UMC Utrecht

Personalizing medical decision making

Recent advances in prediction model research

22 sept 2020

Thomas Debray

About me

Assistant Professor

University Medical Center Utrecht

Research of statistical methods

- Risk prediction
- Evidence synthesis

@TPA_Debray

My talk today

- What is prediction?
- Recent advances in Machine Learning
- Recent advances in Penalization
- Recent advances in Evidence Synthesis
- Recent advances in Treatment effect modelling
- Next Steps

Background

UMC Utrecht

Prediction

Estimate something that is yet unknown

- Presence of a certain disease (diagnosis)
- Future occurrence of a particular event (prognosis)

Prediction

Calculate the absolute risk (probability) for distinct individuals

Why?

- Identify high-risk individuals
- Identify absolute treatment effect
- Target decision making to individuals

Prediction

Calculate the absolute risk (probability) for distinct individuals

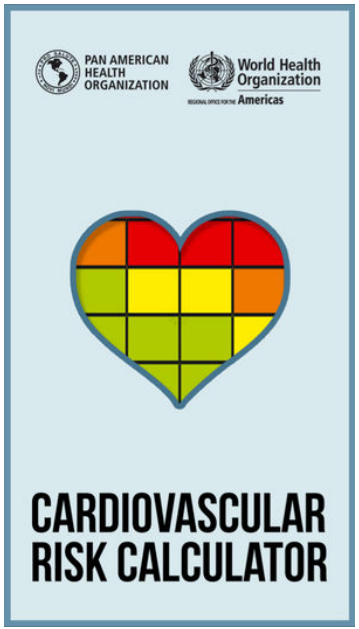
How?

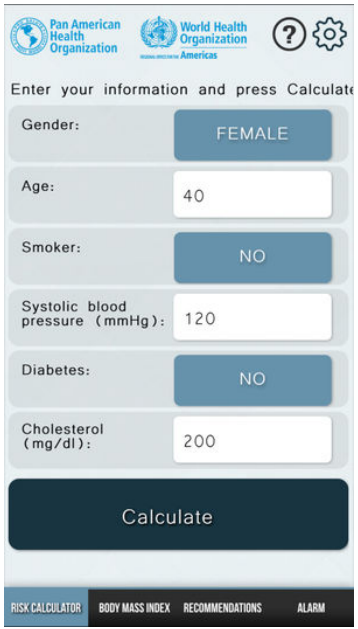
Combine information from multiple predictors

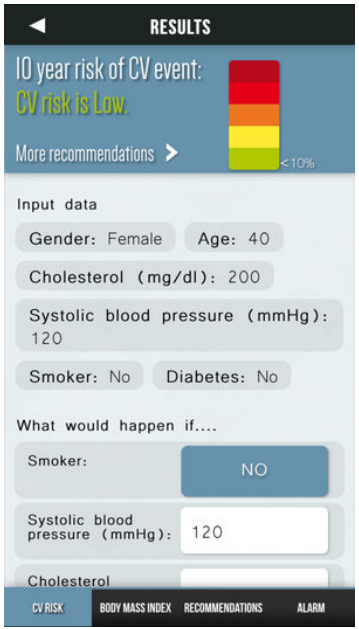
- Subject characteristics (e.g. age, gender)
- History and physical examination results (e.g. blood pressure)
- Imaging results
- (Bio)markers (e.g. coronary plaque)

Prediction

Calculate the absolute risk (probability) for distinct individuals



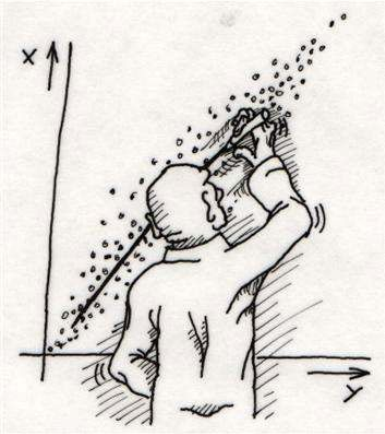


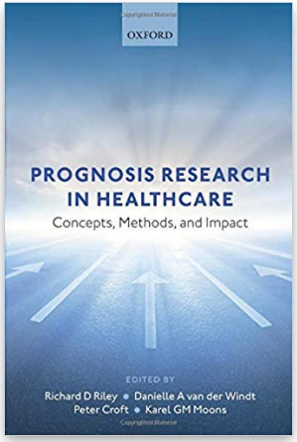


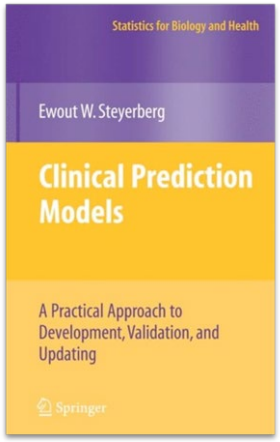
Prediction

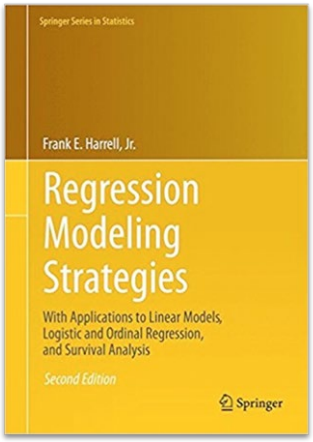
Develop a multivariable statistical model

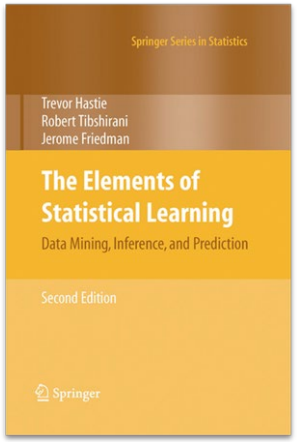
- Need for patient data from large cohort studies
- Many strategies available (Regression, decision trees, neural networks)











What is a good model?

Accurate predictions

Good and consistent performance across different settings and populations

Ability to distinguish between low and high risk patients

Influence decision making

What is a good model?

Reproducitbility versus Transportability

- Performance in same population*
- Evaluated with:
  - Internal validation (resampling methods using random-split)
  - External validation (same population)

- Performance in a different but related population*
- Evaluated with:
  - External validation (different population)
  - Resampling methods with non-random split

@TPA_Debray

[*] Debray, T., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E. and Moons, K. (2015). A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology*, 68(3), pp.279-289.

