UMC Utrecht
Julius Center

Systematic review of multivariable prediction studies: an individual participant data meta-analysis approach

Thomas Debray, Hans Reitsma, Karel Moons, Richard Riley

for the Cochrane Prognosis Review Methods Group
(Co-convenors: Doug Altman, Katrina Williams, Jill Hayden, Sue Woolfenden, Richard Riley, Karel Moons)

UNIVERSITY OF BIRMINGHAM

The Cochrane Collaboration
Working together to provide the best evidence for health care

University Medical Center Utrecht

Conflict of interest

We have no actual or potential conflict of interest in relation to this presentation

Workshop objectives

Provide guidance to conduct individual participant data (IPD) meta-analysis in prediction research

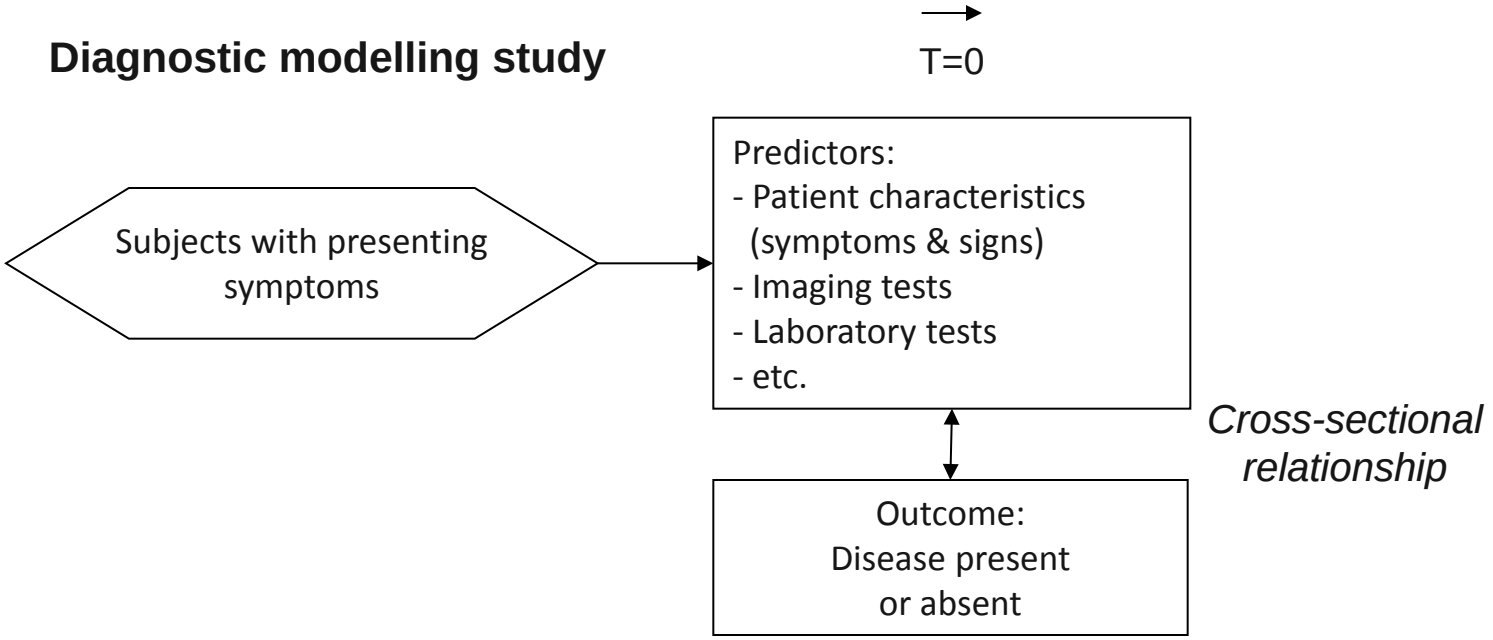
- To explain prediction research
- To describe potential benefits of IPD
- To identify challenges for IPD reviews
- To provide examples of IPD meta-analyses
- To describe appropriate methods
- To illustrate novel methods using real-life case studies

Prediction

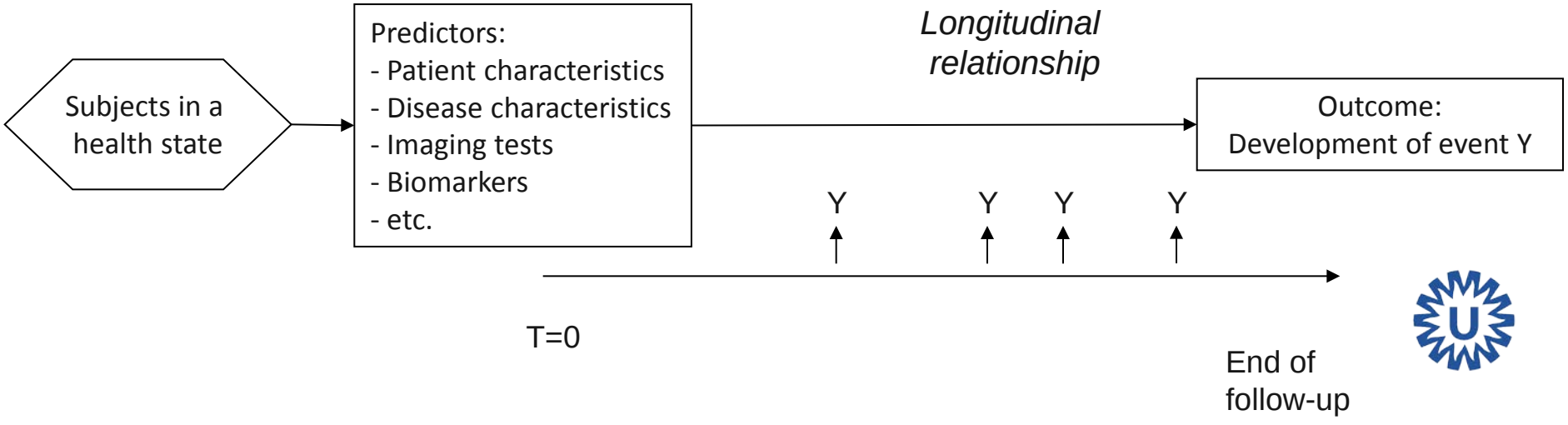
- Risk prediction = foreseeing / foretelling
  … (probability) of something that is yet unknown

- Turn available information (predictors) into a statement about the probability:
  … of having a particular disease -> diagnosis
  … of developing a particular event -> prognosis

Diagnostic modelling study



Prognostic modelling study



Prognosis BMJ series 2009 (Altman, Moons, Royston, Vergouwe)

- Prognosis studies: Examining future outcomes in subjects with a certain health condition in relation to demographic, disease and subject characteristics
  - not necessarily sick people

- Use of prognostic information:
  - to inform patients and their families
  - to guide treatment and other clinical decisions
  - to create risk groups for stratifying severity in clinical studies
  - insight in disease > clues for aetiology and new therapies

Main types of prognosis studies

PROGRESS series 2013: BMJ and Plos Med

Aim of prognostic studies may be:

- Average/overall prognosis: 'What is the most likely course (outcome) of people with this health condition?'
- Prognostic factors: 'What factors are associated with that outcome of interest?
- Prognostic (prediction) models: 'What is the absolute risk in individual subjects, based on multiple risk factors?'
- Model validation: 'What is the best model or how good is a model in particular setting?'

Focus this workshop: IPD-MA of prediction model studies

Prediction in Diagnosis

- Diagnostic studies: Examine the relationship of test results in relation whether a particular condition is present or absent.
  - patients suspected for the condition of interest or screening
  - cross-sectional relationship (here and now)
  - tests can include demographic, signs & symptoms, lab, imaging, etc
- Use of diagnostic information:
  - to start or refrain from treatment
  - further testing

Main types of diagnostic studies

- Technical evaluation studies
- Single test or comparative accuracy evaluation studies
- Multivariable diagnostic prediction models

Focus this workshop: IPD-MA of multivariable prediction studies

Prediction models

Predictors (in both diagnostic & prognostic models) are from:

- history taking
- physical examination
- tests (imaging, ECG, biomarkers, genetic ‘markers’)
- disease severity
- therapies received

Prediction models

Presented as:

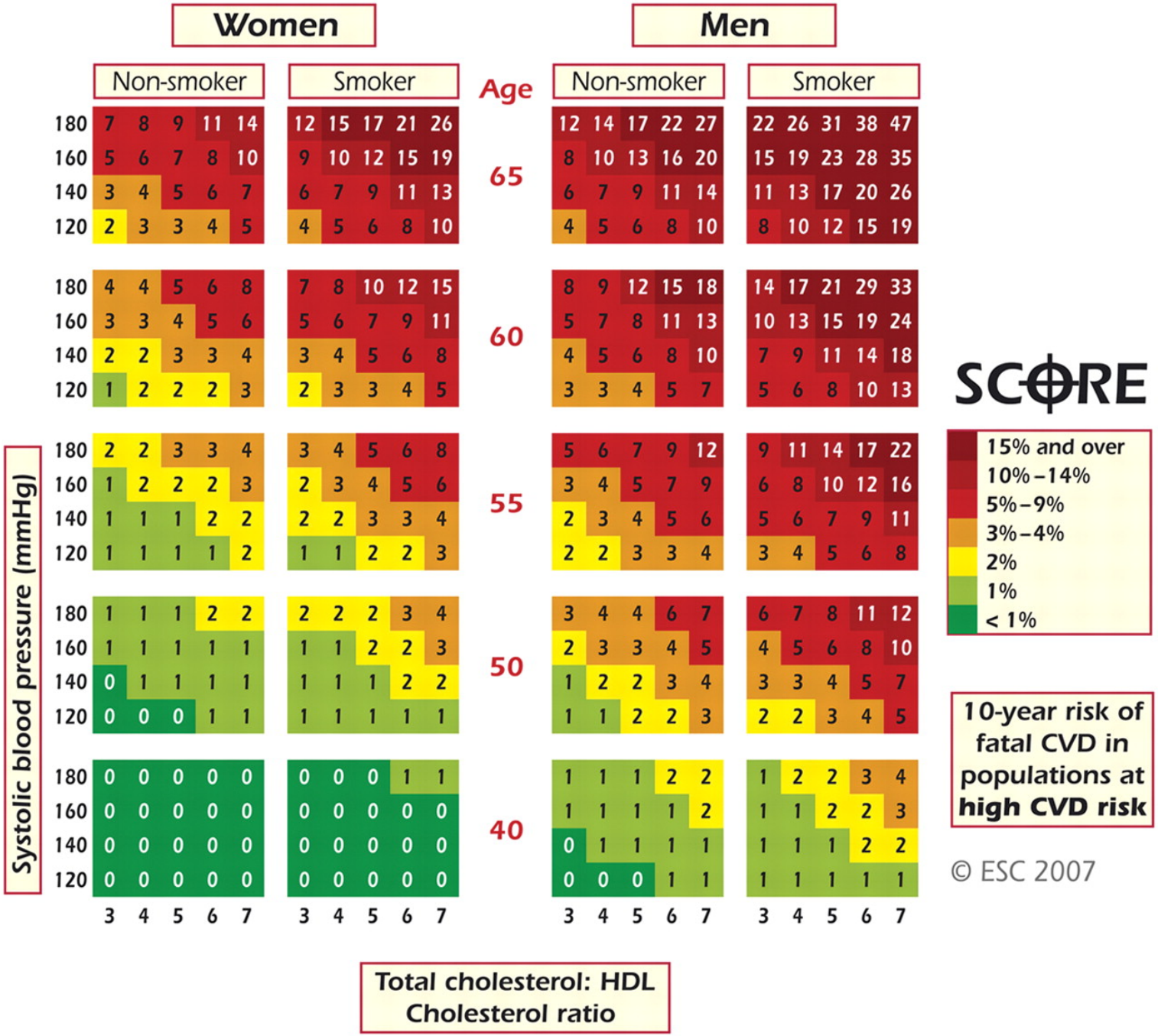
- Mathematical formula requiring computer
- Simple scoring rules
- Score charts / Nomograms

Apgar score in neonates (JAMA 1958)

Table 9–1. Apgar scoring.

