



UMC Utrecht  
Julius Center

# Individual Participant Data (IPD) Meta-analysis of prediction modelling studies

Thomas Debray, Hans Reitsma, Karel Moons, Richard Riley

*for the Cochrane IPD Meta-analysis Methods Group*

*(Co-convenors: Jayne Tierney, Mike Clarke, Lesley Stewart,  
Maroeska Rovers)*

# Conflict of interest

**We have developed and validated several multivariable prediction models.**

**We performed several individual patient data meta-analyses, in addition to methodological work**

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



# Prediction models: dynamic world

- Waves of new biomarkers and prediction models
- Increasing pressure for their evaluation
- Recognition of the importance of external validation
- Performance of models is likely to be variable
- Individual patient data: insight why models vary in performance or to build more robust models
- Improvements in methodology



# Illustration

[https://www.youtube.com/watch?v=OM\\_X\\_Czujrs&feature=player\\_detailpage](https://www.youtube.com/watch?v=OM_X_Czujrs&feature=player_detailpage)



# Workshop objectives

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

- To explain key concepts in prediction research
- To describe potential benefits of IPD
- To identify challenges for IPD reviews
- To provide examples of IPD meta-analyses
- To illustrate basic and novel methods



# 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

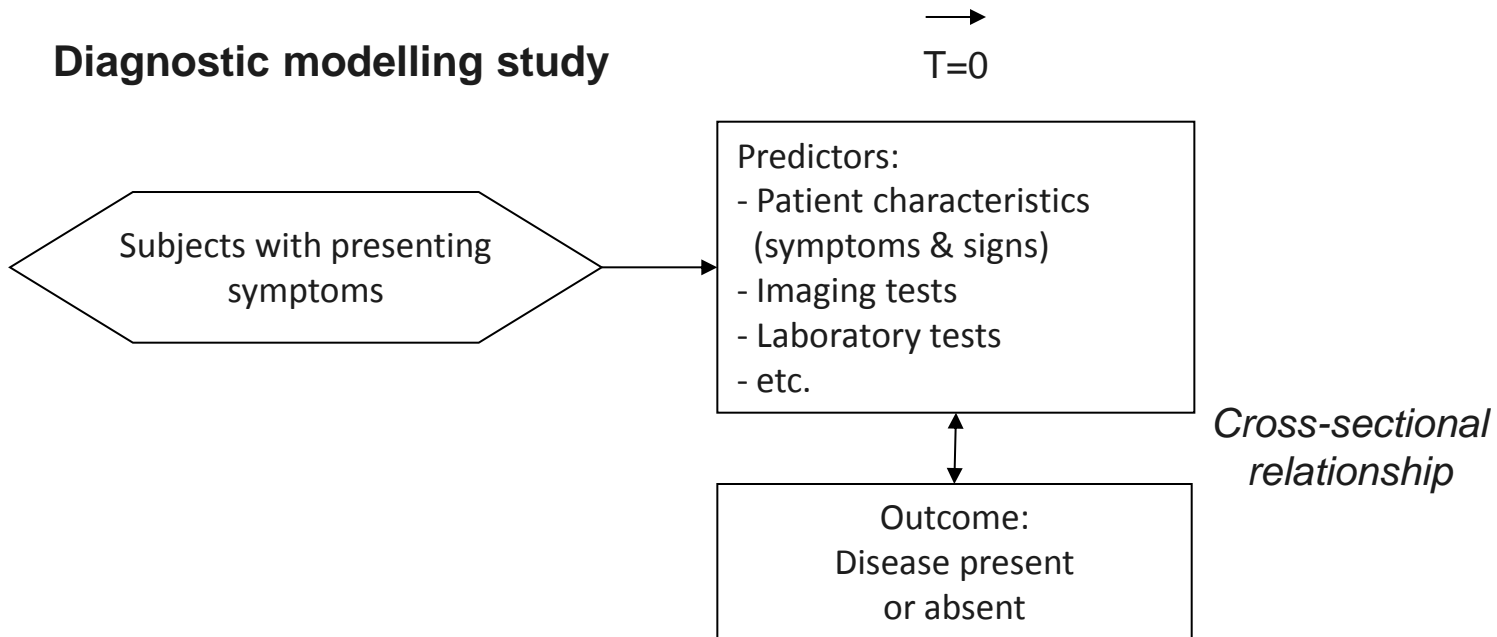


# Multivariable prediction models

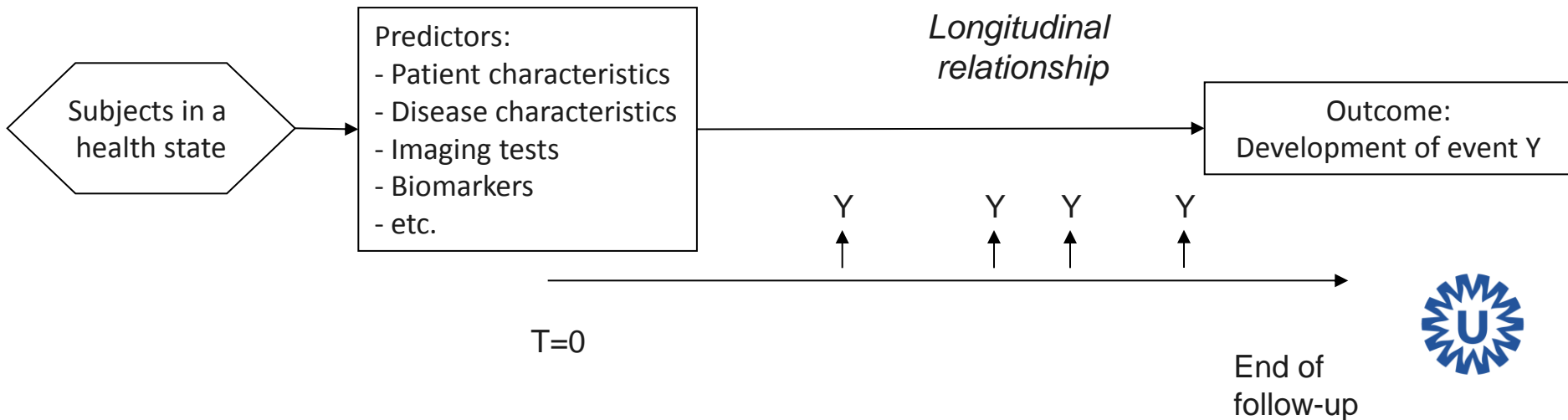
- To calculate absolute risk based on individual profile
- Predict outcome from demographic, patient and disease characteristics (predictors, covariates, risk factors, X variables)
- Use of regression models, two main types:
  - Logistic regression
  - Time-to-event analysis (Kaplan-Meier, Cox)
- Statistical modelling: (1) overlap in information from different predictors; (2) acknowledge strength of each predictor



## Diagnostic modelling study



## Prognostic modelling study





# 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



# Prediction in Prognosis

(Prognosis BMJ series 2009)

- 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



# 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.**

<b>Signs</b>	<b>0</b>	<b>1</b>	<b>2</b>
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

$\Sigma$  = Apgar score (0-10)



## Women

## Men

	Non-smoker					Smoker				
180	7	8	9	11	14	12	15	17	21	26
160	5	6	7	8	10	9	10	12	15	19
140	3	4	5	6	7	6	7	9	11	13
120	2	3	3	4	5	4	5	6	8	10

Age

	Non-smoker					Smoker				
180	12	14	17	22	27	22	26	31	38	47
160	8	10	13	16	20	15	19	23	28	35
140	6	7	9	11	14	11	13	17	20	26
120	4	5	6	8	10	8	10	12	15	19

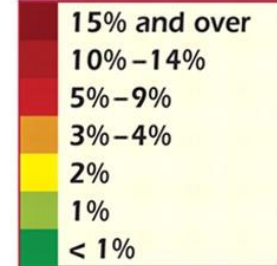
65

	Non-smoker					Smoker				
180	4	4	5	6	8	7	8	10	12	15
160	3	3	4	5	6	5	6	7	9	11
140	2	2	3	3	4	3	4	5	6	8
120	1	2	2	2	3	2	3	3	4	5

60

	Non-smoker					Smoker				
180	8	9	12	15	18	14	17	21	29	33
160	5	7	8	11	13	10	13	15	19	24
140	4	5	6	8	10	7	9	11	14	18
120	3	3	4	5	7	5	6	8	10	13

# SCORE



Systolic blood pressure (mmHg)

	Non-smoker					Smoker				
180	2	2	3	3	4	3	4	5	6	8
160	1	2	2	2	3	2	3	4	5	6
140	1	1	1	2	2	2	2	3	3	4
120	1	1	1	1	2	1	1	2	2	3

55

	Non-smoker					Smoker				
180	5	6	7	9	12	9	11	14	17	22
160	3	4	5	7	9	6	8	10	12	16
140	2	3	4	5	6	5	6	7	9	11
120	2	2	3	3	4	3	4	5	6	8

	Non-smoker					Smoker				
180	1	1	1	2	2	2	2	2	3	4
160	1	1	1	1	1	1	1	2	2	3
140	0	1	1	1	1	1	1	1	2	2
120	0	0	0	1	1	1	1	1	1	1

50

	Non-smoker					Smoker				
180	3	4	4	6	7	6	7	8	11	12
160	2	3	3	4	5	4	5	6	8	10
140	1	2	2	3	4	3	3	4	5	7
120	1	1	2	2	3	2	2	3	4	5

	Non-smoker					Smoker				
180	0	0	0	0	0	0	0	0	1	1
160	0	0	0	0	0	0	0	0	0	0
140	0	0	0	0	0	0	0	0	0	0
120	0	0	0	0	0	0	0	0	0	0

40

	Non-smoker					Smoker				
180	1	1	1	2	2	1	2	2	3	4
160	1	1	1	1	2	1	1	2	2	3
140	0	1	1	1	1	1	1	1	2	2
120	0	0	0	1	1	1	1	1	1	1

