Disclaimer

The views expressed here are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the regulatory agency/agencies or organizations with which the authors are employed/affiliated.

UMC Utrecht
Julius Center

Statistical methods for IPD-MA of prognosis studies

Valentijn de Jong[1,2], PhD
1. Julius Center for Health Sciences and Primary Care
2. Data Analytics and Methods Task Force, European Medicines Agency

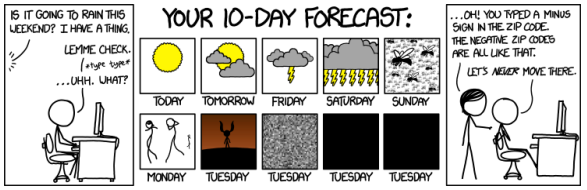
University Medical Center Utrecht

What is a good prediction model?

General requirements

- Generates accurate predictions in individuals from potential population(s) for clinical use
- Ability to discriminate between different risk groups
- Improves patient outcomes by informing treatment decisions

What is a good prediction model?



http://xkcd.com/1245/

The reality

Most prediction models are not as good as we think

- Quality of many prognostic model studies is poor
  - Absence of a study protocol
  - Exclusion of eligible study participants
  - Poor handling of missing data
  - Complex modelling in small samples
  - Incomplete registrations & reporting
- Internal validation too optimistic
- Lack of external validation

Lack of external validity

Poor performance in new patients from different (but related) populations

ELSEVIER CrossMark

Journal of Clinical Epidemiology 68 (2015) 25–34

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination

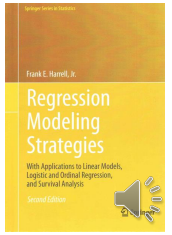
George C.M. Siontis[a,1], Ioanna Tzoulaki[a,b], Peter J. Castaldi[c], John P.A. Ioannidis[d,e,f,*]

The reality

Lack of internal validity

Poor performance in new patients from the development population

- Overfitting
- Inclusion of noise variables (e.g. via stepwise selection)
- Poor handling of missing values



Numerous models for same target population + outcomes

RESEARCH

OPEN ACCESS

Prediction models for cardiovascular disease risk in the general population: systematic review

Johanna A A G Damen,[1,2] Lotty Hooft,[1,2] Ewoud Schuit,[1,2,3] Thomas P A Debray,[1,2] Gary S Collins,[4] Ioanna Tzoulaki,[5] Camille M Lassale,[5] George C M Siontis,[6] Virginia Chiocchia,[4,7] Corran Roberts,[4] Michael Maia Schlüssel,[4] Stephen Gerry,[4] James A Black,[8] Pauline Heus,[1,2] Yvonne T van der Schouw,[1] Linda M Peelen,[1] Karel G M Moons[1,2]

RESEARCH

OPEN ACCESS

Check for updates

FAST TRACK

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Elena Albu,[2] Banafsheh Arshi,[1] Vanesa Bellou,[10] Marc M J Bonten,[8,11] Darren L Dahly,[12,13] Johanna A Damen,[8,9] Thomas P A Debray,[8,14] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,15] Paula Dhiman,[4,5] Joie Ensor,[6] Shan Gao,[2] Maria C Haller,[7,16] Michael O Harhay,[17,18] Liesbet Henckaerts,[19,20] Pauline Heus,[8,9] Jeroen Hoogland,[8] Mohammed Hudda,[21] Kevin Jenniskens,[8,9] Michael Kammer,[7,22] Nina Kreuzberger,[23] Anna Lohmann,[24] Brooke Levis,[6] Kim Luijken,[24] Jie Ma,[5] Glen P Martin,[25] David J McLernon,[26] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[27,28] Chunhu Shi,[29] Nicole Skoetz,[22] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[30] René Spijker,[8,9,31] Ewout W Steyerberg,[1] Toshihiko Takada,[8,32] Ioanna Tzoulaki,[10,33] Sander M J van Kuijk,[34] Bas C T van Bussel,[1,35] Iwan C C van der Horst,[35] Kelly Reeve,[36] Florien S van Royen,[8] Jan Y Verbakel,[37,38] Christine Wallisch,[7,39,40] Jack Wilkinson,[24] Robert Wolff,[41] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

For numbered affiliations see end of the article

Correspondence to: L Wynants laure.wynants@maastrichtuniversity.nl (ORCID 0000-0002-3037-122X)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;369:m1328 http://dx.doi.org/10.1136/bmj.m1328

Originally accepted: 31 March 2020

Final version accepted: 17 July 2022

Abstract

OBJECTIVE

To review and appraise the validity and usefulness of published and preprint reports of prediction models for prognosis of patients with covid-19, and for detecting people in the general population at increased risk of covid-19 infection or being admitted to hospital or dying with the disease.