Current limitations

Many prediction models perform poorly, do not affect clinical practice, or do not improve patient outcomes

- Small & poor quality studies
- Limited variation in studied patients, settings or populations
- Lack of validity and effectiveness assessments

"All models are wrong, but some are useful" – George Box

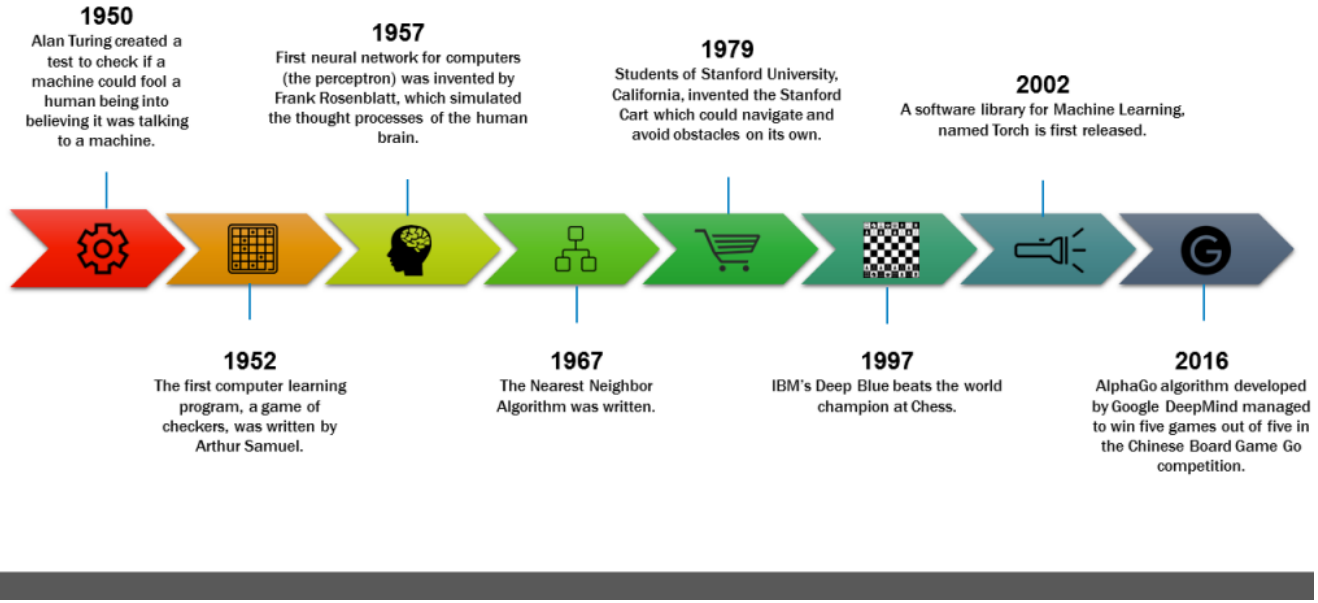
Improving prediction models

- Machine Learning
- Penalization
- Evidence synthesis
- Treatment effect modelling

Machine Learning

UMC Utrecht

Machine Learning



@TPA_Debray

https://medium.com/analytics-vidhya/fundamental-omachine-learning-ada28afa1bd3

Machine Learning

"There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown." – **Leo Breiman**

Statistical Science, Vol. 16, No. 3 (2001), pp. 199-215

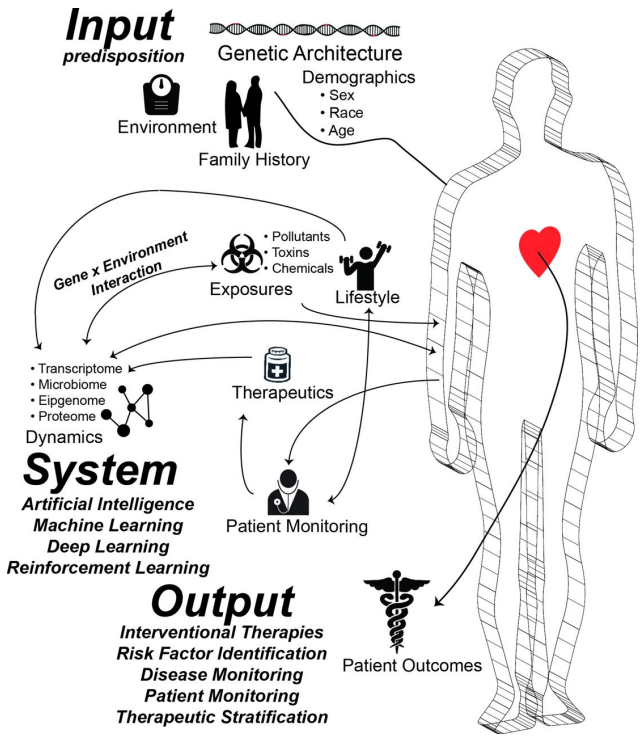
Machine Learning

- Strong focus on prediction and classification
- Combination of data-driven algorithms
  - Nearest Neighbour
  - Recursive Partitioning
  - Neural Network
  - Support Vector Machine
- Avoidance of modeling assumptions (e.g. additivity, linearity), resulting in high flexibility

Machine Learning in Health Care

Data available for prediction:

- Imaging (e.g. CT scan, MRI)
- Text (e.g. medical records)
- High-throughput data (e.g. wearables)
- High-dimensional laboratory data
- Clinical epidemiological data



Heart. 2018 Jul 1;104(14):1156–64.

@TPA_Debray

Machine Learning in Health Care

Major contributions

- Image recognition
- Analysis of unstructured data
- Problems with high signal:noise ratio

Machine Learning in Health Care

Major challenges

- Severe overfitting in “small” samples
- Very limited gains in the analysis of large (structured) epidemiological datasets
- Not designed for causal inference

Machine Learning in Health Care

Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints

Tjeerd van der Ploeg[1,3*], Peter C Austin[2] and Ewout W Steyerberg[3]

Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury

Tjeerd van der Ploeg[a,b,c,*], Daan Nieboer[c], Ewout W. Steyerberg[c]

[a]*Department of Science, Medical Center Alkmaar, Wilhelminalaan 12, Alkmaar 1815 JD, The Netherlands*
[b]*Department of Science, Inholland University, Bergerweg 200, Alkmaar 1817 MN, The Netherlands*
[c]*Department of Public Health, Erasmus MC—University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands*