Signs	0	1	2
Heartbeat per minute	Absent	Slow (<100)	Over 100
Respiratory effort	Absent	Slow, irregular	Good, crying
Muscle tone	Limp	Some flexion of *extremities*	Active *motion*
Reflex irritability	No response	Grimace	Cry or cough
Color	Blue or pale	Body pink, extremities blue	Completely pink

Σ = Apgar score (0-10)


Women
Men
Non-smoker
Smoker
Non-smoker
Smoker
Age
65
60
55
50
40
Systolic blood pressure (mmHg)
180 7 8 9 11 14 | 12 15 17 21 26 | 12 14 17 22 27 | 22 26 31 38 47
160 5 6 7 8 10 | 9 10 12 15 19 | 8 10 13 16 20 | 15 19 23 28 35
140 3 4 5 6 7 | 6 7 9 11 13 | 6 7 9 11 14 | 11 13 17 20 26
120 2 3 3 4 5 | 4 5 6 8 10 | 4 5 6 8 10 | 8 10 12 15 19
180 4 4 5 6 8 | 7 8 10 12 15 | 8 9 12 15 18 | 14 17 21 29 33
160 3 3 4 5 6 | 5 6 7 9 11 | 5 7 8 11 13 | 10 13 15 19 24
140 2 2 3 3 4 | 3 4 5 6 8 | 4 5 6 8 10 | 7 9 11 14 18
120 1 2 2 2 3 | 2 3 3 4 5 | 3 3 4 5 7 | 5 6 8 10 13
180 2 2 3 3 4 | 3 4 5 6 8 | 5 6 7 9 12 | 9 11 14 17 22
160 1 2 2 2 3 | 2 3 4 5 6 | 3 4 5 7 9 | 6 8 10 12 16
140 1 1 1 2 2 | 2 2 3 3 4 | 2 3 4 5 6 | 5 6 7 9 11
120 1 1 1 1 2 | 1 1 2 2 3 | 2 2 3 3 4 | 3 4 5 6 8
180 1 1 1 2 2 | 2 2 2 3 4 | 3 4 4 6 7 | 6 7 8 11 12
160 1 1 1 1 1 | 1 1 2 2 3 | 2 3 3 4 5 | 4 5 6 8 10
140 0 1 1 1 1 | 1 1 1 2 2 | 1 2 2 3 4 | 3 3 4 5 7
120 0 0 0 1 1 | 1 1 1 1 1 | 1 1 2 2 3 | 2 2 3 4 5
180 0 0 0 0 0 | 0 0 0 1 1 | 1 1 1 2 2 | 1 2 2 3 4
160 0 0 0 0 0 | 0 0 0 0 0 | 1 1 1 1 2 | 1 1 2 2 3
140 0 0 0 0 0 | 0 0 0 0 0 | 0 1 1 1 1 | 1 1 1 2 2
120 0 0 0 0 0 | 0 0 0 0 0 | 0 0 0 1 1 | 1 1 1 1 1
3 4 5 6 7 | 3 4 5 6 7 | 3 4 5 6 7 | 3 4 5 6 7
Total cholesterol: HDL Cholesterol ratio
SCORE
15% and over
10%–14%
5%–9%
3%–4%
2%
1%
< 1%
10-year risk of fatal CVD in populations at high CVD risk
© ESC 2007

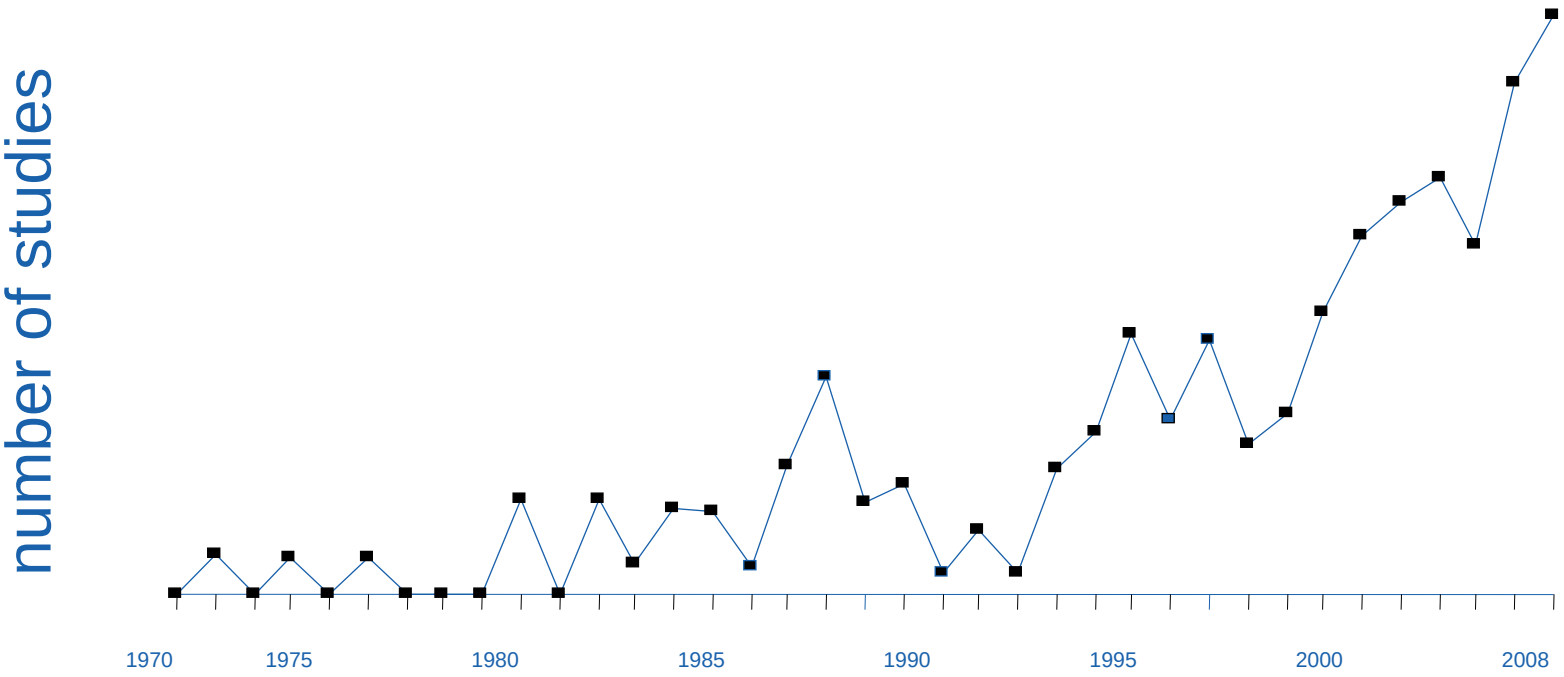
Predicting bacterial cause in infectious conjunctivitis

Indicator	Odds ratio (95% CI)	Regression coefficient	Clinical score*
Two glued eyes	14.99 (4.36 to 51.53)	2.707	5
One glued eye	2.96 (1.03 to 8.51)	1.086	2
Itching	0.54 (0.26 to 1.12)	–0.61	–1
History of conjunctivitis	0.31 (0.10 to 0.96)	–1.161	–2
Area under ROC curve (95% CI)	0.74 (0.65 to 0.82)	–	–

ROC=receiver operating characteristics.
*Clinical scores of every symptom present are added up. For example, a patient with two glued eyes, itch, and no history of conjunctivitis has a clinical score of: 5 + –1 = 4.

Rietveld et al. BMJ 2004;329:206

Why focus on prognostic prediction models? (Steyerberg 2009)



Four phases of prediction modelling

BMJ series 2009 (Altman, Moons, Royston, Vergouwe)

1. Developing a prediction model
2. Validate the model in other subjects
3. Update existing model to local situation
4. Quantify model's impact on doctor's decision making and patient outcome (cost-effectiveness)

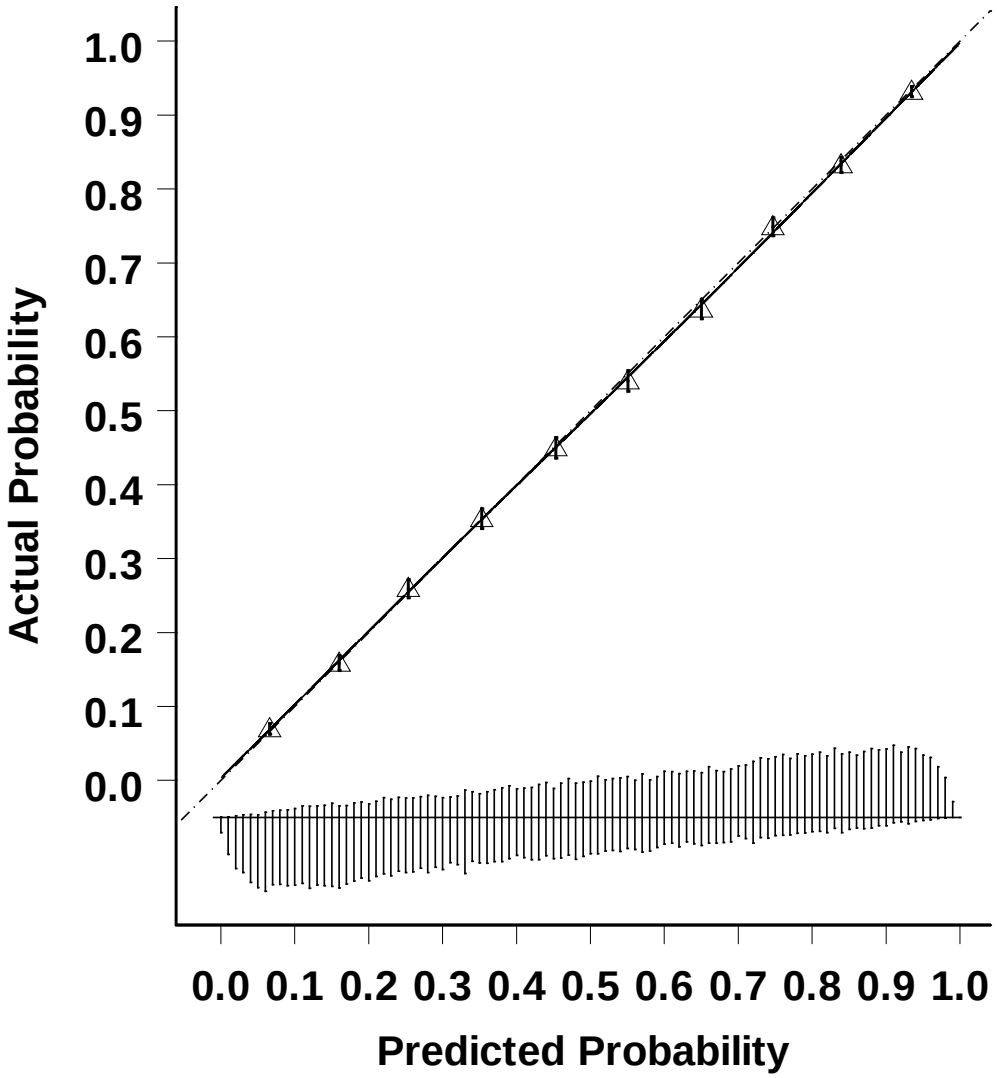
What is big difference between 4 versus 1-3?