DESIGN

Living systematic review and critical appraisal by the covid-PRECISE (Precise Risk Estimation to optimise covid-19 Care for Infected or Suspected patients in diverse sEttings) group.

DATA SOURCES

DATA EXTRACTION

At least two authors independently extracted data using the CHARMS (critical appraisal and data extraction for systematic reviews of prediction modelling studies) checklist; risk of bias was assessed using PROBAST (prediction model risk of bias assessment tool).

RESULTS

126 978 titles were screened, and 412 studies describing 731 new prediction models or validations were included. Of these 731, 125 were diagnostic models (including 75 based on medical imaging) and the remaining 606 were prognostic models for either identifying those at risk of covid-19 in the general population (13 models) or predicting

But wait… this is not the end

There are numerous models for same target population and outcomes

- >500 models alike Framingham, SCORE, QRisk
- >100 models for brain trauma patients
- >100 diabetes type 2 models
- > 60 models for breast cancer prognosis

Numerous models for same target population + outcomes

"Comparing risk prediction models should be routine when deriving a new model for the same purpose" (Collins 2012)

"Substantial work is needed to understand how competing prediction models compare and how they can best be applied to individualize care." (Wessler 2015)

"There is an excess of models predicting incident CVD in the general population. The usefulness of most of the models remains unclear." (Damen 2016)

Prediction models for COVID-19

731 models from 412 studies included in the review:

- **16 models** to identify subjects at risk in the general population
- **125 diagnosis models** for COVID-19 or COVID-19 pneumonia
  - 75 based on medical images (deep learning)
- **593 prognosis models** for predicting mortality risk, progression to severe disease, or length of stay, or other

Prediction models for COVID-19

- Participants domain: **107/606 (18%) at high risk of bias**
  - Non-representative of the target population (e.g., non-consecutive patients)
- Predictors domain: **27/606 (4%) at high risk of bias**
  - Predictors not available at time of intended model use
- Outcome domain: **106/606 (17%) at high risk of bias**
  - Subjective or proxy outcomes
- Analysis domain: **530/606 (86%) at high risk of bias**
  - Small sample size (->overfitting & no adjustment), incomplete reporting of model performance (e.g., no calibration)

RESEARCH

OPEN ACCESS

Check for updates

FAST TRACK

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Elena Albu,[2] Banafsheh Arshi,[1] Vanesa Bellou,[10] Marc M J Bonten,[8,11] Darren L Dahly,[12,13] Johanna A Damen,[8,9] Thomas P A Debray,[8,14] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,15] Paula Dhiman,[4,5] Joie Ensor,[6] Shan Gao,[2] Maria C Haller,[7,16] Michael O Harhay,[17,18] Liesbet Henckaerts,[19,20] Pauline Heus,[8,9] Jeroen Hoogland,[8] Mohammed Hudda,[21] Kevin Jenniskens,[8,9] Michael Kammer,[7,22] Nina Kreuzberger,[23] Anna Lohmann,[24] Brooke Levis,[6] Kim Luijken,[24] Jie Ma,[5] Glen P Martin,[25] David J McLernon,[26] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[27,28] Chunhu Shi,[29] Nicole Skoetz,[22] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[30] René Spijker,[8,9,31] Ewout W Steyerberg,[1] Toshihiko Takada,[8,32] Ioanna Tzoulaki,[10,33] Sander M J van Kuijk,[34] Bas C T van Bussel,[1,35] Iwan C C van der Horst,[35] Kelly Reeve,[36] Florien S van Royen,[8] Jan Y Verbakel,[37,38] Christine Wallisch,[7,39,40] Jack Wilkinson,[24] Robert Wolff,[41] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

For numbered affiliations see end of the article

Correspondence to: L Wynants laure.wynants@maastrichtuniversity.nl (ORCID 0000-0002-3037-122X)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;369:m1328 http://dx.doi.org/10.1136/bmj.m1328

Originally accepted: 31 March 2020

Final version accepted: 17 July 2022

Abstract

OBJECTIVE

To review and appraise the validity and usefulness of published and preprint reports of prediction models for prognosis of patients with covid-19, and for detecting people in the general population at increased risk of covid-19 infection or being admitted to hospital or dying with the disease.

DESIGN

Living systematic review and critical appraisal by the covid-PRECISE (Precise Risk Estimation to optimise covid-19 Care for Infected or Suspected patients in diverse sEttings) group.

DATA SOURCES

DATA EXTRACTION

At least two authors independently extracted data using the CHARMS (critical appraisal and data extraction for systematic reviews of prediction modelling studies) checklist; risk of bias was assessed using PROBAST (prediction model risk of bias assessment tool).

