Development and validation of prediction models

Statistics in Practice II

Ben Van Calster

Department of Development and Regeneration, KU Leuven (B)
Department of Biomedical Data Sciences, LUMC (NL)
Epi-Centre, KU Leuven (B)
1. Model performance and validation

(focus on binary outcomes)
Model validation is important!

What do we mean by validating a prognostic model?

Douglas G. Altman1,*,† and Patrick Royston2

Prognosis and prognostic research: validating a prognostic model

Douglas G Altman,1 Yvonne Vergouwe,2 Patrick Royston,3 Karel G M Moons2

Prognostic models are of little clinical value unless they are shown to work in other samples. Douglas Altman and colleagues describe how to validate models and discuss some of the problems

Assessing the Performance of Prediction Models

A Framework for Traditional and Novel Measures

Ewout W. Steyerberg,a Andrew J. Vickers,b Nancy R. Cook,c Thomas Gerds,d Mithat Gonen,b Nancy Obuchowski,e Michael J. Pencina,f and Michael W. Kattan*g

Practical experiences on the necessity of external validation

I. R. König1, J. D. Malley2, C. Weimar3, H.-C. Diener3 and A. Ziegler1,*,†, on behalf of the German Stroke Study Collaboration
Model validation?

It is not primarily about statistical significance of

1. The predictors

2. A goodness-of-fit test like the Hosmer-Lemeshow test or its variants

Although: a GOF test does give a gist of what is important

→ agreement between predictions and observations

→ CALIBRATION
Calibration is complicated

Calibration of clinical prediction rules does not just assess bias

Werner Vach*

Clinical Epidemiology, Institute of Medical Biometry and Medical Informatics, Freiburg University Medical Center, Freiburg, Germany
Accepted 2 June 2013; Published online 8 September 2013

Abstract

Objectives: Calibration is often thought to assess the bias of a clinical prediction rule. In particular, if the rule is based on a linear logistic model, it is often assumed that an overestimation of all coefficients results in a calibration slope less than 1 and an underestimation in a slope larger than 1.

Study Design and Setting: We investigate the relation of the bias and the residual variation of clinical prediction rules with the typical behavior of calibration plots and calibration slopes, using some artificial examples.

Results: Calibration is not only sensitive to the bias of the clinical prediction rule but also to the residual variation. In some circumstances, the effects may cancel out, resulting in a misleading perfect calibration.

Conclusion: Poor calibration is a clear indication of limited usefulness of a clinical prediction rule. However, a perfect calibration should be interpreted with care as this may happen even for a biased prediction rule.

Keywords: Bias; Calibration; External validation; Prognosis; Prognostic model; Residual variation
Calibration is complicated

A calibration hierarchy for risk models was defined: from utopia to empirical data

Ben Van Calster a,b,*, Daan Nieboer b, Yvonne Vergouwe b, Bavo De Cock a, Michael J. Pencina c,d, Ewout W. Steyerberg b

aKU Leuven, Department of Development and Regeneration, Herestraat 49 Box 7003, 3000 Leuven, Belgium
bDepartment of Public Health, Erasmus MC, 230, 3015 CE Rotterdam, The Netherlands
cDuke Clinical Research Institute, Duke University, 2400 Pratt Street, Durham, NC 27705, USA
dDepartment of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Durham, NC 27719, USA

Accepted 23 December 2015; Published online 6 January 2016
Level 1: Mean calibration

Mean estimated risk = observed proportion of event

“On average, risks are not over- or underestimated.”

O:E ratio: observed events / expected events

If violated: adjust intercept.
Level 2: Weak calibration

Mean AND variability/spread of estimated risk is fine

“On average, risks are not over- or underestimated, nor too extreme/modest.”

Logistic calibration model: \( \log \left( \frac{\pi}{1-\pi} \right) = a + b_L \times LP \)

Calculate - calibration slope: \( b_L \) (cf spread) → should be 1
- calibration intercept: \( a|b_L=1 \) (cf mean) → should be 0

If violated: adjust intercept and coefficients using \( a + b_L \times LP \)
Level 3: Moderate calibration

Estimated risks are correct, conditional on estimated risk.

“Among patients with estimated risk \( x \), the proportion of events is \( x \).”

Flexible calibration model: \( \log \left( \frac{\pi}{1-\pi} \right) = a + f(LP) \)

Calibration curve.

If violated: refit model, perhaps re-evaluate functional form.
Level 4: Strong calibration

Estimated risks are correct, conditional on covariate pattern.

“Per covariate pattern, the proportion of events equals the estimated risk.”

In short, the model is fully correct.

Vach [30] writes: “the idea to identify the “true” model by statistical means is just a great wish which cannot be fulfilled” (p202).
Calibration assessment

### Table 2. A hierarchy of calibration levels for risk prediction models

<table>
<thead>
<tr>
<th>Level</th>
<th>Definition</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Observed event rate equals average predicted risk; “calibration-in-the-large”</td>
<td>* Compare event rate with average predicted risk;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>* Evaluate $a</td>
</tr>
<tr>
<td>Weak</td>
<td>No systematic overfitting or underfitting and/or overestimation or underestimation of risks; “logistic calibration”</td>
<td>Logistic calibration analysis to evaluate $a</td>
</tr>
<tr>
<td>Moderate</td>
<td>Predicted risks correspond to observed event rates</td>
<td>Calibration plot (eg, using loess or splines), or analysis by grouped predictions (including Hosmer-Lemeshow test)</td>
</tr>
<tr>
<td>Strong</td>
<td>Predicted risks correspond to observed event rates for each and every covariate pattern</td>
<td>Scatter plot of predicted risk and observed event rate per covariate pattern; impossible when continuous predictors are involved</td>
</tr>
</tbody>
</table>
Discrimination

Calibration: ~ absolute performance
Discrimination: ~ relative performance

*Do I estimate higher risks for patients with vs without an event?*

C-statistic, equal to AUROC for binary outcomes.
Discrimination

C or AUROC is interesting, ROC not so much.

ROC shows classification results but ‘integrated over’ threshold. But classification critically depends on threshold! ROC = bizarre.