Focus on 1-3

Prediction model performance measures

- **Calibration** plot
  (for specific time point in case of survival models)
- **Discrimination**
  - C-statistic (ROC area for logistic regression)
- **(Re)classification** → requires probability thresholds
  - Assess the potential effect on patient-level outcomes
  - Comparative test accuracy studies
  - Examples: Net Reclassifiation Index, Net Benefit, …

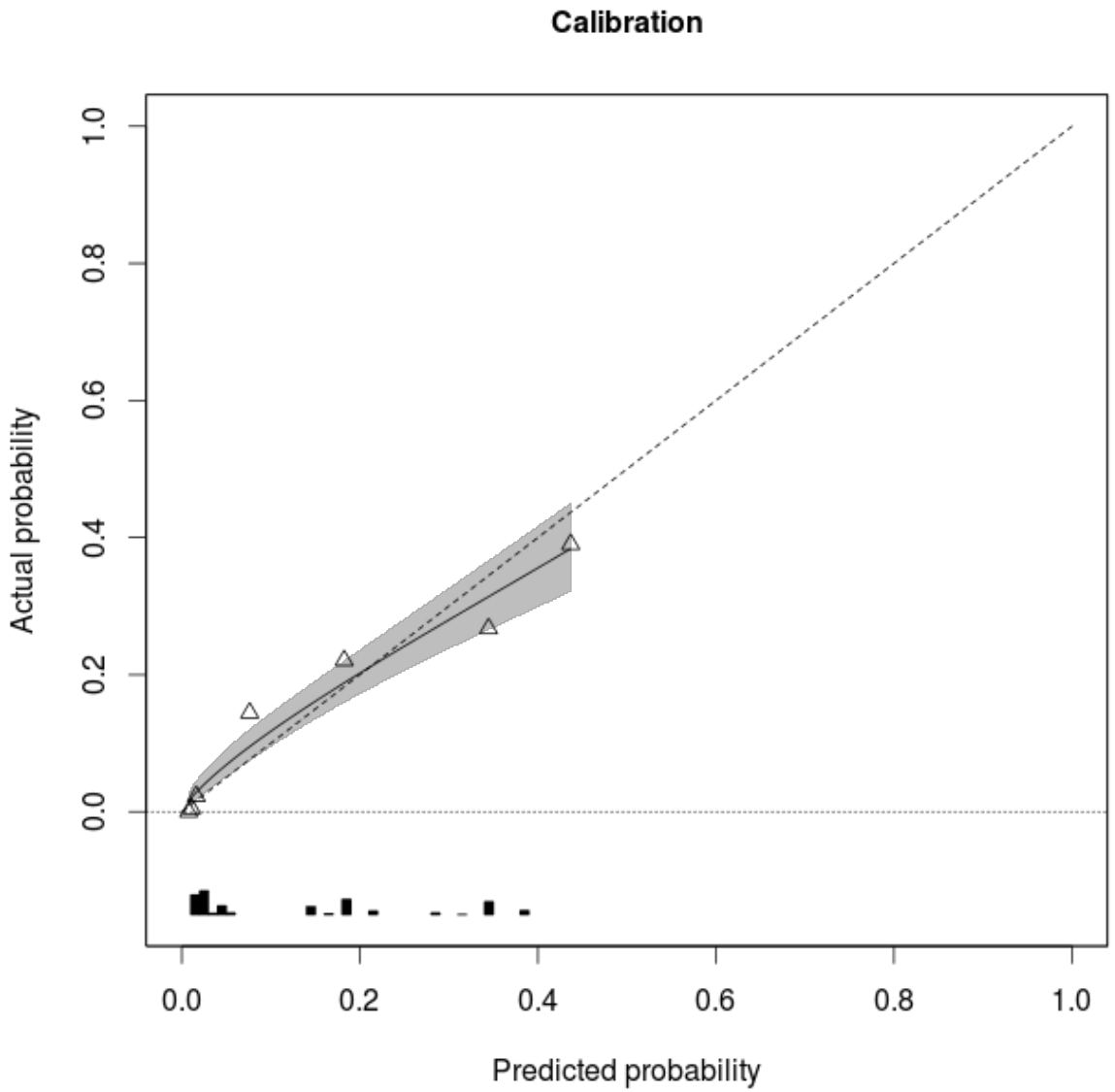
Calibration plot



Ideal calibration
Observed versus expected risk (O/E) = 1

Slope = 1

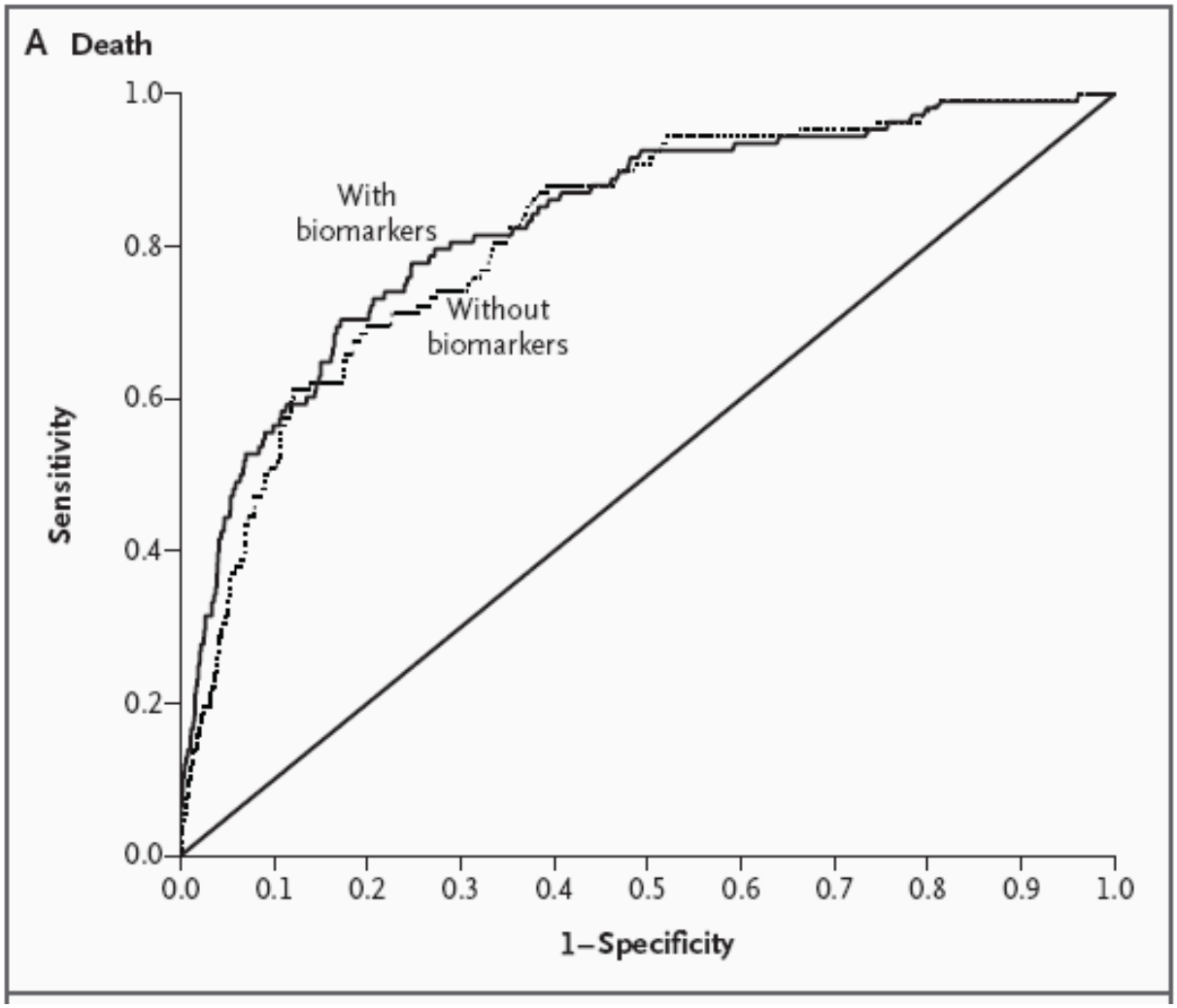
Calibration plot



O:E = 1
Slope = 0.79

Sub-obtimal slope because curve does not follow reference line

Model to predict cardiovascular morbidity/mortality



AUC 0.76

AUC 0.77

Wang TJ, et al. NEJM

What are the main differences between prediction and intervention research?

Intervention research	Prediction research
Aim: Estimate (relative) effects of a specific treatment, across different populations or subgroups	**Aim**: Estimate absolute risk probabilities for distinct individuals across different populations or subgroups
Typical design: Randomized Clinical Trials	**Typical design**: observational studies (e.g. cohort study), RCTs, …
Evaluation: bias and precision of estimated comparative treatment effects	**Evaluation**: model discrimination and calibration

Pitfalls of prediction research

- The **quality** of much prognosis research is poor (incomplete reporting, poor data sharing, incomplete registrations, absent study protocols)
- Development dataset often **too small or too local**
- Most prediction models are never validated in independent data (**external validation**)
- **Heterogeneity** across studies and settings, requiring local adjustments
- Many prediction models **generalize poorly** across different but related study populations, and tend to perform more poorly than anticipated when applied in routine care

Overcoming the problems of heterogeneity and poor reporting

- **Collaboration** of research groups required to seek **consistency** in cut-offs, adjustment factors, outcomes, analysis, measurement methods, etc.
- Improve study design standards -> more **protocol** driven, rather than additional post-hoc analyses of data 'on the shelf'
- Promote better reporting: REMARK and TRIPOD
- Collaborate across research groups to **pool existing IPD** and conduct **IPD meta-analysis**
- Design **large prospective studies** to answer pre-specified questions of clinical interest

Advantages over aggregate data (AD) meta-analysis

- Meta-analysis of reported summary statistics already implemented to …
  - Summarize the performance of an existing model
  - Summarize the (adjusted) association between a marker and outcome of interest
  - Combine existing prediction models
  - **See other workshop**! (Friday)
- AD has limited capabilities to …
  - Combine statistics of interest (e.g. due to variations in modeling approaches and reporting)
  - Account for between-study heterogeneity
  - Investigate modifiers of model performance

The benefit of having IPD from each study

IPD would overcome poor reporting and differences in data analysis approaches by allowing:

- Data checking
- Consistent statistical analysis in each study
- Verification of model assumptions
- Calculation of estimates of interest
- Proper handling of continuous variables

The benefit of having IPD from each study

IPD would limit heterogeneity in

- Type of estimates (adjusted/unadjusted)
- Type of association (dichotomized/linear/nonlinear)
- Type of outcome
- Adjustment factors

The benefit of having IPD from each study

IPD from multiple studies facilitates

- Model development studies
  - Investigation of more complex associations (e.g. nonlinearity, interaction and time-varying effects)
  - Identify added value of novel markers
  - Development and direct validation of models
- Multiple validations of existing prediction model(s)
  - To identify boundaries of model generalizability
  - To investigate differences in model performance across study populations

IPD – are we realistic?

