UMC Utrecht
Julius Center

# Development of prediction models in big datasets

Thomas Debray, PhD
Julius Center for Health Sciences and Primary Care
Cochrane Netherlands

Cochrane

University Medical Center Utrecht

# Model development using big datasets

# Model development using big datasets

**Main opportunities**

- Increase total sample size
- Increase available case-mix variability
- Ability to standardize analysis methods across IPD sets
- Ability to investigate more complex associations
- Ability to directly validate developed prediction models across a wide range of populations and settings
- Ability to evaluate generalizability of the model

# Model development using big datasets

**So, let's pool our IPD and 'launch' the analysis?**

# Model development using big datasets

**Wait… which analysis?**

# Current practice

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

BMC
Medical Research Methodology

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*]

**Abstract**

**Background:** Risk prediction models estimate the risk of developing future outcomes for individuals based on one or more underlying characteristics (predictors). We review how researchers develop and validate risk prediction models within an individual participant data (IPD) meta-analysis, in order to assess the feasibility and conduct of the approach.

**Methods:** A qualitative review of the aims, methodology, and reporting in 15 articles that developed a risk prediction model using IPD from multiple studies.

**Results:** The IPD approach offers many opportunities but methodological challenges exist, including: unavailability of requested IPD, missing patient data and predictors, and between-study heterogeneity in methods of measurement, outcome definitions and predictor effects. Most articles develop their model using IPD from *all* available studies and perform only an internal validation (on the same set of data). Ten of the 15 articles did not allow for any study differences in baseline risk (intercepts), potentially limiting their model's applicability and performance in some populations. Only two articles used external validation (on different data), including a novel method which develops the model on all but one of the IPD studies, tests performance in the excluded study, and repeats by rotating the omitted study.

**Conclusions:** An IPD meta-analysis offers unique opportunities for risk prediction research. Researchers can make more of this by allowing separate model intercept terms for each study (population) to improve generalisability, and by using 'internal-external cross-validation' to simultaneously develop and validate their model. Methodological challenges can be reduced by prospectively planned collaborations that share IPD for risk prediction.

**Keywords:** Meta-analysis, Prognostic factor, Prognosis, Individual participant (patient) data, Review, Reporting

# Current practice

**Analysis methods** (review of 15 IPD-MA)

- 10 articles pooled all the IPD into one big dataset and analysed it ignoring clustering of patients

  most of the studies ignored clustering

- 4 articles used a one-stage approach accounting for clustering (e.g. Stratification of intercept term)

- 1 article used a two-stage approach accounting for clustering

**Ref**: Ahmed I, et al. Developing and Validating Risk Prediction Models in an Individual Participant Data Meta-Analysis. *BMC Medical Research Methodology* 2014.

# Current practice

**Investigation of heterogeneity**

- 12 articles did not consider heterogeneity in predictor effects
- 1 article investigated interaction terms between study and each predictor
- 1 article investigated heterogeneity using the I2 statistic
- 1 article investigated heterogeneity using Chi-square test

# Current practice

Out of 15 IPD-MA

- 10 studies completely ignore potential of heterogeneity
- 4 studies allow for heterogeneity in baseline risk some efforts
- 1 study allows for heterogeneity in predictor effects

However, using random effects methods may not even be appropriate to deal with heterogeneity!

# Problems with meta-analysis methods

random effects allow to account for potential presence of heterogeneity they do not tell you how to use the model in a different setting - which value should you take in next study?

## Random effects summaries are of limited value

- Predictor effects and/or baseline risk 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.

how to handle it in a new study

# Problems with meta-analysis methods

Statistics in Medicine

Explore this journal >

Research Article

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

Thomas P.A. Debray ✉, Karel G.M. Moons, Ikhlaaq Ahmed, Hendrik Koffijberg, Richard David Riley


View issue TOC
Volume 32, Issue 18
15 August 2013
Pages 3158–3180

**First published:** 11 January 2013 Full publication history

**DOI:** 10.1002/sim.5732 View/save citation

**Cited by:** 19 articles Refresh Citing literature

Am score 5

✉ Correspondence to: Thomas P. A. Debray, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Str[illegible] 6.131, PO Box 85500, 3508GA Utrecht, The Netherlands.
E-mail: T.Debray@umcutrecht.nl

# 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**

  to investigate the generalizability

# The framework

**Step 1**: Different choices for combining IPD

- Merge all data into one big dataset and ignore heterogeneity
- Allow heterogeneous baseline risk across studies

by

  - assume random effects distribution for the intercept terms
  - estimate study-specific intercept terms

- Advanced modeling of predictor effects is also possible

consider

  - Nonlinear effects
  - Interaction terms

# The framework

how to implement prediction models for a new population

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

- Average intercept term (e.g. pooled estimate)
- Updating of intercept term (requires patient-level data)
- Use intercept of included study (e.g. based on outcome occurrence)

which intercept term should be used?
Average?

