper

Selecting the optimal diagnostic strategy for patients with stable angina: an evidence report for the Dutch Health Care Insurance Board.

Ann van den Bruel, Jooan Vlayen, Yolba Smit, Mattias Neyt, Rafael Perera

Introduction: Stable angina is chest discomfort or pain, often provoked by physical exertion. Patients with stable angina are investigated for coronary artery disease and subsequently treated with medication or revascularisation procedures if needed.
New imaging techniques such as coronary CT angiography (CCTA), MRI angiography or SPECT scans may be better at identifying coronary artery disease in patients with stable angina. However it is unclear whether, compared to the usual strategy, these tests lead to improved patient outcomes including decreased mortality and acute myocardial infarction during follow-up.
The aim of the evidence report was to summarise all available evidence of different diagnostic strategies on patient outcome.

Methods: The literature was searched in Medline, Embase, Central, the Cochrane Database of Systematic Reviews and DARE. Randomized controlled trials or systematic reviews of RCTs were of primary interest. When these were not identified, diagnostic accuracy studies were sought. Selection, quality assessment and data extraction were done in duplicate. Meta-analyses were performed when appropriate. Finally, the effect of ten different diagnostic strategies, including various combinations of CCTA, MRI, SPECT, stress ECG, stress ultrasound (US) and coronary angiography (CA), on short-term (diagnostic accuracy, costs, radiation exposure, procedure-related mortality) and long-term outcomes (all-cause mortality and acute myocardial infarction) was modelled in a decision analysis.

Results: Fourteen RCTs were identified: none of the RCTs provided high level of evidence. Most studies were of short duration, had important methodological weaknesses and were underpowered to detect effects on mortality or myocardial infarction. In addition, the literature search yielded 15 systematic reviews on diagnostic accuracy; almost all studies had an inappropriate spectrum of disease because patients were recruited at referral for the reference standard.
Based on the modelled cost per true positive, all strategies with MRI and SPECT were dominated and therefore offer less diagnostic value at a higher cost. The remaining strategies are CCTA+CA, stress US+CA, stress US+CCTA+CA, stress ECG+CCTA+CA and CA. The CA only strategy is associated with the highest procedure-related mortality, leaving four potentially promising strategies: stress US+CA, stress US+CCTA+CA, CCTA+CA and stress ECG+CCTA+CA. The lack of reliable input parameters prevented modelling of long-term outcomes.

Conclusions: Existing evidence is insufficient to recommend one particular diagnostic strategy over another. Modelling suggests four strategies could be prioritized for evaluation in RCTs.

Contact: ann.vandenbruel@phc.ox.ac.uk

Notes

Contributed paper

Cost Benefit Analysis for the evaluation for new genetic tests : a pilot feasibility study

Martin Eden, Katherine Payne, Ryan Combs, Georgina Hall, Marion McAllister, Graeme Black

Background: In the UK, the National Institute for Health and Clinical Excellence (NICE) encourages the use of cost effectiveness analysis (CEA) as an input into national decision making. Using CEA means the evaluative framework focuses on maximising health gain, which may not adequately capture the broad range of potential benefits from diagnostics. The need to value more than health gain may be particularly relevant for genetic-based diagnostics, which often do not offer direct health benefits. Consequently, using the current NICE recommended evaluative framework means that genetic diagnostics may be deemed to be an inefficient use of NHS resources. Cost-benefit analysis (CBA) is an alternative evaluative framework, which translates all costs and benefits into monetary values to allow inclusion of health, non-health, process and spillover values into the decision-making process. Technological advances in genetic diagnostics means more individuals can now be offered testing and has NHS resource implications. The need to value the benefits of genetic diagnostics has led to renewed interest in the potential application of CBA but the approach has attracted criticism within the health economics community due to theoretical and practical weaknesses in its application. It was within this context that this pilot study was designed to identify if, and how, CBA could be used in the evaluation of new genetic tests to diagnose inherited eye diseases.

Methods: The contingent valuation method, using willingness to pay, was used to elicit monetary values for two 'diagnostic' scenarios for inherited eye diseases (1) genetic counseling only (2) genetic testing and genetic counselling. The WTP values were elicited using telephone interviews and a semi-structured qualitative interview schedule was used to explore the reasons behind the stated WTP values. A purposive sample of 52 participants was recruited comprising people with experience of inherited eye diseases (n=25) and people with no experience representing the general public perspective (n=27). Individuals' maximum WTP values for both scenarios were elicited using an iterative bidding game and analysed using descriptive statistics. The telephone interviews were recorded, transcribed and analysed using thematic framework analysis. The analytic framework was structured to capture themes on (1) identification & measurement of relevant value types and (2) identification of biased and anomalous WTP values.

Results: The majority of participants stated that they would seek genetic counselling and testing in this context. Stated WTP was higher for genetic counselling plus testing (median=£524 IQR=£174 to £1024) compared to counselling alone (median=£224 IQR=£99 to £524). Respondents offered similar justifications for stated WTP values. Qualitative analysis indicated that health & non-health value and perceived spillover effects had influenced stated WTP amounts (e.g. test-derived information was considered valuable because it could aid important decisions and increase well-being for individuals and their family). Data suggested that certain types of perceived benefits (including non-use values and some spillover effects) would not be reflected in the monetary amounts without modification to the pilot WTP study design used here.

Conclusions: Participants were able to attach a monetary value to the perceived potential benefit that genetic testing offered regardless of prior experience of eye conditions. Specific implications for future work and the potential means of improving study design have been identified. This early pilot work represents an important first step towards the evaluation of genetic tests using CBA.

Contact: martin.eden@manchester.ac.uk

Notes

Contributed paper

Individual participant data meta-analysis to examine the accuracy of serum mesothelin for diagnosing malignant pleural mesothelioma

Johannes Reitsma, Kevin Hollevoet

Background & Objectives: Individual participant data (IPD) meta-analyses have several potential advantages when examining the diagnostic accuracy of a continuous marker. These include the more direct estimation of the summary ROC curve and greater flexibility, validity and power to investigate differences in accuracy between clinical subgroups. No preferred statistical approach exists, and we will illustrate our approach based on ROC regression modelling using standardized marker values.

Methods: Our IPD meta-analysis approach was inspired by the approach described by Janes & Pepe to perform ROC regression within a single study. It consisted of the following steps: (1) determine the cumulative marker distribution among controls stratified by study; (2) use this distribution to standardize marker values for corresponding cases, these are known as placement values; (3) the cumulative distribution of 1 minus the placement of values in all cases will produce the summary ROC curve, this can be estimated using binary regression techniques; (4) the mean value of the placement values of cases is equal to the AUC; (5) covariates can be added to the binary regression model to examine whether covariates have an impact on discrimination. Bootstrap techniques were used to calculate 95% confidence intervals.

Results: The IPD data of 16 studies was available for analysis, including a total of 1,026 patients with mesothelioma (cases) and 4,491 without (controls). Most studies applied a case-control design, often with multiple control groups per study. At a common threshold of 2 nmol/L, the sensitivities and specificities of mesothelin in the different studies ranged widely from 19% to 68% and 88% to 100%, respectively. This heterogeneity can be explained by differences in study population, because type of control group, mesothelioma stage, and histologic subtype significantly affected the diagnostic accuracy. For mesothelioma patients with stage 1 and 2 compared to symptomatic controls, the resulting area under the ROC curve was 0.77 (95% CI, 0.73 to 0.81). At 95% specificity, the corresponding sensitivity of mesothelin was 32% (95% CI, 26% to 40%).

Conclusion: IPD meta-analysis enables more insightful examination of the accuracy of a continuous marker, in particular to investigate differences in accuracy between patient subgroups. Several statistical issues deserve further attention.

References:

1. Hollevoet K, Reitsma JB, Creaney J, et al. Serum mesothelin for diagnosing malignant pleural mesothelioma: an individual patient data meta-analysis. J Clin Oncol. 2012;30:1541-9.
2. Janes H, Pepe MS: Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: An old concept in a new setting. Am J Epidemiol 2008;168:89-97.

Contact: j.b.reitsma-2@umcutrecht.nl

Notes

Contributed paper

Meta-analysis of diagnostic test accuracy studies with multiple thresholds

Richard Riley, Apratim Guha, Atanu Biswas, Yemisi Takwoingi, R. Katie Morris, Jonathan Deeks

Objectives: When meta-analysing diagnostic test accuracy studies, each study may provide results for one or more thresholds; however, the thresholds reported by each study often differ. In this situation researchers typically meta-analyse each threshold independently. Here, we rather consider jointly synthesising the multiple thresholds to gain more information.

Methods: A number of methods are currently available, but they often fail to converge or cannot easily handle missing data. Therefore we take a more practical approach. We use a linear imputation method that, in each study, imputes two by two tables for any missing thresholds that are bounded between two reported thresholds; this enables additional studies to be included in each threshold's meta-analysis, and allows the exact binomial distribution to be modelled, thereby avoiding continuity corrections. If desired, one can meta-regress the obtained meta-analysis estimates against threshold value, to produce a summary ROC that constrains sensitivity/specificity estimates to decrease/increase as threshold value increases.

Results: Application is made to 13 studies evaluating spot protein:creatinine ratio for detecting significant proteinuria in pregnancy, reporting estimates at 23 different thresholds, with a minimum of one and maximum of seven per study. Compared to the current standard approach of analysing each threshold without imputation, our imputation approach provided more information (with 50 additional results available), and generally revealed lower test accuracy results at each threshold and increased precision. An empirical evaluation also shows that the imputation method is superior.

Conclusion: The imputation approach is a practical method for dealing within missing threshold results in meta-analysis. The application shows that meta-analysis results based on the current standard approach may be over-optimistic, a likely consequence of selective reporting of threshold results in primary studies. The imputation method provides more information toward the meta-analysis in order to reduce this bias.

Contact: r.d.riley@bham.ac.uk

Notes

Contributed paper

Remits, roles and working models for Trial Steering Committees and Data Monitoring Committees in studies evaluating diagnostic tests: a survey of current practice

Lavinia Ferrante di Ruffano, Jac Dinnes, Jane Daniels, Richard Riley, Nick Hicks, Rik Kaplan, Jonathan Deeks

Background: All clinical trials that involve the prospective study of patients need some form of independent monitoring. In the UK, public funders of research – including the NIHR and MRC – require independent monitoring committees to be established for clinical trials; in RCTs of drugs or therapeutic interventions an independent trial steering committee (TSC) oversees the management and financial aspects of the study, and a data monitoring committee (DMC) oversees study progress and accumulating study data. It has been suggested that this fully independent TSC/DMC model may not be necessary for all trials of tests1, however there is no guidance currently available that provides advice regarding how and when independent monitoring should be employed in trials of tests, particularly studies evaluating diagnostic accuracy. To address this important deficit in guidance, a project funded by the MRC Midlands Hub for Trials Methodology Research is being undertaken to investigate how DMCs and TSCs should operate in trials evaluating medical tests.

Aims & Objectives: This presentation reports the results of a survey aiming to establish the breadth of current practice in the application of oversight committees for the management and data monitoring of diagnostic test evaluations. The survey is collecting responses to answer three objectives: 1) What TSC/DMC working models are there? 2) What are their roles and responsibilities? 3) What factors should be used to inform the choice of TSC/DMC model?

Methods: Ongoing and recently completed (after Jan 2010) studies evaluating any medical tests and unrestricted by type of study design were retrieved from systematic searches of several international registers, as well as an EMBASE search of all relevant studies published in 2012. Data are being collected by structured self–completed questionnaires and semi–structured telephone interview of trial statisticians.

Results: This research is ongoing, however we expect to provide a characterisation of how different TSC and DMC models operate, why particular structures were selected, how models differed according to study design, what their roles and responsibilities were, and which aspects of the chosen working models were particularly useful/disadvantageous.

Conclusions: Analysis of survey and interview responses will be used to develop insights into how particular models of TSC/DMC should be selected, which will be used to assist in the future development of guidelines for this purpose.

Refernce:

1. Daniels J, Gray J, Pattison H, et al. Rapid testing for group B streptococcus during labour: a test accuracy study with evaluation of acceptability and cost-effectiveness. Health Technol Assess 2009;13(42).

Contact: l.ferrantediruffano@bham.ac.uk

Notes

Contributed paper

Do we need to improve the study design framework to evaluations of test accuracy?

Christopher Hyde, Zhivko Zhelev, Harriet Hunt

Introduction: Our group has investigated the reporting of the study design of test accuracy studies and found there to be marked inconsistency (see related submission). Improving this is likely to be important for efficient reporting of test accuracy studies. Much space in the title and abstract of reports currently appears to be devoted to long hand description of the basic study design. However, this observation immediately invites consideration on whether we already have a clear, widely understood and accepted typology of study designs which authors can refer to in preparing their articles. This article begins to consider whether such a typology exists and to further consider whether any existing typologies need to extended.

Method:
a) We intend to conduct a systematic review of articles which might contain a typology of test accuracy studies. We will use bibliographic databases and expert groups to identify potentially relevant studies. Key text-books will also be examined.
b) We have reviewed 10 test accuracy reviews on a wide variety of subjects in which we have been involved to identify whether there are aspects of study design which may be candidates for inclusion in an enhanced typology in the future.

Results: After scoping, we found that there are surprisingly few attempts to set out, describe and name the different study designs which might be used to assess test accuracy. Rutjes et al 2005 appears to be the best example of the approach in providing a framework which covers variation in study design according to the means by which the patient sample is obtained differentiating particularly one population source ("single gate" designs) and separate diseased and non-diseased population sources ("two gate" designs).

Aspects of study design which we have encountered which may be candidates for inclusion in an enhanced typology are:
a) Study designs which attempt to make direct comparisons of test accuracy
b) Study designs which specifically address equivalence or superiority of test accuracy
c) Variation in reference standard, distinguishing pragmatic reference standards which are more akin to current best practice from those where the reference standard is closer to a true gold standard involving testing which is unlikely to be achievable in routine practice
d) Use of delayed verification and introduction of a longitudinal component to the traditional cross-sectional test accuracy evaluation

Conclusion: Provisionally we believe that there is a need invest further effort in reinvigorating and extending existing study design typologies for test accuracy to improve the reporting of accuracy evaluations. Some suggestions on how to proceed will be offered.

Contact: christopher.hyde@pcmd.ac.uk

Notes

Contributed Paper

Diagnostic accuracy studies: How to report and analyse inconclusive test results

Bethany Shinkins, Matthew Thompson, Susan Mallett, Rafael Perera

Background: Although the majority of test results provide useful information for diagnostic decision-making, there is often a subset of the results that are relatively uninformative and lead to an 'inconclusive' diagnostic outcome. Failure to report inconclusive test results can lead to misleading conclusions regarding the accuracy and clinical usefulness of a diagnostic tool. The STARD statement recommends that authors "report how indeterminate results, missing responses and outliers of the index tests are handled" (item 22). We assessed the extent to which diagnostic accuracy studies conform to this particular reporting recommendation and produced a framework of guidance for the reporting and analysis of inconclusive diagnostic test results based on a review of the literature.

Methods: Twenty-two systematic reviews (published between 2005 and 2011) assessing the reporting quality of diagnostic accuracy studies based on the STARD statement were identified. The proportion of studies that adhered to item 22 was extracted from each review.

Results: Based on 1156 primary studies included in 22 systematic reviews, only one third (35%) of studies reported the presence or absence of inconclusive results adequately.

Conclusion: Inconclusive diagnostic test results are not consistently reported in diagnostic accuracy research. We provide an extensive framework for reporting and analysing these results, emphasising the importance of reporting all test results broken down by the reference standard and highlighting the potential risk of bias in accuracy statistics when analysing valid inconclusive results.

This project has recently been accepted for publication in the 'Reporting and Methods' section of the BMJ.

Contact: bethany.shinkins@phc.ox.ac.uk

Notes

Keynote speaker

Prognosis research: new opportunities in linked electronic health records

Harry Hemingway, University College London, UK.

The Prognosis Research Strategy (PROGRESS) article series in BMJ and PLosMed 2103 makes recommendations for developing four cognate areas: outcomes research, single prognostic factors (biomarkers), prognostic models and stratified medicine. The challenges which have beset the field of prognosis research are familiar and include small sample sizes, significance chasing biases in analysis, insufficient external validation and lack of relation to clinical decision making. Arguably the central challenge has been the failure to recognise, and appropriately fund, prognosis research as a fundamental field of enquiry which informs research across translational pathways from discovery, through trials to clinical decision making and public health.

New opportunities to establish programmes of prognosis research arise using linked electronic health records. Linking records across multiple sources – for example primary care, secondary care, disease registry and death registration – offer opportunities of phenotypic resolution of startpoints and endpionts, statistical scale and clinical relevance. The CALIBER programme is curating cohorts of about 2 million adults followed for the first occurrence and subsequent progression of a wide range of different cardiovascular diseases affecting the cerebral, coronary and peripheral circulations and myocardium. This talk will discuss some of the promise, and pitfalls, of using health records to provide new ways of doing prognosis research at different stages in the translational cycle. Professor Hemingway co-directs the UK Farr Institute of Health Informatics Research, launched in 2013.

Contact: h.hemingway@ucl.ac.uk

Notes

Contributed paper

The interplay of sensitivity and specificity when comparing risk models and classification rules

Ben van Calster, Ewout W. Steyerberg, Ralph B. Sr. D'Agostino, Michael J. Pencina

Background: When comparing prediction models, it is essential to estimate the magnitude of change in performance rather than rely solely on statistical significance. The Net Reclassification Improvement (NRI) is widely used to assess improved classification by adding markers to risk prediction models or by comparing non-nested models. NRI does not consider misclassification costs in contrast with decision-analytic alternatives.

Aims: We aimed to investigate measures that estimate change in classification performance when adding a new biomarker to a risk prediction model, assuming two-group classification based on a single risk threshold.

Methods: We use simulated data to investigate the change in sensitivity and specificity (ΔSe and ΔSp). Second, we study the influence of ΔSe and ΔSp on the NRI (i.e. sum of ΔSe and ΔSp) and decision-analytic measures (Net Benefit or Relative Utility).

Results: We observed that even when a strong marker is added and/or the extended model has a dominating receiver operating characteristic curve, it is possible that either ΔSe (for thresholds below the event rate) or ΔSp (for thresholds above the event rate) is negative. In these cases decision-analytic measures provide more modest support for improved classification than NRI, but in general all measures confirm that adding the marker improved classification accuracy. Exceptions may occur when important interaction effects are omitted, models are miscalibrated or continuous predictors are not modeled appropriately.