- Researchers **protective** over their own data
- Worried about Data Protection Act (**ethics**) – however, no need to include patient ID numbers
- **Cost, time** – when does it become worthwhile?

To conduct better prognostic & diagnostic research we need:

- To be prepared to **collaborate** and share data to make IPD available – in paper, on Web, on request
- To be involved in **prospectively planned** pooled analyses

Reasons to be optimistic

- **IPD can be obtained**, although may be a long process
  - Meta-analyses have been facilitated when IPD was available, e.g. in determining a consistent cut-off level (*Sakamoto et al 1996, Look et al 2003*)
- A review identified **383 IPD meta-analyses** (1991-2009)
  - 48 IPD meta-analyses of prognostic factors

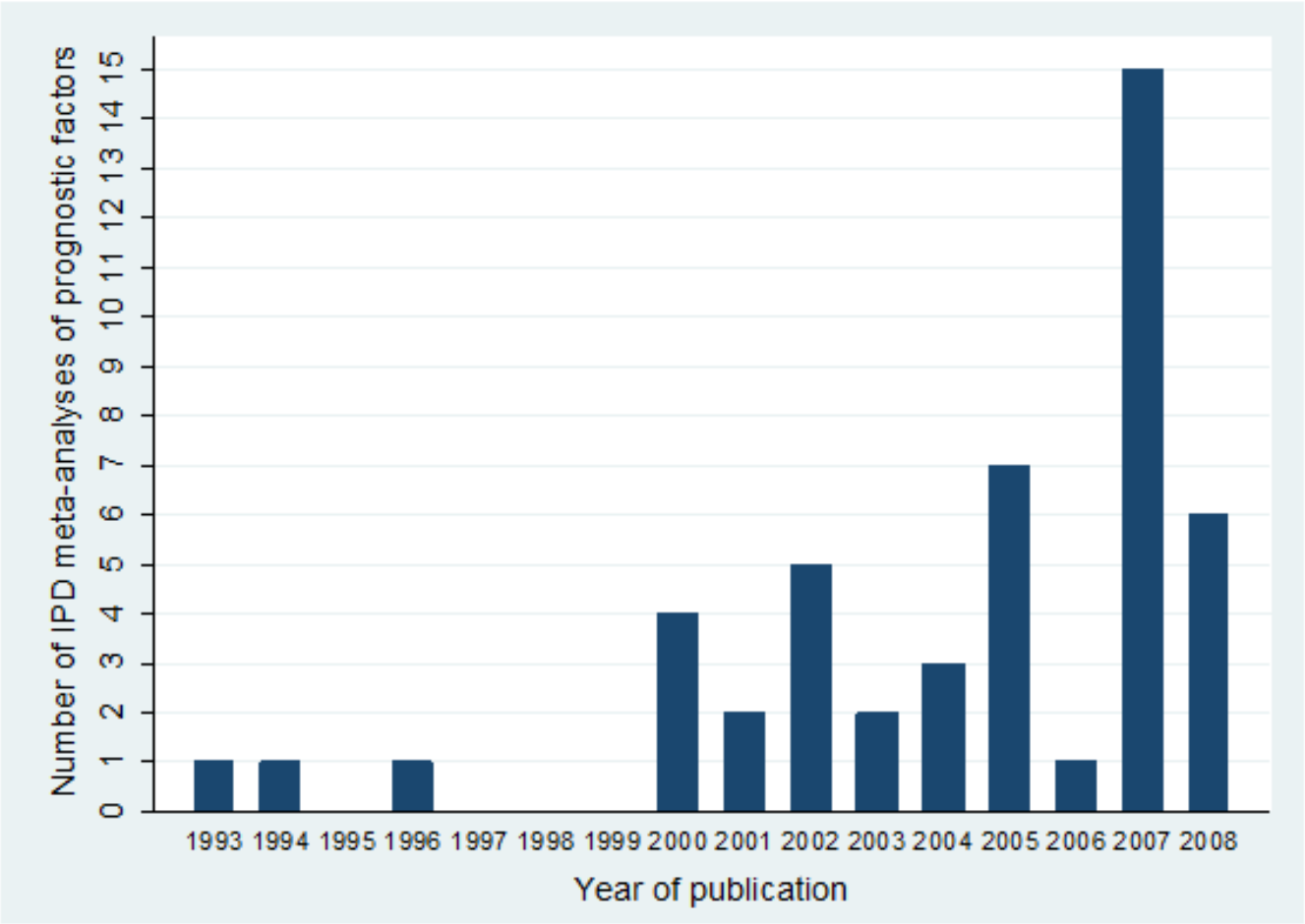
Abo-Zaid *et al. BMC Medical Research Methodology* 2012, **12**:56
http://www.biomedcentral.com/1471-2288/12/56

RESEARCH ARTICLE Open Access

Individual participant data meta-analysis of prognostic factor studies: *state of the art?*

Ghada Abo-Zaid[1], Willi Sauerbrei[2] and Richard D Riley[3*]

Reasons to be optimistic



Number of published IMPF articles over time; the spike in 2007 is due to eight articles from the IMPACT collaboration being published simultaneously.

Ref: Abo-Zaid *et al. BMC Medical Research Methodology* 2012 **12**:56 doi:10.1186/1471-2288-12-56

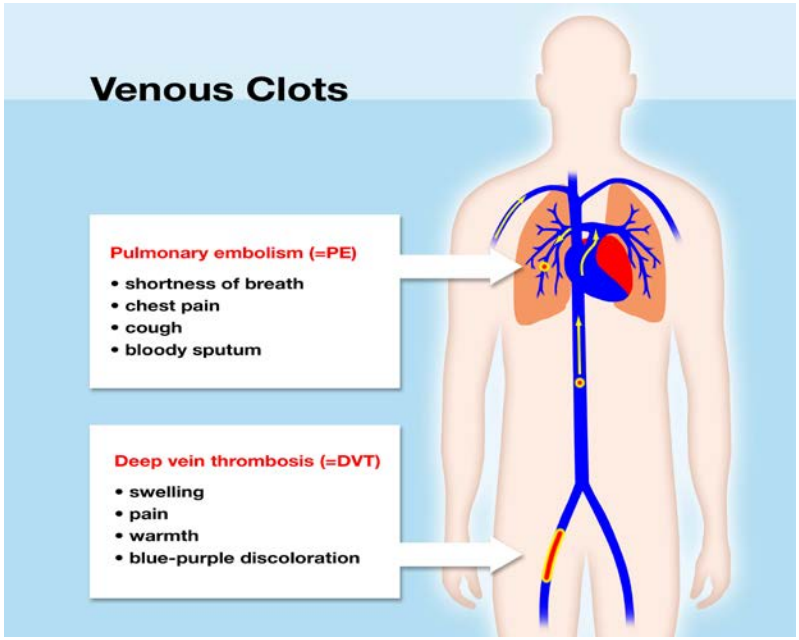
IPD-MA: what aims can be addressed in prediction research?

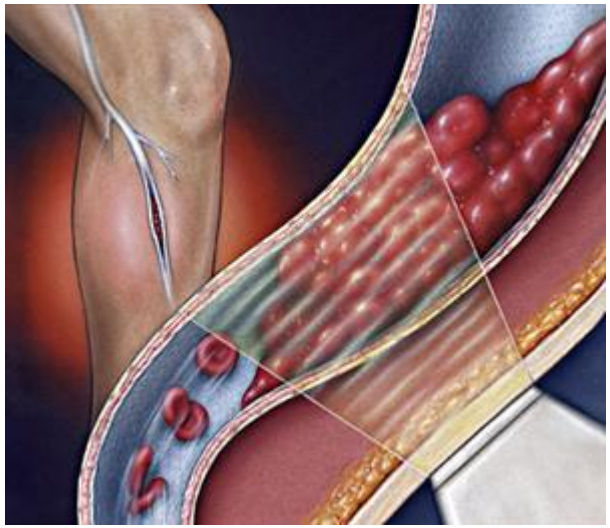
1. **Evaluate the performance of existing model(s)**
   - Which model yields better predictions, under what circumstances?
   - What performance can we expect in a certain study population or setting?
2. **Adjusting an existing model to local settings**
   - Does the model require changes before implementation? (e.g. adjustment for disease prevalence)
3. **Developing a novel prediction model**
   - How can we develop and directly validate a new prediction model?
   - What is the added value of a specific predictor or (bio)marker across different study populations?

Example #1: external validation of an existing prediction model

Diagnosis of deep vein thrombosis (DVT)

- Blood clot that forms in a vein in the body (lower leg/thigh)
- If blood clot breaks off -> blood stream -> lungs -> blockage
- Pulmonary embolism, preventing oxygenation of blood
- Potentially causing death





Example #1: external validation of an existing prediction model

Prediction model for **ruling out DVT in primary care**

- Patient history
- Physical examination
- D-dimer testing (biomarker)

Diagnostic variables	Odds ratio	Regression coefficient*	p-value	Points for the rule
Male gender	1.80 (1.36 – 2.16)	0.59	<0.001	1
Oral contraceptive use	2.12 (1.32 – 3.35)	0.75	0.002	1
Presence of malignancy	1.52 (1.05 – 2.44)	0.42	0.082	1
Recent surgery	1.46 (1.02 – 2.09)	0.38	0.044	1
Absence of leg trauma	1.82 (1.25 – 2.66)	0.60	0.002	1
Vein distension	1.62 (1.19 – 2.20)	0.48	0.002	1
Calf difference ≥ 3 cm	3.10 (2.36 – 4.06)	1.13	<0.001	2
D-dimer abnormal	20.3 (8.25 – 49.9)	3.01	<0.001	6
Constant		–5.47		

DVT= deep vein thrombosis; *=natural logarithm of the odds ratio; D-dimer abnormal for VIDAS ≥ 500 ng/ml and Tinaquant ≥ 400 ng/ml. Probability of DVT as estimated by the final model =1/(1+exp-(-5·47 + 0·59*male gender + 0·75*OC use + 0·42*presence of malignancy + 0·38*recent surgery + 0·60*absence of leg trauma + 0·48*vein distension + 1·13*calf difference ≥ 3cm + 3·01*abnormal D-dimer)).