(Pun intended.)
ROC CURVES FOR CLINICAL PREDICTION MODEL SERIES

ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models

Jan Y. Verbakel\textsuperscript{a,b}, Ewout W. Steyerberg\textsuperscript{c}, Hajime Uno\textsuperscript{d}, Bavo De Cock\textsuperscript{e}, Laure Wynants\textsuperscript{e}, Gary S. Collins\textsuperscript{f,g}, Ben Van Calster\textsuperscript{c,e,h}

ROC curves for clinical prediction models part 2. The ROC plot: the picture that could be worth a 1000 words

A. Cecile J.W. Janssens \textsuperscript{a, e}

ROC curves for clinical prediction models part 3. The ROC plot: a picture that needs a 1000 words

Ben Van Calster\textsuperscript{a,b,c,h}, Laure Wynants\textsuperscript{a,d}, Gary S. Collins\textsuperscript{c,f}, Jan Y. Verbakel\textsuperscript{c,g,h}, Ewout W. Steyerberg\textsuperscript{b}

ROC curves for clinical prediction models part 4. Selection of the risk threshold—once chosen, always the same?

A. Cecile J.W. Janssens \textsuperscript{a, e}
Discrimination

If you want a plot...
Overall measures

E.g. Brier score, R-squared measures (including IDI).

For model validation, I prefer to keep discrimination and calibration separate.
Statistical validity = clinical utility?

Discrimination and calibration: how valid is the model?

That is not the full story. In the end, clinical prediction models are made to support clinical decisions!

To make decisions about giving ‘treatment’ or not based on the model, you have to classify patients as high and low risk.

Classifications can be right or wrong: the classical 2x2 table

<table>
<thead>
<tr>
<th></th>
<th>EVENT</th>
<th>NO EVENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIGH RISK</td>
<td>True Positive</td>
<td>False positive</td>
</tr>
<tr>
<td>LOW RISK</td>
<td>False Negative</td>
<td>True negative</td>
</tr>
</tbody>
</table>
Clinical utility

A risk threshold is needed. Which one?

Myth! The statistician can calculate the threshold from the data.
Myth! A threshold is part of the model.

→ “Model BLABLA has sensitivity of 80%” makes little sense to me.
Clinical utility

**ISSUE 1 – A RISK THRESHOLD HAS A CLINICAL MEANING**

What are the utilities ($U$) of TP, FP, TN, FN?

Then threshold $T = \frac{1}{1 + \frac{U_{TP} - U_{FN}}{U_{TN} - U_{FP}}} = \frac{1}{1 + \frac{Benefit\ of\ TP}{Harm\ of\ FP}} = \frac{Harm}{Harm + Benefit}$

Hard to determine.

Instead of fixing $U$ values, start with $T$: at which threshold are you uncertain about ‘treatment’?
Clinical utility

You might say: I would give ‘treatment’ if risk >20%, not if <20%

You may also say that you are willing to ‘treat’ up to 5 patients to correctly treat one (1 TP)

Then $T = 0.2$

You accept up to 4 FP per TP, so Benefit is 4 times the Harm.
Clinical utility

Net Benefit (Vickers & Elkin 2006), even Charles Sanders Peirce (1884)

When you classify patients, TPs are good and FPs are bad
Net result: good stuff minus bad stuff: $TPs - FPs$.

But: relative importance of TP and FP depends on $T$!
- If $T=0.5$, $Harm = Benefit$; else < or >.
- Harm of FP is $\text{odds}(T)$ times the Benefit of a TP

$$\text{NB} = \frac{TPs - \text{odds}(T) \times FPs}{N}$$
Clinical utility

ISSUE 2 – THERE IS NO SINGLE CORRECT THRESHOLD

Preferences and settings differ, so Harm:Benefit and $T$ differ.

Specify a ‘reasonable range’ for $T$.

Calculate NB for this range, and plot results in a decision curve → Decision Curve Analysis
Clinical utility

DCA is a simple research method. You don’t need anything special. It is useful to assess whether a model can improve clinical decision making. *It does not replace a cost-effectiveness analysis, but can inform on further steps.*
Clinical utility

**ISSUE 3 – CPM ONLY ESTIMATES WHETHER RISK \( T \) OR NOT**

\( T \) is independent of any model.

If the model is miscalibrated, you may select too many (risk overestimated) or too few (risk underestimated) patients for treatment!

Miscalibration reduces clinical utility, and may make the model worse than treating all/none! (Van Calster & Vickers 2015)

NB nicelypunishes for miscalibration.
Clinical utility

Models can be miscalibrated!

→ So pick any threshold with a nice combination of sensitivity and specificity!

That does not solve the problem in my opinion.

You may need a different threshold in a new hospital.

So you may just as well recalibrate the model.
Types of validation: evidence pyramid...

(I love evidence pyramids)

- **APPARENT**
  - Exact same data
  - No real validation
  - Optimistic

- **INTERNAL**
  - Same dataset
  - Hence: same population

- **EXTERNAL**
  - Different dataset
  - Hence: different population
Only apparent validation?

Criminal behavior!

Internal validation

- Only assessment of overfitting: c-statistic and slope sufficient.
- Skepticism: who will publish a paper with poor internal validation?
- Which approach?

PREDICTION MODEL FOR DISEASE: MODEL DEVELOPMENT STUDY.

Bruce Van Callstar, Ronald McDonald.

ABSTRACT

Objective. To develop a prediction model to predict disease.

Methods. Logistic regression.

Results. Apparent AUC 0.80. Internally validated AUC 0.60.

Conclusion. Sorry to have wasted your time.

J Pred Mod Res 2021;1;8.
Internal validation

- Train – test split
  - Inefficient: we split up the data into smaller parts
  - “Split sample approaches only work when not needed” (Steyerberg & Harrell 2016)
  - But: it evaluates the fitted MODEL!

- Resampling is preferred
  - Enhanced bootstrap (statistics) or CV (ML) recommended
  - Efficient: you can use all data to develop the model
  - But: it evaluates the modeling PROCEDURE! (Dietterich 1998)
  - It evaluates expected overfitting from the procedure.
Calibration slope based on bootstrapping

- **Negative correlation with true slope, variability underestimated**
- **Example** (Van Calster 2020)
  - 5 predictors, development N=500, 10% event rate, 1000 repetitions
  - Every repetition: slope estimated using bootstrapping
  - Estimated slope: median 0.91, IQR 0.89-0.92, range 0.78-0.96
  - True slope: median 0.88, IQR 0.78-0.99, range 0.53-1.53
  - Spearman rho -0.72
External validation

• Results affected by overfitting and by differences in setting (case-mix, procedures, ...)
• The real test: including ‘an external validation’ settles the issue, right?