Accepted 5 March 2016; Published online 14 March 2016

Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults

Anita L. Lynam[1], John M. Dennis[1], Katharine R. Owen[2,3], Richard A. Oram[1], Angus G. Jones[1], Beverley M. Shields[1] and Lauric A. Ferrat[1*]

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia Christodoulou[a], Jie Ma[b], Gary S. Collins[b,c], Ewout W. Steyerberg[d], Jan Y. Verbakel[a,e,f], Ben Van Calster[a,d,*]

Machine Learning in Health Care

Machine Learning may not (yet) be suitable for prediction of absolute treatment effects in routine care settings

Penalization

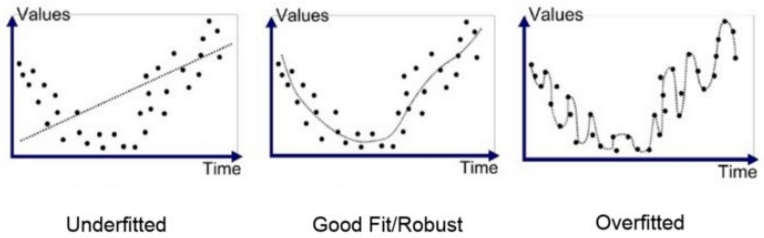
Improved prediction of time-to-event outcomes

UMC Utrecht

The need for penalization

Many prediction models are prone to overfitting

- Noise is (partially) interpreted as signal
- Inaccurate predictions for new patients from the target population
  - Predicted risk is too high for high-risk patients
  - Predicted risk is too low for low-risk patients
- Estimates of out-of-sample performance are over-optimistic

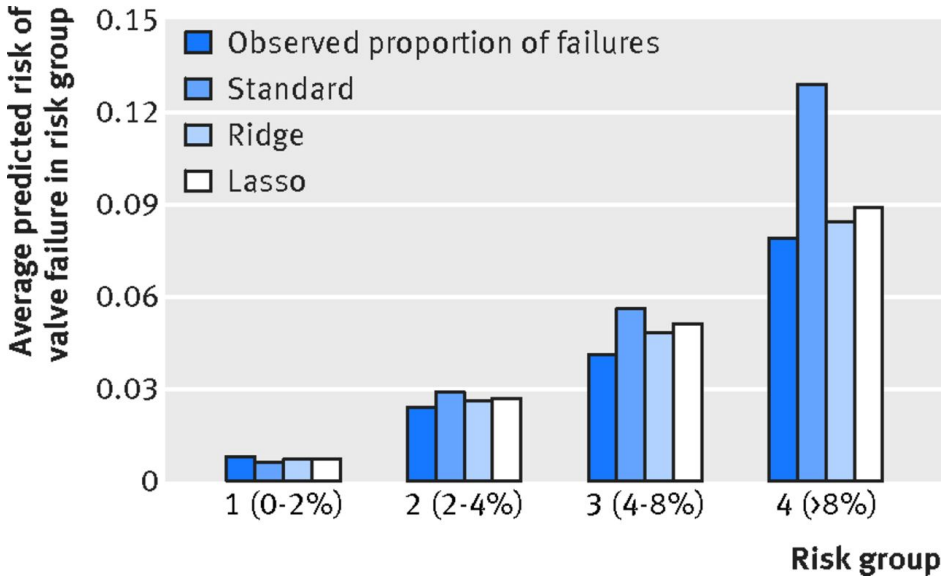


The need for penalization

How to avoid overfitting?

- Regularize model complexity (e.g. via assumptions)
- Shrink poorly calibrated predictions towards the average risk
- Constrain the magnitude of regression coefficients
- Include a penalty term in the log-likelihood
- Examples: LASSO, Ridge, Elastic Net, etc.

Overfitting: an example



Observed proportions versus average predicted risk of the event

Pavlou M, et al. BMJ. 2015;351:h3868.

Penalization in survival models

What about prediction of time-to-event outcomes?

- Need for parametric survival models
- Need for flexible baseline hazard (BH)
- Need for penalization

Model type	Parametric BH	Flexible BH	Penalization
Cox	x	✓	✓
Weibull	✓	x	✓
Royston-Parmar	✓	✓	x

Penalization in survival models

Research by drs. Jeroen Hoogland

- Combine the benefits of the Royston Parmar log cumulative hazards model and penalized maximum likelihood estimation
- Implement an elastic net penalty for the RP model
- Facilitate estimation of non-proportional hazards and other interaction terms

Penalization in survival models

- The log cumulative hazard is modeled as a linear additive combination
- All terms are differentiable w.r.t. (log) time
- Thus, the log-likelihood is available in closed form
- Penalty

$$P_\omega(\boldsymbol{\theta}) = \sum_{i=1}^{d} \omega_i \lambda_{1i} |\theta_i| + (1 - \omega_i) \frac{1}{2} \lambda_{2i} \theta_i^2$$

- The *size* of the penalty can be modified per parameter (lambda)
- The mixture between ridge and lasso can be modified per parameter (omega)

Penalization in survival models

- Full gradient ascent algorithm (based on lasso Cox PH)
- Step size depends on ratio l'_{pen} / l''_{pen}
  - First derivative of the penalized log-likelihood l'_{pen}
  - Second derivative of the penalized log-likelihood l''_{pen}
- Respects discontinuities in the gradient for parameters subject to an absolute value penalty
- When close to the optimum, switches to Newton-Raphson
- Hyper-parameter tuning using out-of-sample log-likelihood