Example #1: external validation of an existing prediction model

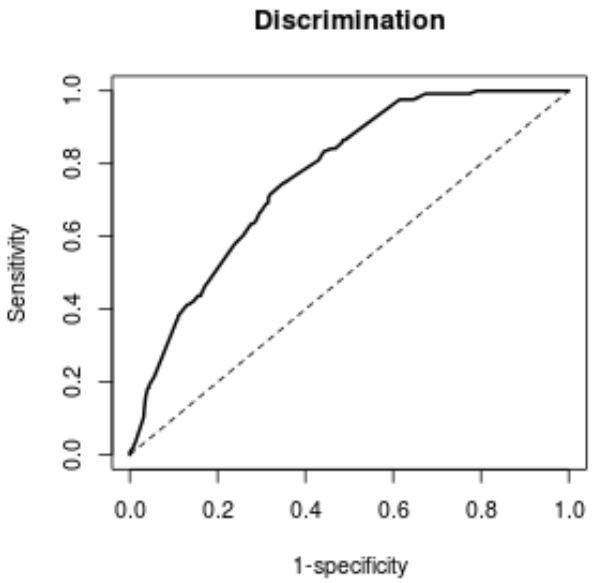
IPD meta-analysis

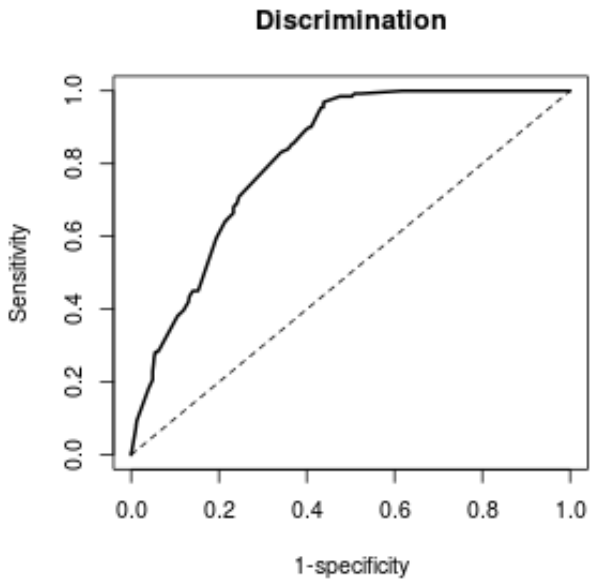
3 studies available for external validation

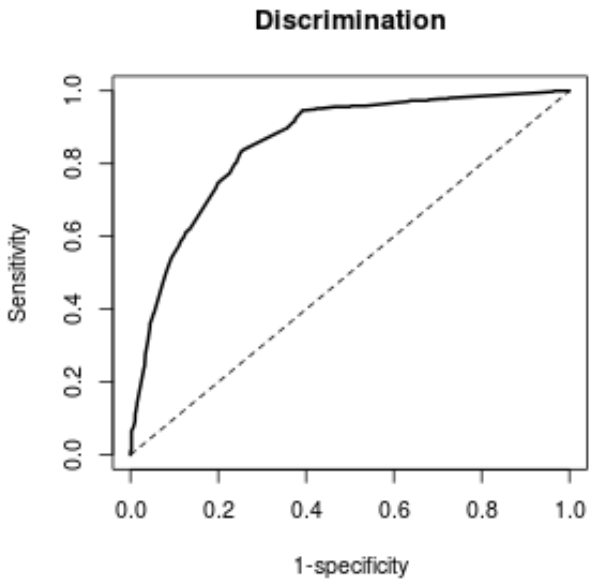
- N=791 (primary care)
- N=1028 (primary care)
- N=1756 (secondary care)

Example #1: external validation of an existing prediction model

ROC curves

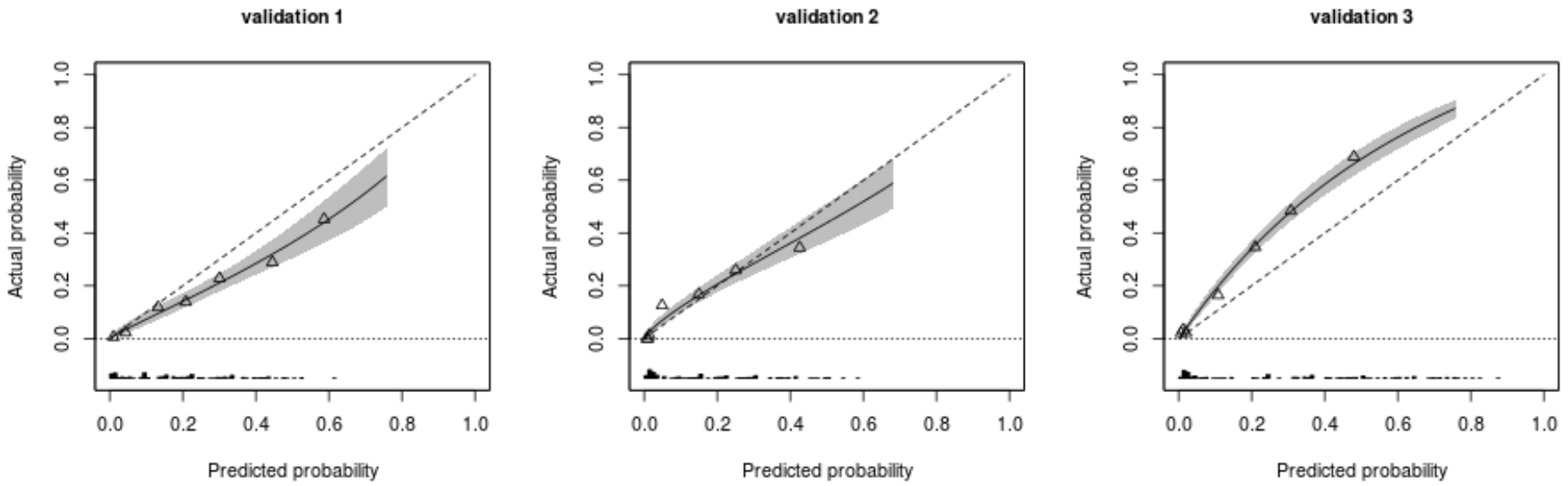






Example #1: external validation of an existing prediction model

Calibration plots



Example #1: external validation of an existing prediction model

Interpretation of model validation results

ELSEVIER

Journal of Clinical Epidemiology ■ (2014) ■

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

A new framework to enhance the interpretation of external validation studies of clinical prediction models

Thomas P.A. Debray[a,*], Yvonne Vergouwe[b], Hendrik Koffijberg[a], Daan Nieboer[b], Ewout W. Steyerberg[b,1], Karel G.M. Moons[a,1]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Str. 6.131, PO Box 85500, 3508GA Utrecht, The Netherlands*

[b]*Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands*

Accepted 30 June 2014; Published online xxxx

Example #2: Systematic review and external validation of existing prediction models

BMJ

BMJ 2012;345:e5900 doi: 10.1136/bmj.e5900 (Published 18 September 2012) Page 1 of 16

RESEARCH

Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study

Ali Abbasi *PhD fellow*[1 2 3], Linda M Peelen *assistant professor*[3], Eva Corpeleijn *assistant professor*[1], Yvonne T van der Schouw *professor of epidemiology of chronic diseases*[3], Ronald P Stolk *professor of clinical epidemiology*[1], Annemieke M W Spijkerman *research associate*[4], Daphne L van der A *research associate*[5], Karel G M Moons *professor of clinical epidemiology*[3], Gerjan Navis *professor of nephrology, internist-nephrologist*[2], Stephan J L Bakker *associate professor, internist-nephrologist/diabetologist*[2], Joline W J Beulens *assistant professor*[3]

Example #2: Systematic review and external validation of existing prediction models

Type 2 Diabetes

- 366 million people worldwide (estimate of 2011)
- Increased morbidity and mortality
- Can be prevented or postponed by early interventions
- Need for risk prediction models!

Systematic review

- 34 basic models (using variables that can be assessed non-invasively) of which 12 presented as final model
- 42 extended models (including data on one to three conventional biomarkers such as glucose)
- **Many models, few validations**!

Example #2: Systematic review and external validation of existing prediction models

IPD meta-analysis

- EPIC-InterAct case-cohort
  - 27,779 participants of whom 12,403 with incident diabetes
  - 8 countries
- External validation of 12 literature models (with non-laboratory based variables)
  - Discrimination: c-statistic
  - Calibration: calibration plot, ratio expected versus observed
  - Other performance measures: Yates slope, Brier score

Example #2: Systematic review and external validation of existing prediction models

Articles

Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models

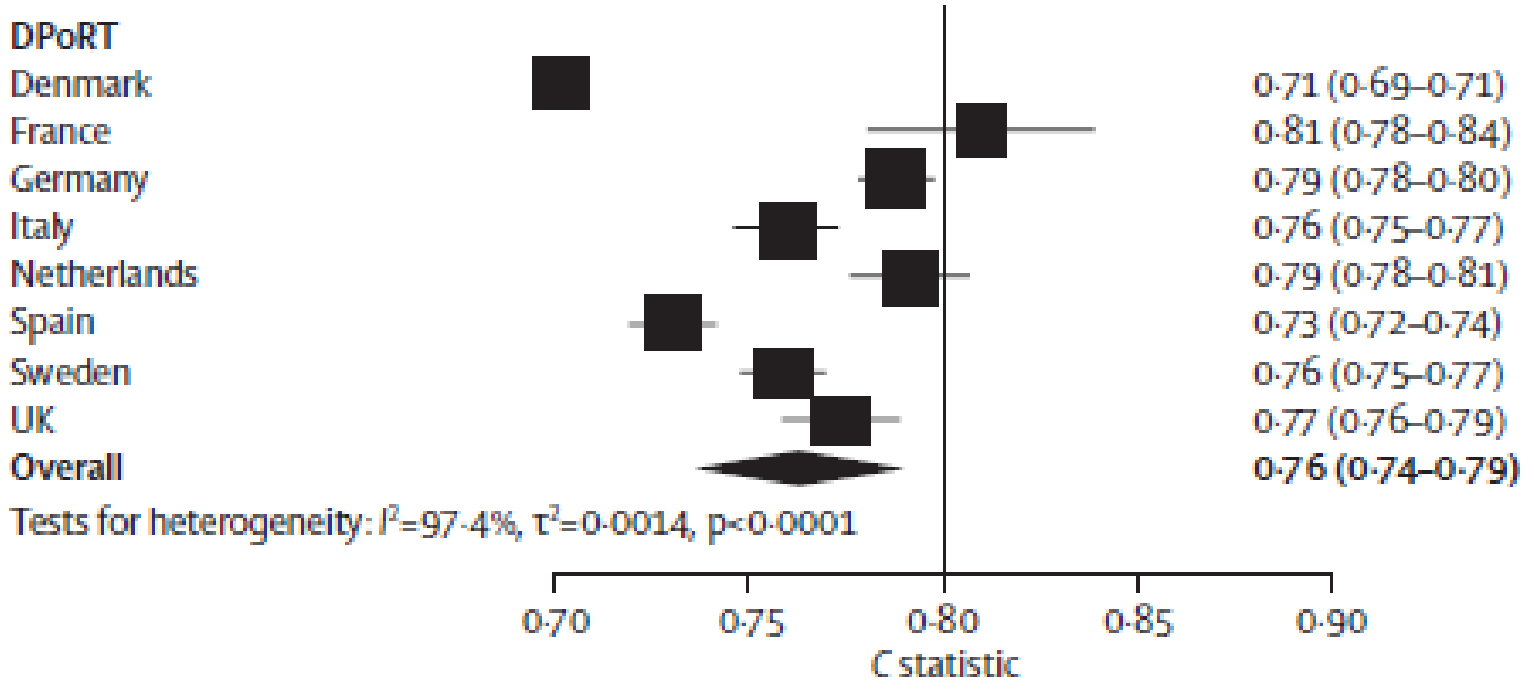
Andre Pascal Kengne, Joline W J Beulens, Linda M Peelen, Karel G M Moons, Yvonne T van der Schouw, Matthias B Schulze, Annemieke M W Spijkerman, Simon J Griffin, Diederick E Grobbee, Luigi Palla, Maria-Jose Tormo, Larraitz Arriola, Noël C Barengo, Aurelio Barricarte, Heiner Boeing, Catalina Bonet, Françoise Clavel-Chapelon, Laureen Dartois, Guy Fagherazzi, Paul W Franks, José María Huerta, Rudolf Kaaks, Timothy J Key, Kay Tee Khaw, Kuanrong Li, Kristin Mühlenbruch, Peter M Nilsson, Kim Overvad, Thure F Overvad, Domenico Palli, Salvatore Panico, J Ramón Quirós, Olov Rolandsson, Nina Roswall, Carlotta Sacerdote, María-José Sánchez, Nadia Slimani, Giovanna Tagliabue, Anne Tjønneland, Rosario Tumino, Daphne L van der A, Nita G Forouhi, Stephen J Sharp, Claudia Langenberg, Elio Riboli, Nicholas J Wareham