locally available data

Propose which intercept term to use in new populations

**!!** More difficult in case of heterogeneous predictor effects

# The framework

evaluation if the model performs well in new population

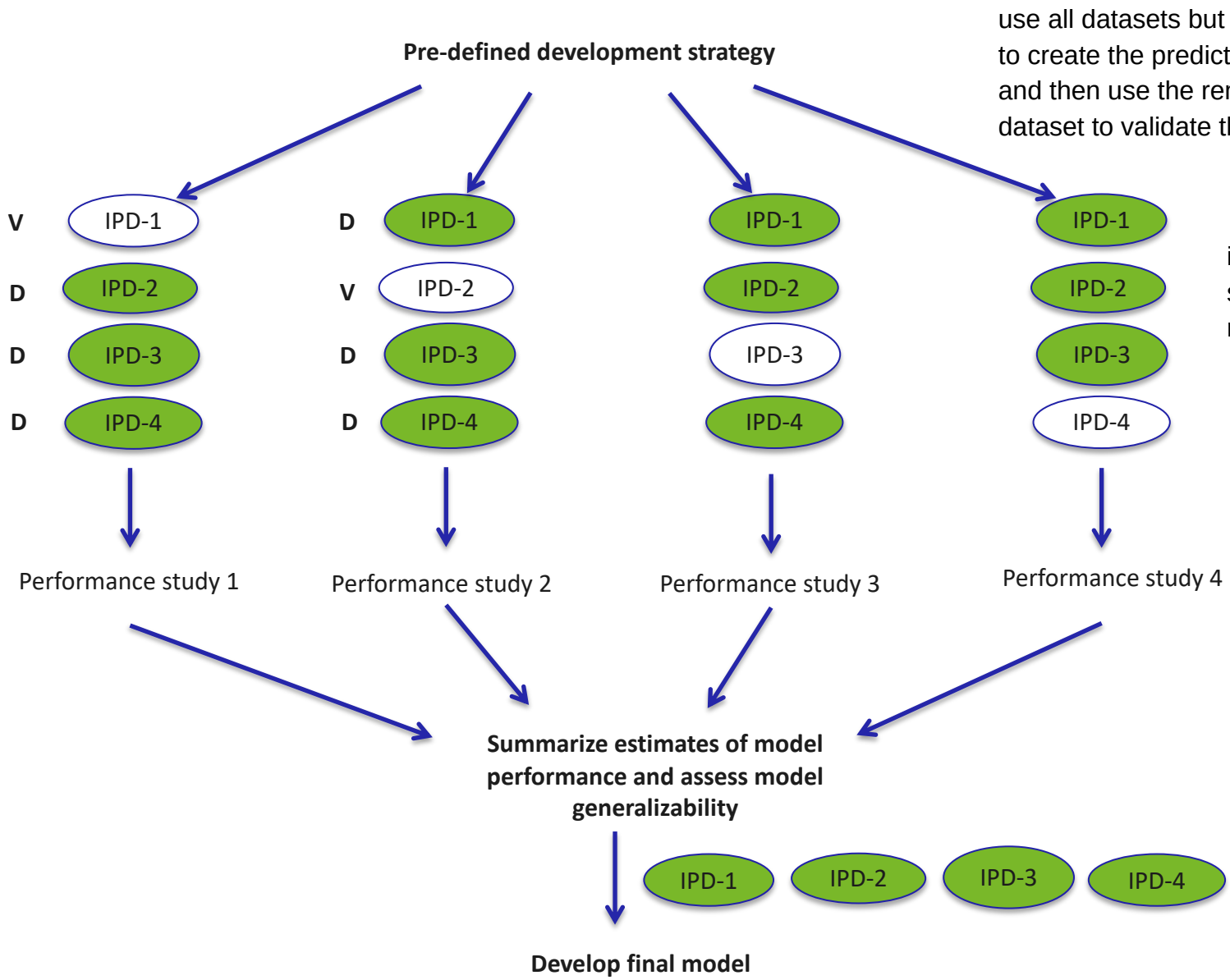
**Step 3:** Model evaluation to check whether…

- Strategy for estimating predictors and intercept is adequate ?
- Strategy for choosing intercept term (and predictor effects) in new study population is adequate ?
- Model performance is consistently well across studies
  - Discrimination
  - Calibration