Conclusions: Our results underscore the necessity of reporting ΔSe and ΔSp (components of the two-group NRI) separately. When a single summary is desired, decision analytic measures allow for a simple incorporation of the misclassification costs.

Contact: ben.vancalster@med.kuleuven.be

Notes

Contributed paper

Net Reclassification Risk: a graph to clarify the potential prognostic utility of new markers

Ewout Steyerberg, Moniek Vedder, Douwe Postmus, Michael Pencina, Ben van Calster

Background: New prognostic markers and tests may improve risk prediction. We aimed to provide graphical support for recently developed reclassification measures to indicate the prognostic utility of new markers, specifically the Net Benefit (NB).

Methods: We consider 3264 subjects from the Framingham study as used before to predict 10-year risk of coronary heart disease (n=183 events, 5.6%, Pencina et al, Stat Med 2008). We evaluate the incremental value of adding HDL to the prediction model, assuming 20% as a threshold to define high risk (n=138 with HDL, n=122 without).

Results: The AUC difference (ΔAUC) between a model with and without HDL was small in numerical value (0.012 for adding HDL with continuous risk; 0.029 with 20% threshold). Novel measures focus on reclassified patients, such as the Net Reclassification Improvement (NRI). The NRI is the sum of the net percentage of correctly reclassified patients with events (6.0% with a 20% threshold) and without events (-0.2% with a 20% threshold). The sum is 0.058, which is higher in numerical value than the difference in ΔAUC (0.029).
For further assessment of prognostic utility, we propose a simple graph that focuses on the number of patients and event rates of the 2 reclassified groups: those reclassified from high to low risk (H/L, n=29, 10% event rate) and those reclassified from low to high risk (L/H, n=45, 31% event rate). This graph shows that the overall event rate may be reduced by at most 45*0.31 – 29*0.10 = 11 events, or 0.34% (11/3264). The burden of overtreatment of those without events is explicitly considered in decision-analytic summary measures such as the Net Benefit (NB). The graph can also be used to understand that we expect 45*(1-0.31) – 29*(1-0.10) = 5 extra overtreatments. These should be weighted by the odds of the decision threshold (0.2/(1-0.2)=0.25. This leads to a penalty for overtreatment of 5*0.25=1.25. The NB hence is (11-1.25)/3264=0.30%, equivalent to potentially preventing 3 events per 1000 without extra overtreatment.

Conclusions: Important insights in the prognostic utility of new markers may be obtained by a simple graph for the Net Reclassification Risk ('NRR graph').

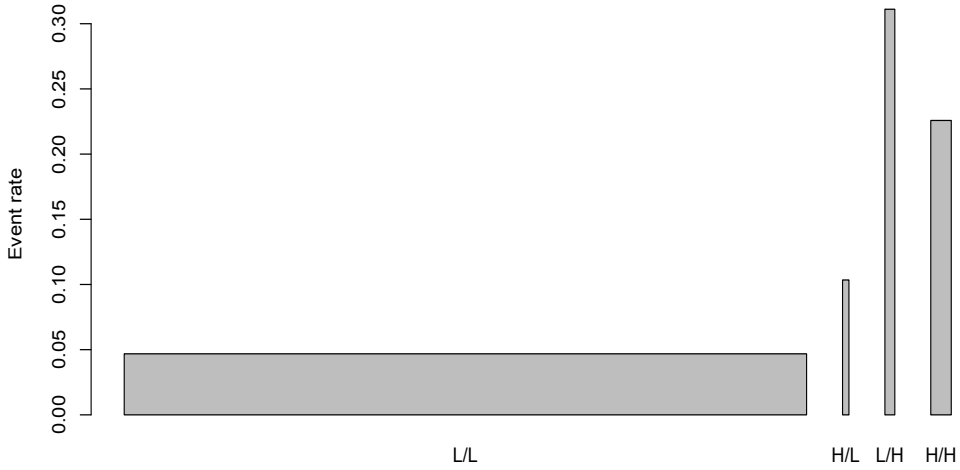


Figure: Net Reclassification Risk ('NRR') graph. L: 'Low risk'; H: 'High risk', defined as >20% event rate (10-year risk of coronary events). Reclassified groups are H/L (High to Low), and L/H (Low to High). Bar width proportional to absolute N (3264 in total).

Contact: e.steyerberg@erasmusmc.nl

Notes

Contributed paper

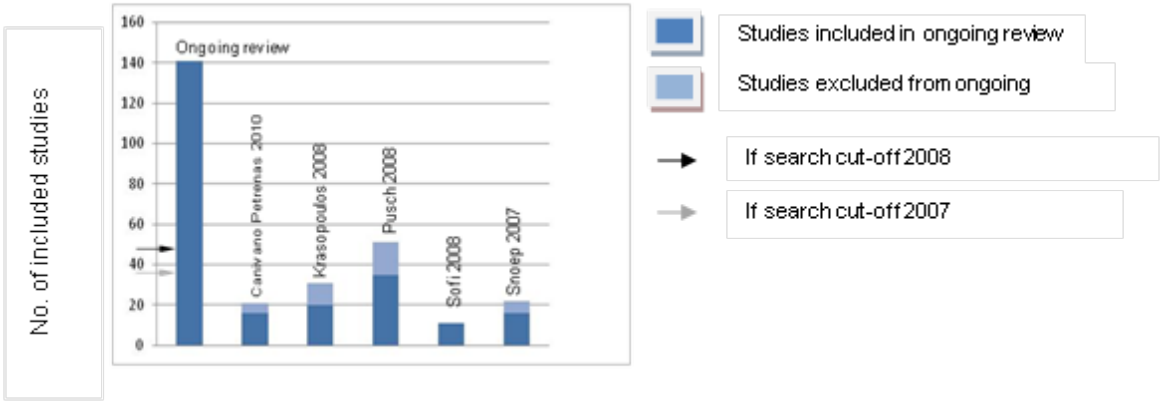
An overview of systematic reviews of the prognostic utility of platelet function tests for predicting vascular event rates in patients on aspirin therapy: implications of discrepancies in study identification and selection

Janine Dretzke, Sue Bayliss, Jennifer O'Donnell, Marie Lordkipanidzé, David Fitzmaurice, David Moore, Smriti Raichand

Introduction: Aspirin is used as an antiplatelet agent to prevent occlusive arterial events in patients with atherothrombotic disease. There is controversy over the existence of the phenomenon of "aspirin resistance", i.e. the failure of aspirin therapy to inhibit platelet function in some patients even when treatment compliance has been verified. Several systematic reviews have attempted to determine the prognostic utility of platelet function tests for identifying patients with "resistance" to aspirin therapy and who thus possess a potentially higher risk of vascular events. An ongoing review is being undertaken by the authors of this abstract (NIHR HTA: 10/36/02, PROSPERO: CRD42012002151).

Aim/Methods: To investigate discrepancies between systematic reviews in terms of (i) search strategies, (ii) study selection process, (iii) inclusion criteria and (iv) included studies, and the implications for the robustness and generalisability of conclusions. Reviews on "aspirin resistance" were systematically sought and included if at least one database was searched and selection criteria were presented.

Results: Five reviews met the inclusion criteria. The numbers of included studies, and overlap with the ongoing review, are shown in the graph below. Greater numbers of included studies were found by the ongoing review even when the search dates of the other reviews are considered. The ongoing review excludes a proportion of studies included in other reviews mainly due to the requirement that sampling for platelet function preceded clinical outcomes.



Discussion: Different time points for searching and study selection criteria were insufficient to explain all discrepancies in terms of included studies. Reasons for missed studies may include: narrower search strategies; reliance on title/abstract only for screening, screening by only one reviewer; or unspecified selection criteria. Optimisation of search strategies in the area of prognostic factors requires further research, but is likely to necessitate screening of larger number of citations than for effectiveness questions. The study selection process is more likely to require a step-wise approach, the screening of full-text articles, independent, duplicate study selection, and very rigorously defined selection criteria. Individual review conclusions must be interpreted cautiously as they are based on different sub-sets of studies.

Contact: j.dretzke@bham.ac.uk

Notes

Keynote Speaker

The Role of Modelling in Test Evaluation: Experience in NICE Diagnostic Appraisals

Christopher Hyde, PenTAG, University of Exeter Medical School, UK

While laboratory scientists, test specialists and clinicians may arguably be able to make adequate decisions based on test accuracy data, it has always been clear that decisions at population level require information on effectiveness (impact on patient outcomes) and cost-effectiveness.

The challenge for policy making thus lies in where the evidence on these features will come from. Ideally it would be produced directly from clinical studies such as test-treat RCTs. However, the practicality of generating RCT evidence on a large scale alone suggests that other approaches are needed.

One alternative is economic modelling, which not only facilitates conclusions on cost-effectiveness but more importantly provides a framework in which evidence on the accuracy of a test and the effectiveness of down-stream treatment, generated in separate studies can be integrated to give an estimate of the clinical effectiveness of the whole test-treatment package. This is much easier said than done, although there have been considerable advances in modelling techniques over the past decade.

NICE has recently become active in making decisions about diagnostics in England and Wales, and has placed great emphasis on the use of economic modelling to inform its guidance. This presentation will reflect on this experience from the perspective of a member of its Diagnostic Assessment Committee and a producer of health technology assessments on diagnostics. It will particularly consider requirements for further methodological research on economic modelling on diagnostics.

Contact: christopher.hyde@pcmd.ac.uk

Notes

Keynote Speaker

Bringing risk prediction models into the 21st century

Patrick Bossuyt, Parvin Tajik
Department of Clinical Epidemiology, Biostatistics and Informatics, Academic Medical Center, University of Amsterdam, the Netherlands

It is now well accepted that decisions about medical interventions should preferably be based on the best available evidence from sound clinical studies. Such studies should document the effects of these interventions on outcomes that matter to patients and society, not just on proxy measures, or in highly selected patients.

Increasingly, a similar logic is applied to decisions about medical tests. Decisions about the introduction, reimbursement and use of medical tests can be guided by their clinical effectiveness: their effects on patient-relevant outcomes. In that sense, testing strategies should be regarded as (complex) interventions, and evaluated as such. Measures of analytical performance of IVD and other medical tests, and measures about the clinical performance of such tests are informative and necessary, but rarely sufficient to demonstrate clinical effectiveness.

Somewhat surprisingly, risk prediction models have not arrived at this stage yet. Despite the growing interest in measures for expressing their performance, the statistics available to evaluate them can be classified as measures of analytical performance and, at best, clinical performance. We will argue that risk prediction models should be treated like other medical tests, should be evaluated as such, and that recommendations to use them should be guided by considerations of their effectiveness: their ability to improve health outcomes and to increase health care efficiency.

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

Notes

Contributed paper

A Systematic Review of Grading Systems for Medical Tests for Guideline Developers

Gowri Gopalakrishna, Miranda Langendam, Rob Scholten, Patrick Bossuyt, Mariska Leeflang

Background: A variety of systems have been developed to grade evidence and develop recommendations based on the available evidence. However, development of guidelines for medical tests is especially challenging given the typical indirectness of the evidence; direct evidence of the effects of testing on patient important outcomes is usually absent. We compared grading systems for medical tests on how they use evidence in guideline development.

Methods: We used a systematic strategy to look for grading systems specific to medical tests in PubMed , professional guideline websites and handsearching back references of key articles. Using the AGREE instrument as a starting point, we defined two sets of characteristics to describe these systems: methodological and process ones. Methodological characteristics are features relating to how evidence is gathered, appraised and used in recommendations. Process characteristics are those relating to the guideline development process. Data was extracted in duplicate and differences resolved through discussion.

Results: Twelve grading systems could be included. All varied in the degree to which methodological and process characteristics were addressed. Having a clinical scenario, identifying the care pathway and/or developing an analytical framework, having explicit criteria for appraising and linking indirect evidence, and having explicit methodologies for translating evidence into recommendations were least frequently addressed. Five systems at most addressed these, to varying degrees of explicitness and completeness. Process wise, features most frequently addressed included involvement of relevant professional groups (8/12), external peer review of completed guidelines (9/12), and recommendations on methods for dissemination (8/12).

Characteristics least often addressed were whether the system was piloted (3/12) and funder information (3/12).

Conclusions: Five systems for grading evidence about medical tests in guideline development addressed to differing degrees of explicitness the need for and appraisal of different bodies of evidence, the linking of such evidence, and its translation into recommendations. At present, no one system addressed the full complexity of gathering, assessing and linking different bodies of evidence.

Contact: g.gopalakrishna@amc.uva.nl

Notes

Keynote Speaker – Interactive Session

The identification and prioritisation of diagnostic research: the experiences of three European Agencies

Nick Hicks, Ann van den Bruel, Mattias Menig
NETSCC, Department of Primary Care Health Sciences, University of Southampton, UK;
University of Oxford, UK;
Institute for Quality and Efficiency in Health Care (IQWiG), Germany

This session presents and contrasts the different ways by which 3 European Health Technology Assessment organisations manage clinical diagnostic research issues. The organisations discussed are the National Institute for Health Research (NIHR) HTA programme in the UK, the Institute for Quality and Efficiency in Health Care (IQWiG) in Germany and the Belgian Health Care Knowledge Centre (KCE).

Three co-presenters with experience of working in these organisations explain the roles and remits of the organisations and how the different health care systems and sources of funding affect how the diagnostic research issues are identified, prioritised and taken forward.

Examples of recent diagnostic projects will be presented to illustrate the types of diagnostic research issues considered by the different organisations and where relevant, how the research findings have subsequently been disseminated for use. We will also present examples where collaboration at a European level has been set up, and where opportunities for collaboration have been missed.

The session will conclude with an open discussion of a number of clinical diagnostic research issues identified by delegates in a survey in advance of the Symposium. Delegates will have the opportunity to exchange views on the importance of these issues and to highlight key aspects for the research questions that need to be resolved if future research on the issues is to be taken forward.

One possible outcome of these discussions is for the diagnostic research issues discussed to be considered by the NIHR HTA programme Diagnostic Technologies Panel with a view to making a call for further diagnostic research in the area.

Contact: hicks@soton.ac.uk

Notes

Keynote speaker

Ajit Lalvani, Imperial College, UK

Scientific, clinical, regulatory and policy aspects of diagnostic test development and deployment will be presented, with specific reference to interferon-gamma release assays (IGRA) for diagnosis of latent TB infection.

The scientific platform underpinning IGRA, T cell-based diagnosis, was a new paradigm in diagnostic medicine presenting raising new challenges for standardisation and assessment of inter-individual, intra-individual and inter-laboratory reproducibility.

The absence of a diagnostic gold standard for latent TB infection was the next major challenge. Several approaches were taken to overcome this practical and conceptual roadblock and will be discussed.

The diagnostic test that IGRA aim to supersede is the century-old tuberculin skin test (TST). Although it has well-recognised limitations, clinical use of TST rests on a huge clinical evidence base generated over several decades. Generating a comparably large and robust database was the next issue that will be discussed.

Positive and negative predictive values of diagnostic tests for latent TB infection vary with the prevalence of TB infection and other factors which in turn vary substantially by geographic area and population. This results in wide heterogeneity in the clinical utility of IGRA that is difficult to capture in systematic reviews or meta-analyses, which limits generalisability of recommendations and guidelines. Within this context, development of national and international guidelines will be reviewed with an emphasis on their limitations and key clinical uncertainties resulting from gaps and controversies in the evidence base. The evolution of guidelines in response to generation of new evidence will also be exemplified.

Application of new more accurate diagnostic tests may be expected to provide new insights into our scientific and clinical understanding of the target disease. The impact of IGRA on our understanding of the spectrum and natural history of TB infection will be described.

Contact: a.lalvani@imperial.ac.uk

Notes

Keynote Speaker

Matthew Thompson, University of Oxford

The pipeline of diagnostic tests from initial discovery and development by industry and bench scientists, to successful implementation in clinical practice seems to be fraught with roadblocks. Currently few medical tests are widely used in primary care settings for example, yet the market worldwide for such tests is enormous – there are almost 1 billion ambulatory care visits per year in the USA alone, many of which involve diagnostic tests. In this presentation I will explore approaches that our NIHR-funded Oxford Diagnostic Evidence Cooperative is using to help solve some of the roadblocks in this 'bench to bedside' pathway, with an emphasis on medical tests in primary care.

Perhaps the most important issue is improving the timing and scope of engagement between clinicians and diagnostics industry. For many in industry it can be difficult to engage meaningfully with clinicians, particularly outside certain clinical specialties. Understanding test needs from end users (e.g. primary care clinicians) seems fundamental to even the early stages of biomarker development. There are many different approaches to 'needs assessment', and I will present initial results from an international survey of primary care clinicians' test needs. A second (and perennial) problem is clarifying the study designs that are essential to generate evidence needed for implementation, particularly outside the settings and populations where tests undergo initial development. This seems to be why many tests never make it along the path to implementation, and falter at the stages of clinical utility and impacts on clinical decision making. I will provide some examples of how we have attempted to approach evidence assessment for medical tests, and the limits of the 'one size fits all' diagnostic test framework.

Contact: matthew.thompson@phc.ox.ac.uk

Notes

Contributed paper

Methods for diagnostic meta-analysis in the absence of a gold-standard reference

Nandini Dendukuri

Absence of a perfect reference test is an acknowledged source of bias in studies of diagnostic test accuracy for many diseases. For example, in the case of community acquired pneumonia, standard reference tests such as blood or sputum culture have poor sensitivity. Yet meta-analyses of new tests for such diseases have typically assumed the reference standard is perfect in individual diagnostic studies, leading to biased estimates of the new test's accuracy. This presentation illustrates recently proposed Bayesian hierarchical models for joint meta-analysis of sensitivity and specificity of the diagnostic test under evaluation, while adjusting for the imperfect nature of the reference standard via a latent class structure. We show how to obtain pooled estimates of sensitivity and specificity, and how to plot a hierarchical summary receiver operating characteristic curve from such models. We describe extensions of the model to situations where multiple reference tests are used, and where index and reference tests are conditionally dependent. The methods are illustrated using data from a meta-analysis of a urinary antigen test for community acquired pneumonia. The estimates of the pooled sensitivity and specificity of the new test were higher compared to estimates from a model that assumed that the reference test was perfect.

Contact: nandini.dendukuri@mcgill.ca

Notes

Contributed paper

Evaluating Diagnostic Accuracy in the Face of Multiple Reference Standards

Christiana Naaktgeboren, Joris de Groot, Loes Bertens, Maarten van Smeden, Johannes Reitsma, Karel Moons

A common challenge in diagnostic studies is the lack of a single error-free ("gold") reference test to which the test under study (the index test) can be compared. When a reference test does not perfectly correspond to disease status, index test accuracy estimates can be biased. One method for dealing with the lack of a single perfect reference test is to combine the results of several tests into one "composite reference standard".