The Lancet, Diabetes & Endocrinology (2014)

Example #2: Systematic review and external validation of existing prediction models

Discrimination of model "DPoRT"

(overall and by country)

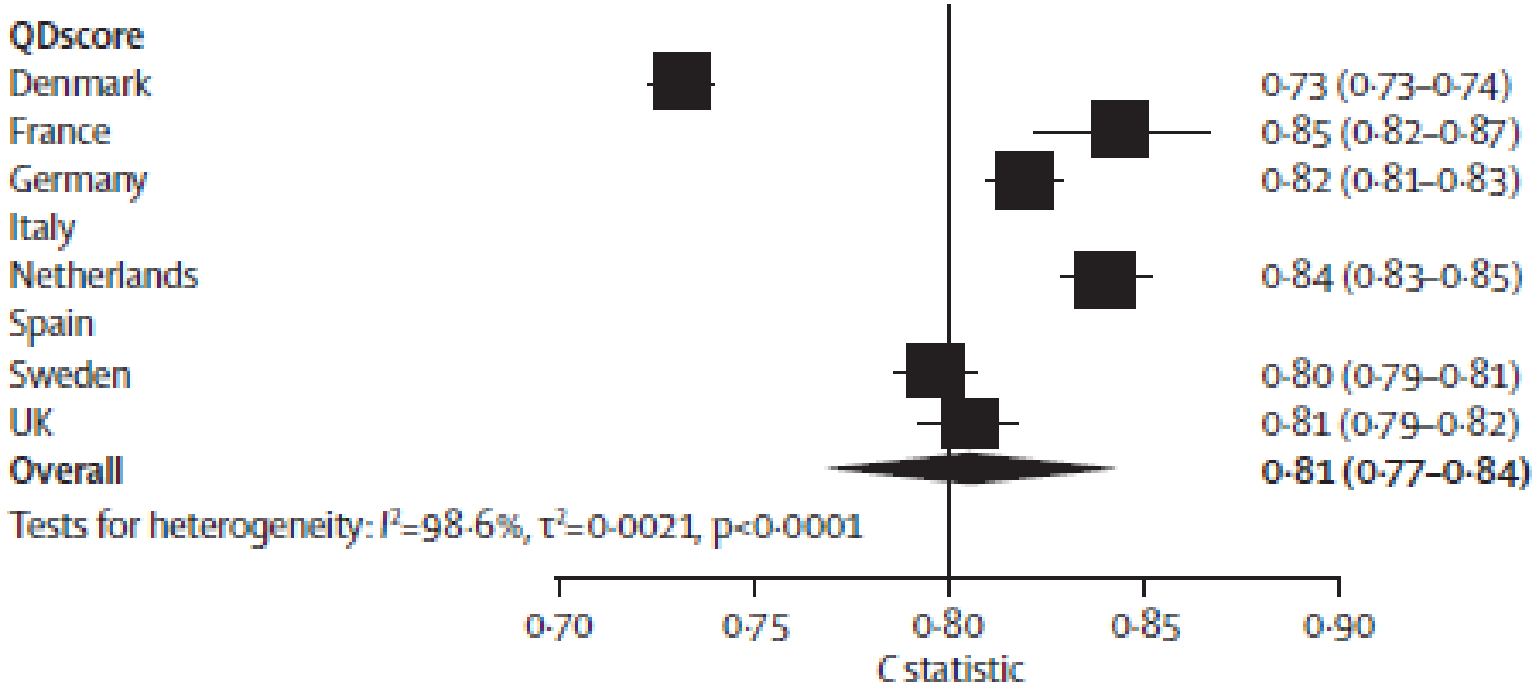


Prediction of incident type 2 diabetes at 10 years of follow-up

Example #2: Systematic review and external validation of existing prediction models

Discrimination of model "QDscore"

(overall and by country)



Prediction of incident type 2 diabetes at 10 years of follow-up

Example #3: Examining the added value of a specific marker

The clinical usefulness of carotid intima-media thickness measurements (CIMT) in cardiovascular risk prediction

Background: problems with Framingham risk score in predicting CVD risk

- No events despite high risk
- Many events in low risk categories

(Hester den Ruijter, Department of experimental cardiology, Julius Center for Health Sciences and Primary Care)

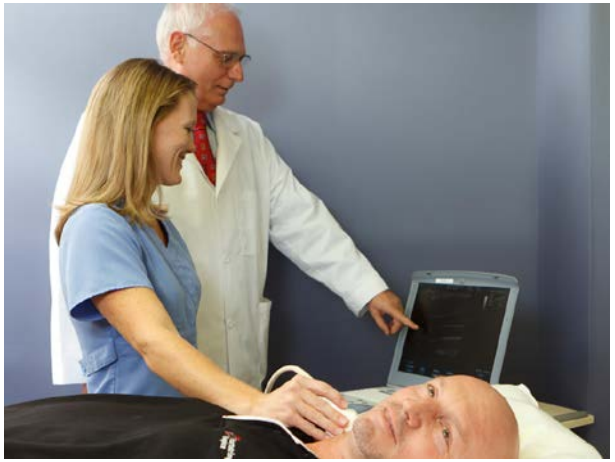
Example #3: Examining the added value of a specific marker

Improvement in CVD risk prediction: incorporation of non-invasive measurement of **atherosclerosis** by means of CIMT measurements

- Reflects long-term exposure to risk factor levels
- Predicts future cardiovascular events
- Modifiable by treatment
- Intermediate between risk factors and events

Example #3: Examining the added value of a specific marker

- B-mode ultrasound measurement of the Carotid Intima Media Thickness (CIMT)



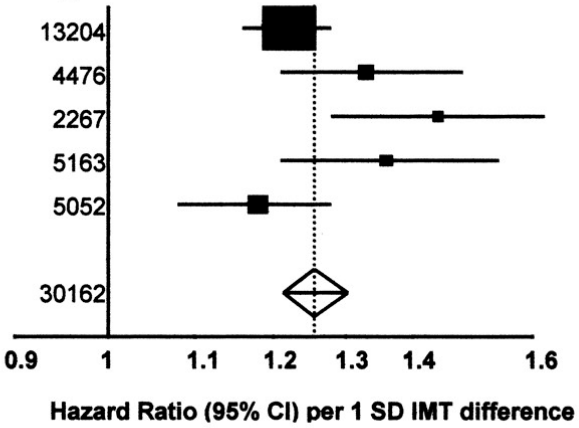
https://www.youtube.com/watch?v=OM_X_Czujrs&feature=player_detailpage

Example #3: So what is the evidence?

Association CIMT-MI: evidence from aggregate data

A **Hazard ratio (HR) for MI per 1 SD difference in CCA-IMT, adjusted for age and sex**

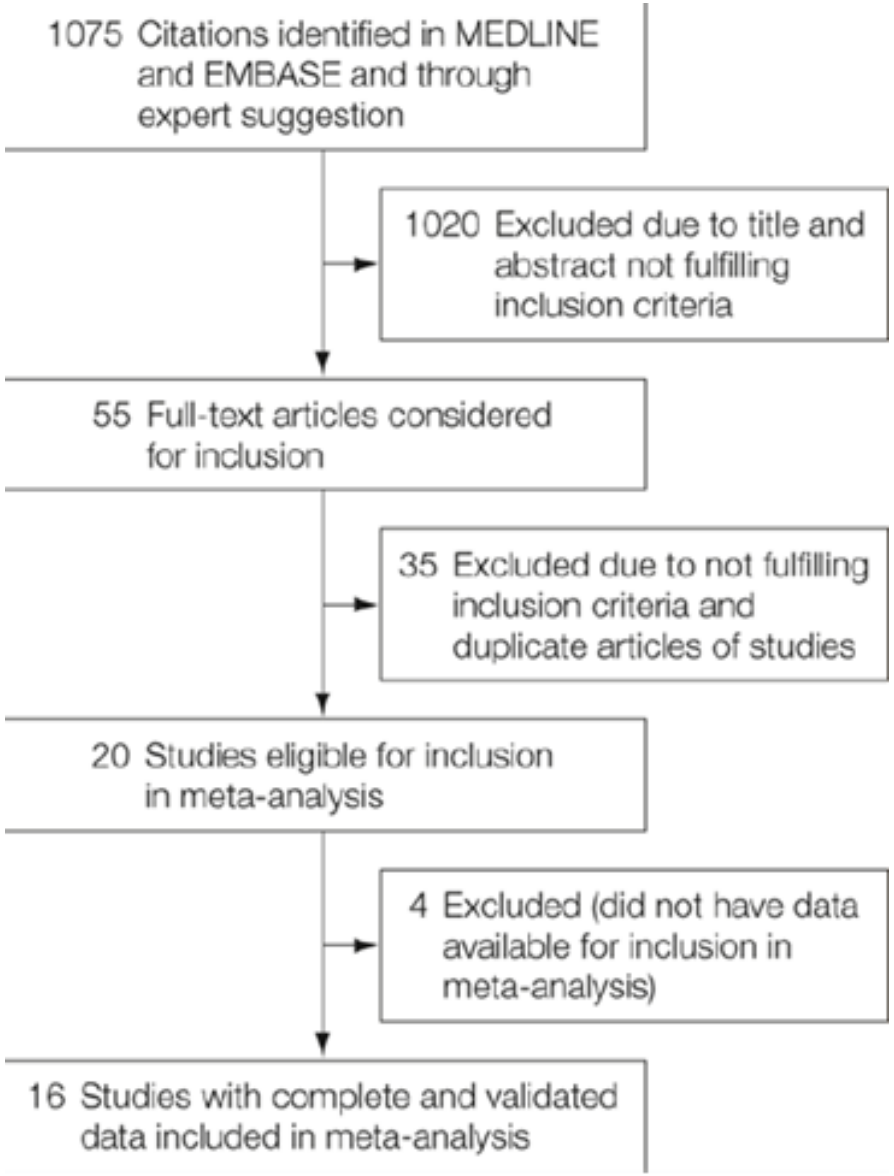
Study	HR	[95% CI]	n	Data source
Atherosclerosis Risk in Communities Study (ARIC)	1.22	[1.16-1.28]	13204	unpublished data
Cardiovascular Health Study (CHS)	1.33	[1.21-1.48]	4476	O'Leary 1999 (5)
Rotterdam Study	1.44	[1.28-1.62]	2267	Del Sol 2002 (7)
Malmö Diet and Cancer Study subcohort (MDCS)	1.36	[1.21-1.54]	5163	Rosvall 2005 (10)
Carotid Atherosclerosis Progression Study (CAPS)	1.18	[1.08-1.28]	5052	Lorenz 2006 (12)
TOTAL	1.26	[1.21-1.30]	30162	
I^2 for heterogeneity	65.2%			