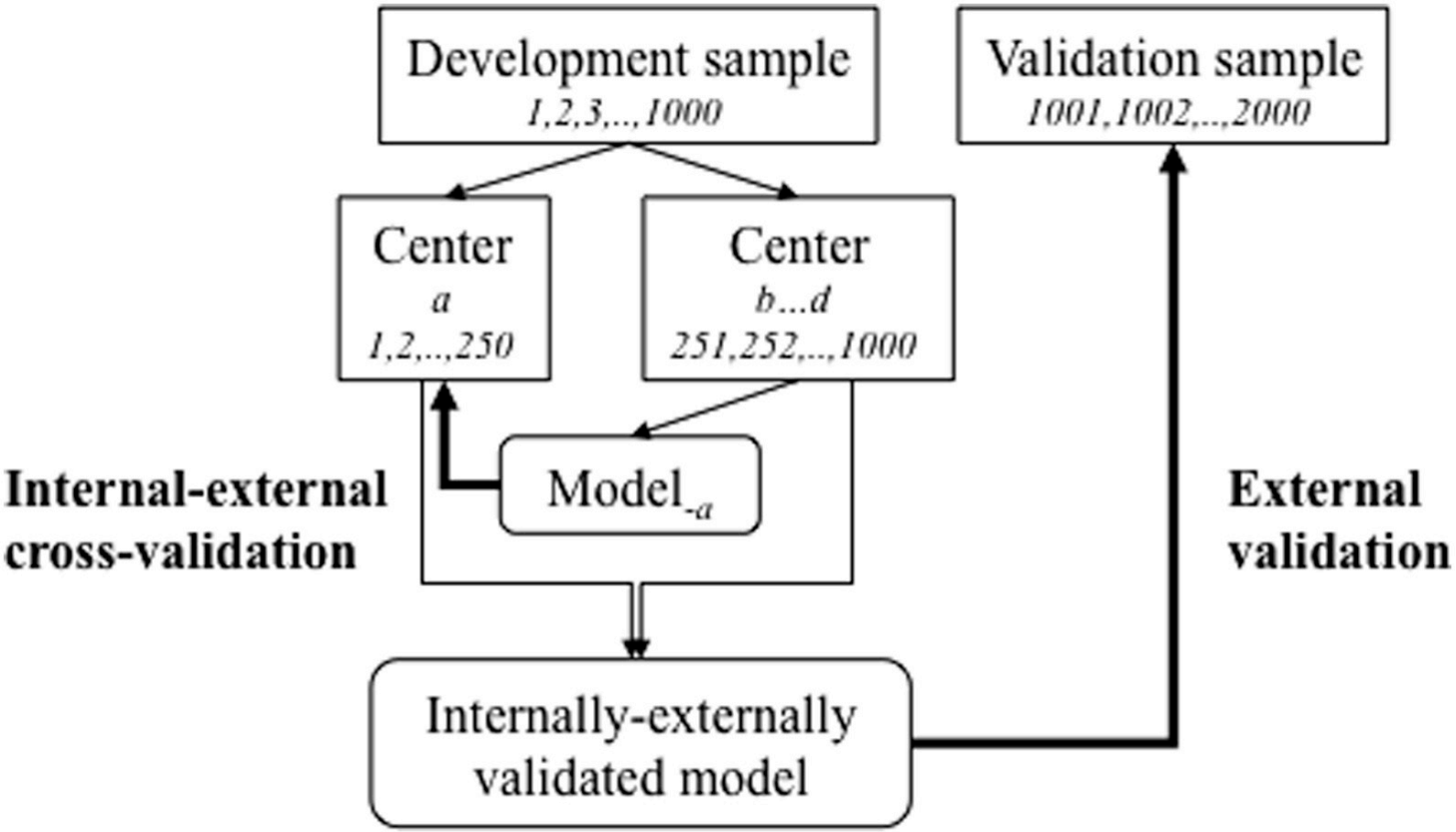
can be implemented by using

=> Use of internal-external cross-validation

# Internal-external cross-validation



# Internal-external cross-validation




**Ref:** 10.1016/j.jclinepi.2015.04.005

# Formally comparing different strategies

end up with multiple validations of a single development strategy

- Meta-analyze estimates of model performance to compare
  (See also MSc course SR&MA of prognostic studies)
  - Compare summary estimates better results?
  - Compare prediction intervals

- Rank different strategies by their overall performance
  - Calculate the joint probability that, in a new population, model performance will meet certain criteria
    (e.g. C-statistic > 0.7 and cal. slope between 0.9 and 1.1)
  - The strategy with the largest probability will be ranked first
  - This requires reliable prediction intervals!!

**Ref:** 10.1016/j.jclinepi.2015.05.009

# Example 1

**Diagnosis of deep vein thrombosis** (N=12)

Strategies evaluated:

logistic regression model with just two predictors

- Inclusion of 2 predictors (gender & recent surgery)
- Modelling of intercept term

compare for modeling

  - Re-estimate in each validation study
  - Random effects modeling
  - Stratified intercept term
- Model implementation
  - Average intercept
  - Select estimated intercept term based on prevalence

Assessment of AUC and calibration-in-the-large (CITL)

compare the different strategies depending on their performance by checking c-statistics & calibration-in-the-large to compare

# Example 1

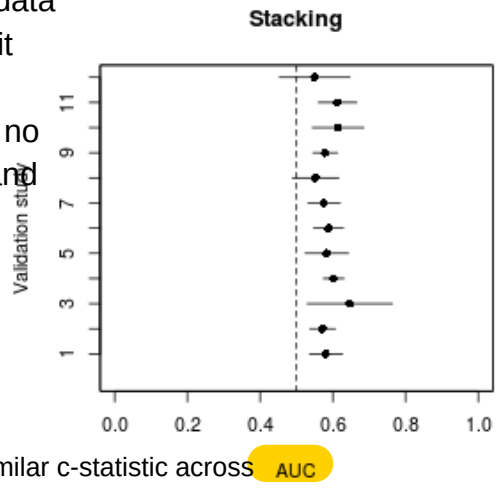
estimated logistic regression model with 2 predictors with logistic regression model in 11 studies and evaluate with the 12 study (iterate will every study was left out once)

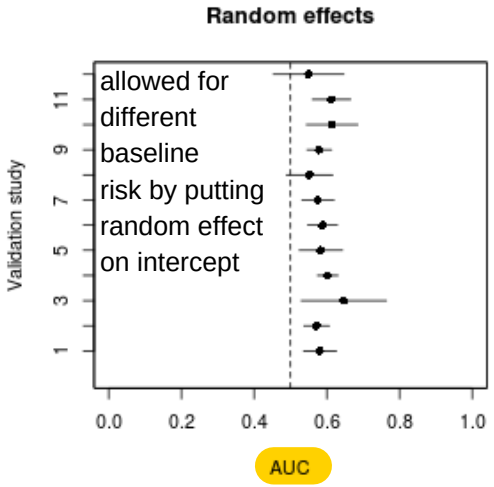
estimated separated intercept term for each study and then for the whole study select intercept term for the included studies depending on the similarity in outcome prevalence

## Diagnosis of deep vein thrombosis (N=12)

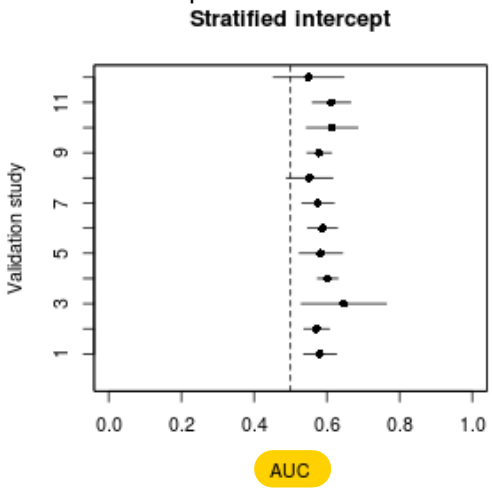
allow for het. intercept terms

merged all the data and pretended it was all one population with no heterogeneity and baseline risk