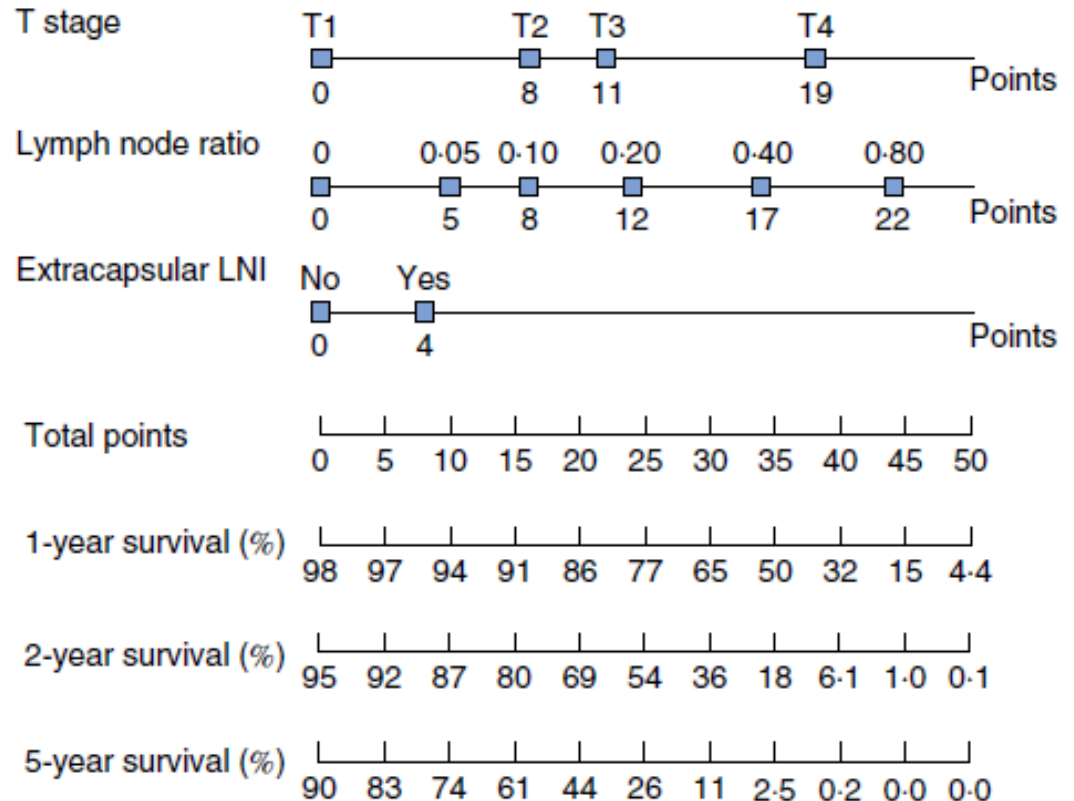
10-year risk of fatal CVD in populations at high CVD risk

© ESC 2007

Total cholesterol: HDL  
Cholesterol ratio



# Nomogram Simplified Model



**Fig. 2** Nomogram for disease-specific survival after surgery for adenocarcinoma of the distal oesophagus or gastro-oesophageal junction, based on the reduced model. From the total points axis, a straight line down through the survival axes shows survival probabilities at 1, 2 and 5 years in the absence of death from another cause. The lymph node ratio is the ratio of the number of positive lymph nodes to the total number of lymph nodes resected. LNI, lymph node involvement

# Survival curves / Kaplan Meier

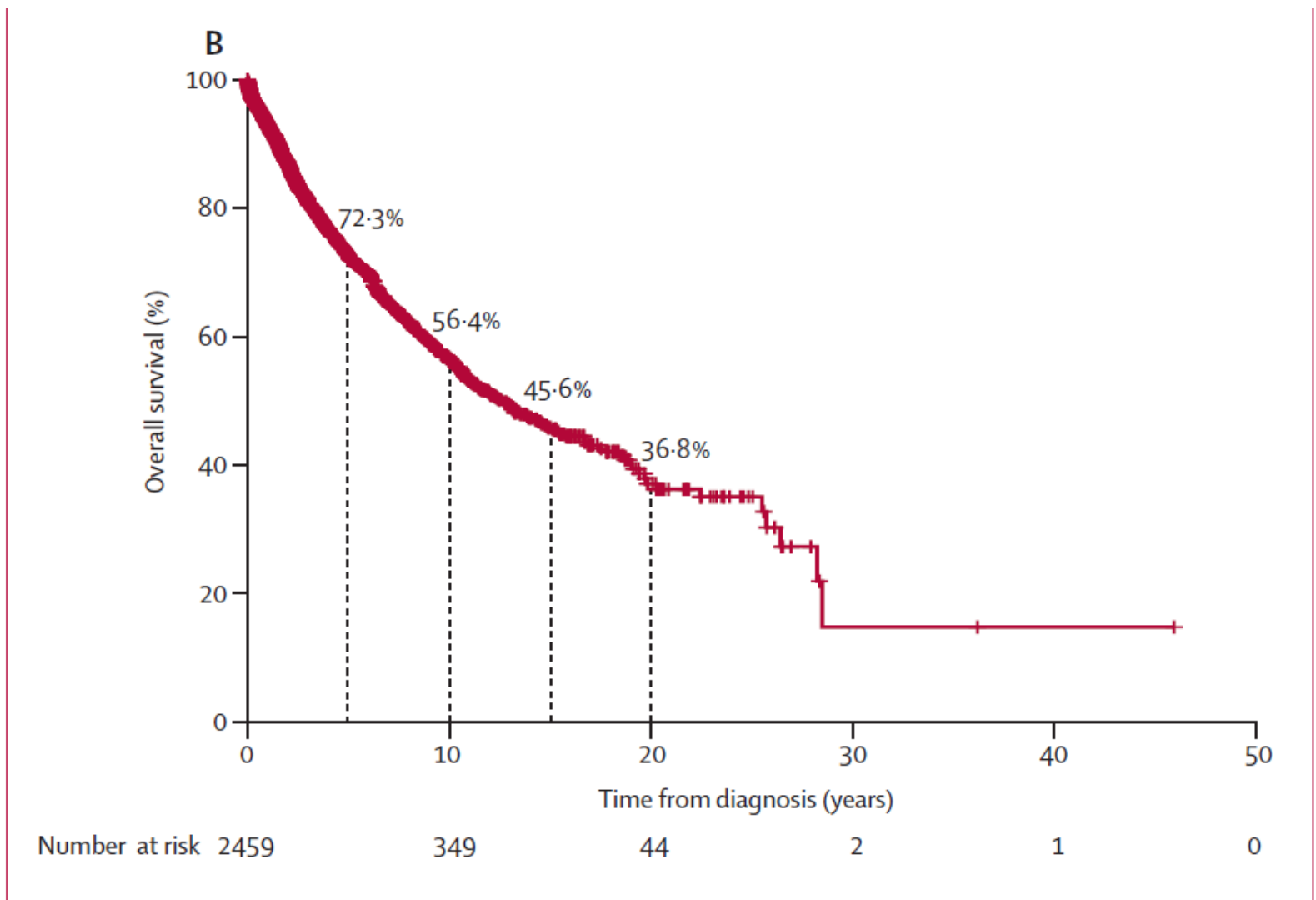
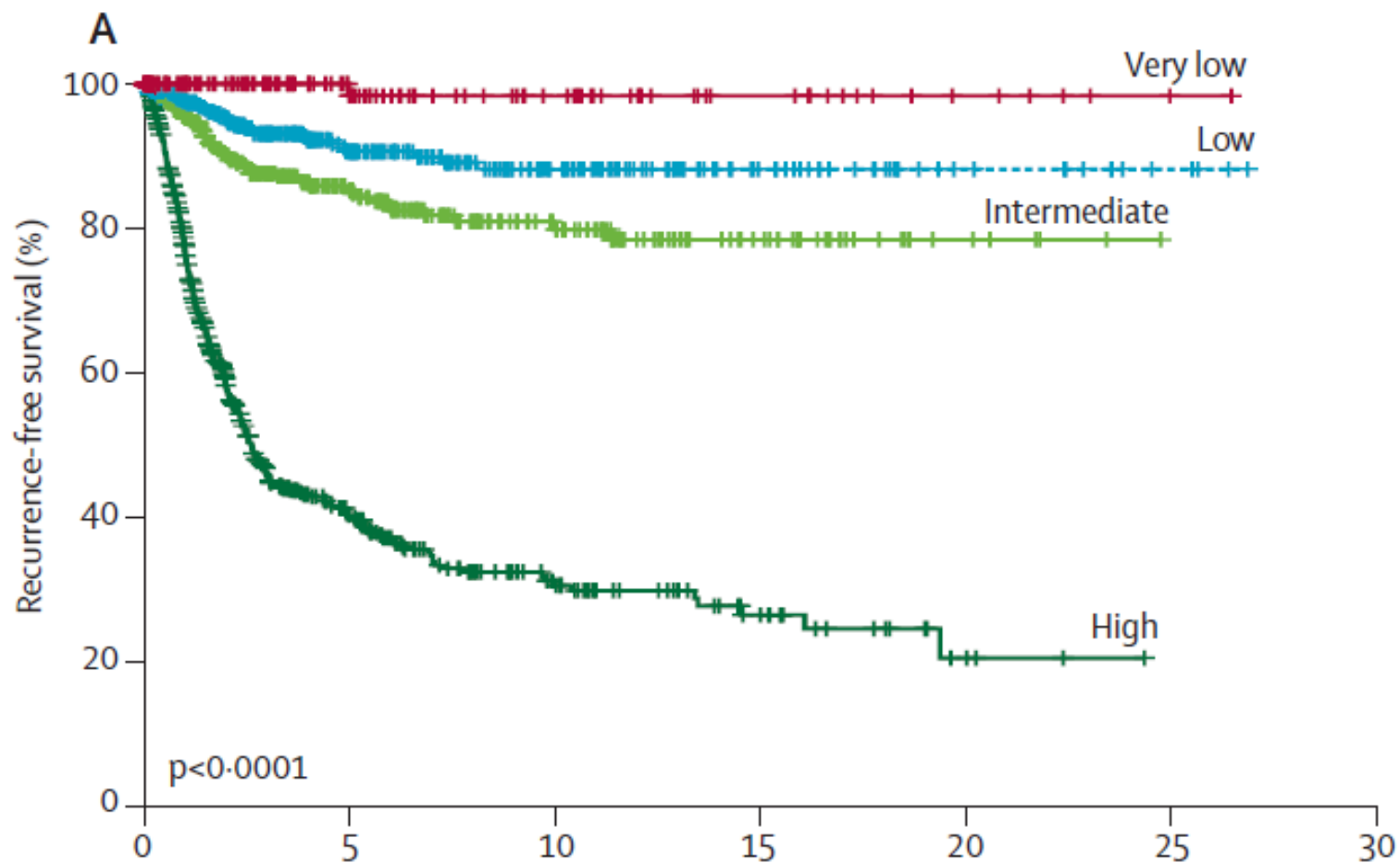


Figure 1: Recurrence-free survival (A) and overall survival (B) in the pooled series





**Number at risk**

Very low risk	130	59	33	16	6	1	0
Low risk	429	164	74	32	12	4	0
Intermediate risk	395	161	71	29	6	0	0
High risk	554	122	47	20	4	0	0

# Predicting bacterial cause in conjunctivitis

**Table 3** Results of logistic regression analysis. Independent indicators of positive bacterial culture and their clinical score

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:  $5 + -1 = 4$ .