These principles, which also inform PLOS Medicine’s editorial priorities for manuscript submissions in this field, require first, that models derived through ML are demonstrably fit for their stated clinical purpose, and second, that researchers undertake and report appropriate efforts to validate these models in external datasets.
2. Expect heterogeneity
Expect heterogeneity

- Between settings: case-mix, definitions and procedures, etc...

*Changing predictor measurement procedures affected the performance of prediction models in clinical examples*

Kim Luijken, Laure Wynants, Maarten van Smeden, Ben Van Calster, Ewout W. Steyerberg, Rolf H.H. Groenwold, Dirk Timmerman, Tom Boume, Chinedu Ukaegbu

![Chart showing the impact of changing predictor measurement procedures on the performance of prediction models in clinical examples.](chart.png)
Expect heterogeneity

- Across time: population drift (Davis et al, 2017)
THERE IS NO SUCH THING AS A VALIDATED MODEL

Validation is important!

But it’s complicated
- Internal and external validations added to model development study may be tampered with (cynical take)
- Internal validation has limitations
- An external validation (whether good or bad) in one setting does not settle the issue
- Performance tomorrow may not be the same as yesterday

So: assess/address/monitor heterogeneity
Address heterogeneity

- Multicenter studies

- IPD
Address heterogeneity

Statistics in Medicine

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

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

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

Thomas P. A. Debray, a*† Karel G. M. Moons, a Ikhlaaq Ahmed, b Hendrik Koffijberg a and Richard David Riley b
Development: random effects models

Address heterogeneity in baseline risk using random cluster intercepts

\[
\log \left( \frac{\pi}{1-\pi} \right) = \alpha + \alpha_j + \mathbf{\beta}^T \mathbf{X},
\]
with \(\alpha_j \sim N(0, \sigma^2)\).

Heterogeneity in predictor effects can be assessed with random slopes

\[
\log \left( \frac{\pi}{1-\pi} \right) = \alpha + \alpha_j + \beta_1 X_1 + \beta_{1,j} + \beta_2 X_2 + \beta_{2,j},
\]
with \[
\begin{bmatrix}
\alpha_j \\
\beta_{1,j} \\
\beta_{2,j}
\end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix}
\sigma_\alpha^2 & \sigma_{\alpha,\beta_1} & \sigma_{\alpha,\beta_2} \\
\sigma_{\alpha,\beta_1} & \sigma_{\beta_1} & \sigma_{\beta_1,\beta_2} \\
\sigma_{\alpha,\beta_2} & \sigma_{\beta_1,\beta_2} & \sigma_{\beta_2}^2
\end{bmatrix} \right)
\]
Add fixed parameters to model that describe the cluster.

E.g. type of hospital, country

Can be done in combination with random effects.
Development: Internal-external CV

k-fold CV

IECV

Model validation in clustered data

External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges

Richard D Riley,1 Joie Ensor,1 Kym I E Snell,2 Thomas P A Debray,3,4 Doug G Altman,5 Karel G M Moons,3,4 Gary S Collins5

A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes

Thomas PA Debray,1,2 © Johanna AAG Damen,1,2 Richard D Riley,3 Kym Snell,3 © Johannes B Reitsma,1,2 Lotty Hooft,1,2 Gary S Collins4 © and Karel GM Moons1,2
Model validation in clustered data

1. Obtain cluster-specific performance
2. Perform random-effects meta-analysis (Snell 2018, Debray 2019)

\[ c \text{ statistic: } \logit(c_j) \sim N(\logit(c), v_j + \tau^2) \]

\[ \text{O:E ratio: } \log(OE_j) \sim N(\log(OE), v_j + \tau^2) \]

Calibration intercept: \( a_j \sim N(a, v_j + \tau^2) \)

Calibration slope: \( b_j \sim N(b, v_j + \tau^2) \)
Model validation in clustered data

One-stage model for calibration slopes:

\[
\text{logit} \left( \frac{P(Y=1)}{1-P(Y=1)} \right) = a + a_j + b \ast \text{logit(risk)} + b_j \ast \text{logit(risk)}, \text{ where}
\]

\[
\begin{bmatrix}
    a_j \\
    b_j
\end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_a^2 & \tau_{ab} \\ \tau_{ab} & \tau_b^2 \end{bmatrix} \right).
\]

\(b\) is overall calibration slope.

\(b_j\) are used to estimate center-specific slopes.

(Bouwmeester 2013; Wynants 2018)
Model validation in clustered data

- For calibration intercepts, the slopes are set to 1:

\[
\text{logit}\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = a' + a'_j + \text{logit}(\text{risk}), \text{ where}
\]

\[
a'_j \sim N(0, \tau^2_{a'j}).
\]

\(a'\) is overall calibration intercept.
\(a'_j\) are used to estimate center-specific intercepts.
Model validation in clustered data

- **Net Benefit at several reasonable thresholds** (Wynants 2018):

  Bayesian trivariate random-effects MA of prevalence, sensitivity, specificity.

\[
\begin{pmatrix}
\logit(p_j) \\
\logit(Se_j) \\
\logit(Sp_j)
\end{pmatrix}
\sim N
\begin{pmatrix}
\gamma_1 \\
\gamma_2 \\
\gamma_3
\end{pmatrix},
\begin{pmatrix}
\tau_1^2 & \rho_{12}\tau_1\tau_2 & \rho_{13}\tau_1\tau_3 \\
\rho_{12}\tau_1\tau_2 & \tau_2^2 & \rho_{23}\tau_2\tau_3 \\
\rho_{13}\tau_1\tau_3 & \rho_{23}\tau_2\tau_3 & \tau_3^2
\end{pmatrix}
\]

**Optimal specification of (vague) priors:** after reparameterization

- Vague normal priors for \( \gamma \)
- Weak Fisher priors for correlations
- Weak half-normal priors for variances
Meta-regression & subgroup analysis

• Largely exploratory undertaking, unless many big clusters

• Meta-regression: add cluster-level predictors to meta-analysis
  o Mean, SD or prevalence of predictors per cluster
  o Can also be added to one-stage calibration analysis

• Subgroup analysis?
  o Conduct the meta-analysis for important subgroups on cluster-level

(Deeks 2008; Debray 2019)
Local and dynamic updating

- Advisable to validate and update an interesting model locally
- And keep it fit by dynamic updating over time (periodic refitting, moving window, ...) (Hickey 2013, Strobl 2015, Su 2018, Davis 2020)
3. Applied example: ADNEX
Case study: the ADNEX model

• **Disclaimer**: I developed this model, but it is by no means perfect.