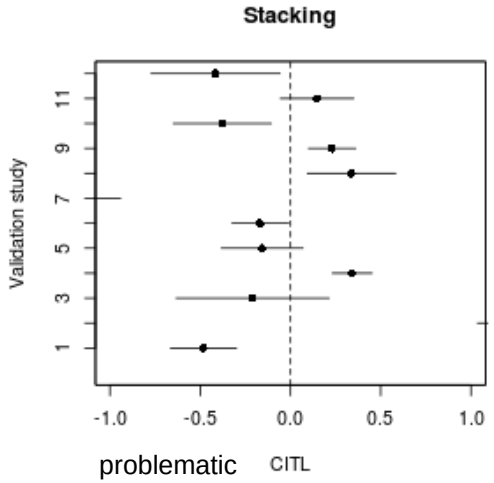


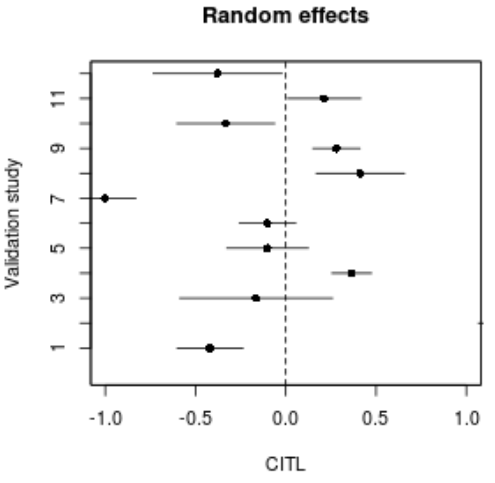



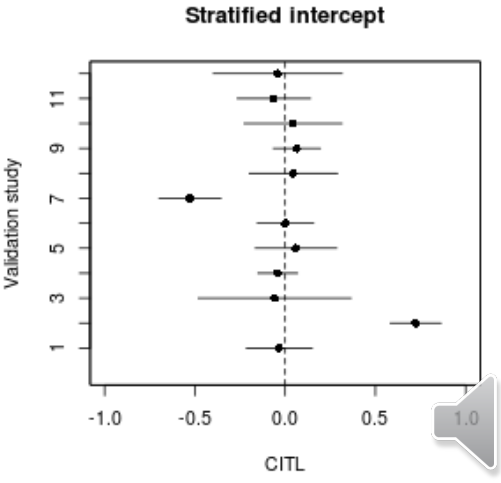

allowed for different baseline risk by putting random effect on intercept




AUC relatively similar c-statistic across the whole studies



# Example 1

impact of different model strategies on heterogeneity

## **Diagnosis of deep vein thrombosis** (N=12)

**Table 1.** Trivariate meta-analysis results[a] for the calibration and discrimination performance of the DVT model for each implementation strategy

| Strategy | Validation statistic | Estimate (95% CI) of mean, $\mu$ | 95% Prediction interval | $I^2$ (%) | $\hat{\tau}$ (95% CI) |
|---|---|---|---|---|---|
| Strategy (1): | Calibration-in-the-large | −0.130 (−0.185, −0.075) | −0.195, −0.065 | 1 | 0.008 |
| Develop using logistic regression and | Calibration slope | 0.975 (0.855, 1.097) | 0.597, 1.353 | 57 | 0.158 |
| implement with intercept estimated | Log(expected/observed) | 0.086 (0.047, 0.124) | 0.041, 0.128 | 0 | 0.0009 |
| in external validation study | C statistic | 0.687 (0.670, 0.704) | 0.645, 0.729 | 34 | 0.017 |
| Strategy (2): | Calibration-in-the-large | −0.004 (−0.313, 0.305) | −1.240, 1.232 | 97 | 0.532 |
| Develop using logistic regression | Calibration slope | 0.980 (0.853, 1.107) | 0.585, 1.375 | 59 | 0.165 |
| and implement with average study | Log(expected/observed) | 0.022 (−0.206, 0.250) | −0.887, 0.931 | 97 | 0.391 |
| intercept taken from developed | C statistic | 0.687 (0.669, 0.705) | 0.640, 0.734 | 37 | 0.019 |
| model | | | | | |
| Strategy (3): | Calibration-in-the-large | 0.047 (−0.120, 0.214) | −0.584, 0.678 | 89 | 0.270 |
| Develop using logistic regression and | Calibration slope | 0.976 (0.851, 1.102) | 0.578, 1.375 | 59 | 0.167 |
| implement with intercept taken from | Log(expected/observed) | −0.029 (−0.150, 0.093) | −0.485, 0.427 | 89 | 0.195 |
| a study used in development data | C statistic | 0.687 (0.669, 0.705) | 0.640, 0.734 | 38 | 0.019 |
| with a similar prevalence | | | | | |

*Abbreviations*: DVT, deep vein thrombosis; CI, confidence interval.

[a] A trivariate meta-analysis was fitted to calibration-in-the-large, calibration slope, and C statistic and then again for log(expected/observed), calibration slope, and C statistic. Perfect negative correlation between calibration-in-the-large and expected/observed within studies prevents all four measures being analyzed together (due to collinearity). Results were practically the same for calibration slope and C statistic, regardless of the trivariate model fitted.

Strat1: substantial estimate of between study heterogeneity in model performance

Strat 2: amount of het. decreases slightly but by choosing (Strat 3) revise prediction model and select an intercept term based on similarity outcome prevalence the heterogeneity is much lower than the other two strategies (like forest plot)

# Example 1

instead of inspecting individual estimates of performance & heterogeneity once can figure out model performance depending on different criteria e.g here a probability that the prediction model for a new study would have a calibration slope within a certain range and c-statistics above a certain value
- compare the probabilities to define which strategy is most appropriate

## **Diagnosis of deep vein thrombosis** (N=12)

**Table 2.** Joint predicted probability of "good" discrimination and calibration performance of the DVT model for each of the three implementation strategies, derived using the multivariate meta-analysis results for the C statistic and calibration slope shown in Table 1

| | | Joint predicted probability of meeting criteria in new population | | |
|---|---|---|---|---|
| Calibration slope required | Minimum C statistic required | Strategy (1): Develop using logistic regression and implement with intercept estimated in external validation study | Strategy (2): Develop using logistic regression and implement with average study intercept taken from developed model | Strategy (3): Develop using logistic regression and implement with intercept taken from a study used in development data with a similar prevalence |
| 0.9–1.1 | 0.70 | 0.027 | 0.037 | 0.037 |
| 0.8–1.2 | 0.70 | 0.146 | 0.158 | 0.156 |
| 0.9–1.1 | 0.65 | 0.427 | 0.413 | 0.409 |
| 0.8–1.2 | 0.65 | 0.728 | 0.712 | 0.707 |

*Abbreviation*: DVT, deep vein thrombosis.

Notice that using the average intercept term is not problematic when the main focus is on calibration slope and C-statistic!

it does not account for calibration in the large!