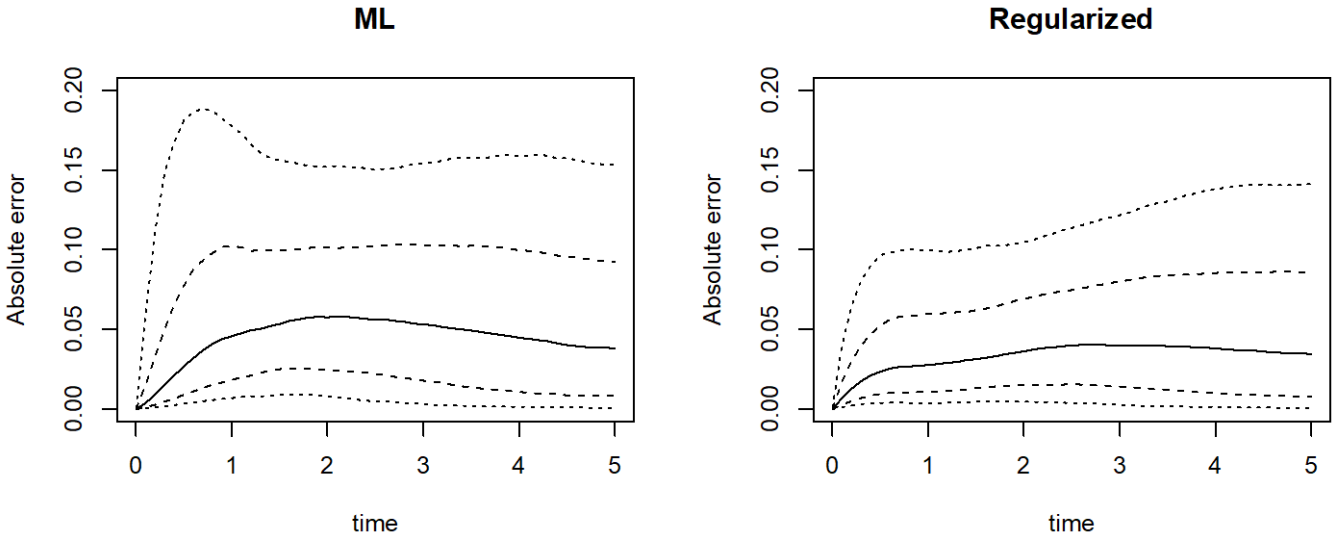
Simulation study

Data simulated from a Weibull mixture with non-proportional hazards

- 20 MVN covariates with mutual correlation 0.25
  - 12 noise variables
  - 8 variables with beta = 0.25
  - 1 (independent) treatment variable with beta = -0.5
- Survival times were right-censored (administrative)
- Event rate ~ 0.75
- 500 patients available for model development
- 5000 patients for model evaluation

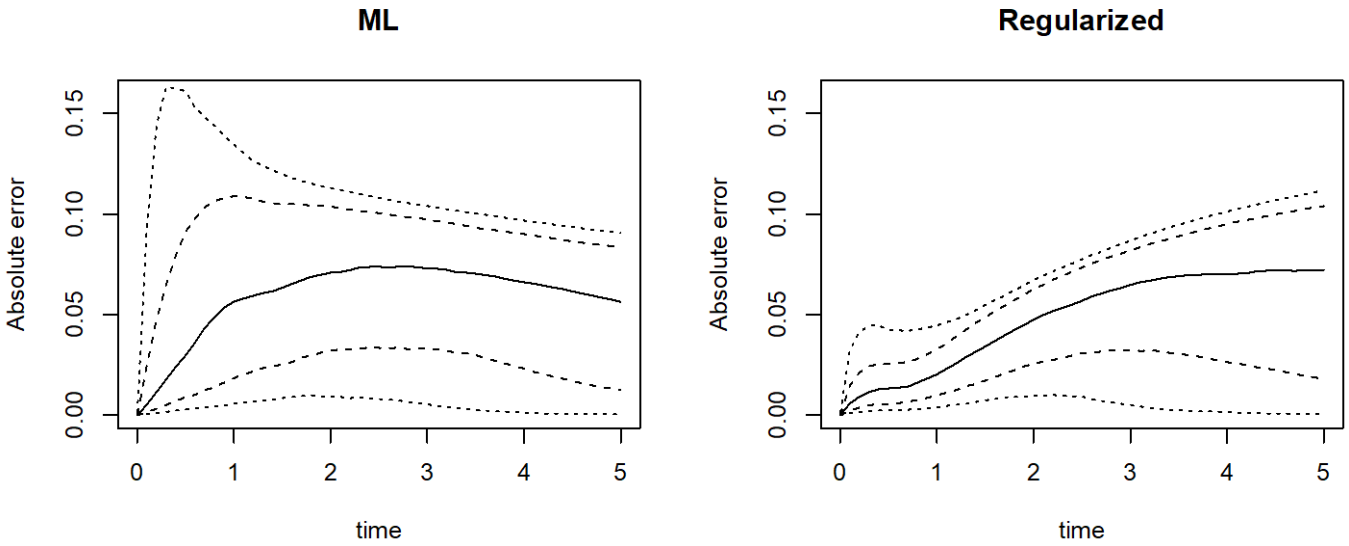
Simulation study results

Error in predicted survival (q .1, .25, .5, .75, .9)



Simulation study results

Error in predicted individual treatment effect



Main findings

- The Royston Parmar log cumulative hazards model is very flexible
- Model complexity often needs to be tuned to the data at hand
- Regularization provides a means to do so

Limitations

- algorithm is sensitive to starting values
- As of yet, it starts from ML and PH based initial values
- Therefore, is does not scale well in case of strongly non-PH models with >> p

Overfitting – a problem solved?

Findings from a recent simulation study

- Despite improved performance on average, shrinkage often worked poorly in individual datasets, in particular when it was most needed.
- Shrinkage methods do not solve problems associated with small sample size or low number of events per variable

Article

Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–13

DOI: 10.1177/0962280220921415
journals.sagepub.com/home/smm
SAGE

Overfitting – a problem solved?

- Traditional penalization methods help to improve performance when the model is applied to new patients from the same target population (i.e. reproducibility)

- Penalization does not aim to improve the model's performance when applied across different (but related) settings and populations (i.e. transportability)

More advanced methods are needed to quantify and improve the generalizability of prediction models

Evidence synthesis

Improving predictions across different settings and populations

UMC Utrecht

Evidence synthesis in prognosis research

Synthesis of prognosis studies may help

- To identify promising markers
- To identify promising prediction models
- To improve the accuracy of prediction models

Evidence synthesis in prognosis research

- Meta-analysis of published aggregate data (AD)
  - Summarize prediction model performance
  - Summarize risk factor-outcome associations
- Meta-analysis of individual participant data (IPD)
  - Develop & validate prediction models
  - Identify prognostic factors
  - Identify predictors of treatment effect
- Meta-analysis of IPD and AD

Meta-analysis of published AD

Research Methods & Reporting

A guide to systematic review and meta-analysis of prognostic factor studies

BMJ 2019 ; 364 doi: https://doi.org/10.1136/bmj.k4597 (Published 30 January 2019)
Cite this as: *BMJ* 2019;364:k4597

Research Methods & Reporting

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

BMJ 2017 ; 356 doi: https://doi.org/10.1136/bmj.i6460 (Published 05 January 2017)
Cite this as: *BMJ* 2017;356:i6460