RESULTS

126 978 titles were screened, and 412 studies describing 731 new prediction models or validations were included. Of these 731, 125 were diagnostic models (including 75 based on medical imaging) and the remaining 606 were prognostic models for either identifying those at risk of covid-19 in the general population (13 models) or predicting

New concept: Living Review

Data supplement

Files in this Data Supplement:

Data Supplement - Web appendix: Supplementary material
Original article PDF - Published 7 April 2020
Update 1 article PDF - Published 3 June 2020
Update 2 article PDF - Published 21 July 2020

Prediction models for COVID-19

Population	Objective	AUC range
General population	Diagnosis or prognosis	0.52 to 0.99
Patients with covid-19	Prognosis	0.49 to 1.00

The rise of big datasets

The rise of big datasets

The **QRESEARCH** database

- Anonymised health records of over 25 million people from 1500 general practices spread throughout the UK
- Linkage to Hospital Episode Statistics, Mortality and Cancer Registration data

Prediction models for COVID-19

- Many COVID-19 prediction models were poorly reported, at high risk of bias, and their reported performance is probably optimistic
- Application of any prediction models that was developed early on in the pandemic was not recommended
- Only 7 out of 606 had a low risk of bias.

The rise of big datasets

Data increasingly available for thousands or even millions of patients from multiple practices, hospitals, or countries.

- Meta-analysis of individual participant data (IPD) from multiple studies
  - Observational studies
  - Randomized controlled trials
- Analyses of databases and registry data containing e-health records

The rise of big datasets

Why do we need big datasets?

- Development of better prediction models
- More extensive testing of model performance

OPEN ACCESS

GUIDELINES AND GUIDANCE

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

Thomas P. A. Debray, Richard D. Riley, Maroeska M. Rovers, Johannes B. Reitsma, Karel G. M. Moons, Cochrane IPD Meta-analysis Methods group

Published: October 13, 2015 • https://doi.org/10.1371/journal.pmed.1001886

UMC Utrecht
Julius Center

External validation of prediction models in big datasets

University Medical Center Utrecht

The rise of big datasets

CALIBER

- EHR data encompassing more than 10 million adults with 400 million person-years of follow-up
- Primary care consultations and hospitalisations
- Clinical examination findings, blood laboratory results, prescriptions and vaccinations
- Diagnoses of diseases and mortality data

CALIBER

The rise of big datasets

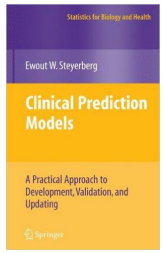
Why do we need external validation?

- The predictive performance of a model estimated on the development data is often too optimistic
- A prognostic model should provide predictions that are valid outside the specific context of the sample that was used for model development
- How a model was derived is of little importance if it performs well.

Measures of prediction model performance

Overall performance

- Amount of explained variation (R^2)
- Brier score (for binary outcomes)



Ref: Steyerberg. Clinical prediction models: a practical approach to development, validation and updating. Springer 2009.

Background

Key references

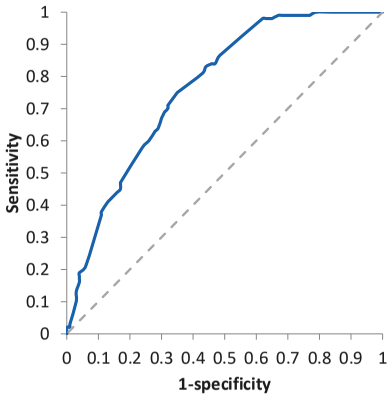
- Riley et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016.
- Debray et al. Individual Participant Data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLOS MED 2015.

What do we mean by external validation?

- Apply the clinical prediction model (CPM) to new individuals
  - Temporal validation
  - Geographical validation
  - Domain validation
- Evaluate the predictive accuracy
  - Overall performance
  - Calibration
  - Discrimination

Measures of prediction model performance

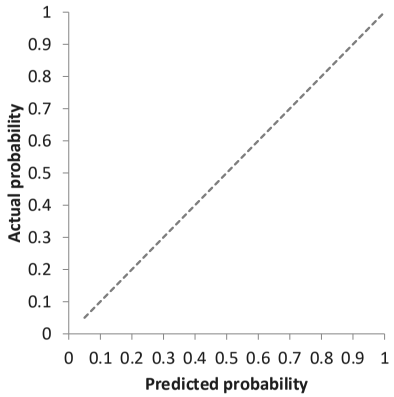
Discrimination



AUC = 0.77

Measures of prediction model performance

Calibration plot – good model?