Lorenz M W et al. Circulation. 2007;115:459-467

Example #3: USE-IMT collaboration

- Ongoing individual participant data meta-analysis of general population
- Studies were invited to participate when they had data on Framingham risk score, CIMT measurements and follow-up to CVD



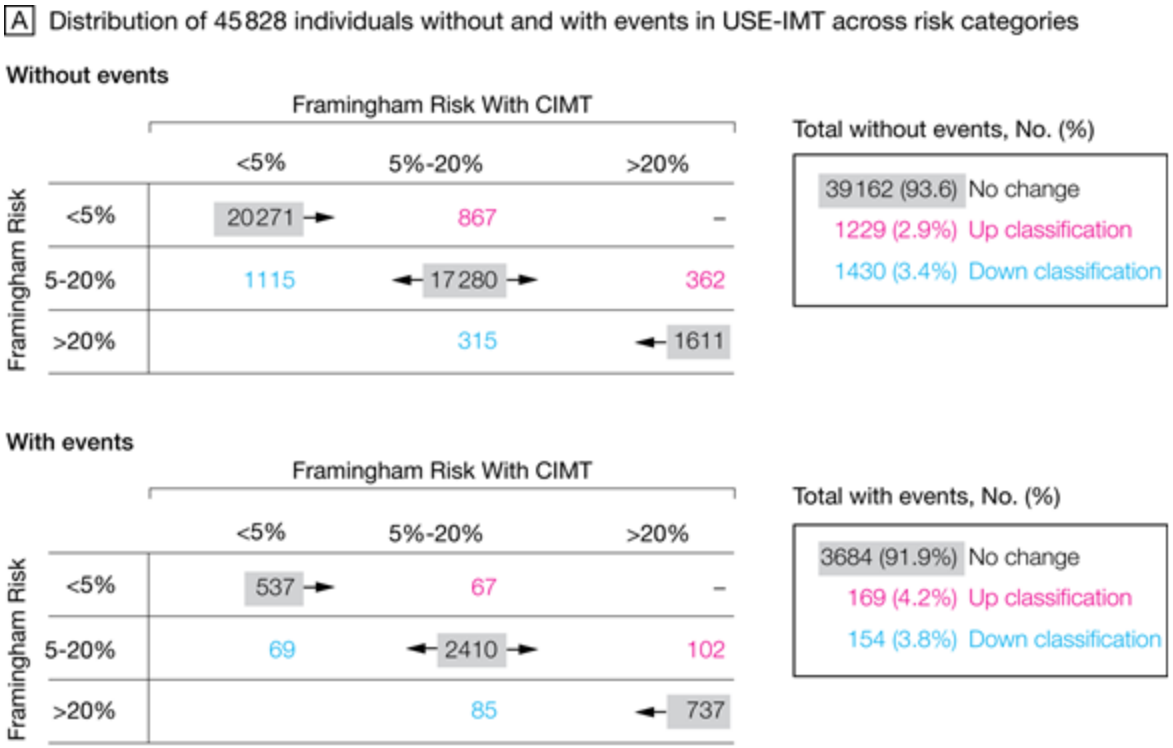
Example #3: models with and without CIMT

- Two Cox proportional hazards models with stroke and MI
  - FRS (refit age, gender, cholesterol, blood pressure, smoking, blood pressure medication)
  - FRS (refit age, gender, cholesterol, blood pressure, smoking, blood pressure medication) + ***CIMT***

- Do these two models reclassify patients differently?

FRS = Framingham Risk Score

Example #3: clinical usefulness



Example #3: conclusion

The **added value of common CIMT** in 10-year risk prediction of cardiovascular events, in addition to the Framingham risk score, **is small and unlikely to be of clinical importance**

Den Ruijter et al. , JAMA 2012

IPD meta-analysis for developing and validating a prediction model

Potential advantages

- Address a wider range of study populations
- Increase variation in subject characteristics
- Increase sample size

However,

- Researchers often simply combine all IPD, and produce a prediction model averaged across all study populations

IPD meta-analysis for developing and validating a prediction model

Simply combining IPD

- Obfuscates the extent to which individual studies were comparable
- Can mask how the model performs in each study population separately
- May lead to prediction models with limited generalizability and poor performance when applied in new subjects

IPD meta-analysis for developing and validating a prediction model

A **qualitative review** was performed to identify…

- … the current **research standards** and techniques
- … the role of IPD meta-analysis **methods** toward development and validation
- … the common **challenges** and methodological problems researchers face

IPD meta-analysis for developing and validating a prediction model

Ahmed *et al. BMC Medical Research Methodology* 2014, **14**:3
http://www.biomedcentral.com/1471-2288/14/3

RESEARCH ARTICLE Open Access

Developing and validating risk prediction models in an individual participant data meta-analysis

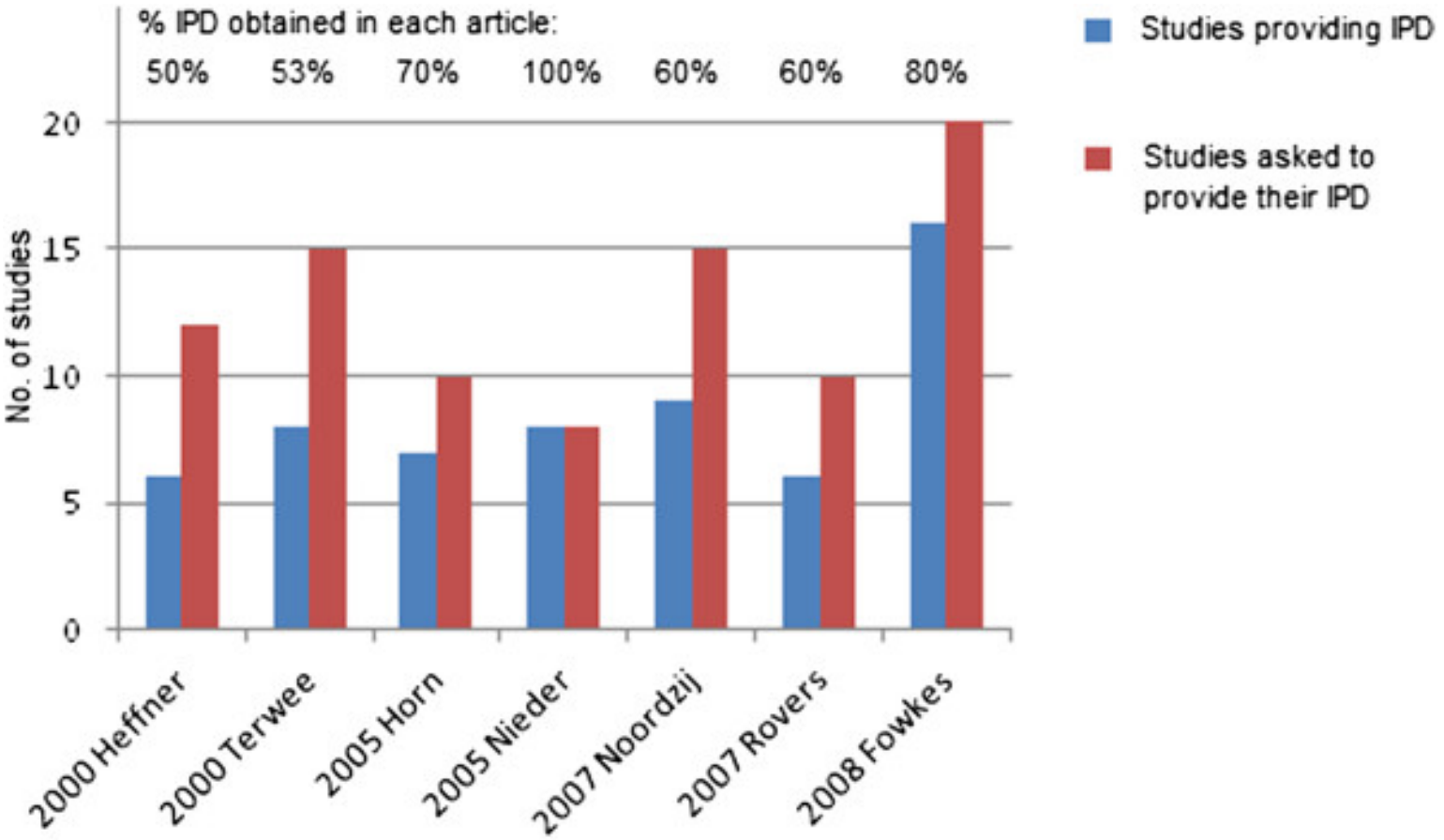
Ikhlaaq Ahmed[1], Thomas PA Debray[2], Karel GM Moons[2] and Richard D Riley[3*]

IPD meta-analysis for developing and validating a prediction model

Systematic review: 15 relevant IPD reviews (1994-2008)

- **Obtaining IPD**
  - (Systematic) literature review (N=7)
  - Collaborative group of selected researchers (N=7)
  - Unclear (N=1)
- **Type of data**
  - Randomized controlled trials (N=7)
    - Data from all treatment groups (N=5)
    - Data from placebo group only (N=2)
  - Observational studies (N=4)
  - Mixture of RCT's and observational studies (N=1)

IPD meta-analysis for developing and validating a prediction model



IPD meta-analysis for developing and validating a prediction model

Systematic review: 15 relevant IPD reviews (1994-2008)

- **Model development**
  - Pool all IPD and ignore clustering of participants (N=10)
  - Pool all IPD and account for clustering, e.g. using dummy variable for study (N=3)
- **Heterogeneity in predictor effects**
  - Not evaluated (N=12)
- **Strategy for inclusion of predictors**
  - P-value driven (N=9 out of 13)
  - Selection procedure (N=4)

IPD meta-analysis for developing and validating a prediction model

Systematic review: 15 relevant IPD reviews (1994-2008)