Meta-analysis of published AD

Guidance for systematic reviews (research by dr. Damen)

- Defining the review question (PICOTS)
- Defining the search strategy
- Quantitative data extraction
- Quality appraisal (PROBAST, QUIPS)
- Meta-analysis (metamisc R package)
- Investigating between-study heterogeneity
- Interpretation (GRADE)
- Reporting (guidelines: REMARK, PRISMA, TRIPOD)

Cochrane Colloquium Edinburgh

16-18 September 2018 | Edinburgh, UK

R-package: metamisc

Meta-analysis of diagnostic and prognostic modelling studies

https://CRAN.R-project.org/package=metamisc

Meta-analysis of published AD

Recent reviews to summarize prediction model performance

- Breast cancer (Meads *et al*; Breast Cancer Res. Treat. 2012)
- Perioperative Mortality (Sullivan *et al*; Am. J. Cardiol. 2016)
- Cardiovascular disease (Damen *et al*; BMC Med 2017)
- Colorectal cancer (Hu *et al*; Surg Oncol 2019)
- Chronic lymphocytic leukemia (Molica *et al*; Leukemia 2020)
- …

Meta-analysis of IPD

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
- Analyses of databases and registry data containing e-health records

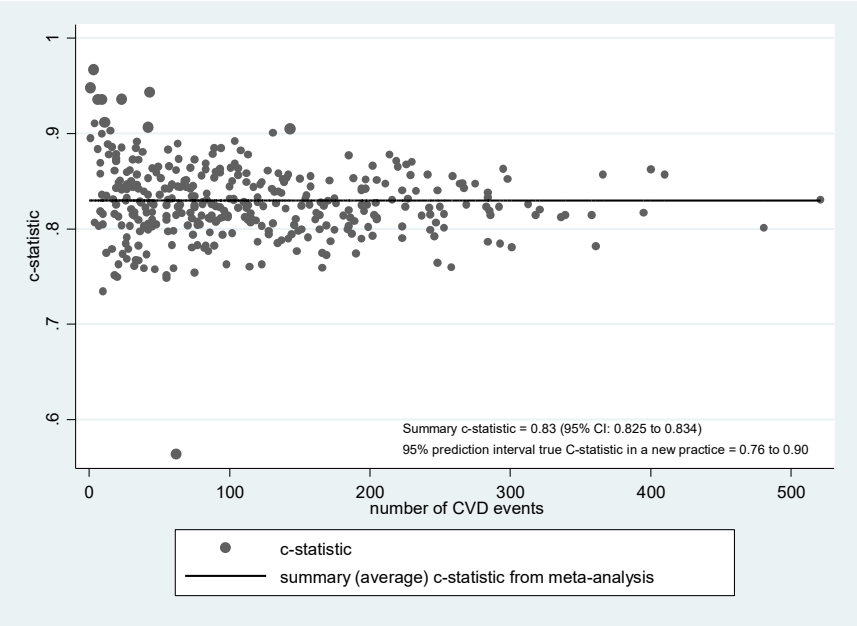
Meta-analysis of IPD

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 evaluate generalizability of the model across different settings and populations

External validation using IPD-MA

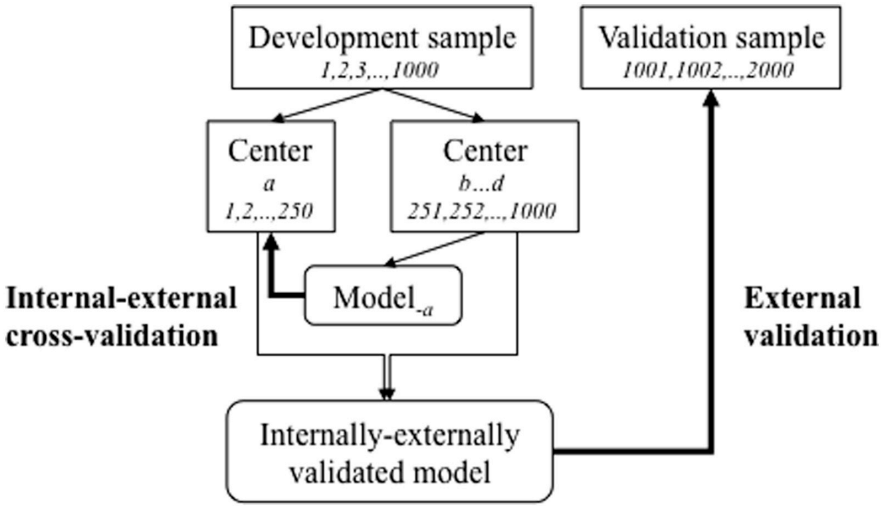
Validation of QRISK 2 in 364 UK practices



Riley RD, et al. BMJ. 2016;353:i3140.

Model development using IPD-MA

Internal-external cross-validation



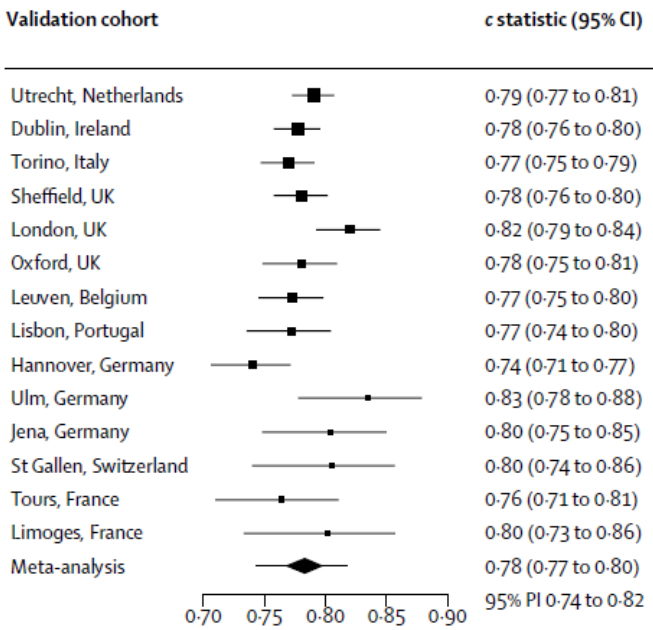
Debray TPA, et al. Stat Med. 2013 Aug 15;32(18):3158–80.
Steyerberg EW, Harrell FE. J Clin Epidemiol. 2015 Apr 18;69:245–7.