# Example 2

consider the development of prediction model to estimate the risk of breast cancer

## Prognosis of breast cancer

- IPD from 8 cohort studies
- Sample size: 69 to 3,242 per study (total N=7,435)
- Event occurrence (total E=2,043)
- Median follow-up: 86.3 months

# Example 2

allowed for heterogeneity by stratifying the baseline hazard by country but assuming the predictor effect to be the same for all

## Model development strategy

- Common predictor effects with proportional hazards
- Backwards variable selection from 8 candidate predictors (using P>0.05 for exclusion)
- Royston-Parmar survival model with country-specific (but proportional) baseline hazard

# Example 2

**Strategies for model implementation**

1. Re-estimate adjustment factor for the baseline hazard in the validation sample

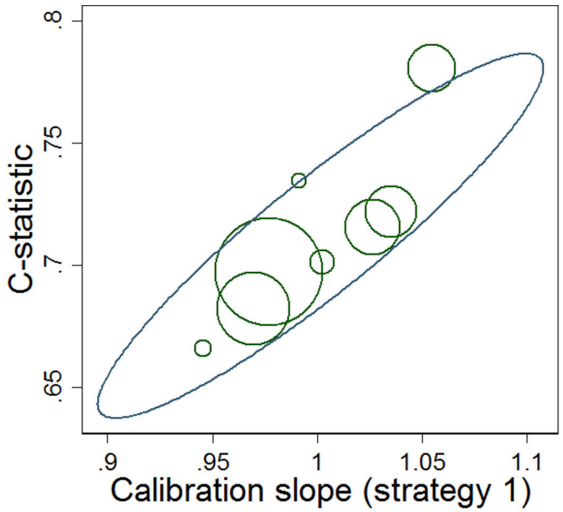
2. Use weighted average of estimated country-specific adjustment factors

3. Select an estimated adjustment factor from a country that is geographically the closest.

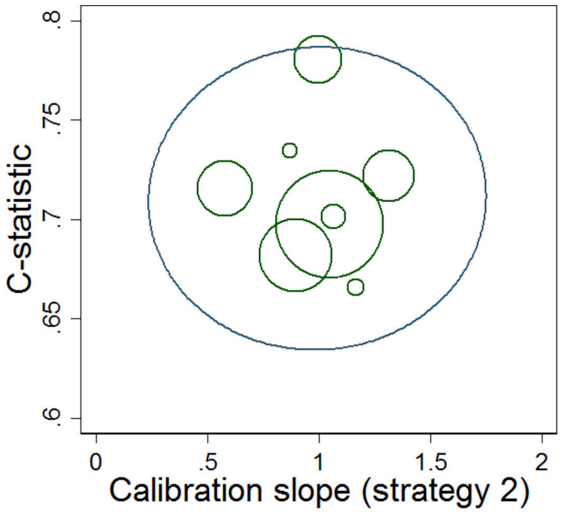
# Example 2

pooled estimates for calibration slope an c-statistic

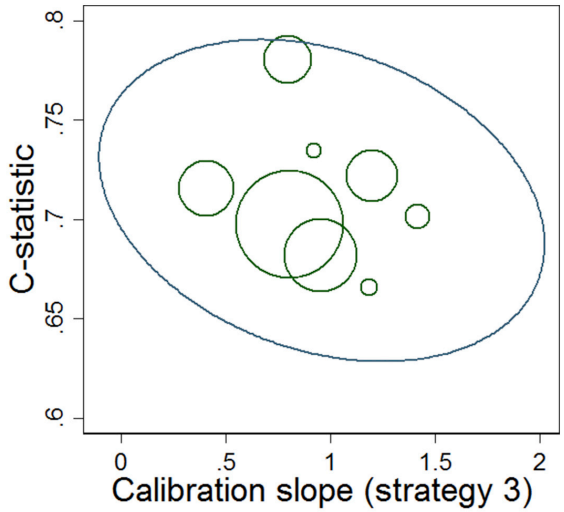
Develop using Royston-Parmar and re-estimate the baseline hazard in each validation study

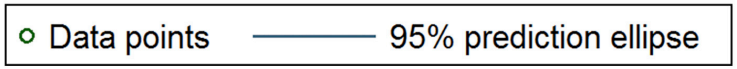



Develop using Royston-Parmar and implement with average baseline hazard from developed model




Develop using Royston-Parmar and implement with the estimated baseline hazard from the closest geographical country



# Example 2

again random effects meta analysis and calculate the probability of good performance
- joined probability of good performance can then be calculated for the overall performance (includes C-stat, D-stat and cal. slope)

**Table 3.** Trivariate random-effects meta-analysis results for calibration and discrimination performance of the breast cancer model for each implementation strategy

| Strategy | Validation statistic | Pooled estimate (95% CI) | 95% Prediction interval | *I* squared (%) | Estimate of $\tau$ | Joint probability of "good"[a] performance in a new population |
|---|---|---|---|---|---|---|
| Strategy (1): | Calibration slope | 1.003 (0.971, 1.036) | 0.927, 1.080 | 35 | 0.026 | 0.67 |
| Develop using Royston–Parmar and | C statistic | 0.711 (0.690, 0.733) | 0.657, 0.766 | 49 | 0.019 | |
| implement with baseline hazard | D statistic | 0.328 (0.215, 0.442) | −0.056, 0.713 | 87 | 0.146 | |
| estimated in validation study | | | | | | |
| Strategy (2): | Calibration slope | 0.994 (0.835, 1.153) | 0.411, 1.577 | 98 | 0.224 | 0.22 |
| Develop using Royston–Parmar model | C statistic | 0.711 (0.691, 0.732) | 0.662, 0.761 | 43 | 0.017 | |
| and implement with the estimated | D statistic | 0.332 (0.212, 0.452) | −0.080, 0.745 | 88 | 0.157 | |
| average baseline hazard from | | | | | | |
| developed model | | | | | | |
| Strategy (3): | Calibration slope | 0.961 (0.741, 1.181) | 0.148, 1.775 | 99 | 0.313 | 0.15 |
| Develop using Royston–Parmar model | C statistic | 0.710 (0.687, 0.734) | 0.653, 0.767 | 50 | 0.020 | |
| and implement with the estimated | D statistic | 0.330 (0.211, 0.450) | −0.068, 0.728 | 87 | 0.151 | |
| baseline hazard from the closest | | | | | | |
| geographical country | | | | | | |

*Abbreviation*: CI, confidence interval.

[a] Defined by a C statistic ≥0.7 and an calibration slope between 0.9 and 1.1.

Again, we should also take CITL into account!