# Predicting bacterial cause in conjunctivitis

Clinical score	Percentage (95% CI) predicted positive cultures†
+5	77 (57 to 90)
+4	65 (47 to 79)
+3	51 (23 to 79)
+2¶	40 (26 to 55)
+1	27 (17 to 39)
0	18 (7 to 38)
-1	11 (4 to 26)
-2	7 (2 to 28)
-3	4 (1 to 15)



# 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



# Meta-analysis of individual participant data

## Opportunities

- Increase total sample size -> reduce risk of overfitting
- Increase available case-mix variability -> enhances the model's potential generalisability
- Ability to standardize analysis methods across IPD sets
- Ability to investigate more complex associations
- Ability to explore heterogeneity in predictive performance
- Ability to evaluate generalisability and usability of prediction models across different situations



# Meta-analysis of individual participant data

## 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



# Meta-analysis of individual participant data

## IPD – 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

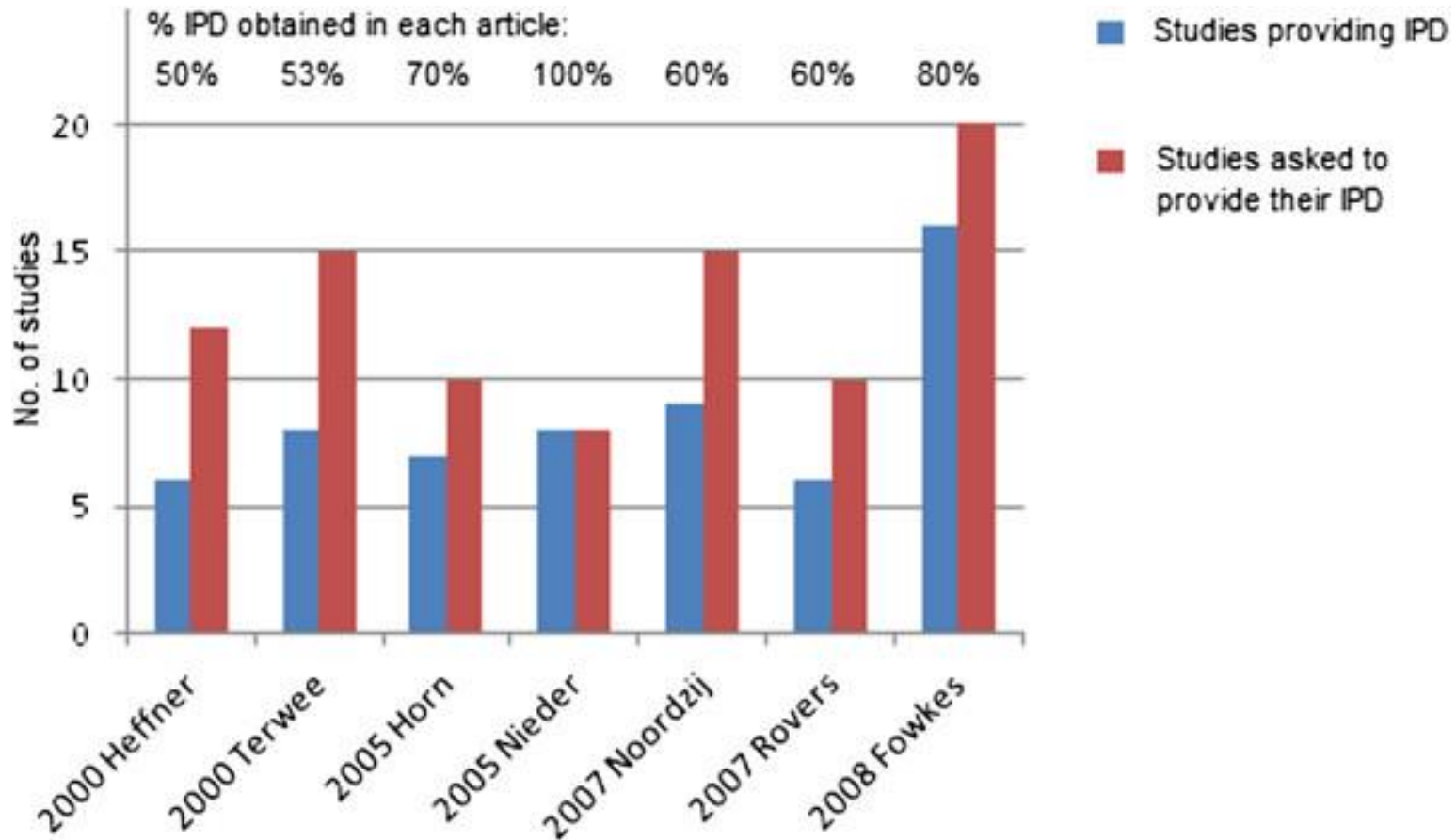
## Individual participant data meta-analysis of prognostic factor studies: *state of the art?*

Ghada Abo-Zaid<sup>1</sup>, Willi Sauerbrei<sup>2</sup> and Richard D Riley<sup>3\*</sup>



# Meta-analysis of individual participant data

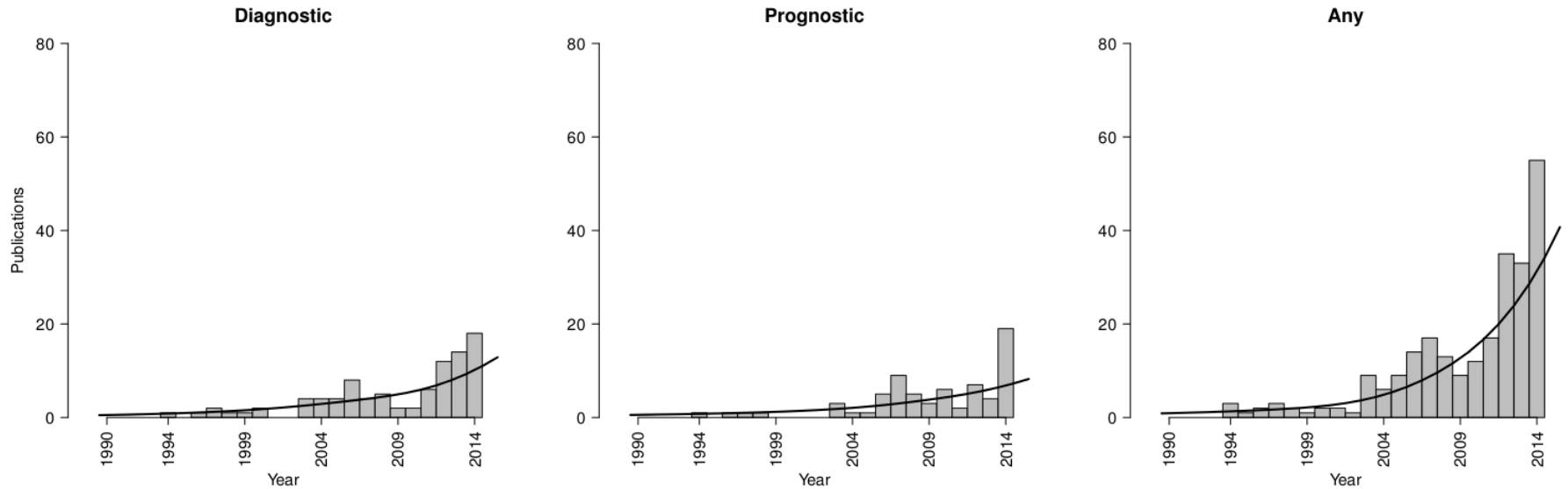
## IPD – Reasons to be optimistic





# Meta-analysis of individual participant data

## IPD – Reasons to be optimistic



Trends in publications of IPD-MA studies focusing on the development and/or validation of diagnostic and prognostic prediction models.



# Meta-analysis of individual participant data

## Why do we need specific guidance?

Evidence synthesis currently gold standard for summarizing relative treatment effects – many methods available!

However,

- Meta-analysis models cannot *mutate mutandis* be applied to prediction modeling studies
- Researchers often simply combine all IPD, and produce a prediction model averaged across all study populations
- There are major differences in the aims, design and analysis of primary studies between prediction modeling and intervention studies!



# Meta-analysis of individual participant data

## Why do we need specific guidance?

### Simply combining IPD

- Obscures 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



# What are the main differences between prediction and intervention research?

Intervention research	Prediction research
<b>Aim(s)</b> <ul style="list-style-type: none"><li>• Estimation of therapeutic effect of a specific treatment</li><li>• Study treatment effect in subgroups</li></ul>	<b>Aim(s)</b> <ul style="list-style-type: none"><li>• Estimation of absolute risk probabilities for distinct individuals across different populations or subgroups</li><li>• Evaluate accuracy of model predictions across subgroups</li></ul>
<b>Association measures:</b> relative risk estimates	<b>Association measures:</b> absolute probability of risk estimates
<b>Study design:</b> Randomized studies	<b>Study design:</b> observational research
<b>Evaluation:</b> bias and precision of estimated comparative treatment effects	<b>Evaluation:</b> model discrimination and calibration



# Types of IPD-MA of prediction modeling studies

1. Validation of existing model(s)
2. Tailoring/combining of existing model(s)
3. Examining added value of a specific marker on top an existing model
4. Developing and directly validating a new model



# Types of IPD-MA

## 1. Validation of existing model(s)

Apply meta-analysis to:

- Summarize estimates of model discrimination and calibration



Use IPD to:

- Investigate sources of heterogeneity in model performance
- Identify which models perform best in what (sub)population, setting or country



# Types of IPD-MA

## 1. Validation of existing model(s)


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

 OPEN ACCESS

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



# Types of IPD-MA

## 1. Validation of existing model(s)

### 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!**





# Types of IPD-MA

## 1. Validation of existing model(s)

After systematic review, IPD was initiated

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, Aurdio 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)