The idea behind a composite reference standard is that the combination of several imperfect tests may provide a better perspective on disease than any individual tests. The key challenge of composite reference standards, however, is selecting the appropriate tests and determining the optimal rule for combining the test results.

There is a lack of consensus in the way the term composite reference standard is used and the reporting of results is generally poor. To address these problems, we provide a thorough explanation of the composite reference standard method, discuss advantages and disadvantages of the method, and make suggestions on how to report results.

Contact: c.naaktgeboren@umcutrecht.nl

Notes

Contributed Paper

Shortcomings in Reporting and Methodology of Latent class models (LCMs) in Diagnostic Research – A Systematic Review

Maarten van Smeden, Joris de Groot, Christiana Naaktgeboren, Karel Moons, Johannes Reitsma

Background & Objectives: Latent class models (LCMs) combine the results of multiple diagnostics tests through a statistical model to obtain estimates of diagnostic accuracy in situations where there is no single, accepted reference test. To explore the methodology and reporting of LCMs in diagnostic research, we performed a systematic review of such studies. Our review generates insight in the quality of reporting of studies that use LCMs and reveal variation in methodology between studies.

Methods: We identified 64 diagnostic studies that reported test accuracy or disease prevalence estimates obtained using LCM parameter estimates. From these studies, information was extracted independently by two reviewers.

Results: The systematic review shows that the use of LCMs in diagnostic studies is increasing, notably in studies which evaluate the accuracy of diagnostic tests to detect the presence of an infectious disease (52%). The majority of studies (64%) reported analyses solely based on 'basic' 2-class LCMs, for which it is assumed that the tests are independent conditional on a binary target disease status.

Discussion: Our review revealed several shortcomings in methodology and reporting of studies which use LCMs. The majority of the inferences made were based on 'basic' LCMs for which the assumptions are easily violated. Evaluations of the tenability of these assumptions, although thoroughly described in statistical literature, a rarely reported.

Contact: m.vansmeden@umcutrecht.nl

Notes

Contributed paper

Reporting and methods in systematic reviews of comparative accuracy

Yemisi Takwoingi, Richard Riley, Jon Deeks

Background: Systematic reviews in which the accuracy of two or more tests are compared can provide evidence to support the clinical validity of each test and aid the process of test selection. Because test evaluation is often limited to the assessment of test accuracy, it is vital that in the rapidly expanding evidence base, reviews and meta-analyses that compare the accuracy of multiple tests are conducted and reported appropriately to avoid misleading conclusions and recommendations.

Objectives: To provide a descriptive survey of current practice with a view to identifying good practice and problems, and to make suggestions for the improvement of future reviews.

Methods: Systematic reviews of test accuracy in the Database of Abstracts of Reviews of Effects published between 1994 and October 2012 were identified. We placed no restrictions on language of publication, test type, purpose of the test (screening, staging, diagnostic, etc), setting, or disease area. We extracted information on the target condition, patient population, tests evaluated, purpose of the tests, analysis methods and reporting characteristics of each review. Descriptive statistics were computed. We also compared reporting characteristics with the most relevant reporting guideline—the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist.

Results: We included 248 reviews that evaluated the accuracy of 2 or more tests. The reviews contained 6915 studies (studies may appear in more than one meta-analysis). Initial results indicate that tests are not often formally compared in the same meta-analysis but instead a separate meta-analysis is performed for each test and comparisons are made informally by comparing summary estimates between meta-analyses. Data analysis is still ongoing and results will be available for presentation at the symposium.

Conclusions: Initial findings highlight the need for better understanding of methods and strategies for comparing tests in meta-analysis and specific guidance for reporting reviews of comparative accuracy.

Contact: y.takwoingi@bham.ac.uk

Notes

Contributed paper

Direct versus Indirect Comparisons in Systematic Reviews of Test Accuracy Studies: An IPD Case Study in Ovarian Reserve Testing

Junfeng Wang, Patrick Bossuyt, Ronald Geskus, Mariska Leeflang

Background: Comparative systematic reviews of diagnostic test accuracy compare relative accuracy of two or more tests. Direct comparisons that evaluate all tests in the same study, even in the same patients, are most valid and regarded as the reference approach. Indirect comparisons are more prone to bias than direct comparisons, but excluding them may lead to a loss in precision in the summary estimates.

Objectives: To compare results from indirect comparisons with results of direct comparisons in meta-analysis; to develop appropriate methods of adjusting indirect comparisons to improve their comparability.

Methods: A dataset from Individual Patient Data (IPD) meta-analysis on the test accuracies of Anti-Müllerian Hormone (AMH), Antral Follicle Count (AFC) and Follicle Stimulation Hormone (FSH) in relation to ovarian response was used in this case study. Test accuracies were measured by sensitivity and specificity for dichotomous tests and the area under the ROC curves (AUCs) for continuous tests and compared in each pair of tests under direct and indirect comparisons. Inconsistency was defined as statistical significant difference in comparative results between the direct and indirect evidence.

Results: 32 studies were included with IPD from 4762 women undergoing IVF. By comparing sensitivities and specificities, the differences in sensitivities between AMH and FSH (-0.1563, p=0.04), AMH and AFC (0.1465, p<0.01) and in specificities between AMH and AFC (-0.0607, p=0.02) are significant; by comparing AUCs, the difference between AFC and FSH (0.0948, p<0.01) is significant in direct comparison but not significant (0.0678, p=0.09) in indirect comparison; while the difference between AFC and AMH is significant (-0.0830, p<0.01) in indirect comparison but not significant (-0.0176, p=0.29) in direct comparison. These differences still existed after adjusting for indirectness by considering covariate effect.

Conclusions: Comparative results of test accuracy obtained through indirect comparisons are not always consistent with those obtained through direct comparisons. There is no straightforward way to make indirect comparisons more comparable. Evidence from indirect comparisons should be assessed carefully and combined with direct comparisons after adequate assessment of the consistency and with adjustment.

Contact: j.wang@amc.uva.nl

Notes

Contributed paper

How to model longitudinal data for the prediction of hemoglobin level in whole blood donors

Yvonne Vergouwe, A. Mireille Baart, Femke Atsma, Karel Moons, Wim L.A.M de Kort

Background and Aim: The biomarker hemoglobin (Hb) measured in blood is used to assess the iron status of blood donors, prior to donation. Donors with low Hb levels are subsequently deferred. Measurements of Hb levels at earlier visits are strong predictors for future Hb level. We investigated how the longitudinal data could be used for the prediction of future Hb level.

Patients and Methods: Data of 87,450 Dutch male blood donors with longitudinal Hb measurements were used to develop and validate three prediction models. Two simple linear models were fitted with either Hb at the last visit plus change in Hb level, or the mean of all previous Hb measurements. The third model was a mixed effect model that included all individual previous Hb measurements and accounted for the dependence of Hb assessments within a donor using a random intercept.

Results: The simple model that included the mean Hb (Model B) showed slightly better overall performance than the model with Hb at the last visit plus change in Hb level (Model A). R2 was 0.44 (B) and 0.41 (A). The mixed effect model including all longitudinal measurements did not outperform the simple models. At best, the performance was similar for the mixed and simple models (Figure), when donor specific information that was incorporated in the random intercept, could be used for model predictions (Model C2).

Conclusion: Longitudinal Hb measurements can be well averaged for the prediction of future Hb level. Longitudinal analysis of Hb measures with a mixed effect model did not improve model performance, probably because of the large measurement error in Hb assessment.

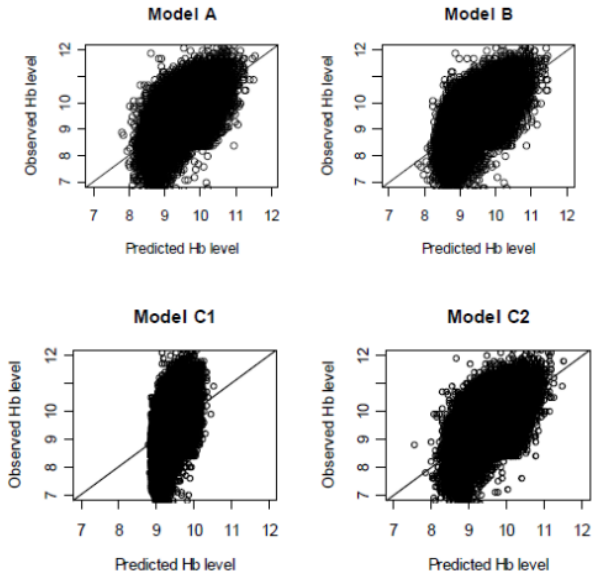


Figure: Validation plots for the simple (A and B) and mixed (C) models applied at a new visit. C1: random intercept not used in prediction of future Hb level; C2: random intercept is used.

Contact: y.vergouwe@erasmusmc.nl

Notes

Contributed paper

Comparing strategies for interpreting longitudinal CEA measures when screening for colorectal cancer recurrence

Bethany Shinkins, Indika Pathiraja, Brian Nicholson, Richard Stevens, Rafael Perera, David Mant

Background: A recent systematic review assessing the accuracy of single CEA measurements concluded that the biomarker was insufficiently sensitive for detecting colorectal cancer recurrence in asymptomatic patients who have undergone surgery with curative intent [1]. It was suggested that trends in longitudinal CEA measurements may improve accuracy, however little research has been carried out to explore this possibility. We are investigating whether longitudinal CEA measurements are useful in determining which patients should undergo further investigation to detect treatable recurrence.

Methods: Using data (n=601, >6000 CEA measurements in total) from two arms of the FACS trial*, we are comparing different models for interpreting trends in longitudinal CEA measurements to the traditional single threshold rule. Algorithms currently being explored include difference from baseline, ratio to baseline and rate of change in CEA. In addition to accuracy, the median leadtime will be compared across interpretation methods.

Results: Preliminary results show that the 'difference from baseline' and 'ratio to baseline' transformations do not appear to offer improvements on the single threshold rule. However, using the rate of change in CEA does improve accuracy. Work is currently being carried out to identify how and if CEA trend information can be usefully incorporated into a follow-up strategy for recurrent colorectal cancer.

The full results will be ready to report at the symposium.

* An RCT comparing the cost-effectiveness of intensive versus no scheduled follow-up in patients who have undergone resection for colorectal cancer with curative intent

Reference:

1. Tan E, Gouvas N, Nicholls RJ, Ziprin P, Xynos E, Tekkis PP. Diagnostic precision of carcinoembryonic antigen in the detection of recurrence of colorectal cancer. Surgical oncology 2009;18(1):15-24.

Contact: bethany.shinkins@phc.ox.ac.uk

Notes

Contributed paper

Evaluation of time-dependent diagnostic accuracy using repeated measurements of a biomarker

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

Often diagnostic accuracy study protocols based on a standard binary diagnostic test design, and require an enormous sample of patients when the disease being detected by the diagnostic test has a low prevalence rate. Recruitment of a large number may be difficult in practice due to limited timelines, availability of funding or the study population being more specific e.g. paediatric. However, in many studies, it is now common for information on repeated (or longitudinal) measurements of potential biomarkers to be available on each individual in a study alongside the clinical endpoint of interest. Therefore, incorporating repeated biomarker measurements rather than a single measurement per patient would reduce the required sample size. However, using approximate statistical methods to combine the information from repeated measurements and clinical endpoint are inappropriate due to time-dependent nature of outcomes and hence are inefficient in estimating the diagnostic accuracy summaries such as sensitivity and specificity. Despite several methods for the combined analysis of longitudinal and event-time data being developed in recent years, incorporating the longitudinal measurements for estimating the diagnostic accuracy has been received limited attention to date. In this study, we propose a two-stage approach. In the first stage, a classification score is estimated by capturing the dependency between the longitudinal biomarker trajectory and subsequent risk of disease onset explicitly, and in the second stage, well-established ROC methodology is deployed to estimate time-dependent diagnostic accuracy summaries of the biomarker. The methodology is illustrated by an application to a prospective observational study of biomarkers to diagnose the onset of sepsis.

References:

Heagerty PJ, Zheng Y. Survival Model Predictive Accuracy and ROC curves. Biometrics 2005; 61, 92 – 105.

Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. Biostatistics 2000; 1, 465-480.

Contact: kdrr@liv.ac.uk

Notes

Contributed paper

Has the randomised controlled trial design been successfully used to evaluate new monitoring strategies? An assessment of experience to date

Jac Dinnes, Alice Sitch, Jenny Hewison, Doug Altman, Jon Deeks

Monitoring is a central activity for patient and disease management, and, as for therapeutic interventions and for diagnostic tests, it is important to try to identify its impact on patient outcomes. Given the advantages of the randomised controlled trial (RCT) design for the evaluation of therapeutic interventions, it is tempting to assume that the same approach must be the gold standard for the evaluation of monitoring strategies. Such trials present considerable challenges however, not least due to the complexity of the intervention under evaluation and the multitude of ways in which testing can affect patients.

Objective: To gain an insight into how successfully the RCT design has been used to evaluate monitoring strategies, particularly monitoring of individuals with (or at risk of) a disease or condition that is likely to progress or recur at some point in the future.

Methods: A review of available RCTs of monitoring regimes was conducted. Trials were identified from CENTRAL, the NHS HTA Programme, Trials journal and reference list review. Eligible trials from the US NIH clinicaltrials.gov database were purposively sampled according to topic area, to complement those already identified. RCTs comparing at least one formal monitoring strategy to no formal monitoring, to an alternative monitoring strategy, or to immediate treatment were included. One author screened the search results for relevant studies. Data extraction was conducted by one author and checked by a second. A narrative synthesis was performed.

Results: Fifty-eight RCTs were included: 34 full trial reports, 10 interim analyses and 14 protocols. Twelve trials were stopped early; usually due to a lower than expected event rate in the control arm (n=5). The median sample size was 272 [IQR 119, 599] and median follow-up 24 months [IQR 12, 59]. Sample size calculations were reported in 78% and 24% were industry-sponsored. The aim of monitoring was usually to improve patient outcomes (71%), through earlier initiation of treatment or better selection of patients requiring treatment, usually through the addition of a new test or replacing a test within an existing monitoring strategy. 'New' or replacement tests were biochemical (34%), imaging (23%), physiological (19%) and implanted devices (11%). Monitoring strategies were not well described for 49% of tests used in control arms and 21% in experimental arms, and there was a lack of citation of evidence to support key features of the monitoring strategies. Few trials reported using test measurements over time to define an abnormal result, and there was limited reporting of repeated testing to confirm abnormal test results (<10%). Trial validity was similar to trials of other types of intervention; blinding was frequently not present or not attempted and in 16% of trials, the monitoring test was also used to assess outcomes. Statistically significant effects for the primary outcome were observed in 35% of full trial reports. Of those also reporting power calculations, the magnitude of the observed effect was less than predicted in 44% (11/25).

Conclusion: Trials were affected by lack of power, issues with study validity and an apparent lack of recognition of both the complexity of the monitoring strategies evaluated and of how the change in strategy might affect patient care pathways. Evidence supporting the strategies was infrequently cited, suggesting that RCT-based evaluation may be premature.

Contact: j.dinnes@bham.ac.uk

Notes

Keynote Speaker

Predicition, Classification and Clinical Decision Rules: Methodological Conduct and Reporting

Gary Collins, Centre for Statistics in Medicine, University of Oxford, UK

Prediction models, combining several patient characteristics or symptoms are increasingly being used to estimate the risk or probability that a specific outcome or disease is present (diagnostic setting) or a specific outcome will occur in the future (prognostic setting). I will provide an overview of the key steps needed in the introduction of a new prediction model, from development to validation through to evaluating the impact of a prediction model. I will highlight practical difficulties and commonly seen flaws in the process of developing and evaluating a new model. In particular, I will discuss available strategies when available data are limited. The talk will be guided by examples from the literature and systematic reviews evaluating the methodological conduct and reporting of studies developing or validating prediction models. Finally, I will present a recent initiative, called TRIPOD (Transparent Reporting of a model for Individual Prognosis Or Diagnosis), to improve the quality of reporting of prediction modelling studies.

Contact: gary.collins@csm.ox.ac.uk

Notes

Contributed Paper

A framework for developing and implementing clinical prediction models across multiple studies

Thomas Debray, Karel Moons, Ikhlaaq Ahmed, Hendrik Koffijberg, Richard Riley

The use of individual participant data (IPD) from multiple studies is an increasingly popular approach when developing a multivariable risk prediction model. Corresponding datasets, however, typically differ in important aspects, such as baseline risk. This has driven the adoption of meta-analytical approaches for appropriately dealing with heterogeneity between study populations. Although these approaches provide an averaged prediction model across all studies, little guidance exists about how to apply or validate this model to new individuals or study populations outside the derivation data.

We consider several approaches to develop a multivariable logistic regression model from an IPD meta-analysis (IPD-MA) with potential between-study heterogeneity. We also propose strategies for choosing a valid model intercept for when the model is to be validated or applied to new individuals or study populations. These strategies can be implemented by the IPD-MA developers or future model validators.

Finally, we show how model generalizability can be evaluated when external validation data are lacking using internal-external cross-validation, and extend our framework to count and time-to-event data. In an empirical evaluation, our results show how stratified estimation allows study-specific model intercepts which can then inform the intercept to be used when applying the model in practice, even to a population not represented by included studies.

In summary, our framework allows the development (through stratified estimation), implementation in new individuals (through focused intercept choice) and evaluation (through internal-external validation) of a single, integrated prediction model from an IPD-MA in order to achieve improved model performance and generalizability.

Contact: t.debray@umcutrecht.nl

Notes

Contributed paper

The Leicester Diabetes Risk Assessment Tools

Laura Gray, Melanie Davies, Kamlesh Khunti

A plethora of risk scores have been developed for identifying those at risk of type 2 diabetes. A recent review concluded that many of these risk scores are rarely used in clinical practice and research into the utility and impact of diabetes risk scores is in its infancy (1). The review also highlighted two promising areas of research, firstly interventions that prompt lay people to check their own diabetes risk and secondly the use of risk scores on population datasets to identify high risk "hotspots" for targeting public health interventions.

We have developed and validated two risk assessment tools for detecting both those at high risk of diabetes and those who have current undiagnosed type 2 diabetes which fulfill both of these promising areas. The Leicester self-assessment tool is a questionnaire based tool to be completed by individuals without intervention from a health care professional. The Leicester practice tool is for use within general practices using routinely stored electronic data.