# Example 3

**Prognosis of amyotrophic lateral disease**

- IPD-MA
  - 14 cohort studies (specialized ALS centres)
- Sample size
  - 190 to 1,936 per study (total N = 11,475)
- Composite endpoint
  - Non-invasive ventilation for more than 23h/day, or death
  - Total number of events E = 8,819
- Median follow-up: 97.5 months

Development of the NCALS model

# Example 3

THE LANCET
Neurology

Volume 17, Issue 5, May 2018, Pages 423-433

Articles

## Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model

Henk-Jan Westeneng MD [a], Thomas P A Debray PhD [b, c], Anne E Visser MD [a], Ruben P A van Eijk MD [a], James P K Rooney MSc [d], Andrea Calvo MD [e], Sarah Martin BSc [f], Prof Christopher J McDermott PhD [g], Alexander G Thompson BMBCh [h], Susana Pinto PhD [i], Xenia Kobeleva MD [j], Angela Rosenbohm MD [k], Beatrice Stubendorff PhD [l], Helma Sommer [m], Bas M Middelkoop [a], Annelot M Dekker MD [a], Joke J F A van Vugt PhD [a], Wouter van Rheenen MD [a] ... Prof Leonard H van den Berg MD [a]

# Example 3

## Prognosis of amyotrophic lateral disease

- Royston-Parmar survival model with country-specific (but proportional) baseline hazard

parameters of estimated regression model

| Variable | Value |
|---|---|
| $\gamma_0$ | -6·409 |
| $\gamma_1$ | 2·643 |
| $\gamma_2$ | -0·546 |
| $\gamma_3$ | 0·585 |
| $\beta_1$ (ALSFRS-R slope) | -1·837 |
| $\beta_2$ (Diagnostic delay) | -2·373 |
| $\beta_3$ (Age at onset) | -0·267 |
| $\beta_4$ (Forced vital capacity) | 0·477 |
| $\beta_5$ (Bulbar onset) | 0·269 |
| $\beta_6$ ('Definite' ALS*) | 0·233 |
| $\beta_7$ (Frontotemporal dementia) | 0·388 |
| $\beta_8$ (*C9orf72* repeat expansion) | 0·256 |

**Supplementary Table S15. Parameters of the final prediction model.** *According to the El Escorial criteria.

# Example 3

again iteratively estimated in all but one study and then validated with remaining study

## Internal-external cross-validation

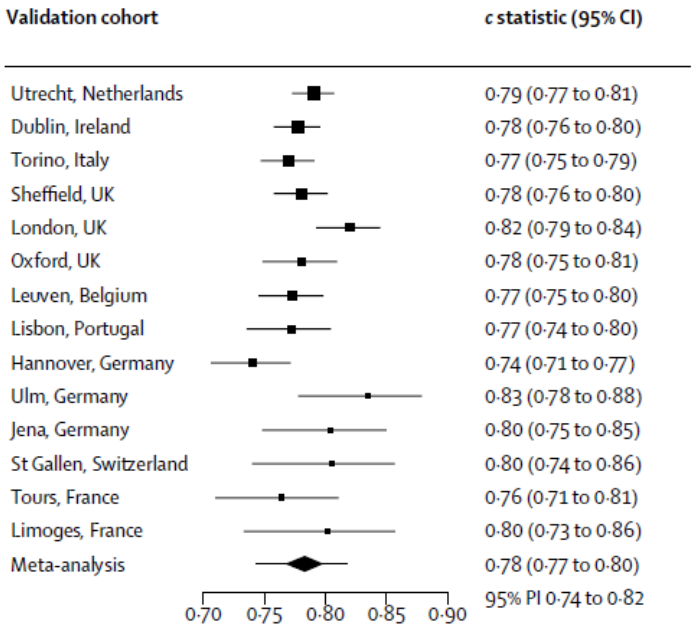
- c-statistic
  - Summary estimate: 0.78 (0.77 to 0.80)
  - 95% PI: 0.74 to 0.82

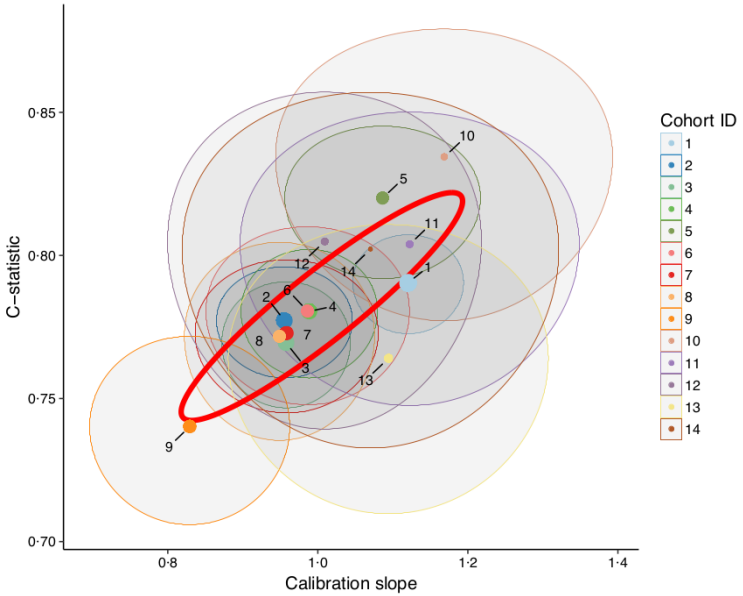
always with prediction intervals for each measure

- Calibration slope
  - Summary: 1.01 (0.95 to 1.07)
  - 95% PI: 0.83 to 1.18
- Calibration-in-the-large
  - Summary: -0.12 (-0.33 to 0.08)
  - 95% PI: -0.88 to 0.63