# Types of IPD-MA

## 1. Validation of existing model(s)

### 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

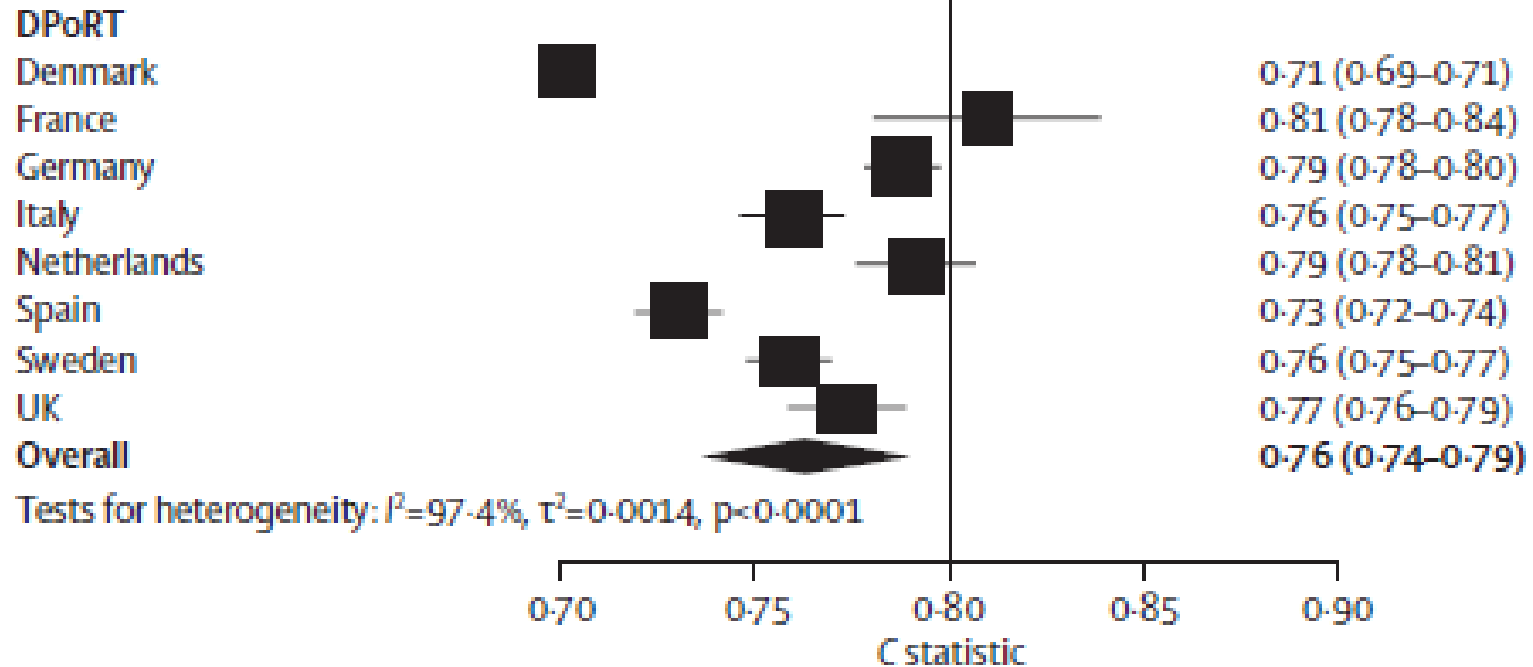


# Types of IPD-MA

## 1. Validation of existing model(s)

### Discrimination of model "DPoRT"

(overall and by country)



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



# Types of IPD-MA

## 2. Tailoring/combining of existing model(s)

Apply meta-analysis to:

- Adjust for between-study heterogeneity in outcome occurrence and/or predictor effects

Use IPD to:

- Combine and tailor the model(s) to specific (sub)populations, settings or countries



# Types of IPD-MA

## 2a. Tailoring of existing model(s)

**Example:** Majed and colleagues evaluated whether the calibration of the Framingham risk equation for coronary heart disease and stroke improved by applying local adjustments.

	E:O ratio			C statistic		
	O	R	L	O	R	L
PRIME-total	1.94	0.98	1.00	0.68	0.68	0.68
PRIME-France	2.23	0.99	1.00	0.67	0.67	0.68
PRIME-Ireland	1.42	0.99	1.00	0.67	0.67	0.67

Outcome: CHD & Stroke, O=original, R=recalibrated, L=local model

Ref: Majed *et al. Preventive Medicine* 2008 **57**.



# Types of IPD-MA

## 3. Examining added value

Apply meta-analysis to:

- Summarize estimates of added value
  - Adjusted predictor effects
  - Improvement in model calibration
  - Improvement in model discrimination
  - Improvement in model reclassification

Use IPD to:

- Investigate sources of heterogeneity in added value
- Identify relevant subgroups that yield different added value



# Types of IPD-MA

## 3. Examining added value

**Example:** 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)



# Types of IPD-MA

## 3. Examining added value

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

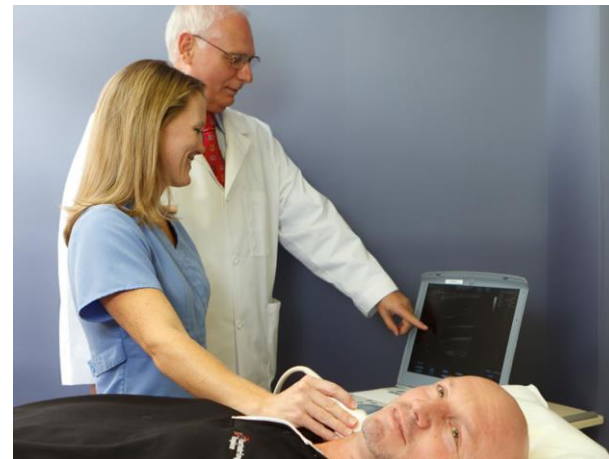
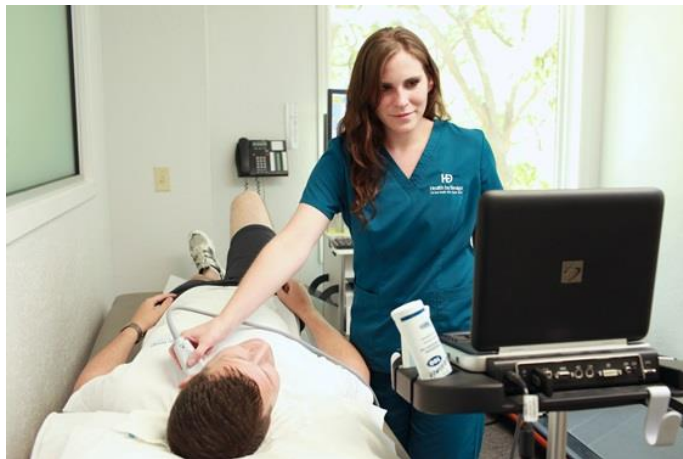




# Types of IPD-MA

## 3. Examining added value

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



[https://www.youtube.com/watch?v=OM\\_X\\_Czujrs&feature=player\\_detailpage](https://www.youtube.com/watch?v=OM_X_Czujrs&feature=player_detailpage)