Ideal calibration
O:E = 1
Slope = 1

Measures of prediction model performance

Discrimination
Quantifies the model's extent to distinguish between events and non-events

- Summary statistics
  - Concordance (c) index
  - Area under the ROC curve (AUC)
  - Discrimination slope
- Visual inspection
  - Receiving Operating Characteristics (ROC) curve

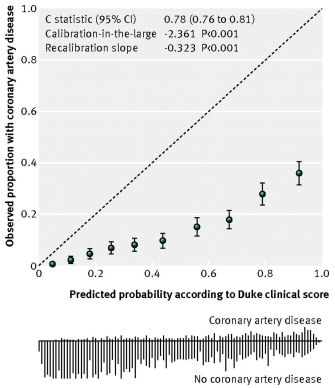
Measures of prediction model performance

Calibration
Agreement between observed outcomes and predictions

- Visual inspection
  - Calibration plot
- Summary statistics
  - O:E statistic (#observed events / #predicted events)
  - Calibration-in-the-large (or calibration intercept, which is related)
  - Calibration slope
  - Hosmer-Lemeshow goodness-of-fit test

Measures of prediction model performance

Calibration plot – good model?

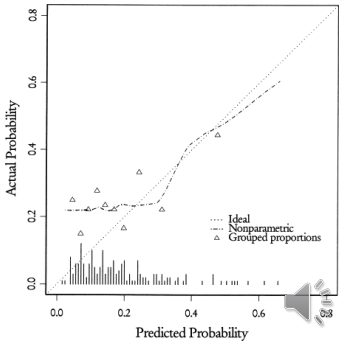


Ref: Genders et al. Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts. BMJ 2012.

Example study

External validation of the model

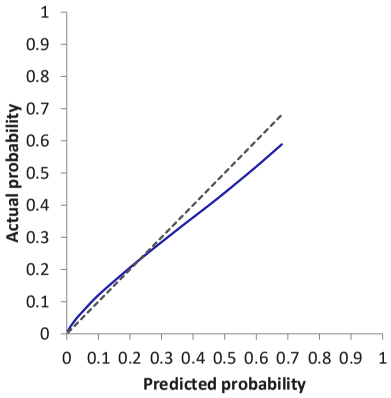
- Population: 179 children from a different time period and another hospital from a different city in NL
- Very similar in- and exclusion criteria and variable def.
- AUC = 0.57 (95% CI: 0.47 – 0.67)



Ref: Bleeker SE et al. External validation is necessary in prediction research: a clinical example. Journal of Clinical Epidemiology 2003.

Measures of prediction model performance

Calibration plot – good model?



O:E = 1
Slope = 0.8

Subobtimal slope because curve does not follow reference line

Example study

Development and validation of a prediction model for the presence of serious bacterial infections in children with fever.

Development of the model

- Population: 379 children between 1 month and 36 months of age referred to the Emergency Department from a hospital in NL (75 events)
- Analysis: logistic regression with forward stepwise variable selection (57 variables -> 9 predictors)
- Internal validation:
  - AUC = 0.825 (95% CI: 0.78 – 0.87)
  - Bootstrap-corrected AUC = 0.756

Current shortcomings of validation studies

Why do we need big datasets for external validation?

- The predictive performance of a model tends to vary across settings, populations and periods
- Multiple external validation studies are needed to fully appreciate the generalizability of a prediction model
- Heterogeneity in model performance is rarely assessed, but investigating its extent is crucial to evaluate the model's potential generalizability and clinical usefulness.

Causes of heterogeneity in model performance

Discrepancies in outcome and predictor assessement

- Different measurement method for predictors (e.g. using equipment from different manufacturers)
- Different recording time of predictors (e.g. before or after surgery)
- Different quantification of predictors (e.g. use of cut-points may vary)
- Different disease and outcome definitions
- Different follow-up lengths

Current shortcomings of validation studies

- External validation requires sufficient data
  - Recommendations: >100 events and >100 non-events
  - Less data available for model development
- Not all validation studies are equally informative.
  - To what extent do individuals from the validation sample represent the target population?
  - To what extent are estimates of model performance affected by flaws in the design and analysis of the validation study?
  - To what extent can the CPM be implemented across different populations and settings?

Causes of heterogeneity in model performance

Invalid predictor effects

- Over-fitting of the prediction model to the development study (sometimes avoided using *penalization*)
- Biased estimates of predictor effects (e.g. due to flaws in the development study)
- Missed interactions or non-linear associations

Causes of heterogeneity in model performance

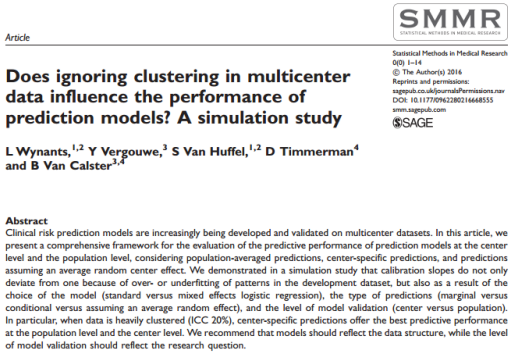
Case-mix variation (spectrum effect)

- Different distribution of predictor values
- Different standards of care and treatment strategies
- Different starting points
  (e.g. earlier diagnosis due to screening program)
- Different outcome prevalence or incidence
- Different participant or setting characteristics

Case-mix variation can lead to genuine differences in the performance of a prediction model, even when the predictor effects remain ''correct'' in the validation study

Interpretation of model performance

Need to adjust for clustering!