This presentation will outline a number of work streams carried out to implement the use of these risk scores and to evaluate their impact, including (i) using the practice risk tool to identify people at risk in primary care for inclusion in a prevention programme; (ii) using a self-assessment risk score in non-English speaking ethnic minority groups; (iii) utilising smart phone technology for assessing risk.

Outcomes from this work which can be used by those implementing risk scores both in the area of diabetes and others will be considered.

Reference:
1. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ. 2011;343(d7163).

Contact: lg48@leicester.ac.uk

Notes

Contributed paper

Prediction study risk of bias assessment tool (PROBAST)

Karel Moons, Robert Wolff, Susan Mallett, Richard Riley, Marie Westwood, Penny Whiting, Jos Kleinen

Background: Quality assessment of included studies is a crucial step in the preparation of any systematic review. Review and synthesis of prediction studies – in general and of prediction modeling studies in particular - is a relatively new and evolving area for which, compared to systematic reviews of e.g. treatments and diagnostic tests, critical appraisal tools and review methods are not well developed. We aim to develop a prediction study risk of bias tool (PROBAST), with a focus on prediction modeling studies.

Methods: We are using a Delphi process to develop PROBAST; this process is currently ongoing. Thirty-five experts in the field of prediction research are taking part in the Delphi process. The procedure is being managed online using the Survey Monkey software. We anticipate that 4 or 5 rounds of this process will be needed before agreement is reached on the content of the final tool, that is in turn based on existing Risk of Bias tools (for e.g. systematic reviews of treatments, diagnostic tests, and prognostic factors), and guidelines for prediction (modeling) research.

Results: Discussion and Conclusions: The presentation will give an overview of the process, the current version of the tool (including the addressed domains and signaling questions) as well as an insight into underlying discussions. Feedback from the first rounds of our survey suggests that a domain-based approach similar to that implemented in e.g. QUADAS-2, with domains being rated as high, low and unclear risk of bias, will be used.

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

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.

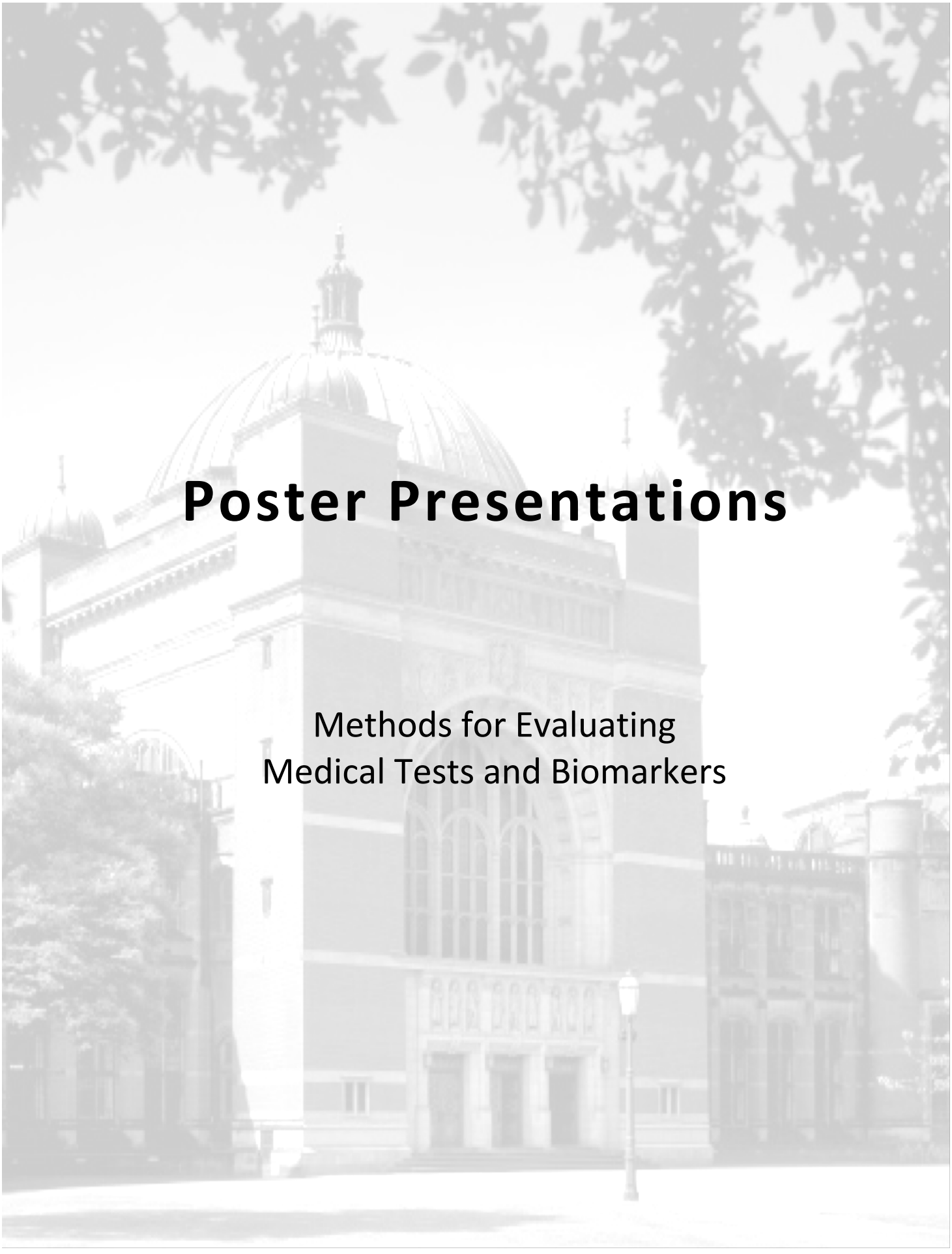
Design: systematic analysis of reported monitoring effects in current CVD prevention guidelines for three major CVD risk factors: cholesterol, smoking and hypertension.
Setting: primary and secondary care. We included guidelines published in English without any limitation on country or region of publication.
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	Evaluating the impact of diagnostic tests on patient health outcomes in the sbsence of randomized trials: a comprehensive decision analytic approach illustrated by two examples.	Erik Koffijberg
P2	How to judge when improved test accuracy will lead to improved treatment outcomes	Lukas Staub
P3	Using first cervical screening data for a non-vaccinated cohort to predict vaccine impact	Robert Newcombe
P4	Urinary Proteomics in Asthma: Search for a biomarker	Scott Elliott
P5	Accuracy of personal alcohol breathalysers with and without regulatory approval claims	Susannah Fleming
P6	Incorporating patients' preferences in the development, evaluation and application of biomarker tests.	Irene Mesa
P7	Meta-analysis and aggregation of multiple published prediction models	Thomas Debray
P8	The need for standardization of predictive biomarker assessment in ongoing clinical trials: an illustrative review of ERCC1 in NSCLC	Kinga Malottki
P9	Integration of biomarkers and clinical characteristics provides the best method of identifying patients with MODY	Beverly Shields
P10	Personalized screening for colorectal cancer; combining risk factors with fecal immunochemical test outcome for selecting CRC screenees for colonoscopy	Inge Stegeman
P11	CA19.9 profile in samples predating pancreatic cancer diagnosis – nested case control study in the UK Collaborative Trial of Ovarian Cancer Screening (UKTOCS)	Wendy Alderton
P12	Concordance in diabetic foot infection (CODIFI): A study protocol	Sarah Brown
P13	Ankle Brachial Index for the diagnosis of smpytomatic peripheral arterial disease in patients managed in a secondary health care setting	Fay Crawford
P14	Systematic review of the diagnostic accuracy of magnetic resonance spectroscopy and enhanced magnetic resonance imaging techniques in aiding the localisation of prostate abnormalities for biopsy	Moira Cruikshank
P15	Statistical analysis of eye tracking data to investigate the effect of computer aided design on diagnostic performance in CT colonography	Tom Fanshawe

P16	Exporation of used methods to detect publication bias in diagnostic test accuracy reviews	Lotty Hooft
P17	No evidence for an effect on the results of a diagnostic test accuracy review when only MEDLINE is searched	Lottty Hooft
P18	The reporting of research design of diagnostic test accuracy (DTA) studies in the abstracts in major medical journals	Harriet Hunt
P19	Registration of studies quantifying the accuracy of diagnostic tests and markers	Daniel Korevaar
P20	Assessing diagnostic impact within diagnostic accuracy study: STREAMLINE-L and STREAMLINE-C Streamlining staging of lung and colon cancer with whole-body MRI	Sue Mallett
P21	A suggested method for developing stopping criteria in large diagnostic studies where independent monitoring by a Data Monitoring Committee is required	Lee Middleton
P22	Blood CEA levels for detecting recurrent colorectal cancer in primary care	Brian Nicholson
P23	Poor interpretation of quality assessment in diagnostic accuracy reviews	Eleanor Ochodo
P24	Estimating the probability of grade misclassification using data from a national screening program for diabetic retinopathy	Jason Oke
P25	Some practical issues in designing diagnostic RCTs	Werner Vach
P26	Simultaneous confidence intervals for AUC, sensitivity and specificity	Antonia Zapf
P27	The ethical implications of prognostic studies of individual patient data	Bob Phillips
P28	Risk factors for hospitalization in children presenting with influenza/influenza-like illness in primary care: a prognostic systematic review	Kay Wang
P29	Panel diagnosis as Reference Standard in diagnostic research: A systematic review and methodological recommendations	Loes Bertens
P30	A Bayesian framework for estimating the incremental value of a diagnostic test in the absence of a gold standard reference	Nandini Dendukuri
P31	Evaluating diagnostic accuracy in the face of multiple reference standards	Christiana Naaktgeboren
P32	Practical methodological strategies for Individual Patient Data (IPD) systematic reviews	Fay Crawford
P33	Application of the GRADE methodology to Cochrane Diagnostic Test Accuracy reviews	Gowri Gopalakrishna

P34	Realist synthesis – what is it and how could it add value to our understanding of the diagnosis of dementia in primary care?	Harriet Hunt
P35	A model-based economic evaluation of test-treatment strategies: the cost-effectiveness of strategies to identify individuals with monogenic diabetes	Jaime Peters
P36	Modelling the long-term cost-effectiveness of using different MRI techniques to localize prostate abnormalities for biopsy in patients with a previous negative biopsy	Graham Scotland
P37	Upfront MRI followe by MRI-targeted biopsy for men with suspected prostate cancer: a decision analysis	Sarah Willis

Poster 1

Evaluating the impact of diagnostic tests on patient health outcomes in the absence of randomized trials: a comprehensive decision analytic approach illustrated by two clinical examples

Erik Koffijberg, Hester Den Ruijter, Ilonca Vaartjes, Michiel Bots, Bas van Zaane, Karel Moons

Background: Proper evaluation of new diagnostic tests is required to reduce overutilization, limit potential negative health effects, and reduce costs. Evidence only on diagnostic accuracy and reclassification no longer suffices as policy makers increasingly demand evidence on actual patient health outcomes. A decision analytic modelling approach may be worthwhile prior to conducting a diagnostic RCT, or when such a RCT is simply infeasible. We demonstrate our approach by assessing the impact of tests on subsequent clinical decision making, patient outcomes and cost-effectiveness of care of 1) a new diagnostic test in cardiac surgery patients, and 2) improved risk prediction in patients at high risk for coronary heart disease (CHD).

Methods: First, we assessed health outcomes and costs of modified transesophageal echocardiography (TEE) for the detection of atherosclerosis prior to cardiac surgery, using a decision analytic model based on evidence from an existing standard diagnostic accuracy study and actual Dutch reimbursement data. Second, we assessed health outcomes and costs of adding carotid intima-media thickness (CIMT) information to Framingham CHD risk prediction in primary prevention. This second model also incorporated detailed reclassification indices and evidence. The robustness of results was evaluated.

Results: Our decision analytic approach allowed estimating the health impacts of modified TEE and of CIMT, for various scenarios and subgroups of individuals. In 65-year-old men and women modified TEE resulted in a 17% lower risk of post-operative stroke, and an additional 0.03 quality-adjusted life years (QALYs). Simultaneously, cost savings were €12, and €60, with an incremental cost-effectiveness ratio of -€420 and -€1940 per QALY gained, for men and women respectively. CIMT-based reclassification induced a 0.01-0.02 increase in QALYs at additional cost of $100 per man, and a 0.03-0.05 increase in QALYs with a cost-saving of $200-300 per woman, over a period of 20-30 years. Results were robust with respect to our assumptions.

Conclusion: Decision analytic modelling to assess the impact of a new diagnostic test based on characteristics, effects and costs of the test itself, reclassification, and changes in subsequent treatment, is both feasible and valuable. Modified TEE appears valuable in patients older than 65 years, providing immediate health benefits, whereas CIMT measurements in individuals aged 60 results in small additional health benefits which only outweigh costs over a long time horizon. We suggest the use of decision analytic approaches before considering a formal large scale RCT.

Contact: h.koffijberg@umcutrecht.nl

Notes

Poster 2

How to judge when improved test accuracy will lead to improved treatment outcomes

Lukas Staub, Sarah Lord

Background: When accuracy studies confirm that a new test is more sensitive than existing tests, one concern is whether treatment harms outweigh benefit in the additional cases detected, particularly when the new test identifies a milder spectrum of disease.

Objective: To demonstrate how the principles used to assess the applicability of treatment evidence to broader populations [1] can be used to assess claims of improved treatment using a new more sensitive test.

Methods: We defined the additional cases detected for treatment by the new test as the 'new test-defined population'. Using published examples, we demonstrate how the principles used to appraise the applicability of treatment evidence to populations not included in treatment trials can help assess claims about treatment benefits using new tests.

Results: The treatment comparison for new test-defined populations is treatment change that will occur for patients based on the new versus existing tests. When this evidence is only available for populations defined by the existing tests, the critical assumptions are:
1. The extra cases detected by the new test are at sufficient risk of disease events to justify treatment. Relevant evidence to make these judgements includes: (i) the disease spectrum for the new test-defined population is similar to that of the existing test-defined population, or if unknown (ii) the baseline risk of disease events (prognosis) is similar across the disease spectrum.
2. Evidence for the effectiveness of treatment in populations defined by existing tests is generalisable to the new test-defined population. Supportive information includes: (i) trials demonstrating the relative effectiveness of treatment in patients with similar disease characteristics, or (ii) trials demonstrating the relative effectiveness of treatment is similar across different disease subgroups, and/or (iii) strong biological plausibility that treatment works in the new population.
3. Treatment harms are generalisable to the new test-defined population.

Conclusions: Using the existing evidence to judge disease risk and treatment benefit and harms will help researchers decide whether a more sensitive test leads to improved treatment outcomes or whether further treatment trials are needed.

Reference:

1. Glasziou,Irwig. BMJ 1995.

Contact: lukas.staub@memcenter.unibe.ch

Notes

Poster 3

Using first cervical screening data for a non-vaccinated cohort to predict vaccine impact

Robert Newcombe

Human papillomavirus (HPV) is the causal agent in cervical malignancy. There are numerous virus types, classed as either high risk (HR) or low risk (LR) of malignancy. Two HPV vaccines are currently offered to schoolgirls. The bivalent Cervarix HPV vaccine targets two HR types, HPV16 and HPV18, with effectiveness approaching 100%. The quadrivalent Gardasil vaccine additionally targets two LR types, HPV6 and HPV11 associated with genital warts. Evaluations of these vaccines in the large PATRICIA and FUTURE I/II trials indicate a possible cross-protection (CP) benefit against some but not all non-vaccine HR types.

Data on HPV status at first cervical screening at age 20-22 are available on a large 'baseline' cohort of women in Wales, the final cohort before vaccination was offered. These results may be combined with relative risks from the above trials to assess the potential benefits arising from direct protection and cross-protection. Several methodological issues arise.

It is most natural to use relative risks, though an odds ratio based approach is also applicable here. For low risks, unsurprisingly the two approaches lead to very similar results. If risks could approach 100%, clearly an odds ratio based model would be more satisfactory.

Confidence intervals for the impact of vaccination need to take account of two sources of imprecision, type-specific prevalence in the baseline series and the RRs (or ORs) for several HR types from the vaccine trial. This may be achieved using two algorithms, the Method of Variance Estimates Recovery (MOVER) adapted to ratios and products, and the Propagating Imprecision (PropImp) algorithm.

Published results indicate statistically significant cross-protection for some non-vaccine HR types, non-significant CP for others, and an apparent non-significant increase in risk for others, although negative cross-protection is considered implausible. Importantly, the methodology used accommodates all available cross-protection data to permit even-handed estimation of the potential cross-protection benefits of each vaccine if applied to our population and compare their resulting benefits.

Contact: newcombe@cardiff.ac.uk

Notes

Poster 4

Urinary Proteomics in Asthma: Search for a biomarker

Scott Elliott, Jonathan Owen, Thomas Brown, Sumita Kerley, Jan Shute, Anoop Chauhan, Dominic Reynish, Vicki Jefferey

Background: The use of inflammatory indices such as sputum eosinophilia to guide anti-inflammatory treatment in asthma has been shown to reduce the frequency and severity of exacerbations.

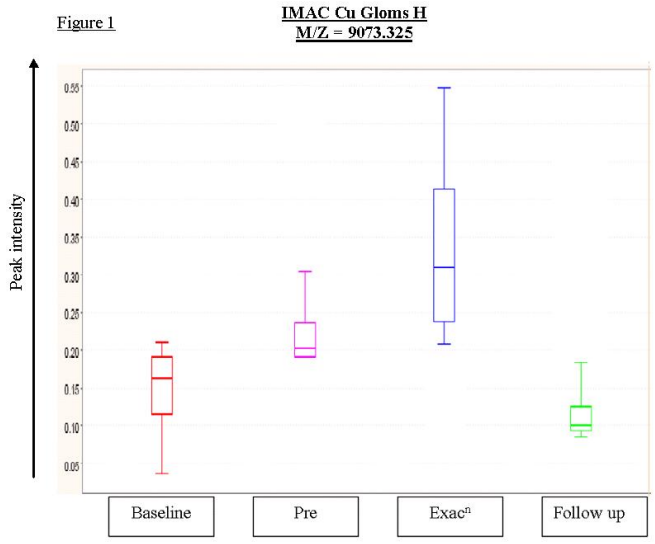
Aims: Sputum induction can be unpleasant for patients and analysis is costly and labour intensive necessitating alternative methods to differentiate inflammatory phenotypes, guide anti-inflammatory treatment and predict exacerbation risk.