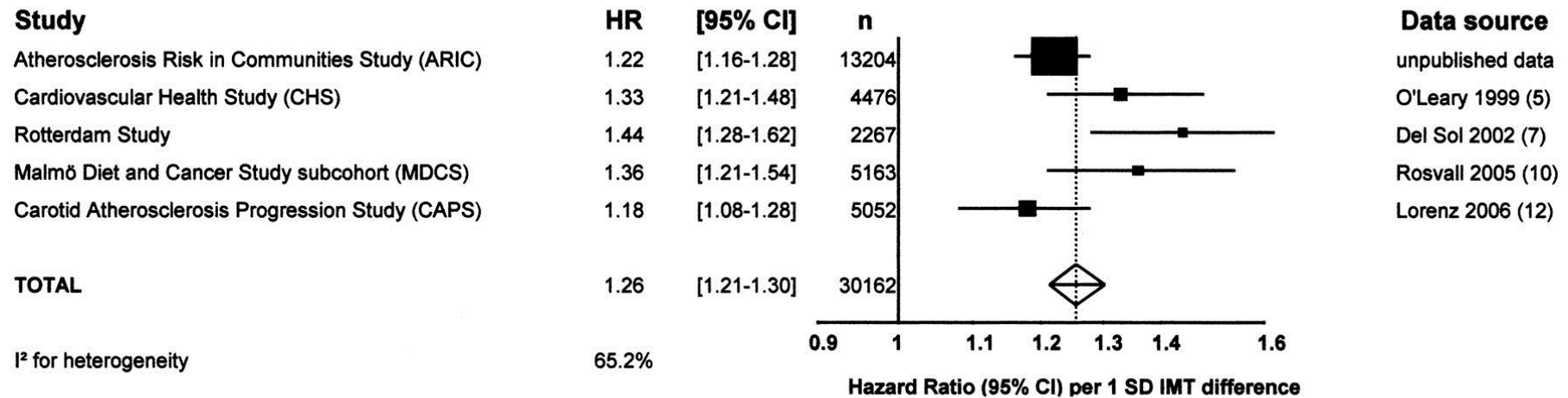


# Types of IPD-MA

## 3. Examining added value

### 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



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

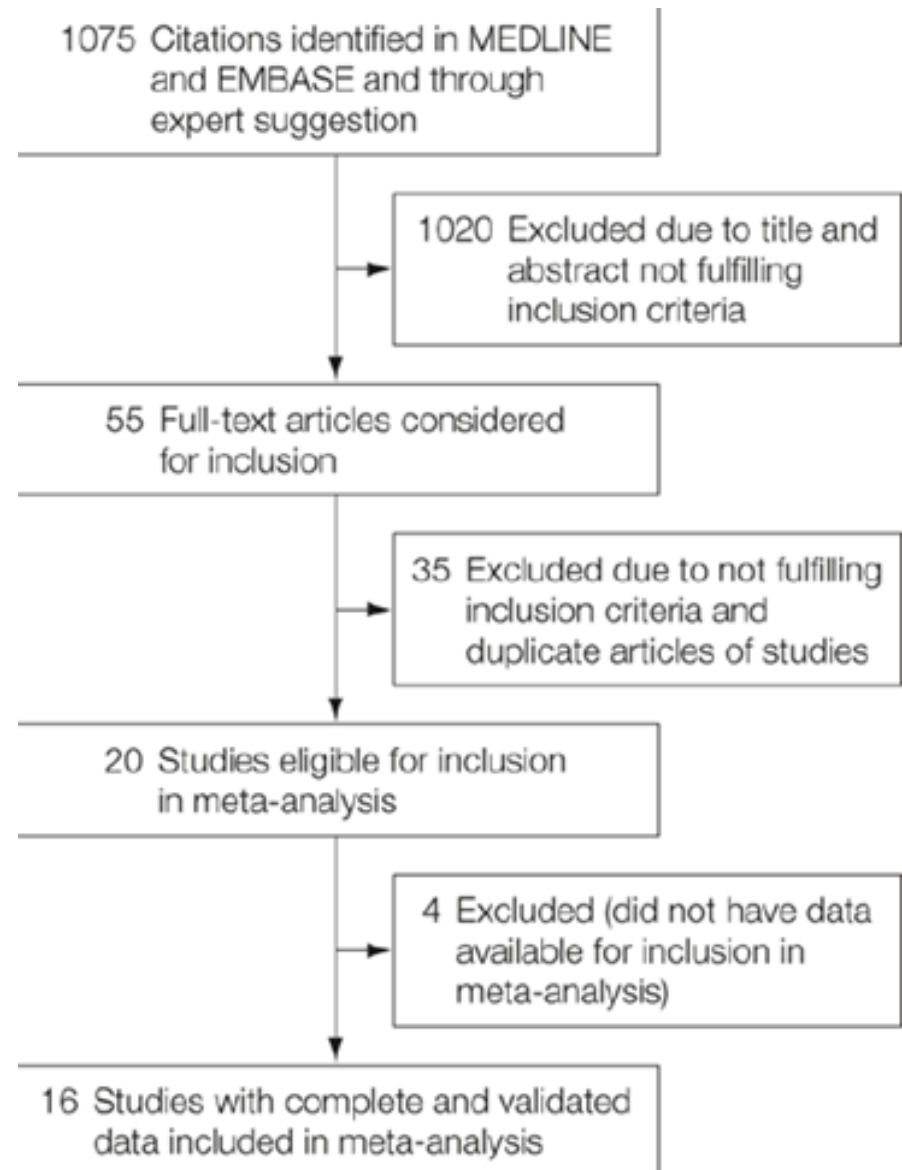


# Types of IPD-MA

## 3. Examining added value

### 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



# Types of IPD-MA

## 3. Examining added value

- 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



# Types of IPD-MA

## 3. Examining added value

**A** Distribution of 45 828 individuals without and with events in USE-IMT across risk categories

Without events

		Framingham Risk With CIMT		
		<5%	5%-20%	>20%
Framingham Risk	<5%	20271 →	867	-
	5-20%	1115	← 17280 →	362
	>20%		315	← 1611

Total without events, No. (%)

39 162 (93.6)	No change
1229 (2.9%)	Up classification
1430 (3.4%)	Down classification

With events

		Framingham Risk With CIMT		
		<5%	5%-20%	>20%
Framingham Risk	<5%	537 →	67	-
	5-20%	69	← 2410 →	102
	>20%		85	← 737

Total with events, No. (%)

3684 (91.9%)	No change
169 (4.2%)	Up classification
154 (3.8%)	Down classification



# Types of IPD-MA

## 3. Examining added value

### 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



# Types of IPD-MA

## 4. Developing and directly validating a new model

Apply meta-analysis to:

- Adjust for between-study heterogeneity in outcome occurrence or predictor effects

Use IPD to:

- Tailor the meta-model to specific (sub)populations, settings or countries

---

**Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies**



*Jacoba P Greving, Marieke J H Wermer, Robert D Brown Jr, Akio Morita, Seppo Juvela, Masahiro Yonekura, Toshihiro Ishibashi, James C Torner, Takeo Nakayama, Gabriel J E Rinkel, Ale Algra*



# Statistical Methods

## Main challenges

- Missing data
  - Partially missing data within studies
  - Systematically missing data within studies
  - Entire study missing (e.g. non-publication)
- Between-study heterogeneity
  - Predictor effects
  - (Change in) model performance
- Combination of IPD and AD
  - Published prediction models
  - Published predictor effects
  - Published estimates of (increased) model performance





# Dealing with missing data

## Recommendations

- Adopt multiple imputation techniques
- Allow for heterogeneity across studies
  - Stratified (two-stage) imputation
  - Multilevel (one-stage) imputation

## References

- Jolani *et al.* Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Statistics in Medicine* 2015.
- Resche-Rigon *et al.* Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine* 2013.
- Burgess *et al.* Combining multiple imputation and meta-analysis with individual participant data. *Statistics in Medicine* 2013.



# Dealing with heterogeneity

## 1. Validation of existing model(s)

### Recommendations

- Investigate whether model performance is adequate and consistent across populations/subgroups/settings
- Investigate the influence of specific study characteristics (e.g. case-mix differences)
- Traditional meta-analysis methods can be implemented

### Example

- EPIC-InterAct IPD-MA
- Validation of existing models to predict the development of type 2 diabetes in general population
  - Evaluation of performance stratified across countries
  - Random effects meta-analysis to summarize performance

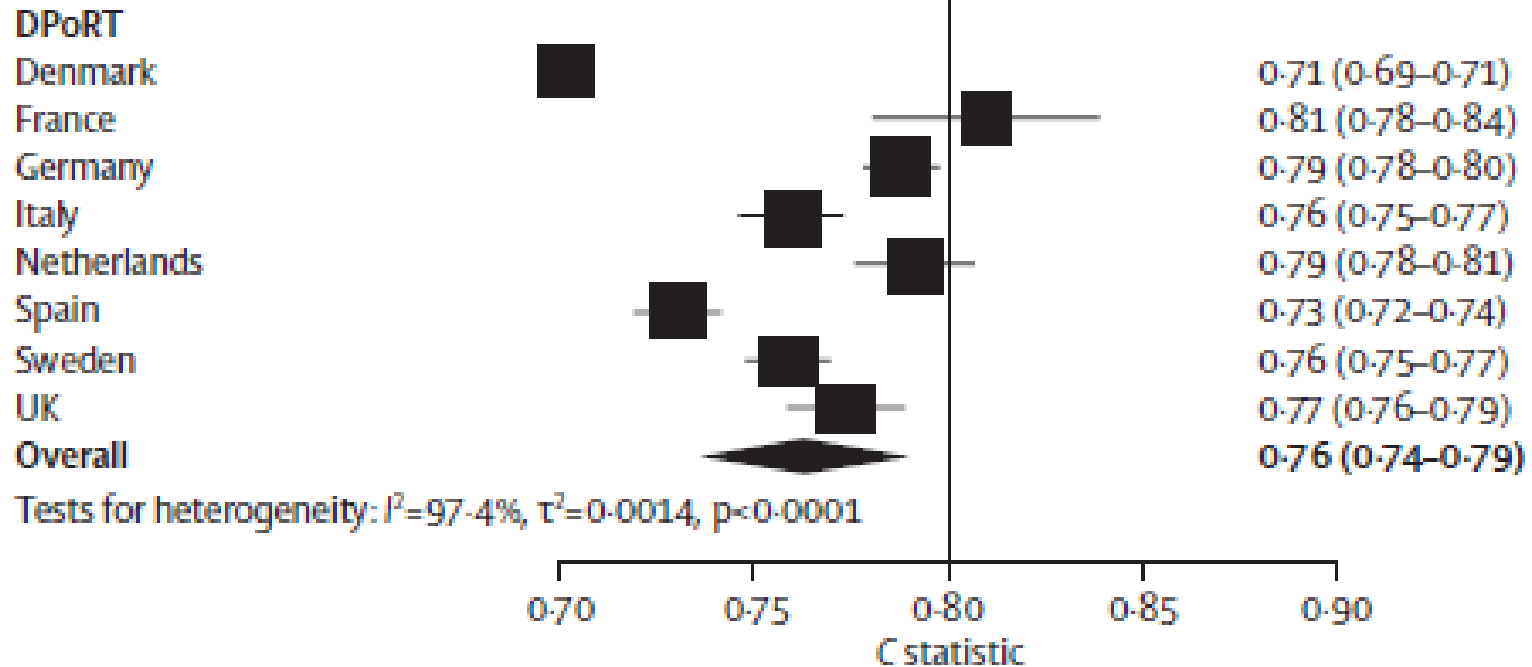


# Dealing with heterogeneity

## 1. Validation of existing model(s)

### Discrimination of model "DPoRT"

(overall and by country)



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

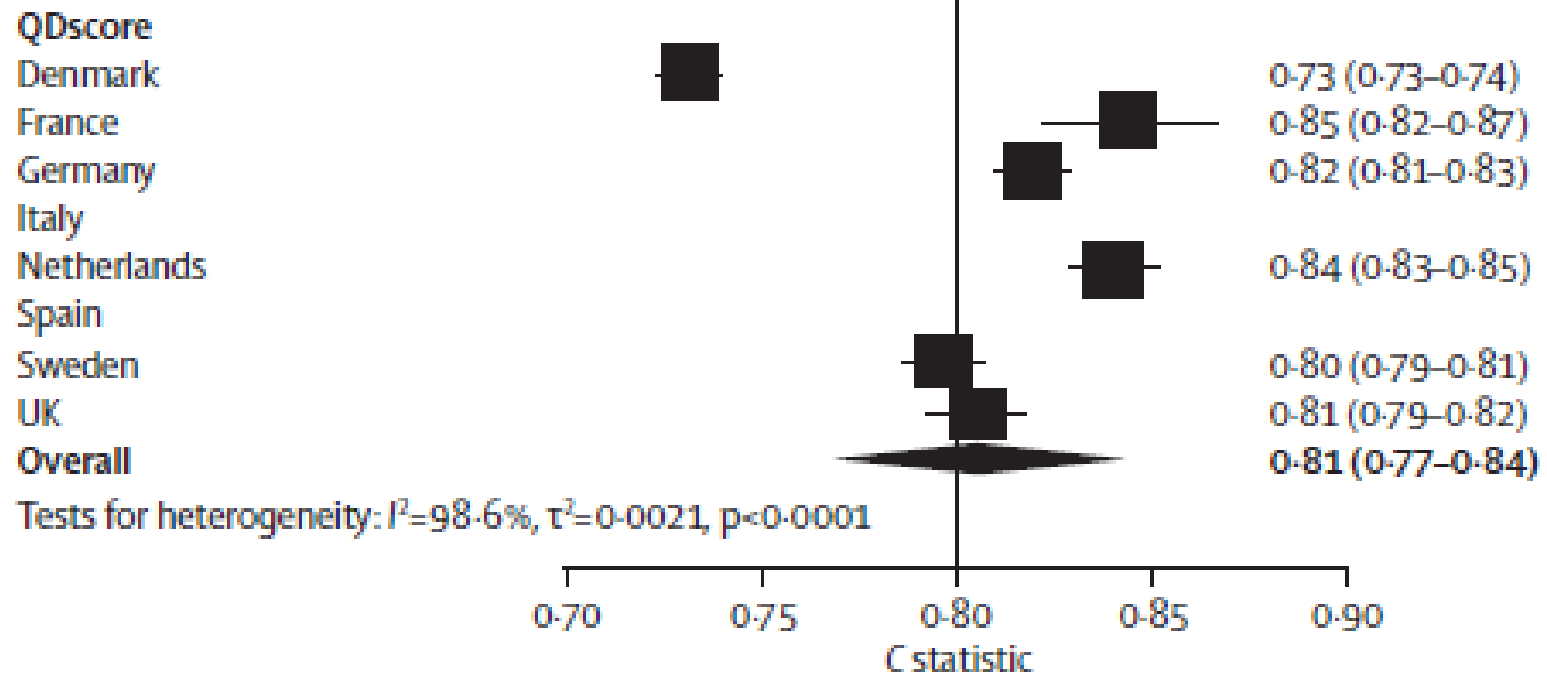


# Dealing with heterogeneity

## 1. Validation of existing model(s)

### Discrimination of model "QDscore"

(overall and by country)