Article

SMMR

Statistical Methods in Medical Research
0(0) 1–14
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280216668555
smm.sagepub.com

SAGE

Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study

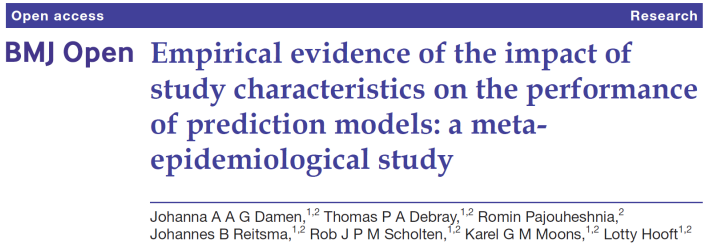
L Wynants,[1,2] Y Vergouwe,[3] S Van Huffel,[1,2] D Timmerman[4] and B Van Calster[3,4]

Abstract

Clinical risk prediction models are increasingly being developed and validated on multicenter datasets. In this article, we present a comprehensive framework for the evaluation of the predictive performance of prediction models at the center level and the population level, considering population-averaged predictions, center-specific predictions, and predictions assuming an average random center effect. We demonstrated in a simulation study that calibration slopes do not only deviate from one because of over- or underfitting of patterns in the development dataset, but also as a result of the choice of the model (standard versus mixed effects logistic regression), the type of predictions (marginal versus conditional versus assuming an average random effect), and the level of model validation (center versus population). In particular, when data is heavily clustered (ICC 20%), center-specific predictions offer the best predictive performance at the population level and the center level. We recommend that models should reflect the data structure, while the level of model validation should reflect the research question.

Causes of heterogeneity in model performance

Differences between study characteristics



Open access Research

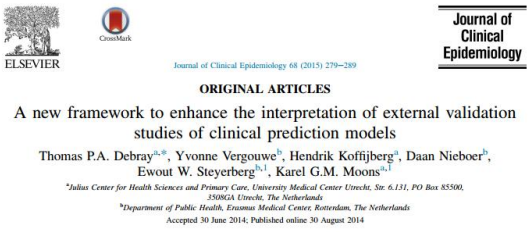
BMJ Open

Empirical evidence of the impact of study characteristics on the performance of prediction models: a meta-epidemiological study

Johanna A A G Damen,[1,2] Thomas P A Debray,[1,2] Romin Pajouheshnia,[2] Johannes B Reitsma,[1,2] Rob J P M Scholten,[1,2] Karel G M Moons,[1,2] Lotty Hooft[1,2]

Interpretation of model performance

Need to disentangle case-mix differences from differences in predictor effects!



ELSEVIER

CrossMark

Journal of Clinical Epidemiology 68 (2015) 279–289

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

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 30 August 2014

Examining heterogeneity and improving model performance

Guidance paper


RESEARCH METHODS AND REPORTING

A guide to systematic review and meta-analysis of prediction model performance

Thomas P A Debray,[1,2] Johanna A A G Damen,[1,2] Kym I E Snell,[3] Joie Ensor,[3] Lotty Hooft,[1,2] Johannes B Reitsma,[1,2] Richard D Riley,[3] Karel G M Moons[1,2]

Example 1 (same as practical)

Diagnosis of deep vein thrombosis in patients suspected of DVT

Previously developed prediction model for diagnosing DVT

- Logistic regression analysis
- Three predictors
  - *Sex*
  - *Surgery*
  - *Calf difference*

$$Pr(\text{DVT}) = \frac{1}{1 + \exp(-(\alpha + \beta_1 \text{sex} + \beta_2 \text{surg} + \beta_3 \text{cdif})}$$

Geersing GJ, Zuithoff NPA, Kearon C, Anderson DR, Cate-Hoek AJ ten, Elf JL, et al. Exclusion of deep vein thrombosis using the Wells rule in clinically important subgroups: individual patient data meta-analysis. BMJ. 2014 Mar 10;348:g1340. Available from: https://www.bmj.com/content/348/bmj.g1340

Examining heterogeneity and improving model performance

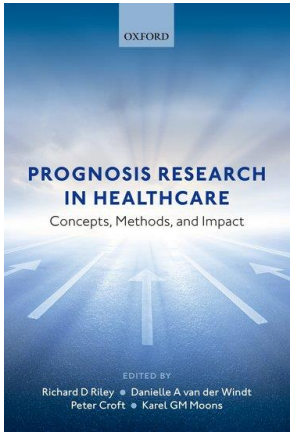
Recommendations

- Calculate key performance statistics in each cluster (e.g. study or hospital)
- Summarize the performance measures by applying (multivariate) random effects meta-analysis
- Quantify between-study heterogeneity in model performance using 95% prediction intervals