Method: Performing Surface Enhanced Laser Desorption/Ionisation Time of Flight Mass Spectrometry utilising 6 different "chips" we analysed the urinary spectra from 3 groups, the first (exacerbation vs recovery (n=16), second (prospective patient samples thrice weekly, before, during and after an exacerbation (n=3), and third (patients with different inflammatory phenotypes (eosinophilic, neutrophilic, mixed granulocytic and paucigranular) (n=10)

Results: Differential protein signatures were found between inflammatory phenotypes (p=<0.05) and between exacerbation and recovery states (p=<0.05). The IMAC Cu chip identified a signature which delineated onset, exacerbation and recovery states [see fig.1]. Protein signatures were able to distinguish patients in each comparative group (P=<0.05)

Conclusion: Further work is warranted with a larger sample size to corroborate our findings and identify the proteins these signatures represent. Multiplexed cytokine panels (Milliplex®) are allowing us to further identify key immunophenotypic differences, detectable in plasma during loss of asthma control. This may ultimately identify a novel biomarker indicating pre-exacerbation states in asthma enabling early intervention.

Figure 1



Reference:

Barton L.M, Patel S.R, Molyneux K et al. SELDI-TOF mass spectrometry to identify urinary biomarkers of deep vein thrombosis Br J of Haem, 2008, 141 (s1), p 64

Contact: scott.elliott@porthosp.nhs.uk

Notes

Poster 5

Accuracy of personal alcohol breathalysers with and without regulatory approval claims

Susannah Fleming, Richard J Stevens, Elizabeth Spencer, Matthew J Thompson, Helen F Ashdown

Introduction: Alcohol-related driving incidents cause over 250 deaths in the UK annually. Alcohol breathalysers are medical devices which measure alcohol in breath and are typically used to estimate whether a person's blood alcohol level is over the legal driving limit. A number of breathalysers are available for sale to the public, some for under £10. Some countries, including France, require that drivers carry breathalysers displaying certain regulatory approval markings.

Method: We performed a diagnostic accuracy study, comparing three personal breathalysers (two single-use and one digital multi-use) with a police breathalyser (reference standard). We recruited participants in a number of licensed establishments in Oxford. Participants were 18 years of age or older, had consumed alcohol, and did not intend to drive within the following six hours. Each participant was asked to report their estimated alcohol intake before testing their breath alcohol levels with (in random order) the police standard breathalyser, a digital multi-use breathalyser, and one randomly allocated single-use breathalyser. The reference device was a breathalyser (Draeger 6510) with UK Home Office Type Approval for preliminary breath testing; it is also listed on the US National Highway Traffic Safety Administration Conforming Products List as an evidential breath tester, and carries a CE mark. Index devices were the Alcosense Elite, a digital multi-use device, and two single-use devices: the Alcosense Single and Draeger Alcocheck.

Results: We recruited 208 participants, of whom 71% were male, with median age of 20 years (range 18-66). According to the police breathalyser, 38 participants (18%) were at or above the current UK driving limit of 35 μg/100ml. Sensitivities were 89.5% (95% CI 75.9-95.8%) for the Alcosense Elite, 94.7% (95% CI 75.4-99.1%) for the Alcocheck, and 26.3% (95% CI 11.8-48.8%) for the Alcosense Single. The Alcosense Elite is CE-marked, but does not conform to any regulatory standards for breathalysers. Neither of the single-use devices carries a CE mark on the packaging. When contacted, the manufacturer of the Alcosense Single confirmed that the specific device tested has not passed any regulatory standards, but that derivative devices have passed standards including the French NF approval required for breathalysers to be carried by drivers in France, and the US National Highway Traffic Safety Administration Conforming Product List approval. We are currently awaiting a response from the manufacturer of the Alcocheck device.

Conclusion: Two devices showed high sensitivity, with a tendency to over-read, as might be expected for devices intended to "fail-safe" by overestimating alcohol levels to discourage intoxicated users from driving. Our results suggest that at least one breathalyser available for sale in the UK has very poor sensitivity for detection of alcohol levels over the driving limit; use of devices with this level of accuracy to guide driving decisions could have potentially catastrophic consequences. We found no apparent correlation between claims of regulatory approval and accuracy for breathalysers marketed to UK consumers.

Contact: susannah.fleming@phc.ox.ac.uk

Notes

Poster 6

Incorporating Patients' Preferences in the Development, Evaluation and Application of Biomarker Tests

Jean Harrington, Steven Sacks, Myfanwy Morgan, Maria Hernandez-Fuentes, Irene Rebollo-Mesa

Aim: The aim of this study is to develop a novel method by which the risk associated with biomarker-led care can be adjusted to individual patient's circumstances.

Background: The study, 'Genetic Analysis and Monitoring of Biomarkers of Immunological Tolerance' (GAMBIT) was designed to validate and develop an algorithm to classify patients as "immunologically tolerant to their transplanted graft", using a set of biomarkers expressed in blood and urine. When standard methods for the evaluation of performance of the biomarker tests are applied, a classification cut-off is selected that maximises equally sensitivity and specificity. Other methods have been suggested that take into account the clinical utility of the marker, however patients' preferences have not yet been incorporated in the process of development of biomarker tests.

Method: We will interview 100 patients currently participating in the GAMBIT study, 10 classified as tolerant, and 90 classified as stable, for whom extensive biomarker and phenotype data is available. We have developed a modified version of the Standard Gamble task to identify the level of risk that patients may be willing to take in order to choose biomarker-led care rather than current standard of care. The risk is varied as a function of the sensitivity and the specificity of the 'hypothetical' biomarker test. We will investigate the association between willingness to take risks and patients' quality of life and symptom burden, as measured by standardized questionnaires. If an association exists between patient's preferences and symptom burden, the latter can be incorporated in the classification algorithm, by varying the classification cutoff as a function of symptom burden.

Results: We will present preliminary results to test the hypothesis that high symptom burden and low quality of life are positively associated with a willingness to take the risks linked with less accurate biomarker tests. Specifically, we expect that patients who experience a greater symptom burden, or whose self-assessed QoL score is relatively low, will be more likely to choose biomarker-led care, in particular that which relies on a test with lower specificity and higher sensitivity. We anticipate this study will develop the methodology to incorporate patient's preference into the translation of biomarker tests into clinical practice

Contact: irene.rebollo_mesa@kcl.ac.uk

Notes

Poster 7

Meta-analysis and aggregation of multiple published prediction models

Thomas Debray, Hendrik Koffijberg, Daan Nieboer, Yvonne Vergouwe, Ewout Steyerberg, Karel Moons

Published clinical prediction models are often ignored during the development of novel prediction models despite similarities in populations and intended usage. Researchers sometimes update existing models using data at hand or develop a novel model when updating yields disappointing results or when differences between the development and updating populations are too large. The plethora of prediction models that arise from this practice tend to perform poorly when applied in new settings. Incorporating prior evidence might improve such models, making them more relevant. Unfortunately, aggregation of prediction models is not straightforward and methods to combine differently specified models are currently lacking.

We propose two approaches, Model Averaging and Stacked Regressions, for aggregating previously published prediction models and updating them when validation data are at hand. These approaches yield user-friendly stand-alone models that are adjusted for the validation population. Both approaches rely on weighting to account for model performance and between-study heterogeneity, but adopt a different rationale (averaging versus combination) to combine the literature models. We illustrate the implementation in two clinical datasets and compare them with established methods for prediction modeling in a simulation study.

Results from the clinical case datasets and simulation studies demonstrate that aggregation yields prediction models with an improved discrimination and calibration in a vast majority of scenarios, and results in equivalent performance (compared to developing a novel model from scratch) when validation samples are relatively large. In conclusion, model aggregation is a promising extension of model updating when several useful models are available from the literature and validation data are at hand.

Contact: t.debray@umcutrecht.nl

Notes

Poster 8

The need for standardisation of predictive biomarker assessment in ongoing clinical trials: an illustrative review of ERCC1 in NSCLC

Kinga Malottki, Lucinda Billingham, Richard Riley, Karen Biddle, Sanjay Popat

Background: Although platinum-based doublet chemotherapy is the standard of care in advanced non small cell lung cancer (NSCLC), it is associated with only a 25-30% response. Therefore the use of a predictive biomarker in this context could result in much more effective treatment selection, and excision repair cross-complementation group 1 (ERCC1) expression has been proposed as such a biomarker in (NSCLC). However a recent systematic review to characterize the prognostic and predictive effect of ERCC1 identified a lack of standardization of ERCC1 assessment methods in primary studies (1). As different methods of biomarker assessment may not provide comparable results(2), conclusions from the literature about ERCC1 and its applicability to clinical practice are therefore limited.

Objectives: We have undertaken a review and survey of ongoing NSCLC trials to investigate the methods used for assessment of ERCC1 and the rationale supporting their choice.

Methods: We have identified ongoing and recently completed clinical trials in NSCLC assessing ERCC1 using the WHO Clinical Trials Search Portal, the ISRCTN register and the ClinicalTrials.gov website. For each identified trial we collected information on trial design and will distribute a questionnaire to the named contacts, asking about the details of methods used for ERCC1 assessment (prospective or retrospective, biopsy type, the laboratory method used, local or central laboratory and the cutoff for positive results) in that trial and the rationale supporting the choice of the method.

Results: Our preliminary results include 35 ongoing studies (of which 5 are phase III, 15 are phase II and 1 is a phase II/III study). The sample size was up to 100 patients for 16 studies, 101-300 patients for 13 studies and over 300 patients for 6 studies. Fourteen studies were conducted in the USA, 9 in Europe, 8 in Asia and 4 in other locations. Of these 12 used ERCC1 for treatment assignment (4 non-RCTs, and 8 randomised biomarker-based strategy studies), 1 to stratify randomisation, 18 were correlative (all non-RCTs) and in 4 studies the role of ERCC1 was unclear. Reported methods of ERCC1 assessment were: polymerase chain reaction (PCR) in 4 studies, immunohistochemistry (IHC) in 3 and osing other methods or not reported in the remaining 28 studies. We will also report the results of our survey describing the methods currently used to assess ERCC1 and the reasons for their choice.

Conclusions: Standardisation of biomarker assessment is necessary for comparison of results of different studies and their implementation in clinical practice. However, often there is little agreement in the methods for assessment of a single biomarker in different trials. The results of our research will enable discussion of the current methods of assessing ERCC1 and the reasons motivating the choice, as well as possible steps leading to the standardisation of its assessment.

References:

1. Hubner, Richard A., et al. "Excision repair cross-complementation group 1 (ERCC1) status and lung cancer outcomes: a meta-analysis of published studies and recommendations." PloS one 6.10 (2011): e25164.
2. Besse, Benjamin, Ken A. Olaussen, and Jean-Charles Soria. "ERCC1 and RRM1: Ready for Prime Time?." Journal of Clinical Oncology (2013).

Contact: k.malottki@bham.ac.uk

Notes

Poster 9

Integration of biomarkers and clinical characteristics provides the best method of identifying patients with MODY

Beverley Shields, Timothy McDonald, Katharine Owen, Maciej Malecki, Rachel Besser, Angus Jones, Sian Ellard, Andrew Hattersley

Aims/Objectives: Maturity-onset diabetes of the young (MODY) is a rare, young-onset form of diabetes, often misdiagnosed as Type 1 diabetes (T1D) or Type 2 diabetes (T2D), resulting in inappropriate treatment and management. A number of biomarkers have been proposed to aid identification of patients with MODY, and clinical characteristics (age at diagnosis, BMI, HbA1c, parent with diabetes, treatment) are useful when combined in a clinical prediction model. We aimed to determine if combinations of biomarkers and clinical criteria improves the diagnostic accuracy for detecting MODY in patients diagnosed under 45.

Methods: We measured plasma C-peptide and GAD and IA-2 antibodies on patients insulin treated from diagnosis (144 T1D,71 MODY), and hsCRP and HDL cholesterol in patients not insulin treated from diagnosis (118 T2D, 216 MODY). Probability of MODY for each patient was derived using the clinical prediction model. Discriminative ability of the biomarkers/criteria was assessed using Receiver Operating Characteristic (ROC) curves.

Results: In patients insulin treated from diagnosis, C-peptide>80pmol/L and absence of GAD/IA-2 antibodies had 90% sensitivity, 93% specificity for MODY. The addition of clinical characteristics provided a minor improvement in diagnostic accuracy (ROC AUC 0.95 v 0.99, $p<0.001$; perfect test AUC=1). In patients not insulin treated from diagnosis, hsCRP<0.75mg/L and HDL>1.12mmol/L had 56% sensitivity and 87% specificity. The addition of clinical characteristics greatly improved diagnostic accuracy (ROC AUC 0.81 v 0.98, $P<0.0001$).

Conclusion/Summary: The use of multiple biomarkers outperforms single biomarkers for predicting MODY. Excellent discrimination between MODY and Type 1/Type 2 diabetes is achieved when biomarkers are used in combination with clinical characteristics.

Contact: beverley.shields@pms.ac.uk

Notes

Poster 10

Personalized screening for colorectal cancer; combining risk factors with fecal immunochemical test outcome for selecting CRC screenees for colonoscopy

Inge Stegeman, Thomas de Wijkerslooth, Esther Stoop, Monique van Leerdam, Evelien Dekker, Marjolien van Ballegooijen, Ernst Kuipers, Paul Fockens, Roderick Kraaijenhagen, Patrick Bossuyt

Background & aims: Fecal immunochemical testing (FIT) is increasingly used in colorectal cancer (CRC) screening but has a less than perfect sensitivity. Combining risk stratification, based on established risk factors for advanced neoplasia with FIT result for allocating screenees to colonoscopy, could increase the sensitivity of FIT-based screening. This way, the diagnostic yield from screening could be increased with a similar number of colonoscopies. We explored the use of a risk prediction model in CRC screening.

Methods: We collected data in the Colonoscopy or Colonography for Screening (COCOS) study, a multicenter screening trial. For this study 6,600 randomly selected, asymptomatic men and women between 50 and 75 years of age were invited to undergo colonoscopy. During colonoscopy, participants were identified with one or more advanced neoplasia (carcinomas and advanced adenomas). Screening participants were asked for one sample FIT (OC-sensor) and to complete a risk questionnaire prior to colonoscopy. Based on the questionnaire data and the FIT results, we developed a multivariable risk model with the following factors: total calcium intake, family history, age and FIT result. We evaluated goodness-of-fit, calibration and discrimination, and compared the risk model to a model based on primary screening with FIT only.

Results: Of the 1,426 screening participants, 1,112 (78%) completed the questionnaire and FIT. Of these, 101 (8.8%) had advanced neoplasia: 7 (0.6%) had CRC and 94 (8.3%) had advanced adenoma. The risk based model significantly increased the goodness-of-fit compared to a model based on FIT only ($p<0.001$). Discrimination improved significantly with the risk-based model (AUC: from 0.68 to 0.75 ($p=0.02$). Calibration was good (Hosmer-Lemershow test; $p=0.94$).

Conclusion: Risk stratification can be used as a tool to improve the effectiveness of screening by replacing age for risk as a threshold for inviting individuals for colorectal cancer screening.

Contact: i.stegeman@amc.nl

Notes

Poster 11

CA19.9 profile in samples predating pancreatic cancer diagnosis – nested case control study in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS)

Wendy Alderton, Sophia Apostolidou, Aiden Flynn, Alex Gentry-Maharai, Chris Hodkinson, Ian Jacobs, Usha Menon, Andy Ryan, Neomal Sandanayake, John Timms, Julie Barnes

Background: Pancreatic cancer is the fifth most common cause of cancer death and has a 5-year survival rate of only 3%. It often has a very poor prognosis since it is commonly not diagnosed until it is at an advanced stage and the cancer has metastasized. CA19.9 is the most widely used biomarker as an aid to the clinical diagnosis of pancreatic cancer. There are currently no screening methods for the early detection of pancreatic cancer. Here we explore CA19.9 levels prior to diagnosis of pancreatic ductal adenocarcinoma in a nested case control study set within UKCTOCS (Menon et al BMJ 2008, 337:a2079). The trial cohort of >202,000 apparently healthy postmenopausal women donated a single serum at recruitment and 50,000 women continued to donate serum samples annually. Cancer registry and postal follow up ensured that all women diagnosed with cancer following trial recruitment were identified.

Methods: UKCTOCS volunteers provided detailed lifestyle and health data on entry, during the trial and further data on their cancer diagnosis was obtained from their treating clinician. Cancer registration data was provided by the UK NHS Information Centre. Serum CA19.9 concentrations were determined in duplicate on a Roche Elecsys 2010 system. A Student's t-test was used to assess the significance of assay results comparing cases and controls ($p<0.05$), whilst Receiver Operating Characteristic (ROC) curves were constructed to determine performance of CA19.9 at pre-diagnosis time points.

Results: 56 women with primary pancreatic ductal adenocarcinoma (cases) with a total of 270 samples annual samples up to 8 years pre-diagnosis were identified and matched 5:1 (controls:cases) with controls who had no history of cancer. The mean CA19.9 value in controls was 11.1 U/ml. The longitudinal values of CA19.9 levels across cases included levels below the clinically accepted cut-off (37 U/ml).The CA19.9 values were significantly different between cases and controls at 0-1 years ($p<0.0001$) and 1-2 years ($p<0.043$) pre-diagnosis, but not at earlier pre-diagnosis time points. The ROC AUC at 0-1 years pre-diagnosis was 0.81 (60% sensitivity; 90% specificity) and 1-2 years pre-diagnosis was 0.71 (43% sensitivity; 90% specificity).

Conclusion: The prospective biobank derived from UKCTOCS represents a valuable collection of pre-diagnosis pancreatic cancer serum samples. The availability of samples in the years leading up to diagnosis, offers a unique opportunity for discovery and validation of novel, screening biomarkers for the early detection of pancreatic cancer. The longitudinal increase in CA19.9 up to 2 years preceding diagnosis highlights the value of such assessments. While CA19.9 used alone may be limited as a screening marker, its combination with other biomarkers may afford some utility for the early diagnosis of pancreatic cancer.

Contact: wendy.alderton@abcodia.com

Notes

Poster 12

Concordance in diabetic foot infection (CODIFI): A study protocol

Sarah Brown, Alexander Wright-Hughes, Michael R Backhouse, Moninder S Bhogal, Janine Gray, Jane Nixon, E Andrea Nelson

Background: Infection in diabetic foot ulcers is a foot-threatening complication, and is diagnosed clinically. Immediate antibiotic treatment is initiated without results from microbiological analysis of the infecting organisms, but samples of the bacteria in the ulcer are collected to inform treatment review. Wound swabs are commonly used to collect samples but guidelines recommend sampling by collecting tissue from the wound bed. We are comparing the microbiological information gained from these two sampling techniques. This will not determine whether swab or tissue sampling has an impact on clinical outcome, but will allow us to determine potential differences between techniques. As infection in chronic wounds is diagnosed clinically, and there is no agreed microbiological definition of infection for these ulcers, there is no 'gold-standard' against which to compare the two techniques. CODIFI is assessing agreement between results from swab and tissue samples from infected diabetic ulcers.