• Model to estimate the risk that an ovarian tumor is malignant

• Multinomial model for the following 5-level tumor outcome: benign, borderline malignant, stage I invasive, stage II-IV invasive, secondary metastatic. We focus on estimated risk of malignancy

• Population: patients selected for surgery, (outcome based on histology)

• Data: 3506 patients from 1999-2007 (21 centers), 2403 patients from 2009-2012 (18 centers, mostly overlapping).
IOTA data:
5909 patients
24 centers
10 countries
Case study: the ADNEX model

- **Strategy**: develop on first 3506, validate on 2403, then refit on all 5909
  - 2-year gap between datasets, makes sense to do the split
  - IECV might have been an option in hindsight?

- **A priori selection of 10 predictors, then backward elimination** (in combination with multivariable fractional polynomials to decide on transformations). (Royston & Sauerbrei 2008)

- **Heterogeneity**:
  - 1. random intercepts
  - 2. fixed binary predictor indicating type of center: oncology referral center vs other center

Type of center as a predictor?

[Image of a graph showing the percentage of malignant tumours across different centres with error bars.]
Case study: the ADNEX model

• SAS PROC GLIMMIX to fit ri-MLR (took me ages)

```
proc glimmix data=iotamc method=rspl;
class center NewIntercept;
model resp=NewIntercept NewIntercept*oncocenter NewIntercept*l2lesdmax
    NewIntercept*ascites NewIntercept*age NewIntercept*propsol
    NewIntercept*propsol*propsol NewIntercept*loc10 NewIntercept*papnr
    NewIntercept*shadows NewIntercept*l2ca125
    / noint dist=binomial link=logit solution;
random NewIntercept / subject=Center type=un(1);
nloptions tech=nrridg;
output out=iotamcpqlpredca&imp pred(noblup ilink)=p_pqlca pred(noblup noilink)=lp_pqlca;
ods output ParameterEstimates=iotamcpqlca;
run;
```

Case study: the ADNEX model

- Shrinkage: uniform shrinkage factor per equation in the model

- 10 variables, 17 parameters (MFP), 120 cases in smallest group
- Multinomial EPP (de Jong et al 2019): $\frac{120}{((5-1)\times17)} = 2$
- But 52 patients per parameter...
- Sample size for multinomial outcomes needs ‘further research’
- Also, many MFP parameters: less ‘expensive’? Christodoulou et al 2021

- Refit on all data: Multinomial EPP $\frac{246}{68} = 3.6$
- 87 patients per parameter
Case study: the ADNEX model

- Validation on 2403 patients (no meta-analysis, only pooled assessment)

Figure S1. Forest plot with centre-specific areas under the receiver operating characteristic curve (AUC) with regard to discrimination between benign and malignant tumours. Results for the validation data are presented. The AUC is consistent over centres notwithstanding the variability observed for centres that contributed few patients. NC, not computed.
First small external validation

<table>
<thead>
<tr>
<th>Center</th>
<th>N</th>
<th>Events (%)</th>
<th>AUC</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>London (UK)</td>
<td>318</td>
<td>104 (33)</td>
<td>0.942</td>
<td>0.913-0.962</td>
</tr>
<tr>
<td>Southampton (UK)</td>
<td>175</td>
<td>56 (32)</td>
<td>0.900</td>
<td>0.841-0.938</td>
</tr>
<tr>
<td>Catania (IT)</td>
<td>117</td>
<td>22 (19)</td>
<td>0.990</td>
<td>0.959-0.998</td>
</tr>
</tbody>
</table>

Large external validation

Validation of models to diagnose ovarian cancer in patients managed surgically or conservatively: multicentre cohort study

Large external validation

- 4905 patients, 2012-2015, 17 centers

- First large study on the REAL population: any new patient with an ovarian tumor, whether operated or not (2489 were operated)

- If not operated, patients are followed conservatively. Then, outcome was based on clinical and ultrasound information.
  - Differential verification!
  - If insufficient or inconsistent FU information: missing data (MI)
Large external validation: discrimination

• **AUC**
  - Analysis based on logit(AUC) and its SE (Snell 2019) – auRoc package
  - Random effects meta-analysis – metafor package, rma and predict functions
  - 95% prediction interval

• **Calibration**
  - One-stage calibration model for calibration intercept and slope
  - We used this model for calibration curves (i.e. no flexible calibration curves)
  - Overall curve: \( a_j \) and \( b_j \) set to 0
  - Center-specific curves: random terms used
Large external validation: utility

- Decision: referral for specialized oncological care
- Reasonable range of risk thresholds: 5 to 50%

- Per threshold, Net Benefit per center was combined using Bayesian trivariate random effects meta-analysis
- Weak realistic priors
- WinBugs
Large external validation: results

**Supplementary Figure 5.** Forest plot with centre-specific areas under the receiver operating characteristic curve (AUC) of Assessment of Different NEoplasias in the adneXa (ADNEX) with CA125. CI, confidence interval. “Other” includes the following small non-oncology centres with low prevalence of malignancy: London and Nottingham from the UK, and Milan 3 and Florence from Italy.