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



# Dealing with heterogeneity

## 2. Tailoring of existing model(s)

### Recommendations

Tailor the validated model(s) if their performance substantially differs across studies/populations/settings

### Example

- Validation and updating of seks-specific Framingham risk equation for coronary heart disease and stroke
  - Adjust the model for baseline survival
  - Adjust the model for mean predictor values
  - Re-estimate country-specific predictor effects
- Results updated model
  - Poor discrimination
  - Improved calibration in a European population of middle-aged men



# Dealing with heterogeneity

## 2. Tailoring of existing model(s)

Validation and updating of seks-specific Framingham risk equation for coronary heart disease and stroke

	E:O ratio			C statistic		
	O	R	L	O	R	L
PRIME-total	1.94	0.98	1.00	0.68	0.68	0.68
PRIME-France	2.23	0.99	1.00	0.67	0.67	0.68
PRIME-Ireland	1.42	0.99	1.00	0.67	0.67	0.67

Outcome: CHD & Stroke, O=original, R=recalibrated, L=local model

Ref: Majed et al. *Preventive Medicine* 2008 **57**.



# Dealing with heterogeneity

## 3. Examining added value

### Recommendations

- Verify whether the added predictive value substantially differs across the included studies of the IPD-MA
- Evaluate under which circumstances and in which types of individuals/settings the predictor can be used as an addition to existing predictors or models

### Example

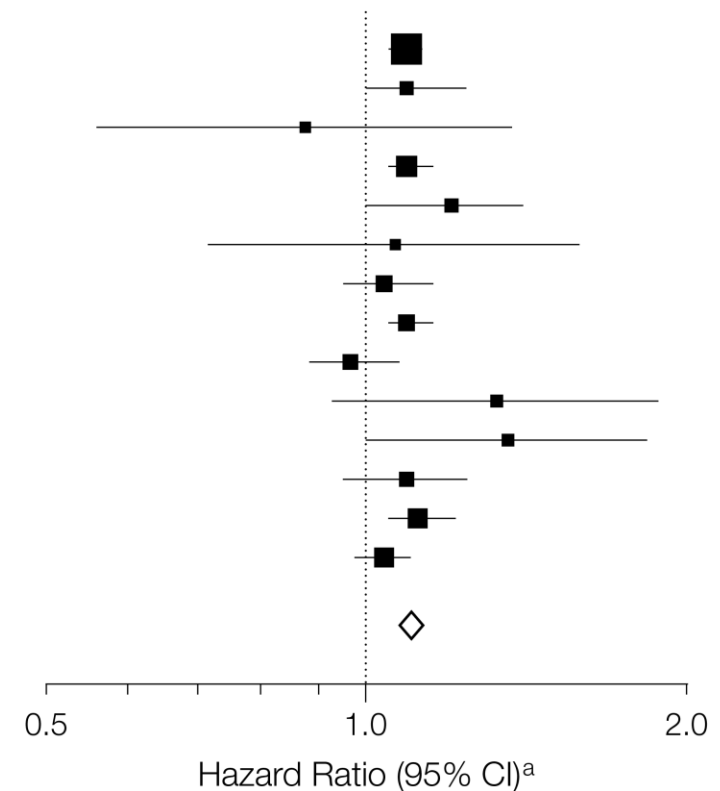
- Prediction of 10-year risk of first-time MI or stroke
- Investigation of added value CIMT above Framingham Risk Score



# Dealing with heterogeneity

## 3. Examining added value

Source	Contribution to Total USE-IMT Population, % of Total	Hazard Ratio (95% CI) <sup>a</sup>
ARIC, <sup>25</sup> 1994	31	1.11 (1.08-1.14)
CAPS, <sup>26</sup> 2006	8	1.10 (0.99-1.23)
Charlottesville, <sup>27</sup> 2006	1	0.88 (0.56-1.36)
CHS, <sup>28</sup> 2007	7	1.11 (1.06-1.16)
FATE, <sup>8</sup> 2011	3	1.20 (1.01-1.42)
Hoorn Study, <sup>29</sup> 2003	1	1.07 (0.72-1.59)
KIHD, <sup>30</sup> 1991	2	1.05 (0.96-1.16)
Malmö, <sup>31</sup> 2000	10	1.10 (1.04-1.17)
MESA, <sup>32</sup> 2007	13	0.98 (0.89-1.08)
Nijmegen Study, <sup>33</sup> 2009	3	1.34 (0.94-1.90)
NOMAS, <sup>34</sup> 2007	2	1.36 (0.99-1.85)
OSACA2 Study, <sup>35</sup> 2007	1	1.09 (0.96-1.24)
Rotterdam Study, <sup>36</sup> 1997	8	1.13 (1.06-1.20)
Tromsø Study, <sup>37</sup> 2000	9	1.04 (0.98-1.10)
$I^2 = 12.30\%$ ; Q test for heterogeneity, $P = .24$		1.09 (1.07-1.12)





# Dealing with heterogeneity

## 3. Examining added value

### Results

- No evidence for heterogeneity
- Small improvement in 10-year risk prediction

### Conclusion

- The addition of CIMT on top of FRS is unlikely to be of clinical importance



# Dealing with heterogeneity

## 4. Developing and directly validating a new model

### Caveats

- Model parameters may 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.



# Dealing with heterogeneity

## 4. Developing and directly validating a new model

### Recommendations from Ahmed et al.

- **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**



# Dealing with heterogeneity

## 4. Developing and directly validating a new 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,<sup>a,\*†</sup> Karel G. M. Moons,<sup>a</sup> Ikhlaaq Ahmed,<sup>b</sup>  
Hendrik Koffijberg<sup>a</sup> and Richard David Riley<sup>b</sup>

**Step 1:** modeling of intercept and predictor effects

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

**Step 3:** model evaluation



# Dealing with heterogeneity

## 4. Developing and directly validating a new model

### Step 1: model development

Different choices to combine IPD

- Merge all data into one big dataset and ignore heterogeneity
- Allow heterogeneous baseline risk across studies
  - by assuming random effects distribution for the intercept terms
  - By estimating study-specific intercept terms
- Advanced modeling of predictor effects is also possible
  - Nonlinear effects
  - Interaction terms



# Dealing with heterogeneity

## 4. Developing and directly validating a new model

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

- Average intercept versus population-specific intercept
- Propose which intercept term to use in new populations

### **Step 3: model evaluation to check whether**

- Modeling of predictors and intercept is adequate
- Strategy for choosing intercept term in new study population is adequate
- Model performance is consistently well across studies
  - Discrimination
  - Calibration



# Dealing with heterogeneity

## 4. Developing and directly validating a new model

### Example

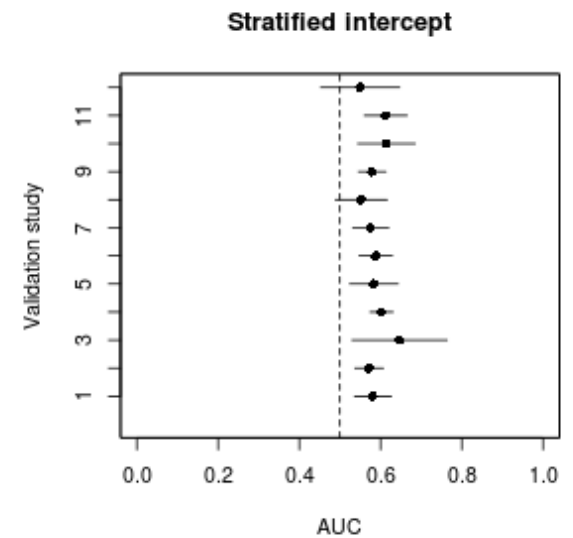
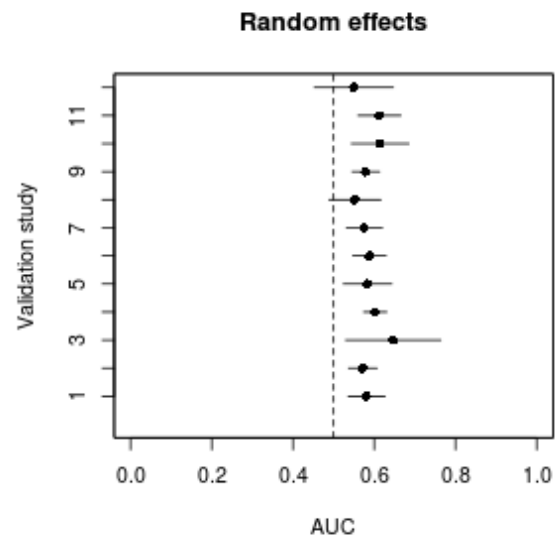
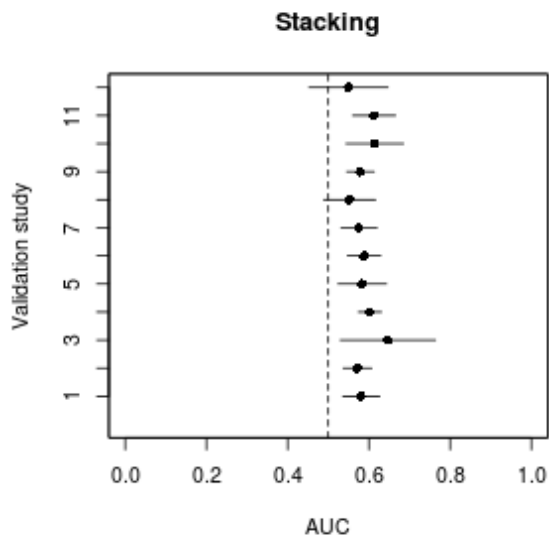
- Diagnosis of deep vein thrombosis (DVT)
  - IPD-MA of 12 studies
  - 10,014 patients (1,897 with DVT)
  - Focus on 2 homogeneous predictors: sex & recent surgery
- Comparison of 3 strategies
  - **Stacking**, ignore clustering of subjects within studies
  - **Random effects modeling** on intercept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