Methods: CODIFI aims to recruit 400 patients with a diabetic foot ulcer and suspected infection requiring antibiotic therapy from 25 sites across England (349 at March 2013). Both swab and tissue samples are collected according to the UK Health Protection Agency (HPA) standards.
Co-primary endpoints for the study will assess agreement between the two techniques for three microbiological parameters: reported presence of likely pathogens identified by the HPA, identification of antimicrobial resistance, and the number of isolates reported per specimen. Evaluation of the co-primary endpoints will use statistical methods including the Kappa statistic to measure agreement, McNemar's test to investigate the pattern of disagreement, and multinomial logistic and ordinal regression to determine whether agreement is influenced by baseline factors.

Secondary endpoints are an evaluation of the clinical significance of any differences in reports (by ascertaining from a panel of clinicians whether the specimen results would have led to a change in clinical management or not), adverse events, and a comparison of the pathogens reported by conventional culture techniques against molecular techniques (in 20 patients).
A sample size of 400 provides 80% power for detecting a difference of 3% in the primary outcome, assuming an overall prevalence of 10%, discordance of 5%, and a two-sided test at the 5% level of significance. This is based on less prevalent isolates such as Pseudomonas. Acceptable agreement is defined a priori as Kappa larger than 0.6.

Conclusions: CODIFI will produce robust data to evaluate the two most commonly used sampling techniques in infected diabetic foot ulcers. This has relevance for all clinicians working with diabetic foot ulcers.

This project was funded by the NIHR Health Technology Assessment programme (project number 09/75/01) and will be published in full in Health Technology Assessment. ISRCTN: 52608451. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the HTA programme, NIHR, NHS or the Department of Health.

Contact: medsbro@leeds.ac.uk

Notes

Poster 13

Ankle Brachial Index for the diagnosis of symptomatic peripheral arterial disease in patients managed in a secondary health care setting

Fay Crawford, Francesca Chappell, Karen Welch, Alina Andras, Julie Brittenden

Background: Peripheral arterial disease (PAD) is common, with a prevalence of both symptomatic and asymptomatic disease estimated at 13% in the over 50 age group. Symptomatic PAD affects about 5% of Western populations between the age of 55 and 74 years. Arterial stenosis and occlusion in the lower limb leads to inadequate blood flow to the muscles on exercise causing muscle pain which is relieved by rest. This is known as intermittent claudication (IC). The ankle brachial index (ABI), or ankle brachial pressure index (ABPI) as it is also known, is used to diagnose PAD. The ABI is widely used to assess peripheral vascular disease by a wide variety of health care professionals including specialist nurses, physicians, surgeons, and podiatrists working in secondary care settings. Other imaging tests are used, duplex ultrasound (DUS) allows the assessment of blood flow in the arteries The ABI is used as a triage or add-on test in clinical practice.

Types of studies: our review includes cross sectional studies of fully-paired direct comparisons between ABI and duplex ultrasound or ABI alone. Only studies which report that all patients received a reference standard and present 2x2 data will be eligible for inclusion.

Types of participants: Adults with leg pain, which is worse on walking and alleviated by rest who are tested in secondary care settings (hospital-out patients) and who are classified as category 1, 2, or 3 based on the Rutherford index or category II based on the Fontaine classification system are included in the review.

Index tests: Ankle brachial index, sometimes called the ankle brachial pressure index. Data collected using manual sphygmomanometers (both manual and aneroid) as well as digital equipment using manual or automatic inflation will be included in the review.

Target conditions: Peripheral arterial disease: the presence, absence and severity of disease as classified by the Rutherford or Fontaine Classification indices (Rutherford classification 1,2,3 or Fontaine II) . We have defined positive test results as those below 0.91 for peripheral arterial disease and are aware that readings of 1.30 and above may indicate calcification of the arteries.

Reference standard: Duplex ultrasound.

Search strategy: The Trials Search Co-ordinator of the Cochrane Peripheral Vascular Diseases Group will carry out electronic searches of several databases using specifically designed electronic search strategies.

Data extraction: Two review authors (FC, AA) will independently apply the exclusion criteria to the full papers and resolve any disagreements by discussion.

Quality assessment: After a pilot phase involving two reviewers working independently, we will use the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) to develop a quality assessment tool, incorporating the review question.

Results: we will present the preliminary findings of our review of the diagnostic test accuracy of ABI for the diagnosis of symptomatic peripheral arterial disease in patients managed in a secondary health care setting

Contact: fay.crawford@nuth.nhs.uk

Notes

Poster 14

Systematic review of the diagnostic accuracy of magnetic resonance spectroscopy and enhanced magnetic resonance imaging techniques in aiding the localisation of prostate abnormalities for biopsy

Moira Cruikshank, Graham Mowatt, Graham Scotland, Charles Boachie, John Ford, Cynthia Fraser, et al

Background: Prostate cancer is the most common cancer in men in the UK. Diagnosis can be confirmed only by prostate biopsy. However, men with a negative biopsy often continue to have a raised prostate specific antigen (PSA) level and the optimal re-biopsy strategy is uncertain. New imaging techniques have therefore been introduced in order to reduce unnecessary follow-up biopsies. Conventional standard magnetic resonance imaging (MRI) can be performed with add-ons including three-dimensional magnetic resonance spectroscopy (MRS), dynamic contrast enhanced MRI (DCE-MRI) and diffusion weighted MRI (DW-MRI).

Aims/objectives: To assess the diagnostic accuracy of MRS and enhanced MRI techniques (DCE-MRI, DW-MRI) in aiding the localisation of prostate abnormalities for biopsy in men with suspected prostate cancer and elevated PSA but previously negative biopsy.

Methods: Electronic searches of 15 databases and websites were undertaken. Types of studies considered included direct studies/randomised trials reporting diagnostic outcomes. Index tests were MRS, DCE-MRI and DW-MRI while comparator tests were standard (T2-weighted) MRI and transrectal ultrasonography (TRUS). The reference standard was histopathological assessment of biopsied tissue. Meta-analysis models were fitted using hierarchical summary receiver operating characteristic (HSROC) curves.

Results: Fifty-one studies (39 full text, 12 abstracts) were included, involving over 10,000 men. In pooled estimates, sensitivity (95% CI) was highest for MRS at 92% (86 to 95%), followed by T2-MRI at 86% (74 to 93%) and DCE-MRI at 79% (69 to 87%), while specificity (95% CI) was highest for TRUS (used as an imaging test) at 81% (77 to 85%), followed by MRS at 76% (61 to 87%). Only one small study involving 43 participants reported DW-MRI, with sensitivity of 100% (specificity not reported).

Conclusions: MRS had higher sensitivity and specificity than T2-MRI. Evidence relating to DCE-MRI and DW-MRI was limited. TRUS used as an imaging test had low sensitivity but high specificity.

Contact: mcruickshank@abdn.ac.uk

Notes

Poster 15

Statistical analysis of eye tracking data to investigate the effect of computer aided design on diagnostic performance in CT colonography

Tom Fanshawe, Susan Mallett, Emma Helbren, Peter Phillips, Stuart Taylor, Douglas Altman, Alastair Gale, 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 colorectal polyps. To assist readers, potential polyps may be marked on the video by the addition of a visual indicator ('CAD mark') automatically generated via computer aided design software. Interest lies in how the addition of a CAD mark affects the visual search and identification of polyps by readers, or whether it acts a distractor that hinders correct polyp identification.

Study design and methods: In this study, the visual search of 42 radiologists (25 novices and 17 experts) was recorded using a Tobii X50 eye-tracker as they each viewed the same 15 30-second 3D CT colon videos. Each reader viewed each video twice, once with and once without the addition of CAD marks. Readers indicated with a mouse click potential polyps that they would scrutinize further if seen in daily practice. From the recorded longitudinal eye tracking data (typically consisting of over 1000 coordinate pairs per reader:video combination), several metrics were developed to characterise the search patterns and diagnostic accuracy of readers and the effect of the CAD mark. These include time to first visual pursuit of polyp, time of first correct identification of polyp and time spent looking at CAD mark.

Analysis: Multilevel modelling, taking into account the cross-classified nature of the design, was used to determine differences in the pre-specified metrics between novice and expert readers, and in videos with CAD compared to those without CAD. Methodological issues encountered include a high proportion of missing data (addressed using multiple imputation) and zero-inflation in the time to first pursuit metric, caused by the introduction of the CAD mark.

Summary: This collaborative programme is the first to develop the necessary technical capacity and methodology for assessing diagnostic accuracy in medical imaging using eye tracking data obtained from moving 3D images. Results showed clear differences in most of the metrics analysed between videos with CAD and those without CAD. Further work is needed to obtain insight into the nature of errors in diagnosis and the capabilities of CAD software to prevent them.

Contact: thomas.fanshawe@phc.ox.ac.uk

Notes

Poster 16

Exploration of used methods to detect publication bias in diagnostic test accuracy reviews

Annefloor van Enst, Eleanor Ochodo, Rob Scholten, Mariska Leeflang, Lotty Hooft

Background: The effect of publication bias can seriously distort the results of systematic reviews. Therefore it is advocated to graphically explore the presence of publication bias in a funnel plot or to statistically test for it. However, the results of reviews of diagnostic test accuracy (DTA) are more heterogeneous than the results of intervention reviews, making the tests less reliable (1), and to date it's not well understood how those methods could be applied in DTA reviews. In this study, we explored if and how authors are currently addressing publication bias in DTA reviews.

Methods: A systematic search was executed in MEDLINE between September 2011 and January 2012 to identify recently published quantitative DTA reviews. Data was extracted on whether and how publication bias was mentioned in the text of the review, results were graphically displayed and formal tests were being used.

Results: Out of 1335 references, we included113 reviews. In 73/113 the authors discussed publication bias whereof in 44 they actually planned to test for publication bias. The three most used methods were the Egger test (1997) in 25 reviews (61%), the Deeks funnel plot asymmetry (2005) in 11 (27%) and Begg's funnel plot asymmetry (1994) in 8 (20%) reviews. In 12 reviews the authors concluded to have identified publication bias based on the Egger test (9 times), Deeks test (2 times) and Begg's and Egger test (1 time); in 20 the test results were negative and for 6 reviews the results of the tests were interpreted as inconclusive. When authors discussed publication bias without testing for it, the reason most often provided was that the current methods to investigate publication bias are insufficient.

Discussion: Diagnostic meta-analyses may also be threatened by publication bias. However, little is known about the actual presence and the potential impact of publication bias and selective outcome reporting in DTA reviews. Statistical methods to test for publication bias in diagnostic meta-analyses have their limitations. More guidance and empirical studies on the use and interpretation of these tests are needed.

Contact: l.hooft@amc.uva.nl

Notes

Poster 17

No evidence for an effect on the results of a diagnostic test accuracy review when only MEDLINE is searched

Annefloor van Enst, Rob Scholten, Aeilko Zwinderman, Lotty Hooft

Background: In every systematic review comprehensive searching in multiple databases to identify primary studies is of major importance. However, this process is time-consuming and costly. Methods for efficient searching for diagnostic test accuracy (DTA) studies are therefore needed, but should not introduce biased results. Limiting the number of databases could reduce the number needed to screen (NNS). However, empirical research has shown that excluding EMBASE to find randomised controlled trials (RCTs) will affect the results of intervention reviews. To date, little research has been done on the need for extensive searches in multiple databases for DTA reviews. The aim of this meta-epidemiological study is to analyse whether bias is introduced when only DTA studies that are indexed in MEDLINE are included in the meta-analyses of a DTA review.

Methods: A systematic search in MEDLINE (PubMed) was performed to identify high quality DTA reviews that searched multiple databases, published between January 2010 and August 2011.
First, for each included review and meta-analysis, the proportion of Not-in-MEDLINE (NiM) studies was calculated. Secondly, the impact on the results of the meta-analyses of including only studies indexed in MEDLINE (iM) was measured quantitatively by redoing the meta-analysis where possible. Analyses were executed according to the bivariate random effects model (11) in Stata version 10.0 (12). The RDOR of iM compared to all included studies was calculated with corresponding confidence intervals.

Results: In total 42 reviews were included which addressed 1277 primary studies; 1225 (95.9%) of those were indexed in MEDLINE. In 28 reviews (67%) all included primary studies were indexed in MEDLINE. For the remaining 14 reviews the percentage of iM studies ranged from 77.8 to 99.2%. Seven meta-analyses could be replicated. The pooled RDOR was 0.97 (95% CI 0.95 – 1.00) indicating that limiting the search to MEDLINE could lead to an underestimation of the 'true' accuracy (non-significant).

Discussion/Conclusion: A limited search could lead to biased results. However, 67% of the DTA systematic reviews will not have hampered results because all included primary studies could be found in MEDLINE.

Contact: l.hooft@amc.uva.nl

Notes

Poster 18

The reporting of research design of diagnostic test accuracy (DTA) studies in the abstracts in major medical journals

Zhivko Zhelev, Harriet Hunt, Christopher Hyde

Background: The abstract of a scientific publication helps readers to gain an overall idea of the study and to decide whether or not they should read the full text. Readers will often not read beyond the abstract due to limited time and interest, or when screening on 'title and abstract' if conducting a systematic review. However, within diagnostic test accuracy studies, the reporting of research designs within the abstract section is commonly inconsistent or absent. This has implications for screening on the basis of study design as well as for study identification in terms of keywords and indexing. Our aims were to catalogue different ways in which the research design terminology of DTA studies is reported in their abstracts, and to explore the potential inconsistency in the terminology used to refer to different DTA designs.

Methods: A convenience sample of DTA studies published in English within EMBASE between 2012 and 2013 was identified and their abstracts analysed to identify a) the different ways in which the research design is reported in the abstracts; b) the range of terms used to refer to research design across the studies; and c) any inconsistency in their use. For the purposes of this study "research design" is defined as "the procedures and methods, predetermined by an investigator, to be adhered to in conducting a research project" (IEA p.114, 1988).

Results: OvidSP Embase 1996 to 2013 Week 01 was searched on 17 January 2013 using 'diagnostic accuracy' (subject heading) as a search term, and 714 publications were identified. Search results were imported into EndNote X5, and the first investigator screened the first 200 records. Of these, 82 were primary diagnostic test accuracy studies and were included in the current analysis. 64 (76%) did not use shorthand terminology within the title or abstract to identify the design of the study. The remaining 18 abstracts varied in their use of design terminology, with terms including prospective study, retrospective cohort study, concurrent validity study and controlled study. 21 (25%) references used no keywords referring to the specific diagnostic accuracy design in the 'Keywords' section of the reference.

Conclusions: We created a summary map of frequency and distribution of the research design reporting within titles and abstracts of our small sample (N=200) of papers published in major medical journals. Whilst this is a limited sample, this shows that there is a great deal of inconsistency within the terminology used to refer to different DTA designs. We would suggest that this inconsistency of reporting could be overcome with greater standardisation and wider understanding of study design terminology. The research reported could be extended by investigating whether any study design term used in the abstract matches the description of the study method in the body of the journal article.

Reference:

International Epidemiological Association (1988) Editor: John Last. A Dictionary of Epidemiology (2nd Edition). Oxford University Press: New York, Oxford, Toronto.

Contact: zhelev.zhivko@pcmd.ac.uk

Notes

Poster 19

Registration of Studies Quantifying the Accuracy of Diagnostic Tests and Markers.

Daniel Korevaar, Lotty Hooft, Patrick Bossuyt

Background: Since September 2005, the International Committee of Medical Journal Editors (ICMJE) obliges researchers to prospectively register essential information about the design of their randomised controlled trials (RCT's) in a publicly available trial registry (1). The reasons for prospectively registering RCT's also apply to studies quantifying the accuracy of diagnostic and predictive tests and markers. By facilitating transparency and completeness of reporting, publication bias, outcome reporting bias and duplication of research efforts can be prevented. In addition, knowledge and research gaps can be identified, and a more efficient allocation of research funds can be promoted. Currently, several trial registries already contain test accuracy studies but the extent of these studies that is registered is unknown.

Aim: Our objective was to identify the proportion of test accuracy studies reporting that their study was registered in a trial registry.

Methods: We searched PubMed for studies published in May and June 2012 in journals with an impact factor of 5 or higher. Articles were included if they reported on original studies evaluating the accuracy of one or more diagnostic or predictive tests or markers against a clinical reference standard in human subjects. We excluded commentary/ies, discussion articles, reviews and meta-analyses. The full text of every included article was checked for a trial registration number.

Results: We found 1,952 references of which 351 met the inclusion criteria: 239 (68%) studies reported on diagnostic tests, 112 (32%) studies reported on predictive tests. A trial registration number was provided by 20 (6%) studies: 13 (5%) in diagnostic studies and 7 (6%) in predictive studies. Of the registered studies, only 6 (30%) had been prospectively registered. The others had been registered somewhere between the start and completion date as provided in the registry (n=11, 55%), or after the completion date (n=3, 15%). Registered data varied considerably between studies.

Conclusion: The number of test accuracy studies published in journals with an impact factor of 5 or higher reporting that their study has been registered is low. The majority of reported studies has not been prospectively registered. Further promotion of the necessity of trial registration among authors of test accuracy studies is needed.

Reference:

1. DeAngelis CD, Drazen JM, Frizelle FA et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. JAMA. 2004;292(11):1363-1364.

Contact: d.a.korevaar@amc.uva.nl

Notes

Poster 20

STREAMLINE-L and STREAMLINE-C Streamlining staging of lung and colon cancer with whole body MRI

Susan Mallett, Steve Halligan, Steve Morris, Doug Altman, Sam Janes, John Bridgewater, Anne Miles, Shonit Punwani, Tobias Schaeffter, Andrea Rockall, Vicky Goh, Anwar Padhani, Rob Glynne Jones, Lee Siow-Ming, Sandy Beare, Neal Navani, Ashley Groves

Introduction: STREAMLINE trials are designed to compare diagnostic accuracy for detection of metastases during staging of newly diagnosed lung and colon cancer, using early whole-body magnetic resonance imaging (WB-MRI) compared to standard NICE approved imaging pathways. Accurate detection of presence or absence of metastatic spread facilitates appropriate treatment. In lung cancer, 20% of patients undergoing "curative" thoracotomy rapidly relapse with metastatic spread undetected by conventional staging tests. In colon cancer initially undetected liver metastases result in reduced use of surgical/ablative therapy.