<table>
<thead>
<tr>
<th>Centre</th>
<th>AUC (95% CI)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malmö (Sweden)</td>
<td>0.93 (0.89 to 0.95)</td>
<td>794</td>
</tr>
<tr>
<td>Rome (Italy)</td>
<td>0.96 (0.93 to 0.97)</td>
<td>681</td>
</tr>
<tr>
<td>Athens (Greece)</td>
<td>0.89 (0.84 to 0.92)</td>
<td>567</td>
</tr>
<tr>
<td>Leuven (Belgium)</td>
<td>0.96 (0.93 to 0.98)</td>
<td>501</td>
</tr>
<tr>
<td>Genk (Belgium)</td>
<td>0.95 (0.89 to 0.98)</td>
<td>406</td>
</tr>
<tr>
<td>Milan 1 (Italy)</td>
<td>0.97 (0.95 to 0.99)</td>
<td>367</td>
</tr>
<tr>
<td>Stockholm (Sweden)</td>
<td>0.93 (0.89 to 0.95)</td>
<td>363</td>
</tr>
<tr>
<td>Monza (Italy)</td>
<td>0.95 (0.90 to 0.97)</td>
<td>267</td>
</tr>
<tr>
<td>Cagliari (Italy)</td>
<td>0.96 (0.89 to 0.99)</td>
<td>166</td>
</tr>
<tr>
<td>Katowice (Poland)</td>
<td>0.94 (0.82 to 0.98)</td>
<td>139</td>
</tr>
<tr>
<td>Pamplona (Spain)</td>
<td>1.00 (0.83 to 1.00)</td>
<td>111</td>
</tr>
<tr>
<td>Trieste (Italy)</td>
<td>0.96 (0.85 to 0.99)</td>
<td>111</td>
</tr>
<tr>
<td>Milan 2 (Italy)</td>
<td>0.97 (0.91 to 0.99)</td>
<td>98</td>
</tr>
<tr>
<td>Other</td>
<td>0.81 (0.63 to 0.91)</td>
<td>334</td>
</tr>
</tbody>
</table>

**Meta-analysis**

<table>
<thead>
<tr>
<th>AUC (95% CI)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC (95% CI)</td>
<td>0.94 (0.92 to 0.96)</td>
</tr>
<tr>
<td>95% Prediction interval</td>
<td>(0.83 to 0.98)</td>
</tr>
</tbody>
</table>

![Forest plot](attachment:image.png)
Large external validation: results

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC (95% CI)</th>
<th>AUC (95% CI)</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMI</td>
<td>0.89 (0.85 to 0.92)</td>
<td></td>
<td>(0.74 to 0.96)</td>
</tr>
<tr>
<td>LR2</td>
<td>0.92 (0.89 to 0.94)</td>
<td></td>
<td>(0.82 to 0.96)</td>
</tr>
<tr>
<td>SRRisk</td>
<td>0.94 (0.91 to 0.95)</td>
<td></td>
<td>(0.83 to 0.98)</td>
</tr>
<tr>
<td>ADNEX without CA125</td>
<td>0.94 (0.91 to 0.95)</td>
<td></td>
<td>(0.82 to 0.98)</td>
</tr>
<tr>
<td>ADNEX with CA125</td>
<td>0.94 (0.92 to 0.96)</td>
<td></td>
<td>(0.83 to 0.98)</td>
</tr>
</tbody>
</table>

Fig 2 | Summary forest plot with overall area under the receiver operating characteristic curve (AUC) for each model. ADNEX=assessment of different neoplasias in the adnexa; LR2=logistic regression model 2; PI=prediction interval; RMI=risk of malignancy index; SRRisk=simple rules risk model
Large external validation: results

Supplementary Figure 16. Histogram of estimated risks of malignancy given by Assessment of Different NEoplasias in the adnexa (ADNEX) with CA125. Results are based on a stacked dataset of the 100 completed datasets following multiple imputation.
Large external validation: results

**Table 1**

<table>
<thead>
<tr>
<th>Model</th>
<th>Intercept (95% CI)</th>
<th>Slope (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR2</td>
<td>0.30 (-0.02 to 0.63)</td>
<td>0.84 (0.72 to 0.96)</td>
</tr>
<tr>
<td>SRRisk</td>
<td>0.14 (-0.19 to 0.47)</td>
<td>1.00 (0.87 to 1.13)</td>
</tr>
<tr>
<td>ADNEX without CA125</td>
<td>0.19 (-0.04 to 0.42)</td>
<td>1.12 (0.98 to 1.26)</td>
</tr>
<tr>
<td>ADNEX with CA125</td>
<td>0.19 (-0.01 to 0.40)</td>
<td>1.11 (0.98 to 1.25)</td>
</tr>
</tbody>
</table>

**Fig 3** | Summary figure with overall calibration curves for risk prediction models.
ADNEX=assessment of different neoplasias in the adnexa; Intercept=calibration intercept; LR2=logistic regression model 2; RMI=risk of malignancy index; slope=calibration slope; SRRisk=simple rules risk model
Large external validation: results

Supplementary Figure 10. Centre-specific calibration curves of Assessment of Different NEoplasias in the adneXa (ADNEX) with CA125. “Other” includes the following smaller non-oncology centres with low prevalence of malignancy: London and Nottingham from the UK, and Milan 3 and Florence from Italy.
Large external validation: results

Fig 4 | Overall decision curves for risk prediction models and RMI. Higher net benefit implies higher clinical utility (the higher the curve, the better the clinical utility at the chosen risk threshold). ADNEX = assessment of different neoplasias in the adnexa; LR2 = logistic regression model 2; RMI = risk of malignancy index; SRRisk = simple rules risk model
Large external validation: subgroups

Operated (n=2489)
38% malignant

At least 1 FU visit (n=1958)
1% malignant
Large external validation: subgroups

Oncology centers (n=3094)
26% malignant

Other centers (n=1811)
10% malignant

(Interestingly, LR2 (blue line) is the only model without type of center as a predictor)
Large external validation: meta-regression

\[
\text{rma.uni(logit\_auc, sei = se\_logit\_auc, mods = logit(EventRate), data = iota5cs, method = "REML", knha=TRUE)}
\]

- R\(^2\) 42\%  
  - p 0.02

- R\(^2\) 34\%  
  - p 0.04

- R\(^2\) 11\%  
  - p 0.24

- R\(^2\) 16\%  
  - p 0.09
Large external validation: meta-regression

\[
rma.uni(lnOE, sei = se_lnOE, mods = logit(EventRate), data = iota5cs, method = "REML", knha=TRUE)
\]