Development of ENCALS

Prognosis of amyotrophic lateral disease

- 14 cohort studies (specialized ALS centres)
  - N = 190 to 1,936 per study (total N = 11,475)
  - Median follow-up: 97.5 months
  - Composite endpoint
    (Non-invasive ventilation for more than 23h/day, or death)

Development of ENCALS



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]

Performance	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%	

Developing generalizable prediction models

Stepwise estimation procedure (research by dr. de Jong)

- Fitting of a pre-specified GLM in each study
- Evaluation of performance using IECV
- Loss = *f*(overall performance in hold-out studies, between-study variation)
- Expand (or reduce) model until the overall loss no longer decreases
- Implementation in “metamisc”

ISCB ASC18
26-30 AUGUST 2018
MELBOURNE, AUSTRALIA

Developing generalizable prediction models

Further extensions

- Methods to adjust for measurement error in IPD-MA
- Methods to disentangle case-mix variation from invalid predictor effects
- Methods to account for missing participant-level data in IPD-MA

https://recodid.eu/

Treatment effect modelling

Improving predictions of absolute treatment effect

UMC Utrecht

Background

Individualized absolute treatment effects provide a natural starting point to engage in shared decision making

Requirements

- Move to the absolute risk scale
- Adjust for individual patient characteristics
- Consider counterfactual outcomes

Background

Individualized absolute treatment effects provide a natural starting point to engage in shared decision making

Two important sources of information (in RCTs):

- Prognostic variables
  predicting outcome risk on reference treatment

- Treatment variables
  with potential for effect modification

Background

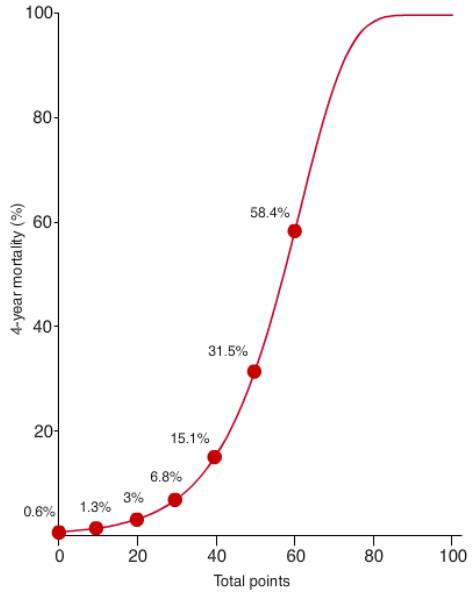
An example: **The SYNTAX score II**

"The SYNTAX score II is a clinical tool that combines clinical variables with the anatomical SYNTAX score, providing expected 4-year mortality for both coronary artery bypass grafting (CABG) and percutaneous coronary intervention (PCI) — thus recommending either PCI only, CABG only or equipoise in treatment based on long-term mortality."

DOI: 10.21037/acs.2018.07.02

Background

SYNTAX SCORE II 4-year mortality



Nomogram depicting predicted 4-year mortality as a function of the SYNTAX II Score for patients proposed to undergo myocardial revascularization (CABG or PCI).

Adapted from Farooq et al., The Lancet. 2013 Feb 23;381(9867):639-50

SYNTAX Score II questions

SYNTAX Score I

Age (years)

CrCl mL/min

LVEF (%)

Left Main ○ no ○ yes

Gender ○ male ○ female

COPD ○ no ○ yes

PVD ○ no ○ yes

SYNTAX Score II Calculate

Background

SYNTAX Score II

SYNTAX II

Decision making -between CABG and PCI- guided by the SYNTAX Score II to be endorsed by the Heart Team.

PCI	
SYNTAX Score II:	46.6
PCI 4 Year Mortality:	24.7 %
CABG	
SYNTAX Score II:	26.8
CABG 4 Year Mortality:	5.3 %
Treatment recommendation ⓘ*:*	*CABG or PCI*

Absolute treatment effect is 19.4% in favor of CABG

Problem definition

- Randomized Clinical Trials are designed for estimating relative treatment effects (e.g. RR, OR)
- Can we use RCT data to predict more individualized absolute treatment effects?

Development of treatment effect models

Aim: To compare regression modeling methods on their ability to predict individual absolute treatment effect

ISCB40
LEUVEN, BELGIUM | 2019

Investigation of treatment effect models

Based on logistic regression

- Global: a single model for the whole population
- Partitioning: multiple simple models for partitions of the population

Global models

- Absence of HTE (risk magnification)

$$\text{logit}(P(Y_i = 1|\boldsymbol{x}_i, t_i = 0)) = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} = \eta_i$$

#df = p+2

$$\hat{\delta}_i = P(Y_i = 1|\eta_i)) = \frac{1}{1 + e^{-(\eta_i + \beta_t)}} - \frac{1}{1 + e^{-\eta_i}}$$

- Presence of HTE (individual treatment-covariate interactions)

$$\text{logit}(P(Y_i = 1|\boldsymbol{x}_i, t_i)) = \beta_0 + t_i\beta_1 + \boldsymbol{x}_i^\top \boldsymbol{\beta_m} + t_i\boldsymbol{x}_i^\top \boldsymbol{\beta_z}$$

#df = 2p+2

- Presence of HTE (interaction between treatment and baseline risk)

$$\text{logit}(P(Y_i = 1|\eta_i, t_i)) = \gamma_0 + t_i\gamma_1 + \eta_i + t_i f(\eta_i)$$

#df = p+3+(**1+**)

Partitioning models

- Model-based recursive partitioning
- Start with a simple global model $\text{logit}(P(Y_i = 1|)) = \beta_0 + t_i\beta_1$
- Form partitions $\mathcal{B}_b$ in the space of $\mathcal{X} = X_1 \times \ldots \times X_p$ such that $\text{logit}(P(Y_i = 1|\mathcal{B}_b)) = \beta_{0b} + t_i\beta_{1b}$ holds
- Implemented as
  1. Variable-by-variable subgroup selection (single split)
  2. Single tree
  3. Random forest