# Example 3

used estimates of meta-analysis to calculate the prob. of good performance



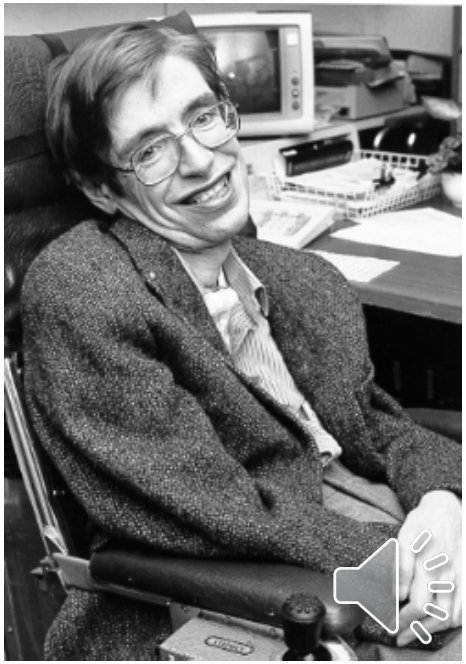

| Measure | Criteria | Prob. of "good" performance | Joint probability |
|---|---|---|---|
| C-statistic | > 0.70 | 100% | 98.3% |
| Calibration slope | 0.80 to 1.20 | 97.1% | |
| Calibration-in-the-large | -0.587 to 0.587 | 85.5% | |

# Example 3

**The life expectancy of Stephen Hawking, according to the ENCALS model**

*"Using publicly available data, we examined whether Professor Hawking's survival was as rare as his intellectual performance, or could be predicted solely based on his disease characteristics at diagnosis in 1963."*

- Predicted 10-year survival probability: 94%
- The IQR for his predicted survival lay between 1981 and 2011
- Young age of onset was the most important factor for his long survival



# Example 3

**The life expectancy of Stephen Hawking**, according to the ENCALS model

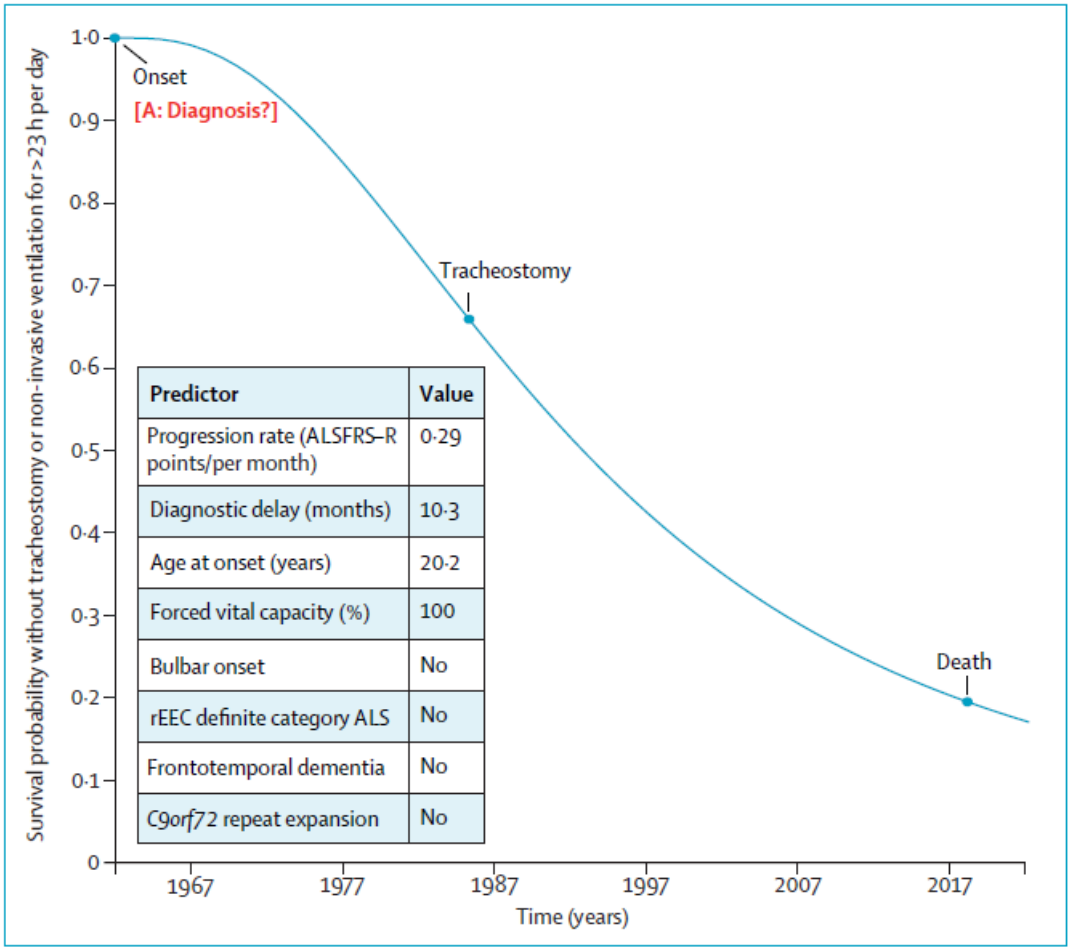



***Figure*: Personalised survival curve for Stephen Hawking**
Using publicly available data of disease characteristics at diagnosis in 1963, we were able to estimate the probability for survival without tracheostomy or non-invasive ventilation for more than 23 hours per day. The predictor values used for calculating the survival probability are summarised in the inset. ALS=amyotrophic lateral sclerosis. ALSFRS-R=revised ALS functional rating scale. rEEC=revised El Escorial criteria.