- \( R^2 \) 41\%  
  - \( p \) 0.29

- \( R^2 \) 0\%  
  - \( p \) 0.50

- \( R^2 \) 0\%  
  - \( p \) 0.52

- \( R^2 \) 0\%  
  - \( p \) 0.31
# Independent external validation studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Location</th>
<th>N</th>
<th>ER (%)</th>
<th>AUC</th>
<th>Calibration</th>
<th>Utility</th>
<th>Missing Data</th>
<th>Reporting guideline mentioned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joyeux 2016</td>
<td>Dijon/Chalon (FR)</td>
<td>284</td>
<td>11</td>
<td>0.938</td>
<td>No</td>
<td>No</td>
<td>CCA</td>
<td>None</td>
</tr>
<tr>
<td>Szubert 2016</td>
<td>Poznan (PL)</td>
<td>204</td>
<td>34</td>
<td>0.907</td>
<td>No</td>
<td>No</td>
<td>CCA</td>
<td>None</td>
</tr>
<tr>
<td>Pamplona (SP)</td>
<td></td>
<td>123</td>
<td>28</td>
<td>0.955</td>
<td>No</td>
<td>No</td>
<td>CCA</td>
<td>None</td>
</tr>
<tr>
<td>Araujo 2017</td>
<td>Sao Paulo (BR)</td>
<td>131</td>
<td>52</td>
<td>0.925</td>
<td>No</td>
<td>No</td>
<td>CCA</td>
<td>STARD</td>
</tr>
<tr>
<td>Meys 2017</td>
<td>Maastricht (NL)</td>
<td>326</td>
<td>35</td>
<td>0.930</td>
<td>No</td>
<td>No</td>
<td>MI</td>
<td>STARD</td>
</tr>
<tr>
<td>Chen 2019</td>
<td>Shanghai (CN)</td>
<td>278</td>
<td>27</td>
<td>0.940</td>
<td>No</td>
<td>No</td>
<td>CCA</td>
<td>None</td>
</tr>
<tr>
<td>Stukan 2019</td>
<td>Gdynia (PL)</td>
<td>100</td>
<td>52</td>
<td>0.972</td>
<td>No</td>
<td>No</td>
<td>CCA</td>
<td>STARD, TRIPOD</td>
</tr>
<tr>
<td>Jeong 2020</td>
<td>Seoul (KR)</td>
<td>59</td>
<td>17</td>
<td>0.924</td>
<td>No</td>
<td>No</td>
<td>No info</td>
<td>None</td>
</tr>
<tr>
<td>Viora 2020</td>
<td>Turin (IT)</td>
<td>577</td>
<td>25</td>
<td>0.911</td>
<td>No</td>
<td>No</td>
<td>CCA</td>
<td>None</td>
</tr>
<tr>
<td>Nam 2021</td>
<td>Seoul (KR)</td>
<td>353</td>
<td>4</td>
<td>0.920</td>
<td>No</td>
<td>No</td>
<td>No info</td>
<td>None</td>
</tr>
<tr>
<td>Poonyakanok 2021</td>
<td>Bangkok (TH)</td>
<td>357</td>
<td>17</td>
<td>0.975</td>
<td>No</td>
<td>No</td>
<td>CCA</td>
<td>None</td>
</tr>
</tbody>
</table>
Independent external validation studies

```r
metagen(logitc, logitc_se, data = adnex, studlab = Publication, sm = "PLOGIT", backtransf = T, level = 0.95, level.comb = 0.95, comb.random = T, prediction = T, level.predict = 0.95, method.tau = "REML", hakn = TRUE)
```
Heterogeneity?
4. Machine learning
Regression vs Machine Learning

Statistical Modeling: The Two Cultures
Leo Breiman

The Data Modeling Culture
The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from
response variables = \( f(\text{predictor variables, random noise, parameters}) \)
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:

\[ y \quad \text{linear regression} \quad \text{logistic regression} \quad \text{Cox model} \quad x \]

The Algorithmic Modeling Culture
The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function \( f(x) \)—an algorithm that operates on \( x \) to predict the responses \( y \). Their black box looks like this:

\[ y \quad \text{unknown} \quad x \]

\[ \text{decision trees} \quad \text{neural nets} \]
Regression vs Machine Learning

Monopolies in Different Fields

Explain
Social Sciences
Describe
Statistics
Predict
Machine Learning

Statistics versus machine learning

Statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns.
Reason for popularity

“Typical machine learning algorithms are highly flexible
So will uncover associations we could not find before
Hence better predictions and management decisions”

→ One of the master keys, with guaranteed success!
1. Study design trumps algorithm choice

Do you have a clear research question?
Do you have data that help you answer the question?
What is the quality of the data?

One simple example: EHR data were not collected for research purposes
Uses ANN, Random Forests, Naïve Bayes, and Logistic Regression. (Logistic regression performed best, AUC 0.92)
Example (contd)

The model also uses postoperative information.
David J Cohen, MD: “The model can’t be run properly until you know about both the presence and the absence of those complications, but you don’t know about the absence of a complication until the patient has left the hospital.”
Example 2

Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition

Julia K. Winkler, MD; Christine Fink, MD; Ferdinand Toberer, MD; Alexander Enk, MD; Teresa Deinlein, MD; Rainer Hofmann-Wellenhof, MD; Luc Thomas, MD; Aimillos Lallas, MD; Andreas Blum, MD; Wilhelm Stolz, MD; Holger A. Haenssle, MD

A Unmarked C Marked

E Unmarked G Marked

Eric Topol @EricTopol Aug 14
How surgical skin markings faked out a deep learning #AI neural net-- a commercially approved product for algorithm-aided melanoma diagnosis. Highly instructive. Machines can be dumb. jamanetwork.com/journals/jamad ... @JAMA Derm by UniHeidelberg

KU LEUVEN

Winkler et al. JAMA Dermatol 2019; in press.
Example 3

Automated characterisation of ultrasound images of ovarian tumours: the diagnostic accuracy of a support vector machine and image processing with a local binary pattern operator

Objectives: In this study, we developed and validated a computerised model to characterise ovarian masses as benign or malignant.