Methods for treatment effect modelling

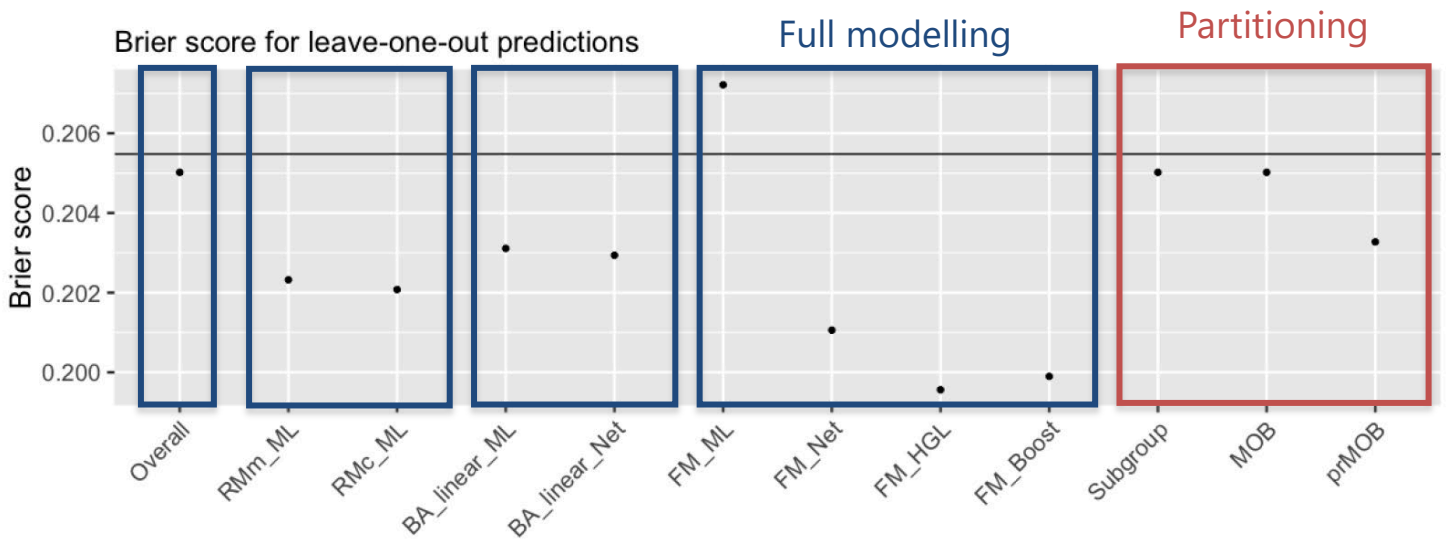
Methods	Equations	Estimation
Global		
Overall absolute treatment effect (Overall)	–	ML
Risk magnification marginal treatment effect (RMm)	(2),(3)	ML, Elastic net
Risk magnification conditional treatment effect (RMc)	(1),(2),(3)	ML, Elastic net
Baseline risk modifier approach linear treatment interaction (BA_linear)	(2),(5)	ML, Elastic net
Full modeling (FM)	(4)	ML, Elastic net, HGL, Boosting
Partitioning		
Single subgroup	(6)	MOB stump
Single tree	(6)	MOB
Random forest	see [17]	pMOB

ML = Maximum Likelihood, HGL = Hierarchical Group Lasso

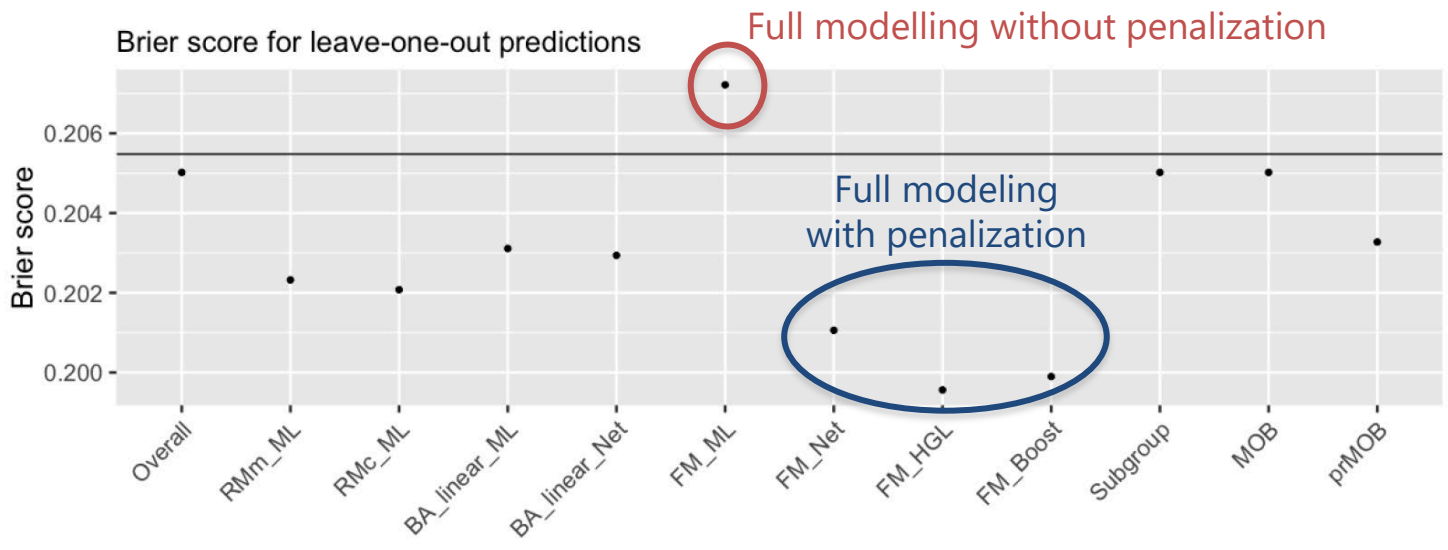
Empirical example

- RCT with 1:1 allocation ratio (N = 512)
- Population: clinically diagnosed acute otitis media (AOM) in children 6 months to 5 years of age
- Intervention: amoxicillin
- Outcome: fever or ear pain was after 3 days' follow-up
- Baseline data on: treatment received, sex, presence of recurrent AOM, fever, bilateral occurrence, ear pain, presence of a runny nose, cough, tympanic membrane abnormality, and age