# Assessing heterogeneity in predictor effects

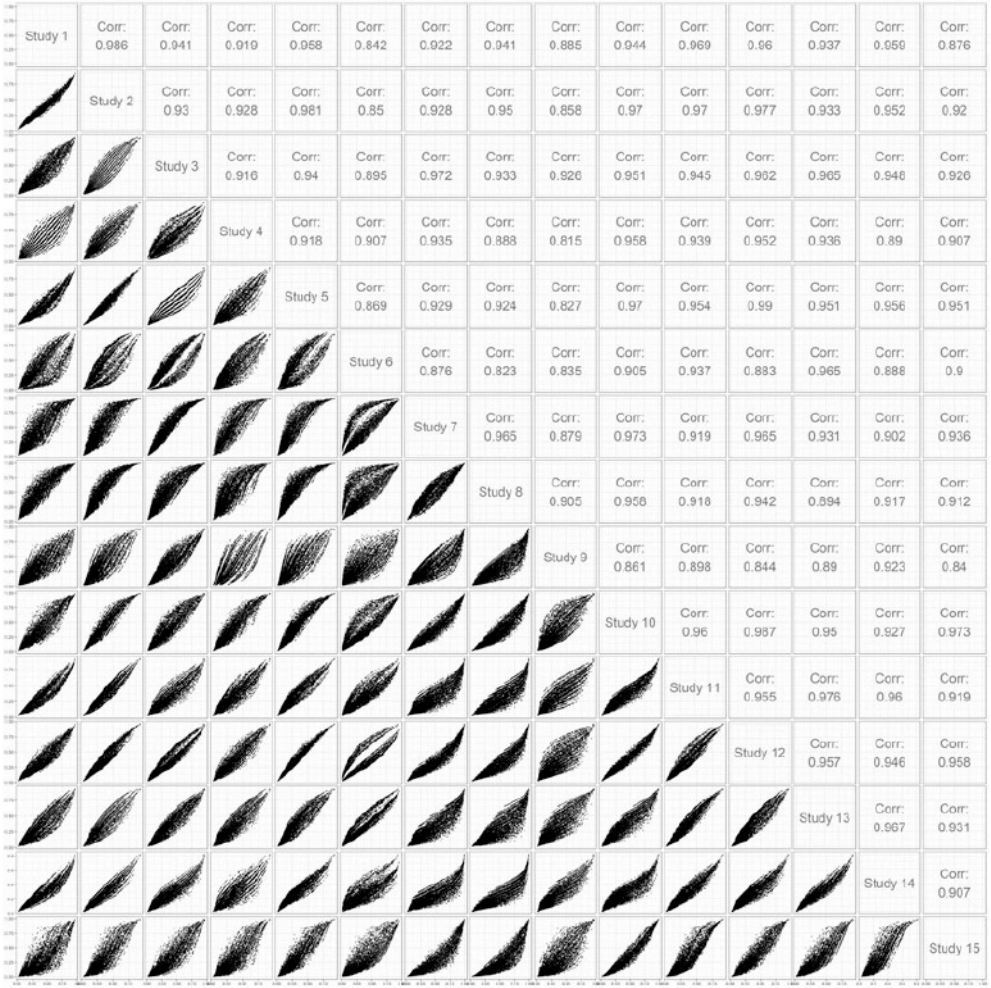
more challenging when possible heterogeneity in predictor effects

- Generalizable prediction models have little heterogeneity in predictor effects
- However, heterogeneity may sometimes appear because of collinearity
- Heterogeneity in absolute risk is what matters most
- This can be explored by plotting the predictions of study-specific models in a pairwise comparison

to gener. there should be no het. in predictors

# Assessing heterogeneity in predictor effects

1:1 comparison




https://onlinelibrary.wiley.com/doi/full/10.1002/sim.8296

# Assessing heterogeneity in predictor effects

Received: 4 June 2017 | Revised: 23 March 2019 | Accepted: 6 June 2019

DOI: 10.1002/sim.8296

RESEARCH ARTICLE

WILEY Statistics in Medicine

## Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration

Ewout W. Steyerberg[1,2] | Daan Nieboer[2] | Thomas P.A. Debray[3,4] | Hans C. van Houwelingen[1]

https://onlinelibrary.wiley.com/doi/full/10.1002/sim.8296

# Summary points

Major advantages of large clustered datasets

many opportunities for predictor model development

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

predictors

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

# Summary points

Debray *et al. Diagnostic and Prognostic Research* (2019) 3:13
https://doi.org/10.1186/s41512-019-0059-4

Diagnostic and Prognostic Research

METHODOLOGY Open Access

# Evidence synthesis in prognosis research

Thomas P.A. Debray[1,2*†], Valentijn M.T. de Jong[1†], Karel G.M. Moons[1,2] and Richard D. Riley[3]

**Abstract**

Over the past few years, evidence synthesis has become essential to investigate and improve the generalizability of medical research findings. This strategy often involves a meta-analysis to formally summarize quantities of interest, such as relative treatment effect estimates. The use of meta-analysis methods is, however, less straightforward in prognosis research because substantial variation exists in research objectives, analysis methods and the level of reported evidence.

We present a gentle overview of statistical methods that can be used to summarize data of prognostic factor and prognostic model studies. We discuss how aggregate data, individual participant data, or a combination thereof can be combined through meta-analysis methods. Recent examples are provided throughout to illustrate the various methods.

**Keywords:** Prediction, Meta-analysis, Prognosis, Validation, IPD

# Remaining challenges in IPD-MA

- IPD-MA no panacea against poorly designed primary studies no solution for poor design
  - 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? or a model that predicts well locally
  - Methods for tailoring still underdeveloped

**New methods are on their way!**

# R software

## metamisc: Diagnostic and Prognostic Meta-Analysis

Meta-analysis of diagnostic and prognostic modeling studies. Summarize estimates of prognostic factors, diagnostic test accuracy and prediction model performance. Validate, update and combine published prediction models. Develop new prediction models with data from multiple studies.

| | |
|---|---|
| Version: | 0.1.9 |
| Depends: | R (≥ 3.2.0), stats, graphics |
| Imports: | metafor (≥ 2.0.0), mvtnorm, ellipse, lme4, plyr, ggplot2 |
| Suggests: | runjags, rjags, testthat (≥ 1.0.2) |
| Published: | 2018-05-13 |
| Author: | Thomas Debray [aut, cre], Valentijn de Jong [aut] |
| Maintainer: | Thomas Debray <thomas.debray at gmail.com> |
| License: | GPL-3 |
| URL: | http://r-forge.r-project.org/projects/metamisc/ |
| NeedsCompilation: | no |
| In views: | MetaAnalysis |
| CRAN checks: | metamisc results |

### Downloads:

| | |
|---|---|
| Reference manual: | metamisc.pdf |
| Package source: | metamisc_0.1.9.tar.gz |
| Windows binaries: | r-devel: metamisc_0.1.9.zip, r-release: metamisc_0.1.9.zip, r-oldrel: metamisc_0.1.9.zip |
| OS X binaries: | r-release: metamisc_0.1.9.tgz, r-oldrel: metamisc_0.1.8.tgz |
| Old sources: | metamisc archive |

### Linking:

Please use the canonical form https://CRAN.R-project.org/package=metamisc to link to this page.