Examining heterogeneity and improving model performance

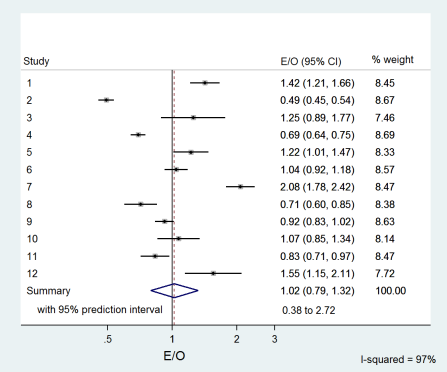
More guidance

- Prognostic model research
- Systematic reviews and meta-analysis of prognosis research studies
- Individual participant data meta-analysis of prognosis studies
- Electronic healthcare records and prognosis research



Example 1

External validation in 12 studies



Example 2

Prognosis of cardiovascular disease in patients from general practices using QRISK2

- 364 practices from the UK
- Total sample size N=2,084,445
- Total number of events E=93,564

Again, each cluster might be viewed as a different external validation study!

Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008 Jun 28;336(7659):1475–82. Available from: https://doi.org/10.1136/bmj.39609.449676.25

Example 1

External validation in 12 studies

- Sample size: 153 – 1768 (total N=10014)
- Event occurrence: 8% - 39% (total E=1897)
- Results (95% confidence interval)
  - Calibration-in-the-large: -0.004 (-0.313; 0.305)
  - Calibration slope: 0.980 (0.853; 1.107)
  - E:O ratio: 1.02 (0.81; 1.28)
  - C-statistic: 0.687 (0.669; 0.705)

Does the model predict well? What about generalizability?

Example 1

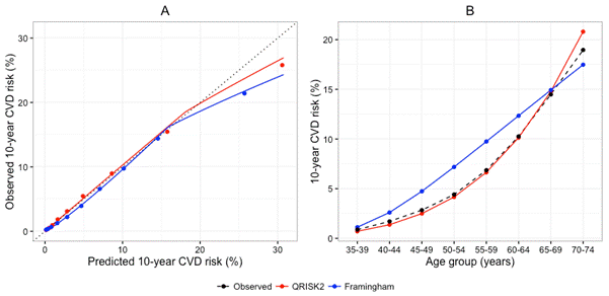
External validation in 12 studies

- Substantial between-study heterogeneity!
- Approximate 95% prediction intervals:
  - Calibration slope = 0.59 to1.38
  - E:O ratio= 0.38 to 2.72
  - c-statistic = 0.64 to 0.73

The model requires improvements to improve discrimination and to be clinically useful!

Example 2

External validation in 364 practices



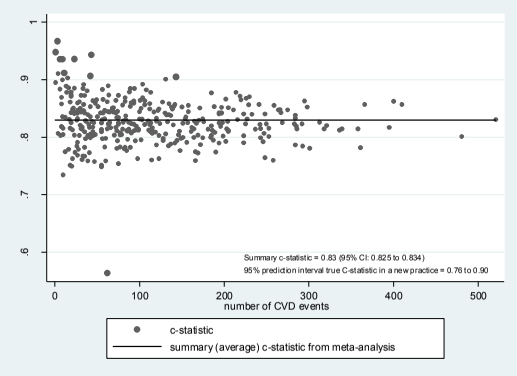
Example 2

Investigating between-study heterogeneity

- Recall that variation in case-mix severity and case-mix heterogeneity may affect model performance.
  - Larger case-mix variation is related to larger discrimination performance
  - Populations with a narrower case-mix tend to have worse discrimination performance
- Recall that discrimination of QRISK2 heterogeneous
  - Population mean age
  - Percentage smokers

Example 2

External validation in 364 practices



Example 2

External validation in 364 practices

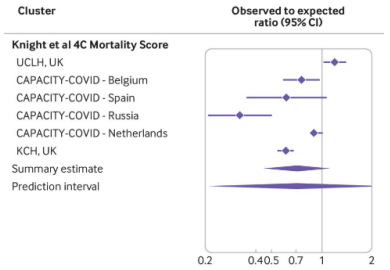
- Discrimination
  - Summary c-statistic: 0.83
  - 95% confidence interval: 0.825 to 0.834
  - 95% prediction interval: 0.76 to 0.90
- Calibration
  - E:O summary estimate: 1.01
  - Slight over-prediction in women at higher CVD risk
  - QRISK2 appears to accurately predict 10-year CVD risk across all age groups.