- **Evaluation of model performance**
  - **None** (N=4)
  - **Internal validation** (N=11): same data are used to develop and validate the model
  - **External validation** ( ): different datasets are used for development and validation
  - **Internal-external cross-validation** (N=2): rotating external validation by iteratively omitting studies during development

IPD meta-analysis for developing and validating a prediction model

Recommendations

- **Allow for different baseline risks in each of the IPD studies**
  - Account for differences in outcome prevalence (or incidence) across studies
  - Examine between-study heterogeneity in predictor effects and prioritize inclusion of (weakly) homogeneous predictors
  - Appropriate intercept for a new study can be selected using information on outcome prevalence (or incidence)
- **Implement a framework that uses internal-external cross-validation**

IPD meta-analysis for developing and validating a prediction model

Statistics
in Medicine

Research Article

Received 20 June 2012, Accepted 18 December 2012 Published online 11 January 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5732

A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis

Thomas P. A. Debray,[a*†] Karel G. M. Moons,[a] Ikhlaaq Ahmed,[b] Hendrik Koffijberg[a] and Richard David Riley[b]

IPD meta-analysis for developing and validating a prediction model

Dealing with **heterogeneity** in an IPD-MA

- Due to differences in study design, inclusion and exclusion criteria, disease severity, interventions undergone, …
- Differences in baseline risk
  - Outcome prevalence (diagnostic models): intercept term
  - Outcome incidence (prognostic models): baseline hazard
- Differences in predictor-outcome associations
  - Regression coefficients

IPD meta-analysis for developing and validating a prediction model

Dealing with **heterogeneity** in an IPD-MA

- Typically accounted for by **random effects** modeling (intervention research). However:
  - Model parameters take different values for each included study
  - Which parameters to use when validating/implementing the model in new individuals or study populations?
  - When do study populations differ too much to combine?

- Need for a framework that can identify the extent to which aggregation of IPD is justifiable, and provide the optimal approach to achieve this.

Step 1: model development

Different choices to combine IPD

- **Stacking**: ignore clustering of subjects within studies, merge all data into one big dataset
- **Random effects modeling** (of intercept term): account for differences in baseline risk across studies by assuming a certain distribution of intercept terms
- **Stratified modeling** (of intercept term): account for differences in baseline risk across studies, without assuming a certain distribution of intercept terms.

Step 2: choosing an appropriate model intercept when implementing the model to new individuals

- **Average intercept**: can directly be used in a new study population; dangerous when there is much heterogeneity in baseline risk across studies
- **Intercept selection**: choose intercept term from study with most similar outcome prevalence.
- **Intercept estimation (option 1)**: directly estimate most appropriate intercept term for the new study population from **outcome prevalence**
- **Intercept estimation (option 2)**: re-estimate the model intercept from **locally collected IPD**

Step 3: model evaluation

Check whether

- Modeling of predictors is adequate (e.g. choice of predictors, nonlinear terms, interactions, …)
- Intercept term is adequately modeled (e.g. random effects versus stratified intercept term)
- Strategy for choosing intercept term in new study population is adequate (e.g. average intercept versus intercept selection)
- Model performance is consistently well across studies
  - Discrimination
  - Calibration

Internal-external cross-validation

Procedure

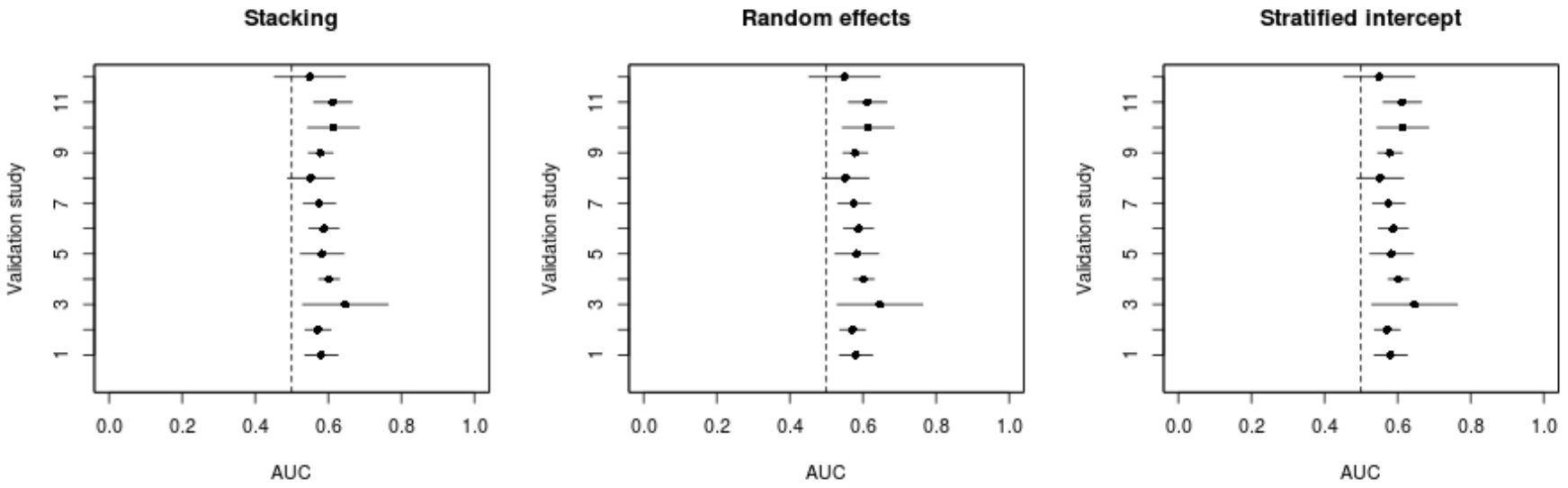
1. Check whether baseline risk (intercept term) is heterogeneous across studies
2. Iteratively develop model using M-1 studies, and externally validate model in remaining study
3. Evaluate whether derived models have good performance in independent studies
4. Derive a single final model from all available IPD

Example #4: developing and directly validating a prediction model

- Diagnosis of **deep vein thrombosis** (DVT)
  - IPD-MA of 12 studies
  - 10,014 patients (1,897 with DVT)
  - Focus on 2 homogeneous predictors: sex & recent surgery
- Comparison of 3 strategies
  - **Stacking**, ignore clustering of subjects within studies
  - **Random effects modeling** on intercept term (use average intercept in new study)
  - **Stratified intercept terms** (select intercept term based on outcome prevalence)
- Evaluate discrimination and calibration

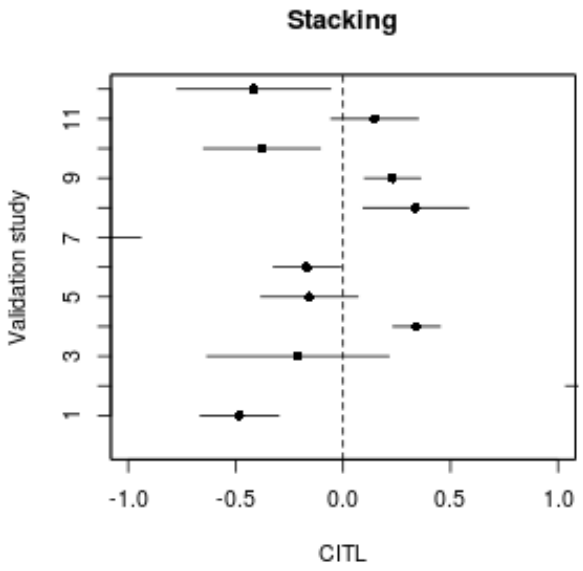
Example #4: developing and directly validating a prediction model

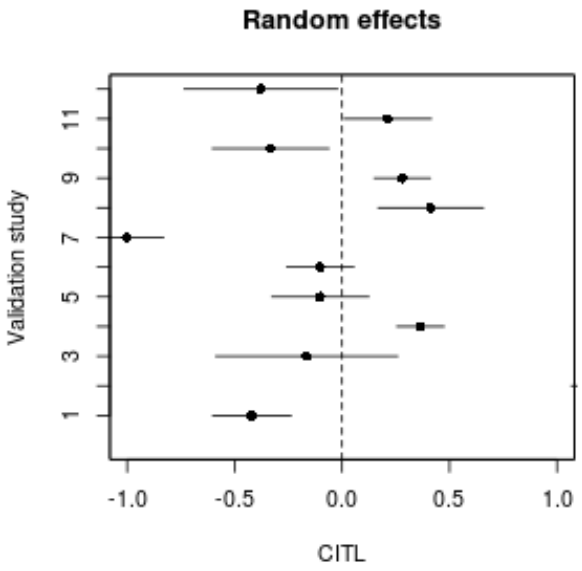
Model discrimination

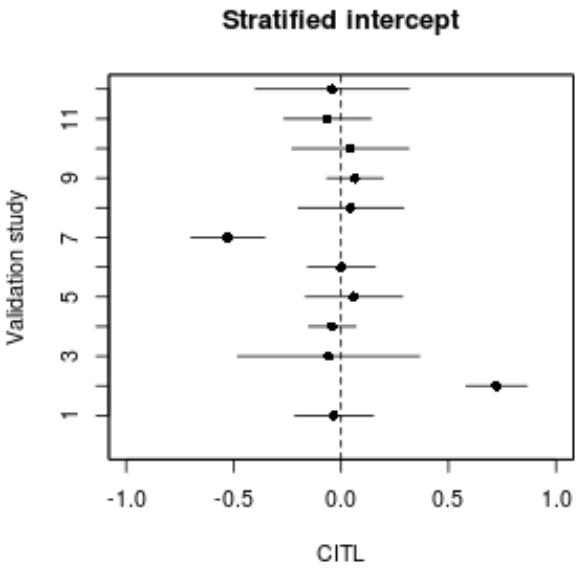


Example #4: developing and directly validating a prediction model

Model calibration







Example #4: overall conclusions

Outcome prevalence = reliable proxy for selecting an appropriate intercept term…

- Leads to consistent performance across studies

… as long as predictor effects are homogenous

- Outcome prevalence no longer reliable proxy (affects *calibration-in-the-large*)
- Predictor effects no longer consistent across studies (affects *calibration slope*)
- Other predictors may, however, improve discrimination!!
  - Sex & surg : AUC varies between 0.55 to 0.65
  - malignancy, recent surgery, calf difference and D-dimer test: AUC varies between 0.73 to 0.92

Take home messages

IPD meta-analysis in prediction research

- Improving the performance of novel prediction models across different study populations
- Attain a better understanding of the generalizability of a prediction model
- Exploring heterogeneity in model performance and the added value of a novel (bio)marker

Unfortunately, most researchers analyze their IPD as if representing **a single dataset**!

Take home messages

Remaining challenges in IPD meta-analysis

- Synthesis strategies from intervention research cannot directly be applied in prediction research (due to focus on absolute risks)
- Adjustment to local circumstances often needed
  - One model fits all?
  - Methods for tailoring still underdeveloped

New methods are on their way!

Take home messages

Reasons to be optimistic

- Cochrane Prognosis Methods Group
  - Aims to facilitate evidence-based prognosis research
  - Improve design, quality & reporting of primary studies
  - Facilitate systematic reviews & meta-analysis in long-run
  - Bring together prognosis researchers, and guide Cochrane reviewers facing prognostic information
  - Develop handbook