# Dealing with heterogeneity

## 4. Developing and directly validating a new model

### Model discrimination



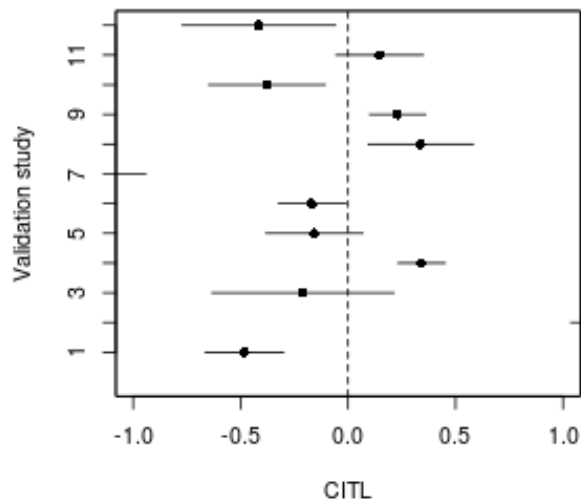


# Dealing with heterogeneity

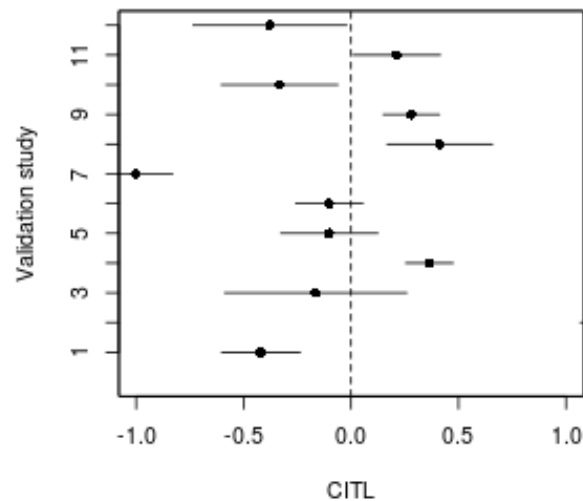
## 4. Developing and directly validating a new model

### Model calibration

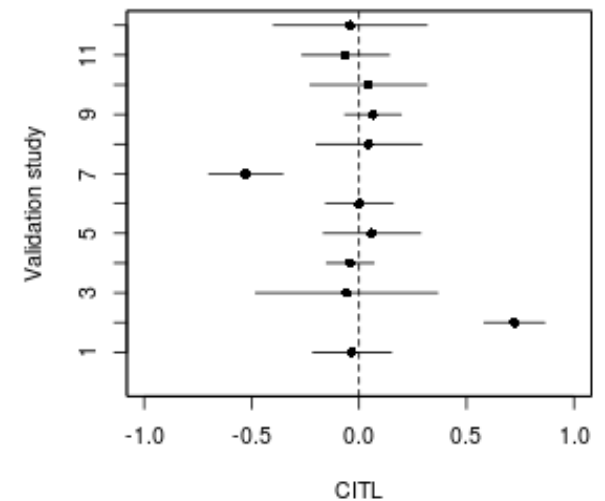
Stacking



Random effects



Stratified Intercept



# Dealing with heterogeneity

## 4. Developing and directly validating a new model

**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



# Combining IPD and AD

Beyond the scope of this workshop!

## References

- Debray *et al.* Meta-analysis and aggregation of multiple published prediction models. *Statistics in Medicine* 2014
- Debray *et al.* Aggregating published prediction models with individual participant data: a comparison of different approaches. *Statistics in Medicine* 2012
- Debray *et al.* Incorporating published univariable associations in diagnostic and prognostic modeling. *BMC Medical Research Methodology* 2012
- Steyerberg *et al.* Prognostic models based on literature and individual patient data in logistic regression analysis. *Statistics in Medicine* 2000.



# Take home messages

## Major advantages IPD-MA

- 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

- IPD-MA no panacea against poorly designed primary studies
  - Prospective multi-center studies remain important
- 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



# Take home messages

## Reasons to be optimistic



GUIDELINES AND GUIDANCE

# Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use

**Thomas P. A. Debray<sup>1,2\*</sup>, Richard D. Riley<sup>3</sup>, Maroeska M. Rovers<sup>4</sup>, Johannes B. Reitsma<sup>1,2</sup>, Karel G. M. Moons<sup>1,2</sup>, Cochrane IPD Meta-analysis Methods group<sup>†</sup>**

**1** Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands, **2** The Dutch Cochrane Centre, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands, **3** Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, The United Kingdom, **4** Radboud Institute for Health Sciences, Radboudumc Nijmegen, The Netherlands

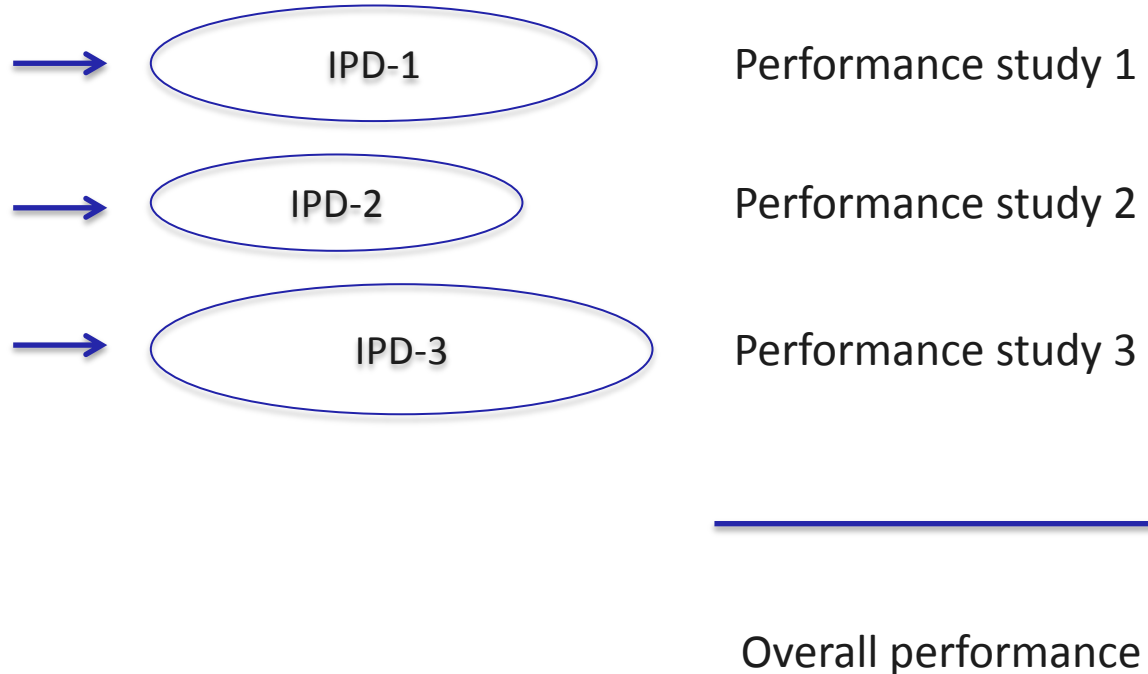


CrossMark



# TYPE I: VALIDATION OF EXISTING MODEL(S)

Existing (published) model(s)  
 $y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$



## Output:

What is the overall performance?

How large is the heterogeneity?

What are drivers of heterogeneity?

Competing models: difference in performance?





## TYPE II: TAILORING EXISTING MODEL

Existing (published) model(s)  
 $y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$



+ updating 1  
+ refitting 1



+ updating 2  
+ refitting 2



+ updating 3  
+ refitting 3

---

Updating needed?  
Refitting needed?

### Output:

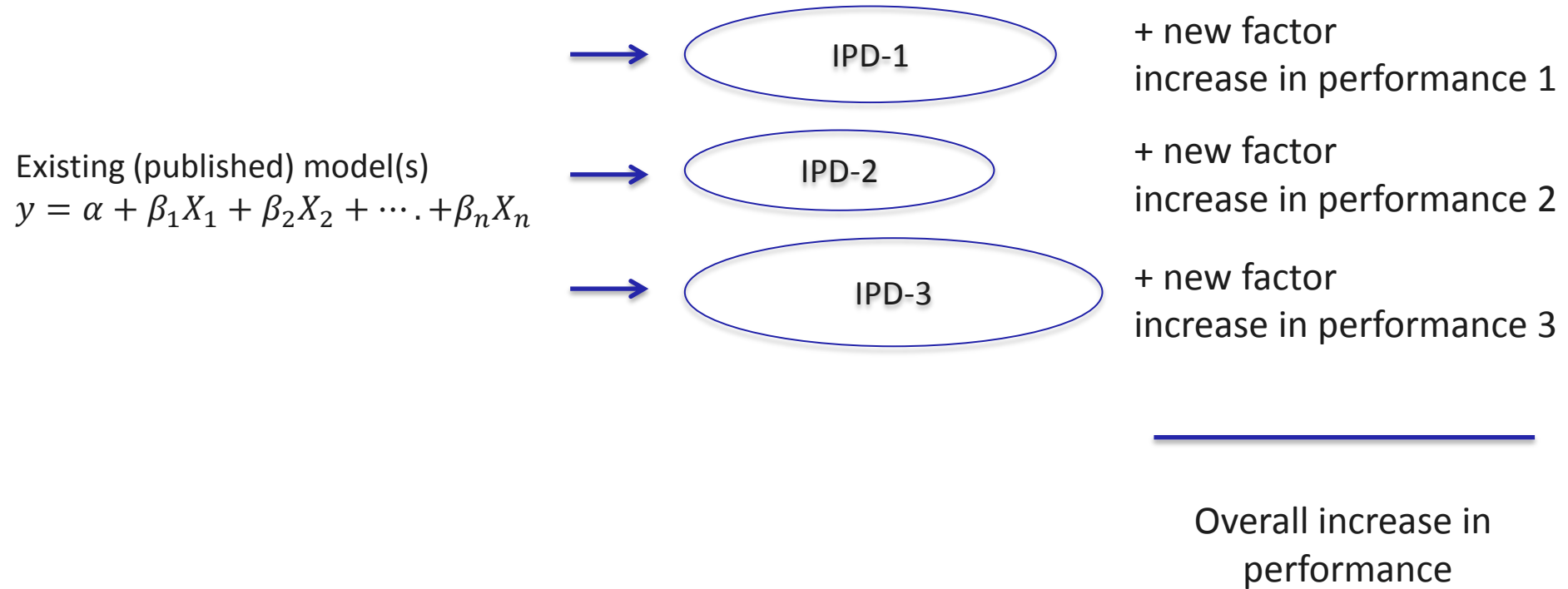
Updating needed?