Study design: Two parallel but separate cohort studies of patients with newly diagnosed lung and colorectal cancer are being recruited from six NHS Imaging hubs. All patients will receive both early WB-MRI and standard NICE approved pathways for staging. Interpretation of tests are blinded. At multidisciplinary team meetings treatment decisions are assigned sequentially based on the standard pathway, a theoretical scenario of only early WB-MRI pathway results being known and finally based on all tests. WB-MRI results are revealed only after treatment decisions based on standard imaging tests are documented. Reference standard is by expert panel using imaging from up to 12 month followup and includes both a reference standard for diagnostic accuracy and treatment decision. Recruitment is about to start with a trial duration of 54 months.

Analysis: The primary outcome is the difference in sensitivity per patient to detect metastases with early WB-MRI as a replacement test to standard NICE pathway. Secondary outcomes include analysis of diagnostic impact in terms of how WB-MRI affects patient treatment decisions. Other secondary outcomes include the time to first major treatment decision, the number of tests in pathway, comparison of specificity per patient, sensitivity and specificity measured per organ and per metastasis. We will also analyse reader perception errors.

Summary: This study design is a prospective diagnostic accuracy study that includes design features to allow within trial measurement of diagnostic impact on patient treatment.

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

Notes

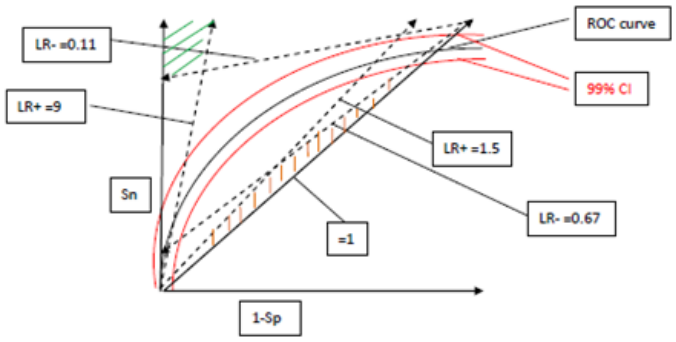
Poster 21

A suggested method for developing stopping criteria in large diagnostic studies where independent monitoring by a Data Monitoring Committee is required

Lee Middleton, Patrick Chien, Jonathan Cook, Karen Ward, Pallavi Latthe, Jon Deeks

Background: Independent monitoring by a Data Monitoring Committee (DMC) of large, publically funded (NIHR HTA, MRC, etc) test accuracy (TA) studies is considered mandatory but there is a lack of guidance as to how DMC members should use results of interim accuracy estimates (e.g. sensitivity, specificity, likelihood ratios) to make recommendations regarding the possible stopping or altering the study they are monitoring. It is advised that stopping rules are agreed by a DMC prior to a randomised controlled trial1 with respect to interim analysis and the same could apply in a TA study. In this example, DMC members for the BUS study (accuracy of bladder wall thickness [BWT] by ultrasound in the diagnosis of Detrusor Overactivity [DO]) formulated a novel approach for developing early stopping criteria prior to the examination of any data.

Methods: All three independent members of the DMC agreed that provision should be made for the possibility that BWT may be shown to be overwhelmingly good or bad at ruling in or out DO in an interim analysis. Levels of post-test probabilities were agreed upon that would be considered convincing or not convincing in test positive and test negative cases. These were to be converted to likelihood ratios (using the formula: likelihood ratio=post-test probability/(pre-test odds x (1-post-test probability))) and plotted alongside the confidence intervals (CI) of a receiver operating curve (ROC) in any interim results for visual comparison. If the lower CI of the ROC curve were to enter the shaded green area or similarly the upper CI of the ROC curve were to enter the brown shaded area (see figure) then the members would consider stopping the study prematurely. The estimates used for likelihood ratio thresholds were dependent on the presumed disease prevalence at the outset so interim estimates were to be examined to check they were broadly accurate. The use of ROC curves was agreed to be preferable as there was some doubt about the optimum cut-off of BWT prior to the study starting. 99%CI curves were used as a conservative measure.



Conclusions: Members of DMCs are advised to inspect interim accuracy estimates in publically funded TA studies but may not always use pre-agreed thresholds of accuracy estimates for stopping criteria. Thresholds for post-test probabilities and corresponding likelihood ratios can be visually compared against the confidence intervals of ROC curves. References: 1DAMOCLES Study Group. A proposed charter for clinical trial data monitoring committees: helping them do their job well. Lancet 2005; 365:711-22.

Contact: l.j.middleton@bham.ac.uk

Notes

Poster 22

Blood CEA levels for detecting recurrent colorectal cancer in primary care

Brian Nicholson, Bethany Shinkins, Indika Pathiraja, Tim James, Rafael Perera-Salazar, David Mant

Background: Preliminary analysis from the Follow-up After Colorectal Surgery (FACS) trial suggests that measurement of blood Carcino-embryonic Antigen (CEA) may be the most cost-effective way of detecting recurrence of colorectal cancer. CEA therefore has potential as a triage test in the primary care setting where it could be used alone to prompt referral for further investigations in secondary care. However, there is no consensus on the interpretation of CEA test results, with substantial variability in everyday clinical practice. The most recent review focused only on the diagnostic value of the absolute level from a single test and suggested using a cut-off of 2.2ug/L; this level generates a high level of false-alarms and is implemented by few clinicians1.

Methods: We plan to conduct a Cochrane Diagnostic Test Accuracy Review to further define the optimal CEA threshold for use as a triage test in primary care to prompt further investigation in secondary care for colorectal cancer recurrence, and to identify sources of between and within study heterogeneity to guide a subsequent Individual Patient Data (IPD) analysis. We will conduct a systematic literature search followed by QUADAS-2 assessment of all included papers, and use changes in laboratory analysis techniques over time to inform the interpretation of meta-analysis.

Results: Initial evaluation shows a large degree of heterogeneity between available studies with regards to study design, the predefined CEA threshold, the biochemical analysis technique, gold standard investigation.

We will present our most up-to-date results at the symposium.

Reference:

1. Tan E, Gouvas N, Nicholls RJ, Ziprin P, Xynos E, Tekkis PP. Diagnostic precision of carcinoembryonic antigen in the detection of recurrence of colorectal cancer. Surgical oncology 2009;18(1):15-24.

Contact: brian.nicholson@phc.ox.ac.uk

Notes

Poster 23

Poor interpretation of quality assessment in diagnostic accuracy reviews

Eleanor Ochodo, Patrick Bossuyt, Mariska Leeflang

Background: Interpreting and presenting results in systematic reviews without taking into account the outcome of quality assessment has been shown to be common in systematic reviews of interventions. This may also play a role in reviews of test accuracy studies. Drawing conclusions or making recommendations without considering the risk of bias and limited applicability of included studies may lead to unwarranted optimism about the value of the corresponding test. We sought to identify and compare approaches used to make conclusions out of the results of quality assessment in diagnostic accuracy reviews, and to provide guidance on recommended methods.

Methods: We searched MEDLINE for test accuracy reviews published between May and September 2012. We examined the abstracts and main texts of these reviews to check whether and how the results of quality assessment were taken into account when drawing conclusions. Data was extracted by one author; a sample was checked by a second author.

Results: Our search identified 53 eligible reviews. Of these, 49 (92%) had formally assessed the methodological quality of included studies; Twenty-two (45%) distinguished high quality from low quality studies using summary scores (n=13), summary graphs (n=5) and other methods. Overall, only 5 articles (10%) incorporated quality assessment results in the conclusions of the abstracts; in the main texts, only 14 (29%) reviews considered results of quality assessment in the recommendations or conclusions. Examples of approaches used to mention and interpret quality assessment in the sample of reviews can be found in the table.

Conclusion: About 7 out of 10 recent reviews of test accuracy do not take into account quality assessment when drawing conclusions or making recommendations. We recommend the results of quality assessment be always factored into the conclusions of test accuracy reviews, to limit the misleading presentation of the performance of the diagnostic tests.

Contact: e.a.ochodo@amc.uva.nl

Notes

Poster 24

Estimating the probability of grade misclassification using data from a national screening program for diabetic retinopathy.

Jason Oke, Irene Strattor, Richard Stevens, Peter Scanlon

Introduction: Annual screening for retinopathy is recommended for people with diabetes, with referral to ophthalmology clinics for patients with sight-threatening diabetic retinopathy. As part of the screening process, digital photographs of the retina are categorised into one of eight grades according to the degree of retinopathy in both eyes. As grading is not an exact science there is variation between graders and between the same grader on a different day. The sensitivity and specificity (or misclassification rates) of the screening program relies on accurate grading of photographs, but direct estimation of misclassification is not always possible. We used longitudinal data from an established screening programme with good quality assurance and quality control procedures and a stable well trained workforce to model the progression of diabetic retinopathy allowing for grading error.

Methods: We used a time-inhomogenous hidden Markov model with four states (No retinopathy, one eye affected, two eyes affected and referable level) to estimate the probability of true progression or regress and the conditional probability of an observed grade given the true grade (misclassification). The stage of retinopathy was assumed to progress as a function of the duration of diabetes and transitions were adjusted for baseline glycated blood sugar and type of diabetes.

Results: In 12,004 patients with up to eight years of annual screening with 49,110 person-years of follow-up, 1069 (9%) people had photographs that were graded as being of referable level. Conditioning on a true grade of no retinopathy, the probability (95 C.I.) of misclassification as background retinopathy in one eye was 0.18 (0.17 – 0.19), two eyes 0.081 (0.077 – 0.084) and referable grade retinopathy 0.000006. Conditioning on a true state of one eye with background retinopathy, the probability of misclassification as two eyes with background retinopathy was 0.2 (0.15 – 0.27) and no retinopathy 0.001 (0 – 0.1). For a true state of two eyes with background retinopathy, the probability of misclassification as one eye affected was 0.05 (0.03 – 0.08) and no background retinopathy 0.0003 (0 - 0.0.3).

Conclusion: The misclassification of a photograph to a more advanced stage (false positive result) is more common than misclassification into a lower grade. The probability of misclassifying a photograph as a referable level is very small from all other grades but this may be more reflective of the limitations of the method and data. A modelling approach to estimating misclassification rates is feasible using data from a screening program but may be limited to progression up to and no further than referral.

Reference:

1. Freund and Schapire, (1997). Adaboost.M1.

Contact: jason.oke@phc.ox.ac.uk

Notes

Poster 25

Some practical issues in designing diagnostic RCTs

Werner Vach, Izabela Kolankowska, Bettina Schnitter

To assess the additional clinical benefit of a new diagnostic method, well conducted RCTs with patient relevant outcomes may provide information at a high evidence level. They can overcome limitations of comparative accuracy studies, which can only study the improvement in sensitivity and specificity, but do not allow directly to assess the benefit of improved patient management due to better diagnoses. In the last years many papers have advocated the use of RCTs in diagnostic research and a variety of designs have been proposed. However, little has been said about practical issues in designing comparative diagnostic RCTs. In our talk we try to address some of these issues:

Trial justification and patient information: How to use results from accuracy studies to justify an RCT without losing equipoise?

Choice of patient-relevant outcomes: How to avoid just measuring accuracy or management decisions? Dimensionality of outcome: How to catch the consequences of both false positive and false negative decisions?

Choice of comparator: How to ensure a relevant comparison?

Standardization of diagnostic procedures: Efficacy or effectiveness?

Standardization of management: Necessary, useful and feasible?

Sample size: How to ensure sufficient power?

Trial-implementation gap: How to ensure later implementation of study results?

Trial-relevance gap: How to ensure that study results are still relevant at the end of the study?

We present general considerations about these issues and illustrate them using results from a systematic review of RCTs comparing PET with another modality for primary diagnosis, staging or follow up evaluation.

Contact: wv@imbi.uni-freiburg.de

Notes

Poster 26

Simultaneous confidence intervals for AUC, sensitivity and specificity

Antonia Zapf

Background: In early diagnostic trials the aim is often to select the most promising markers or tests. The selection criterion in such studies is usually the area under the ROC curve (AUC). In confirmatory studies sometimes an experimental test is performed under several conditions, for example an imaging system with different contrast agents. In this case, sensitivities and specificities are compared with pre-defined thresholds.

In both situations the confidence intervals for the outcome measures are much more interesting than the p-values. But to control the global type one error, the intervals have to be adjusted for multiplicity. Simple corrections, for example the Bonferroni method, are often too conservative and lead to low power. Therefore my objective was to develop simultaneous confidence intervals for the AUC, sensitivity and specificity.

Methods: In the study design investigated, each test is performed on each individual. For the construction of the simultaneous confidence intervals I used rank based estimators and the quantile of a multivariate normal distribution. In addition, I applied the Logit transformation to get range-keeping intervals. For the comparison of the simultaneous, unadjusted and Bonferroni-adjusted intervals I performed a simulation study and analyzed the data of a diagnostic study with different contrast agents. The influence of the following parameters was investigated in the simulation study: number of patients, effect size, prevalence, and correlation.

Results: As expected, in the example dataset the unadjusted intervals were the broadest, and the simultaneous intervals were narrower than the Bonferroni-adjusted ones. The simulation study demonstrated that the simultaneous confidence intervals maintain the type I error α best. Furthermore there is a positive correlation between empirical α and true AUC, and between α and sample size. The correlation between number of factor levels and α, and between prevalence and α is negative. In contrast there is no relationship between the correlation coefficient and α.

Key reference: Konietschke et al. (2011). Rank-based multiple test procedures and simultaneous confidence intervals. EJS, 6:737-758.

Contact: antonia.zapf@med.uni-goettingen.de

Notes

Poster 27

The Ethical Implications of Prognostic Studies of Individual Patient Data

Bob Phillips, Neil Ranasinghe, Lesley Stewart

Introduction: Meta-analysis of aggregate prognostic data from studies produces greater difficulties than similar approaches with trial data, it has been suggested that individual participant data (IPD) meta-analyses should be undertaken. We formed an international collaborative (PICNICC) to find clinical variables that predict the outcome of children and young people presenting with febrile neutropenia, a complication of cancer therapy, using anonymous un-linked patient data. As part of this group we undertook an investigation into the ethical and regulatory considerations involved in sharing such information for our projects.

Methods and materials: The "Predicting Infectious ComplicatioNs In Children with Cancer" (PICNICC) Collaborative IPD review collaborators, and those who expressed an interest in the project but did not submit data, were surveyed by email to find their experiences of the ethical and regulatory issues they had faced.

Results: The collaborators came from North and South America, Europe and the UK. . In the UK, some other European countries and USA, specific applications were made and consent obtained to share the information from previously conducted studies. Other groups were able to share their data from previous investigations without further formal approval. To our knowledge, no potential collaborative group had their request to share such data declined.

Discussion: Despite a perception from some clinicians that regulatory and ethical frameworks are blocking the work of clinical academics, it is important to report that large-scale international collaborations can effectively share data without obstruction or significant delay. Most service users wish to see care improved, and science advanced, and desire that the broadest possible use be made of their information. With a clear protocol, a sound ethical argument and appropriate requests to regulatory authorities, we believe that other studies should also be able to progress their objectives and may seek to use our experience and data to support their work.

Contact: bob.phillips@york.ac.uk

Notes

Poster 28

Risk factors for hospitalisation in children presenting with influenza/influenza-like illness in primary care: a prognostic systematic review

Peter Gill, Kay Wang, Helen Ashdown, Carl Heneghan, Anthony Harnden, Susan Mallett

Background: Influenza and influenza-like illness (ILI) create considerable burden on NHS resources each winter. 'At risk' children are more prone to influenza-related complications and hospitalisation than otherwise healthy children. The Department of Health currently lists several 'at risk' groups as broad clinical disease categories. However, this list is not child-specific and is based on expert opinion rather than evidence. This systematic review aims to identify risk factors and evaluate existing prognostic risk scores for influenza-related hospitalisation in children who present with influenza/ILI in primary care.

Methods: We performed an electronic literature search of Medline, EMBASE, Web of Knowledge and CINAHL (June 2012). We identified additional articles by searching 'related articles' in PubMed, reviewing the reference lists of included studies and relevant narrative and systematic reviews, and by snowballing. We did not apply any language restrictions to our search. We included cohort and case-control studies which reported data on risk factors and/or prognostic models for influenza/ILI-related complications in children up to 18 years of age who presented in primary care. One review author (PG) screened the titles of articles retrieved by our search to exclude any obviously irrelevant studies. Two review authors (PG and KW) independently screened the abstracts of the remaining articles and assessed full-text versions of articles included after abstract screening. The authors of articles whose study design suggested that appropriate data had been collected but not reported were contacted for these data. Two authors (PG and HA) assessed the risk of bias using a modified QUADAS-2 form and extracted data from included studies. Odds ratios with 95% confidence intervals were calculated for risk factors reported in the included studies.

Results: Of 9,444 titles, 2,296 abstracts were screened; 112 full text articles were evaluated and 9 articles included. We contacted 31 authors for additional information, of which 16 responded. This enabled us to increase our number of included articles to 24. Most studies were conducted in Europe and focused on pandemic influenza. The number of children defined as being 'at risk' ranged from 3% to 82%. The most common risk factors studied
include respiratory, cardiac, neurological, renal and immunological conditions. Obesity, reported as a new risk factor in adults, was evaluated as a risk factor in children in three studies. The number of children hospitalised for influenza/ILI ranged from 9% to 62%. Results of meta-analysis for risk factors predicting hospitalisation due to influenza/ILI will be presented.

Conclusions: This prognostic systematic review will lead to a comprehensive evidence-based definition of which children are at risk of influenza-related hospitalisation. The results will guide appropriate risk stratification and clinical management of children with influenza/ILI during influenza epidemics and pandemics.

Contact: kay.wang@phc.ox.ac.uk

Notes

Poster 29

Panel diagnosis as Reference Standard in diagnostic research: A systematic review and methodological recommendations

Loes Bertens, Berna Broekhuizen, Christiana Naakteboren, Frans Rutten, Arno Hoes, Yvonne van Mourik, Karel Moons, Johannes Reitsma

Background: In diagnostic studies, a single and error-free test to make the final diagnosis (reference standard) often does not exist. One solution is to use the results from different tests as the reference standard. In panel diagnosis, the various test results are assessed by multiple experts to reach a final diagnosis in each patient. Although panel diagnosis, also known as consensus diagnosis, is frequently used, guidance on preferred methodology is lacking. Therefore we conducted a systematic review to summarize the methodology currently applied and identify areas for improvement.