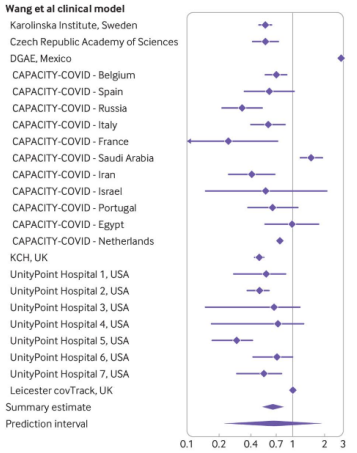
Example 3: prognosis in COVID-19 patients

- External validation in 46,914 patients from 18 countries
- Design: Two stage individual participant data meta-analysis.
- Setting: Secondary and tertiary care.
- Patients: diagnosed with covid-19
- Outcome: 30-day mortality or in-hospital mortality.

De Jong VMT, Rousset RZ, Antonio-Villa NE, Buenen AG, Calster BV, Bello-Chavolla OY, et al. Clinical prediction models for mortality in patients with covid-19: external validation and individual participant data meta-analysis. BMJ 2022 Jul 12;378:e069881. Available from: https://doi.org/10.1136/bmj-2021-069881

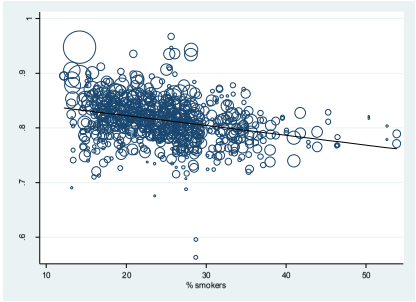
Example 3: prognosis in COVID-19 patients





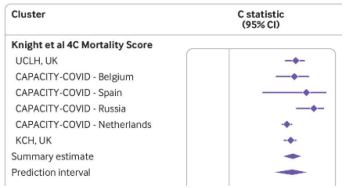
Example 2

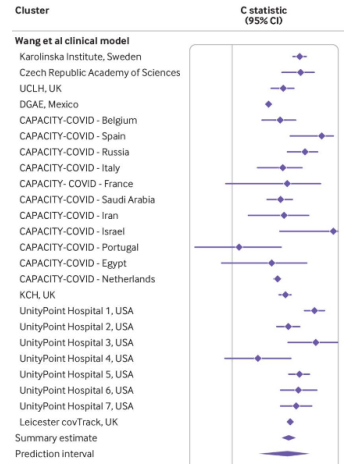
Investigating between-study heterogeneity



NB Circle size is weighted by the precision of the c-statistic estimate (i.e. larger circles indicate c-statistic estimates with smaller standard errors, and thus more weight in the meta-regression)

Example 3: prognosis in COVID-19 patients





UMC Utrecht
Julius Center

Development of prediction models in big datasets

University Medical Center Utrecht

Model development using big datasets

Practical and methodological challenges

Caution is warranted when interpreting summary estimates of model performance and between-study heterogeneity.

- **Data quality**
  - Missing predictor values
  - Non-standardised definitions of diagnoses and outcomes
  - Incomplete follow-up times and event dates
  - Lack of recording of novel/costly predictors
  - Risk of double entries
- **Data dredging**

Need for study protocols and quality appraisal tools!

Practical and methodological challenges

Caution is warranted when interpreting summary estimates of model performance and between-study heterogeneity.

- **Data quality**
  - Missing predictor values
  - Non-standardised definitions of diagnoses and outcomes
  - Incomplete follow-up times and event dates
  - Lack of recording of novel/costly predictors
  - Risk of double entries
- **Data dredging**

Need for study protocols and quality appraisal tools!

Model development using big datasets

So, let's pool our IPD and 'launch' the 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

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

Wait… which analysis?

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

Problems with meta-analysis methods

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.

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
- 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

Out of 15 IPD-MA

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

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
- **Implement a framework that uses internal-external cross-validation**

The framework

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)

Propose which intercept term to use in new populations

!! More difficult in case of heterogeneous predictor effects

Problems with meta-analysis methods

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

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

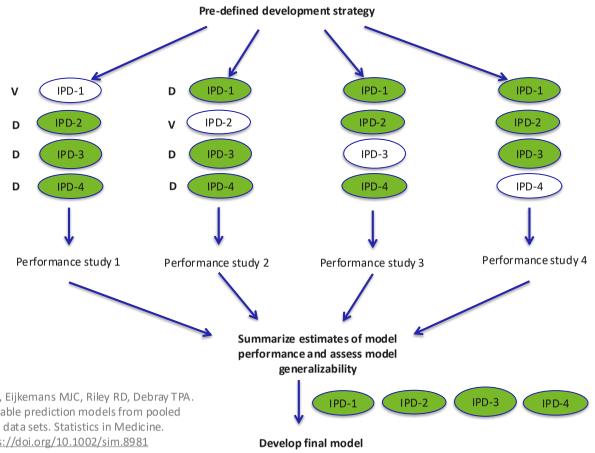
The framework

Step 1: Different choices for combining IPD

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

*For info on this, see: Falconieri N, Calster BV, Timmerman D, Wynants L. Developing risk models for multicenter data using standard logistic regression produced suboptimal predictions: A simulation study. Biometrical Journal. 2020;62(4):932–44. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201900075