For which setting / populations

Updated model(s)



## TYPE III: EXAMINING ADDED VALUE



### Output:

What is the overall added value?

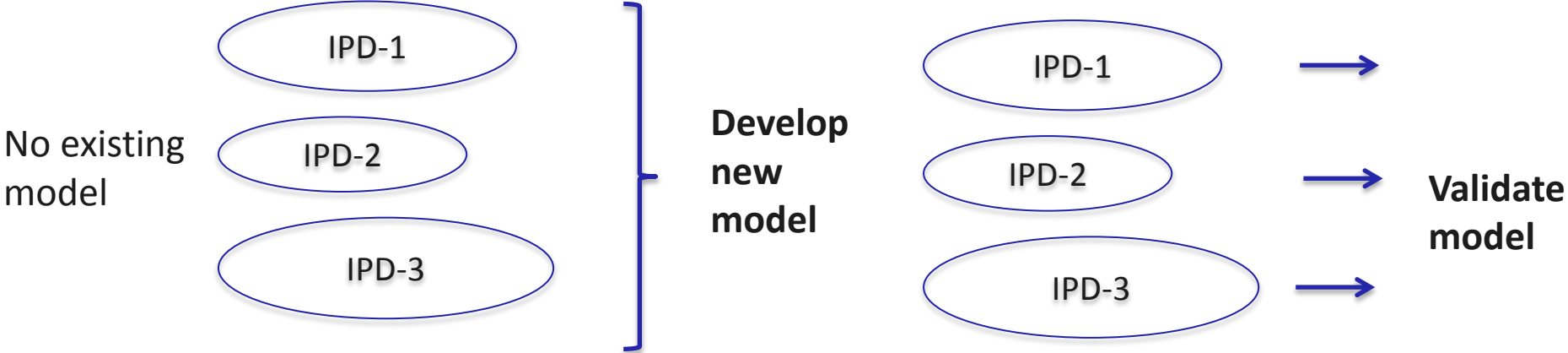
Heterogeneity in added value?

Drivers of heterogeneity?

What is the updated model?



# TYPE IV: DEVELOPMENT NEW MODEL AND VALIDATION



**Output:**  
New model / tailored models

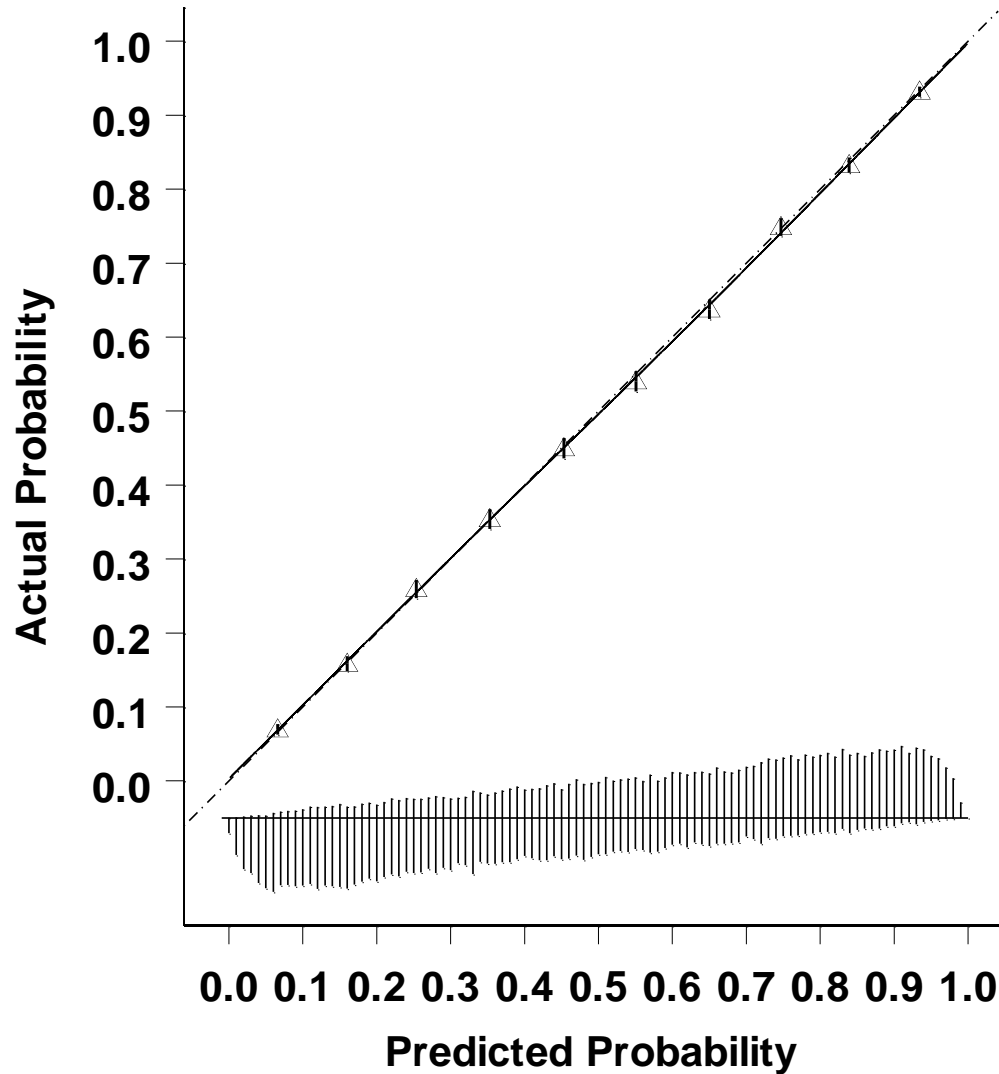


# Prediction model performance measures

- **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 Reclassification Index, Net Benefit, ...



# Calibration plot

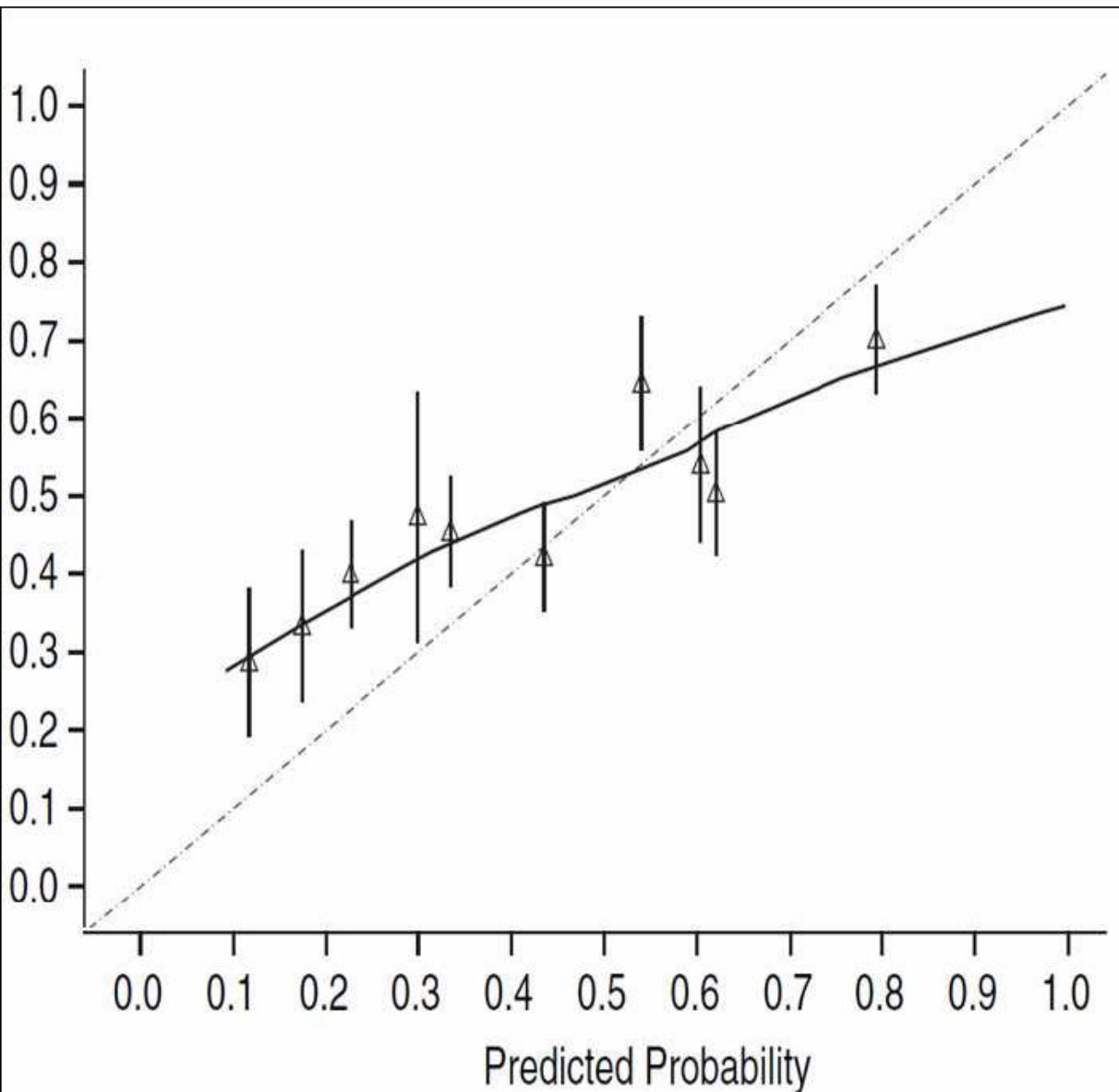


**Ideal calibration**  
Observed versus  
expected risk (O/E) = 1

Slope = 1



# External validation: typical result



- Slope plot  $< 1.0$ 
  - Low prob too low
  - High prob too high
    - Overfitted
- AUC = 0.63 (was 0.75)