Methods and findings: PubMed was searched for diagnostic studies in which the final diagnosis was made by two or more experts based on results from multiple tests. General study characteristics and details of panel methodology were extracted. Eighty-one studies were included, of which most reported on psychiatry (37%) and cardiovascular (26%) diseases. Data extraction was hampered by incomplete reporting; one or more pieces of critical information about panel reference standard methodology was missing in 83% of studies. In most studies (65%), the panel consisted of three or less members. Panel members were not blinded to the results of the test under evaluation in 69% of the studies. Reproducibility of the decision process was assessed in only 21% of the studies. Choices in panel constitution, the number of test results and other information available to the panel, and methods of decision-making varied largely between studies.

Conclusions: Methods of panel diagnosis varied substantially across studies, but unfortunately many key features were insufficiently reported. Panel constitution, information available to the panel members and the methods of decision making were identified as the key items of panel reference standard methodology. Based on our review we developed a checklist and flow chart that can assist researchers in designing a diagnostic study in which panel diagnosis is used.

Contact: l.c.m.bertens-2@umcutrecht.nl

Notes

Poster 30

A Bayesian framework for estimating the incremental value of a diagnostic test in the absence of a gold standard reference

Nandini Dendukuri, Daphne Ling, Madhukar Pai

We will illustrate methods to estimate the incremental value of a new, imperfect test when the reference standard is also imperfect. Using a Bayesian approach we estimate the latent disease status via a latent class model and extend two commonly-used measures of incremental value based on predictive values [area under the ROC curve (AUC) and integrated discrimination improvement (IDI)] to the context where no gold standard exists. The methods are illustrated using simulated data and applied to the problem of estimating the incremental value of a novel interferon-gamma release assay (IGRA) over the tuberculin skin test (TST) for latent tuberculosis screening. We also show how to estimate the incremental value of IGRAs when decisions are based on observed results. We found that the incremental value statistics have the greatest magnitude when both sensitivity and specificity of the new test are better and that conditional dependence between the tests reduces the incremental value. The incremental value of the IGRA depends on the sensitivity and specificity of the TST and may thus vary in different populations. Even in the absence of a gold-standard reference, incremental value statistics may be estimated and can aid decisions about the practical value of a new test.

Contact: nandini.dendukuri@mcgill.ca

Notes

Poster 31

Evaluating Diagnostic Accuracy in the Face of Multiple Reference Standards

Christiana Naaktgeboren, Joris de Groot, Loes Bertens, Maarten van Smeden, Johannes Reitsma, Karel Moons

A common challenge in diagnostic studies is the lack of a single error-free ("gold") reference test to which the test under study (the index test) can be compared. When a reference test does not perfectly correspond to disease status, index test accuracy estimates can be biased. One method for dealing with the lack of a single perfect reference test is to combine the results of several tests into one "composite reference standard".
The idea behind a composite reference standard is that the combination of several imperfect tests may provide a better perspective on disease than any individual tests. The key challenge of composite reference standards, however, is selecting the appropriate tests and determining the optimal rule for combining the test results.

There is a lack of consensus in the way the term composite reference standard is used and the reporting of results is generally poor. To address these problems, we provide a thorough explanation of the composite reference standard method, discuss advantages and disadvantages of the method, and make suggestions on how to report results.

Contact: c.naaktgeboren@umcutrecht.nl

Notes

Poster 32

Practical Methodological Strategies for Individual Patient Data (IPD) Systematic Reviews

Fay Crawford, Chantelle Anandan, Francesca Chappell, Jackie Price, Gordon Murray, Aziz Sheikh, Colin Simpson, Gerard Stansby, Matthew Young

Background and context: Diabetes–related lower limb amputations are associated with considerable morbidity and mortality and are usually preceded by foot ulceration. Annual assessment procedures are recommended to identify those people with diabetes who are at risk of foot ulceration and there is some evidence to support the use of certain diagnostic tests, symptoms, signs and elements from the patients' history but the role of other contributory factors is less clear. Current foot screening clinical guidelines are largely based on consensus and the findings from individual studies rather than any systematic integration of all available data. Systematic reviews to integrate evidence of predictive factors exist but are compromised because both adjusted and unadjusted estimates are found in the primary studies. As adjusted meta analyses of aggregate data can be challenging the best way to standardise the analytical approach is to use individual patient data (IPD).There are many challenges associated with this type of systematic review including its time-consuming and costly nature. During this presentation we will share the key methodological strategies which underpin our IPD systematic review of prognostic factors for foot ulceration in diabetes.

Aim:
- To systematically review individual patient data from cohort studies in a meta-analysis to estimate the predictive value of clinical characteristics and diagnostic tests for diabetic foot ulceration (DFU).
- To develop a prognostic model of the risk factors for DFU based on data collected worldwide.
- To test the robustness of the model in different demographic profiles - for example, age, duration of diabetes, control of diabetes (insulin, diet or oral medication), type of diabetes Type I, Type II.

Methods: MEDLINE and EMBASE databases from inception until 2012 were searched and the corresponding authors of eligible primary studies were contacted and invited to contribute the data to an IPD analysis. The review protocol can be found at PROSPERO;
http://www.crd.york.ac.uk/PROSPERO

Results: Fourteen eligible cohort studies involving more than 20,000 patients worldwide were identified and all corresponding authors have given signed agreements to share their data. These datasets share common variables from which meta analyses of predictive factors for foot ulceration in diabetes are feasible. The key methodological strategies which have positively contributed to our review are: Prior conduct and publication of an aggregate systematic review allowed us to develop search strategies and communicate with authors about their primary studies and resultant data; the absence of industry involvement meant that authors were in possession of their dataset; the on-going involvement of the authors in the field of diabetes foot research sustained their interest in the review question; a detailed protocol including items for an assessment of quality and ethical considerations helped to establish a "contract" between the collaborators. Importantly, the project has received funding to support research activity and International collaborations.

Interpretation: Practical methodological strategies have positively contributed to the conduct of this IPD systematic review which is central to the development of a global evidence-based strategy for the risk assessment of the diabetic foot. IPD systematic reviewers should carefully consider practical influences in the preparatory stages of their work.

Contact: fay.crawford@nuth.nhs.uk

Notes

Poster 33

Application of the GRADE Methodology to Cochrane Diagnostic Test Accuracy Reviews

Gowri Gopalakrishna, Reem Mustafa, Clare Davenport, Rob Scholten, Chris Hyde, Jan Brozek, Holger Schünemann, Miranda Langendam, Mariska Leeflang, Patrick Bossuyt

Background: The GRADE criteria for rating evidence can also be used for evaluating the results of diagnostic accuracy studies, but experience is limited. We applied the five GRADE domains to published Cochrane diagnostic test accuracy reviews (DTAR) with the aim of better understanding the application of these domains to diagnostic test accuracy reviews.

Methods: We selected three DTARs based on diversity of clinical areas and methodological issues. At least 3 reviewers with expertise in the GRADE approach and/or in diagnostic test accuracy studies independently rated the evidence in each review according to the five "GRADE domains". Reviewers strived to explain each judgment made by documenting all considerations. Two teleconferences were held to exchange experiences.

Results: Some reviewers assessed the evidence from a patient important outcome perspective while others assessed the quality of the evidence from an accuracy standpoint. Having a clear key question before starting the grading process was particularly pertinent in DTARs that compared multiple index tests and different patient spectrums. There was no consensus on the criteria and thresholds to use when assessing the GRADE domains "inconsistency", "imprecision" and "publication bias". In the comparative test accuracy review, reviewers assessed the quality of evidence of each index test against the reference standard before making an indirect comparison of the two tests.

Discussion and Conclusions: The perspective from which the evidence is graded can influence the judgment of the quality of evidence. Worked examples illustrating the application of GRADE domains: "inconsistency", "imprecision" and "publication bias" would facilitate the operationalization of GRADE for diagnostics. Explicit guidance on how to rate the quality of evidence for a comparative test review is needed.

Contact: g.gopalakrishna@amc.uva.nl

Notes

Poster 34

Realist synthesis – what is it and how could it add value to our understanding of the diagnosis of dementia in primary care?

Harriet Hunt, Mark Pearson, Chris Hyde

In 2012, the United Kingdom Government launched the Dementia Challenge aiming to make major improvements in dementia care and research by 2015. Primary care has been identified as a key area of weakness in dementia diagnosis, with 68% of people in a recent survey waiting longer than a year between noticing symptoms and getting a diagnosis of dementia (Alzheimer's Society 2012). This is not only a national issue; globally, it has been predicted that most people with dementia have not received a formal diagnosis. At presentation, there are many difficulties for primary care practitioners, patients and carers in understanding the diagnostic information presented and translating the diagnosis into meaningful prognosis.

Within primary care, conventional diagnostic tools such as 2x2 tables may be used less often by general practitioners than heuristics-driven clinical decision-making with limited purposive patient input. Recently, Patrick Bossuyt and colleagues have coined the term "clinical utility" to quantify the extent to which diagnostic testing improves health outcomes relative to the current best alternative, and this enables a broader evaluation of a diagnostic test beyond technical performance and accuracy. However, this assessment is limited to the diagnostic test rather than the broader diagnostic process. Issues around disclosure, over-diagnosis, presymptomatic presentation, and the rhetoric of therapeutic interaction combine to make dementia diagnosis a multifaceted and thorny area for study, yet in order to improve rates of dementia diagnosis and develop prognosis it is an area which needs critical investigation.

To investigate how the needs and expectations of GPs, patients and their carers match up to the current reality of diagnostic and prognostic information available, Realist synthesis allows analysis of what diagnostic approaches work, who these approaches benefit and where these approaches are most effective. As far as we are aware, this method has not been applied to the fields of clinical diagnosis and prognosis yet offers the potential for new insights into dementia diagnosis. Realist synthesis is a systematic approach which can provide more richly-contextualised findings than a conventional systematic review of the current evidence. More than this, as 'policy-friendly' approaches to evidence synthesis are increasingly sought after, the Realist method is suited to investigating the complexities of modern health service delivery using a broad evidence base. This presentation aims to introduce the technique of Realist synthesis and demonstrate the features that might particularly add value to existing evaluations. We plan to cover the following:

- The current challenge of dementia diagnosis and prognosis within primary care
- GP need and patient expectation: "useability" versus availability
- Why Realist synthesis may be suitable for addressing the needs gap
- Discussion

Reference:

Alzheimer's Society (2012) Dementia 2012: A national challenge. alzheimers.org.uk/dementia2012

Contact: h.a.hunt@exeter.ac.uk

Notes

Poster 35

A model-based economic evaluation of test-treatment strategies: the cost-effectiveness of strategies to identify individuals with monogenic diabetes

Jaime Peters, Rob Anderson, Chris Hyde

Background: Economic evaluations of diagnostic strategies are challenging when many realistic test-treatment strategies and care pathways are to be considered. A clinical study is rarely able to assess all realistic comparisons and collect the totality of evidence required, and so model-based economic evaluations have great value to inform decisions on the cost-effectiveness of diagnostic strategies. We present the development and results of an economic evaluation of test-treatment strategies to identify and change current treatment for patients with monogenic diabetes. Patients with monogenic diabetes are often misdiagnosed as having type 1 or type 2 diabetes, commonly receiving more invasive and costly treatment than is necessary. Identifying these individuals and changing their treatment has the potential to be a cost-effective use of NHS resources.

Methods: Based on expert opinion and a systematic appraisal of existing long term diabetes models, a hybrid model to evaluate 5 test-treatment strategies was developed. The strategies included no genetic testing, a representation of current practice, use of clinical prediction models and biochemical tests to target genetic testing, and blanket genetic testing of all individuals diagnosed with diabetes <30 years old. Due to the rarity of monogenic diabetes, the clinical effectiveness and quality of life data, as measured in the clinical study, are based on a small number of individuals. Using the model-based economic evaluation, investigation of various assumptions and scenarios was undertaken, even in the presence of limited data availability.

Results: Preliminary results suggest that, when compared to the no testing strategy, targeted testing of some form might be considered cost-effective according to the threshold used by the National Institute for Health and Clinical Excellence (<£20-£30,000 per quality-adjusted life year). However, this depends on the clinical benefits of a change in treatment and their long-term impacts, and this is where a great deal of uncertainty exists. Results of the economic evaluation will be presented, demonstrating the ability of the decision model to explore various assumptions and uncertainties.

Discussion: To conduct economic evaluations of testing strategies, clinical studies are rarely able to assess all realistic strategies. Model-based economic evaluations alongside clinical studies allow a framework for investigation of the cost-effectiveness of multiple strategies, even if the data are lacking. Thus, model-based economic evaluations are powerful tools to aid decisions on the cost-effectiveness of testing strategies.

Contact: j.peters@exeter.ac.uk

Notes

Poster 36

Modelling the long-term cost-effectiveness of using different MRI techniques to localise prostate abnormalities for biopsy in patients with a previous negative biopsy

Graham Scotland, Emma Tassie, Graham Mowatt, Charles Boachie, Moira Cruikshank, John Ford, Cynthia Fraser, Lufti Kurban, Thomas B Lam, Anwar R Padhani, Justine Royle

Background: The diagnosis of prostate cancer (PCa), the commonest cancer affecting men, is based on prostatic biopsy. However, some men with a negative biopsy remain at risk of harbouring PCa, and the most effective and cost-effective way of managing these patients remains uncertain.

Objectives: To assess the cost-effectiveness of using T2-weighted magnetic resonance imaging (MRI) and enhanced MRI techniques (eMRI) - including diffusion weighted (DW), dynamic contrast enhanced (DCE) and MRI spectroscopy (MRS) - to direct trans-rectal ultrasound-guided (TRUS) prostate biopsies in men with a previous negative biopsy. Methods: Estimates of the sensitivity and specificity of the alternative MRI techniques were obtained form a systematic review and indirect comparison meta-analysis. These were incorporated in a semi-Markov model simulating the progression of undiagnosed PCa and the impact of diagnosis and treatment on health service costs, survival and quality adjusted life years (QALYs). The MRI/eMRI techniques were compared with TRUS alone to guide biopsies. The model incorporated the costs and consequences (beneficial and adverse) of the alternative diagnostic approaches and consequent future treatment, capturing a potential trade-off between diagnostic yield and over-diagnosis/treatment. The model was analysed for cohorts defined by age and PCa prevalence, and the net-benefit approach was applied to interpret cost-effectiveness findings, using a ceiling ratio of £30,000 per QALY to identify the optimal strategy. Deterministic and probabilistic sensitivity analyses were conducted to characterise the uncertainty surrounding estimated cost-effectiveness.

Results: Fifty-one studies were included in the review of diagnostic accuracy, involving over 10,000 men. In meta-analyses of the MRI/eMRI techniques, sensitivity (95% CI) was highest for MRS at 92% (86 to 95%), followed by T2-MRI at 86% (74 to 93%) and DCE-MRI at 79% (69 to 87%), while specificity (95% CI) was highest for MRS at 76% (61 to 87%). Only one small study involving 43 participants reported on DW-MRI, with sensitivity of 100% (specificity not reported). There was also limited evidence suggesting that, potentially, MRS and enhanced MRI techniques might have greater sensitivity for clinically significant disease (i.e. Gleason 7 and above). T2-MRI was generally cost-effective in comparison with TRUS alone, though QALY gains were negligible. The cost-effectiveness of MRS/eMRI was highly sensitive to the prevalence of PCa, cohort age, approach to utility weighting, severity of disease, risk of progression, and ability to discriminate between high and low risk PCa.

Discussion/conclusions: Models evaluating the use of alternative diagnostic tests within on-going patient monitoring pathways can yield very small QALY differences between strategies. The evaluation of diagnostic strategies for prostate cancer is complicated by concerns about over-diagnosis and uncertainty surrounding the benefits of radical treatment for patients with low to moderate risk disease. If eMRI techniques can be shown to have high sensitivity for high risk cancer, while negating the need for biopsy in patients with no/low risk disease, they could offer a cost-effective approach to diagnosis. Scope also exists for future studies to elicit and apply patient/public preferences for process of care or informational outcomes associated with alternative diagnostic pathways, to better enable judgment on the relative value of these.

Contact: g.scotland@abdn.ac.uk

Notes

Poster 37

Upfront MRI followed by MRI-targeted biopsy for men with suspected prostate cancer: a decision analysis

Sarah Willis, Jan van der Meulen, Hashim Ahmed, Yipeng Hu, Caroline Moore, Ian Donaldson, Mark Emberton

Introduction: Men with localised prostate cancer have no significant survival benefit when undergoing radical surgery (Wilt, 2012). Focussing attention on better selecting patients for such radical treatment may therefore be the key to improving patient outcomes.

One strategy would be to image all men with MRI and carry out an MRI-targeted biopsy (MRI-TB) in those with a lesion on MRI. We carried out a simple decision analysis to estimate how many biopsies could be avoided and the proportion of patients correctly identified with this strategy, compared to the current standard of care; transrectal ultrasound-guided (TRUS) biopsy in all men.

Methods: We drew together the best available evidence from different sources - in a format that is relevant for decision making. Outcomes were evaluated for an un-screened and biopsy-naive population, using a conservative definition of clinically significant disease (a cancer core length of 3mm) and assuming a prevalence of 50%. We investigated the impact of employing different definitions of disease in additional analyses.

Results: Under certain conditions, set out in our base case, our results imply the new diagnostic strategy may reduce the number of men undergoing biopsy by a quarter. The use of a triage MRI followed by MRI-targeted biopsy in a hypothetical cohort of 1000 men would identify 360 men correctly ('true positives') and 60 men erroneously as having prostate cancer ('false positives'). Corresponding figures for TRUS biopsy are 250 men identified correctly as having cancer and 100 identified erroneously.

Conclusion: Triage MRI followed by MRI-targeted biopsy is likely to correctly identify more men with cancer and correctly identify more men without disease than TRUS, whilst at the same time avoiding biopsy in a quarter of men. A more efficient identification of men who might benefit from treatment may hold the key to improving longer-term health outcomes, although ultimately the optimal strategy will be based on the impact on costs and quality-adjusted life expectancy.

Key Reference:

Wilt TJ, Brawer MK, Jones KM, Barry MJ, Aronson WJ, Fox S, et al. (2012) Radical prostatectomy versus observation for localized prostate cancer. N Engl J Med. Jul 19;367(3):203-13.

Contact: sarah.willis@lshtm.ac.uk

Notes

This page is intentionally blank for notes

This page is intentionally blank for notes

This page is intentionally blank for notes

Conference Dinner Information

Tickets are included in the Registration Fee

This year's dinner will take place at 'The Jam House', which was opened by Jools Holland in 1999. Renowned for its live music and relaxed atmosphere, The Jam House occupies a grand Georgian building located in the heart of Birmingham's historic Jewellery Quarter:

The Jam House
3-5 St Pauls Square
Birmingham, B31QU