Empirical example



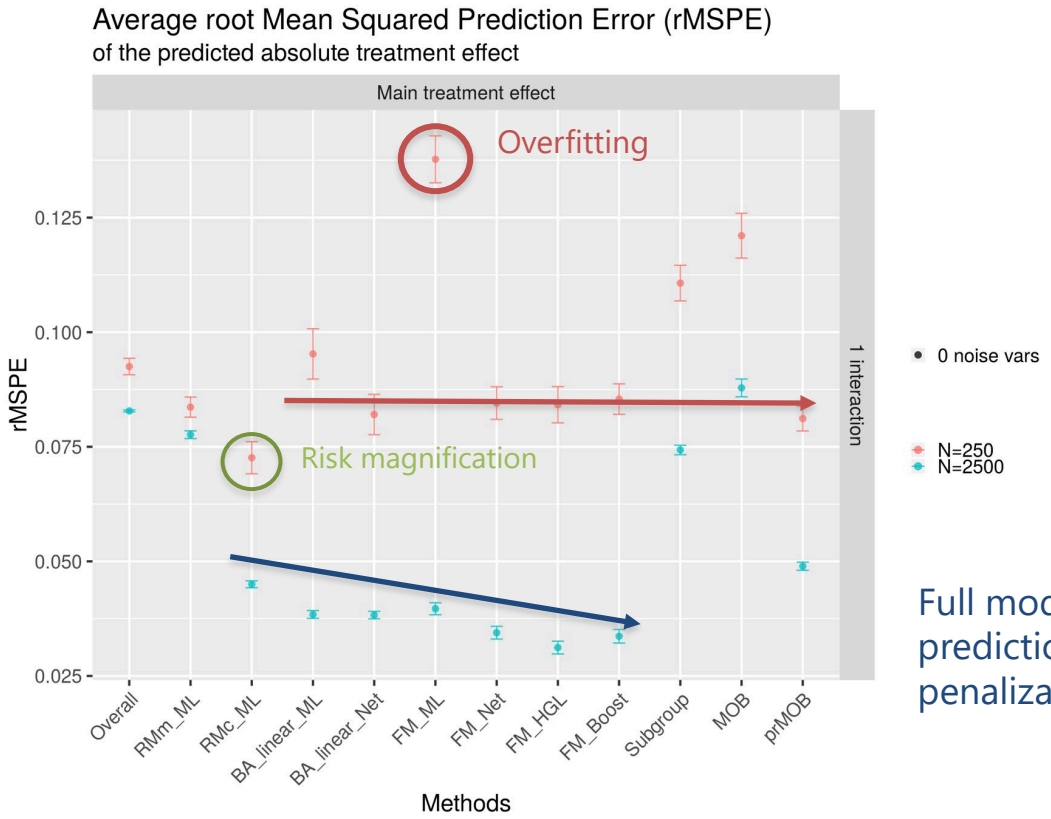
Empirical example



Simulation study

- Logistic data generating mechanism
  - 1:1 allocation ratio
  - 20% event rate
  - 6 covariates with a main effect (MVN with rho = 0.3)
- Variable simulation parameters
  - sample size 250 or 2500
  - presence/absence of average relative treatment effect
  - number and size of treatment-covariate interactions
  - absence/presence of (6) noise variables

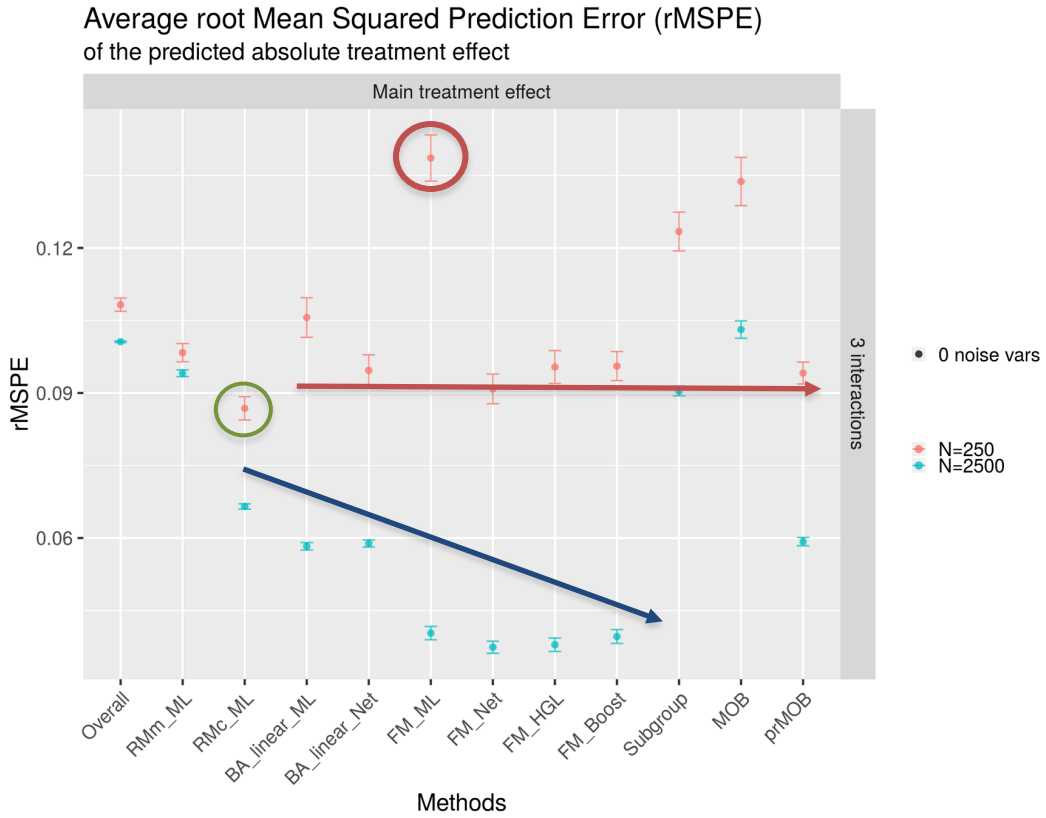
Simulation study results (1 interaction)



Full modelling only improves prediction in large RCTs, but penalization is needed

@TPA_Debray

Simulation study results (3 interactions)



@TPA_Debray

Conclusions

- Small RCTs
  - Hard to improve beyond risk-magnification
  - However, the price to pay to allow for treatment-covariate interactions was small when using both shrinkage and selection, especially for the hierarchical group lasso (HGL)
- Large RCTs
  - Shrinkage and selection still needed
  - Allowing for all interactions was beneficial

Conclusions

- Baseline risk modifier approach, variable-by-variable subgroup, and single MOB were always outperformed
- Random forest MOB performed relatively well given the simulation settings

Next steps

- Improving development and validation in single study
  - Penalization of absolute treatment benefit
  - Machine Learning with assumptions
  - Adjusting for competing risk
  - Quantifying accuracy of absolute treatment benefit
- Evidence synthesis
  - Meta-analysis of individual participant data and published AD
  - Meta-analysis of randomized and observational studies