Internal-external cross-validation



See:
De Jong VMT, Moons KGM, Eijkemans MJC, Riley RD, Debray TPA. Developing more generalizable prediction models from pooled studies and large clustered data sets. Statistics in Medicine. 2021; Available from: https://doi.org/10.1002/sim.8981

Example 1

Diagnosis of deep vein thrombosis (N=12)

Strategies evaluated:

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

The framework

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

=> Use of internal-external cross-validation

Formally comparing different strategies

- Meta-analyze estimates of model performance
  Compare summary estimates
  - 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)

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, μ	95% Prediction interval	$\hat{P}$ (%)	$\hat{\tau}$ (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.

Example 2

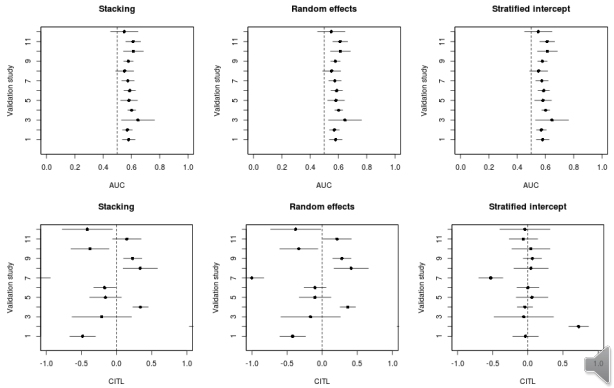
Prognosis of amyotrophic lateral disease

- 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 1

Diagnosis of deep vein thrombosis (N=12)



Example 1

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!

Example 2

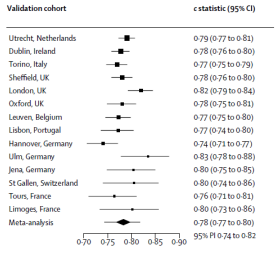
Prognosis of amyotrophic lateral disease

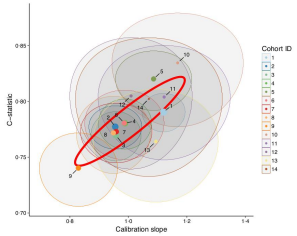
- Royston-Parmar survival model with country-specific (but proportional) baseline hazard

Variable	Value
γ_0	-6·409
γ_1	2·643
γ_2	-0·546
γ_3	0·585
β_1 (ALSFRS-R slope)	-1·837
β_2 (Diagnostic delay)	-2·373
β_3 (Age at onset)	-0·267
β_4 (Forced vital capacity)	0·477
β_5 (Bulbar onset)	0·269
β_6 ('Definite' ALS*)	0·233
β_7 (Frontotemporal dementia)	0·388
β_8 (*C9orf72* repeat expansion)	0·256

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

Example 2





Measure	Criteria	Probability 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 2

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]

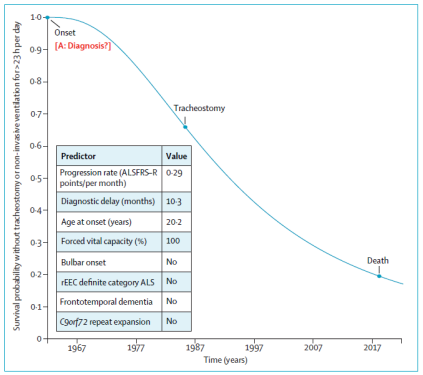
Example 2

Internal-external cross-validation

- c-statistic
  - Summary estimate: 0.78 (0.77 to 0.80)
  - 95% PI: 0.74 to 0.82
- 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 2

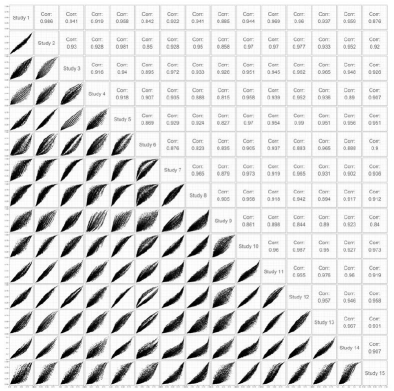
The life expectancy of Stephen Hawking,
according to the ENCALS model



Predictor	Value
Progression rate (ALSFRS-R points/per month)	0·29
Diagnostic delay (months)	10·3
Age at onset (years)	20·2
Forced vital capacity (%)	100
Bulbar onset	No
rEEC definite category ALS	No
Frontotemporal dementia	No
C9orf72 repeat expansion	No

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



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

Example 2

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

Assessing 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

Summary points

Major advantages of large clustered datasets

- 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

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

Remaining challenges in IPD-MA

- IPD-MA not a solution for 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!

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

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

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.