Materials and methods: Transvaginal 2D B mode static ultrasound images of 187 ovarian masses with known histological diagnosis were included. Images were first pre-processed and enhanced, and Local Binary Pattern Histograms were then extracted from 2 × 2 blocks of each image. A Support Vector Machine (SVM) was trained using stratified cross validation with randomised sampling. The process was repeated 15 times and in each round 100 images were randomly selected.

The image used for each mass was selected by author JK that on subjective impression was most representative of the final histopathology.
2. Flexible algorithms are data hungry
Flexible algorithms are data hungry

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

Tjeerd van der Ploeg¹, Peter C Austin² and Ewout W Steyerberg³

Conclusions: Modern modelling techniques such as SVM, NN and RF may need over 10 times as many events per variable to achieve a stable AUC and a small optimism than classical modelling techniques such as LR. This implies that such modern techniques should only be used in medical prediction problems if very large data sets are available.
Flexible algorithms are data hungry

Don't use deep learning your data isn't that big

The other day Brian was at a National Academies meeting and he gave one of his usual classic quotes:

![CMU Stats & DS](https://twitter.com/CMU_Stats/status/861300548878449920)

Best quote from NAS DS Round Table: "I mean, do we need deep learning to analyze 30 subjects?" - B Caffo @simplystats

In problems where data are limited, deep learning often is not an ideal solution.
3. Signal-to-noise ratio (SNR) is often low

There is support that flexible algorithms work best with high SNR, not with low SNR.

*How can methods that look everywhere be better when you do not know where to look or what you look for?*
Algorithm flexibility and SNR

...adaptive non-linear methods are most useful in problems with high signal-to-noise ratio, sometimes occurring in engineering and physical science. In human medical studies, the signal-to-noise ratio is often quite low (as it is here), and hence the modern methods may have less to offer.

One possible explanation for why the simpler analytic approaches work better is that clinical data in general, and EHR data particularly, can be quite noisy. It is noted that the standard deviation of the longitudinal measurements was often a strong predictor. Therefore, analytic approaches that require a degree of regularity or smoothness may be ill suited. However, even in our simulation, where regularity was imposed, we did not find that more complex methods performed better. Instead, all methods performed equally well. The lack of variability in prediction performance, compared with the empirical data, highlights that under regular conditions, all of these methods obtain the same targets. Therefore, it is in noisy, real data settings that the added value of simplicity is realized.

Table 10 is our attempt to show that not all applications can be modeled equally well using AI algorithms. Games are the easiest as the rules are known and do not change, the environment is also known and stable, the predictions cannot influence the future and there is no uncertainty. The exact opposite is true for forecasting applications where not only the rules are not known but can also change, there are structural instabilities in the data, while there is plenty of uncertainty and noise, that can confuse the search for the optimal weights. Moreover,
So I find this hard to believe

If a model is expected to automatically analyze noisy data without any intervening human curation or normalization, then the task becomes complex, and complex models become generally more useful.

When they have large enough data sets, modern models can be successfully trained to map noisy inputs to noisy outputs. The use of a smaller set
What if you don’t know?

If you have no knowledge on what variables could be good predictors (and what variables not), are you ready to make a good prediction model?

I do not think that using flexible algorithms will bypass this.

I like a clear distinction between
- predictor finding studies
- prediction modeling studies
Machine learning: success guaranteed?

• Question 1: which algorithm works when?
  o Medical data: often limited signal:noise ratio
  o High-dimensional data vs CPM
  o Direct prediction from medical images using DL: different ballgame?

• Question 2: ML for low-dimensional CPMs?
A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, Ben Van Calster

*Department of Development & Regeneration, KU Leuven, Herestraat 49 box 805, Leuven, 3000 Belgium
**Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, OX3 7LD UK
***Oxford University Hospitals NHS Foundation Trust, Oxford, UK
****Department of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, Leiden, 2333 ZA The Netherlands
*****Department of Public Health & Primary Care, KU Leuven, Kapucijnenvoer 33 box 7001, Leuven, 3000 Belgium
******Nuffield Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford, OX2 6GG UK

Accepted 5 February 2019; Published online 11 February 2019
### Table A.6. Descriptive statistics, of papers and study characteristics.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Unknown or NA</th>
<th>Median</th>
<th>Interquartile range</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal impact factor</td>
<td>71</td>
<td>6</td>
<td>2.8</td>
<td>2.5-4.2</td>
<td>0.6-10.1</td>
</tr>
<tr>
<td>Number of centers if multicenter</td>
<td>27</td>
<td>10</td>
<td>5</td>
<td>4-15</td>
<td>2-1,137</td>
</tr>
<tr>
<td>Total sample size(^a)</td>
<td>71</td>
<td>1</td>
<td>1,250</td>
<td>353-188,861</td>
<td>72-3,994,872</td>
</tr>
<tr>
<td>Number of predictors(^b)</td>
<td>71</td>
<td>3</td>
<td>19</td>
<td>11-32</td>
<td>5-563</td>
</tr>
<tr>
<td>Event rate(^c)</td>
<td>102</td>
<td>14</td>
<td>0.18</td>
<td>0.09-0.35</td>
<td>0.002-0.50</td>
</tr>
<tr>
<td>Events per predictor, training data(^d)</td>
<td>128</td>
<td>26</td>
<td>8</td>
<td>4-34</td>
<td>0.3-6,697</td>
</tr>
</tbody>
</table>

\(^a\) Total sample size is not reported in all studies.

\(^b\) Number of predictors is not reported in all studies.

\(^c\) Event rate is not reported in all studies.

\(^d\) Events per predictor, training data is not reported in all studies.
ML vs LR: systematic review

<table>
<thead>
<tr>
<th>Type of algorithm</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression (LR) methods</td>
<td>71 (100%)</td>
</tr>
<tr>
<td>- Standard LR only</td>
<td>54</td>
</tr>
<tr>
<td>- Standard and penalized LR</td>
<td>9</td>
</tr>
<tr>
<td>- Penalized LR only</td>
<td>6</td>
</tr>
<tr>
<td>- Standard LR and boosted LR</td>
<td>1</td>
</tr>
<tr>
<td>- Bagged LR</td>
<td>1</td>
</tr>
<tr>
<td>Alternative machine learning methods</td>
<td></td>
</tr>
<tr>
<td>- Classification tree (e.g., CART, C4.5)</td>
<td>30 (42%)</td>
</tr>
<tr>
<td>- Random forest (RF)</td>
<td>28 (39%)</td>
</tr>
<tr>
<td>- Support vector machine (SVM)</td>
<td>24 (34%)</td>
</tr>
<tr>
<td>- Artificial neural network (ANN)</td>
<td>26 (37%)</td>
</tr>
<tr>
<td>- Other algorithms</td>
<td>30 (42%)</td>
</tr>
<tr>
<td>- Boosted tree methods (e.g., gradient boosting machines)</td>
<td>16</td>
</tr>
</tbody>
</table>
## ML vs LR: systematic review

Table 2. Overview of methods for model validation at study level ($n = 71$)

<table>
<thead>
<tr>
<th>Type of validation</th>
<th>Validation: risk of bias classification</th>
<th>$N$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Unclear/yes</td>
</tr>
<tr>
<td>None</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Single random split</td>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td>Resampling</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>Repeated random splits</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Cross-validation</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Bootstrapping</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>External</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Chronological split</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Split by center</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Internal-external CV</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Different data set</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Type depends on algorithm</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Total, $n$ (%)</td>
<td>23 (32%)</td>
<td>48 (68%)</td>
</tr>
</tbody>
</table>

Counts refer to articles. Risk of bias in model validation refers to the first of five bias signaling items that were used in this study.

No risk of bias: the item was scored as “no” for all models in the study; unclear: the item was scored as “unclear” for at least one model; yes: the item was scored as “yes” (bias present) for at least one model.

Note: Table A.2 describes the five bias items. For bias in model validation, we repeat the description here: We discern two general criteria to assess the validation: first, it should be clear that models are developed using training data only; second, if validation is performed using resampling (repeated data splitting, cross-validation, bootstrapping), it should be clear that all model building steps are repeated in every training data set; ad hoc flaws are documented and tabulated.
Poor modeling and unclear reporting

What was done about missing data? 45% fully unclear, 100% poor or unclear

How were continuous predictors modeled? 20% unclear, 25% categorized

How were hyperparameters tuned? 66% unclear, 19% tuned with information

How was performance validated? 68% unclear or biased approach

Was calibration of risk estimates studied? 79% not at all, HL test common

Prognosis: time horizon often ignored completely

ML vs LR: comparison of AUC

<table>
<thead>
<tr>
<th>Overall</th>
<th>Diff logit(AUC) (95% CI)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any ML vs LR</td>
<td>0.25 (0.12;0.38)</td>
<td>282</td>
</tr>
<tr>
<td>Tree vs LR</td>
<td>0.00 (-0.15;0.15)</td>
<td>42</td>
</tr>
<tr>
<td>RF vs LR</td>
<td>0.33 (0.18;0.49)</td>
<td>59</td>
</tr>
<tr>
<td>SVM vs LR</td>
<td>0.24 (0.10;0.39)</td>
<td>43</td>
</tr>
<tr>
<td>ANN vs LR</td>
<td>0.47 (0.32;0.62)</td>
<td>52</td>
</tr>
<tr>
<td>Other ML vs LR</td>
<td>0.22 (0.07;0.37)</td>
<td>86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Low risk of bias</th>
<th>Diff logit(AUC) (95% CI)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any ML vs LR</td>
<td>0.00 (-0.18;0.18)</td>
<td>145</td>
</tr>
<tr>
<td>Tree vs LR</td>
<td>-0.34 (-0.65;-0.04)</td>
<td>16</td>
</tr>
<tr>
<td>RF vs LR</td>
<td>0.06 (-0.15;0.26)</td>
<td>39</td>
</tr>
<tr>
<td>SVM vs LR</td>
<td>0.03 (-0.20;0.26)</td>
<td>17</td>
</tr>
<tr>
<td>ANN vs LR</td>
<td>-0.12 (-0.35;0.12)</td>
<td>27</td>
</tr>
<tr>
<td>Other ML vs LR</td>
<td>-0.09 (-0.30;0.12)</td>
<td>46</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High risk of bias</th>
<th>Diff logit(AUC) (95% CI)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any ML vs LR</td>
<td>0.34 (0.20;0.47)</td>
<td>137</td>
</tr>
<tr>
<td>Tree vs LR</td>
<td>0.05 (-0.10;0.20)</td>
<td>26</td>
</tr>
<tr>
<td>RF vs LR</td>
<td>0.41 (0.22;0.60)</td>
<td>20</td>
</tr>
<tr>
<td>SVM vs LR</td>
<td>0.33 (0.19;0.48)</td>
<td>26</td>
</tr>
<tr>
<td>ANN vs LR</td>
<td>0.71 (0.55;0.88)</td>
<td>25</td>
</tr>
<tr>
<td>Other ML vs LR</td>
<td>0.31 (0.15;0.47)</td>
<td>40</td>
</tr>
</tbody>
</table>
A few anecdotic observations

| One study matched participants with and without the outcome condition on age and gender, and then used these variables as predictors for the outcome. |
| One paper deletes the top and bottom 1% of values for continuous predictors to avoid a large influence of outliers, but then imputes these values using mean imputation. |
| One paper deletes nearly all data in order to obtain a ‘balanced’ data set (i.e. 50% event rate). The observed event rate is 1%, such that nearly all non-events had to be excluded. |
Conclusions

Validation is important, but there is no such thing as a validated model.

Expect and address heterogeneity in model development and validation.

Ideally, models should be locally and dynamically updated. Realistic?

Study methodology trumps algorithm choice.

Machine learning: no success guaranteed.

Reporting is key. TRIPOD extensions are underway (Cluster and AI).
TG6 - Evaluating diagnostic tests and prediction models

Chairs: Ewout Steyerberg, Ben Van Calster

Members: Patrick Bossuyt, Tom Boyles (clinician), Gary Collins, Kathleen Kerr, Petra Macaskill, David McLernon, Carl Moons, Maarten van Smeden, Andrew Vickers, Laure Wynants

Papers:

- A few others in preparation.