Development and evaluation of prediction models: pitfalls and solutions

Statistics in practice I

Maarten van Smeden, PhD

Biometrisches Kolloquium

16 March 2021
Aims of the lectures

• To give our views on the state of the art of clinical prediction modelling

• To highlight common pitfalls and potential solutions when developing or evaluating clinical prediction models

• To argue the need for less model development and more model evaluation

• To plead for increased attention to quantifying differences in model performance over time and place
Agenda

• PART I (11:00 – 12:40)
  • State of the medical prediction modeling art
  • Just another prediction model
  • Methods against overfitting
  • Methods for deciding on appropriate sample size

• PART II (13:30 – 15:10)
  • Model performance and validation
  • Heterogeneity over time and place: there is no such thing as a validated model
  • Applied example: ADNEX model
  • Future perspective: machine learning and AI
Agenda

• PART I (11:00 – 12:40)
  • State of the medical prediction modeling art
  • Just another prediction model
  • Methods against overfitting
  • Methods for deciding on appropriate sample size

• PART II (13:30 – 15:10)
  • Model performance and validation
  • Heterogeneity over time and place: there is no such thing as a validated model
  • Applied example: ADNEX model
  • Future perspective: machine learning and AI
State of the medical prediction modeling art
What do we mean with a prediction model?

“… summarize the effects of predictors to provide individualized predictions of the absolute risk of a diagnostic or prognostic outcome.”
Steyerberg, 2019

”Model development studies aim to derive a prediction model by selecting the relevant predictors and combining them statistically into a multivariable model.”
TRIPOD statement, 2015

“… may be developed for scientific or clinical reasons, or both”
Altman & Royston, 2000
What do I mean with a prediction model?

- Mathematical rule or equation derived from empirical data describing the relation between a health outcome \(Y\) and input \(X\).

- The aim is to find a function \(\hat{f}(X)\) that yields accurate individual level predictions, \(\hat{y}_i\).

- Finding the optimal \(\hat{f}(X)\) is usually not only prediction error minimization problem, but also takes into account requirements regarding transparency, transportability, costs, sparsity etc.

- Interpretation and contribution of individual components in \(\hat{f}(X)\) are not of primary interest (exception: issues related to fairness).
Figure: Schematic representation of diagnostic and prognostic model studies, adapted from TRIPOD reporting guideline¹

van Smeden et al., JCE, in press
**Method of Scoring in Evaluation of Newborn Infant**

<table>
<thead>
<tr>
<th>Sign</th>
<th>Score†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart rate                  Absent</td>
<td>0</td>
</tr>
<tr>
<td>Respiratory effort Absent</td>
<td>1</td>
</tr>
<tr>
<td>Muscle tone Limp</td>
<td>2</td>
</tr>
<tr>
<td>Reflex irritability No response extremities</td>
<td></td>
</tr>
<tr>
<td>Color Blue; pale</td>
<td></td>
</tr>
<tr>
<td>Body pink; extremities blue</td>
<td></td>
</tr>
</tbody>
</table>

* Evaluation 60 seconds after complete birth of infant (disregarding the cord and placenta).
† Score of 10 indicates infant in best possible condition.

Apgar et al. JAMA, 1958
## Wells Prediction Rule for DVT

<table>
<thead>
<tr>
<th>Clinical variable</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active cancer (treatment ongoing or within previous 6 months, or palliative)</td>
<td>1</td>
</tr>
<tr>
<td>Paralysis, paresis, or immobilization of lower extremity</td>
<td>1</td>
</tr>
<tr>
<td>Bedridden for &gt; 3 days or major surgery within previous 12 weeks requiring general or regional anesthesia</td>
<td>1</td>
</tr>
<tr>
<td>Localized tenderness along distribution of deep veins</td>
<td>1</td>
</tr>
<tr>
<td>Swelling of entire leg</td>
<td>1</td>
</tr>
<tr>
<td>Calf swelling at least 3 cm larger than asymptomatic leg (10 cm below tibial tuberosity)</td>
<td>1</td>
</tr>
<tr>
<td>Pitting edema confined to symptomatic leg</td>
<td>1</td>
</tr>
<tr>
<td>Collateral superficial veins (non-varicose)</td>
<td>1</td>
</tr>
<tr>
<td>Previously documented DVT</td>
<td>1</td>
</tr>
<tr>
<td>Alternative diagnosis as least as likely as DVT</td>
<td>-2</td>
</tr>
</tbody>
</table>

### Clinical risk (probability)

<table>
<thead>
<tr>
<th>Clinical risk (probability)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>&gt;2</td>
</tr>
<tr>
<td>Moderate</td>
<td>1-2</td>
</tr>
<tr>
<td>Low</td>
<td>&lt;1</td>
</tr>
</tbody>
</table>

Wells et al., Lancet, 1997
Prognostic prediction model

Simple model

LDH < 365
- yes: hsCRP < 41.2
  - yes: Survival
    - num: 177
      - T: 177
      - F: 0
  - no: num: 202
    - no: Lymphocyte(%) > 14.7
      - yes: Survival
        - num: 12
          - T: 9
          - F: 3
      - no: Death
        - num: 25
          - T: 146
          - F: 3
- no: Death
  - num: 149
    - T: 11
    - F: 2
Deep learning (= Neural network)

- 128,000 images
- Transfer learning (preinitialization)
- Sensitivity and specificity > .90
  - Estimated from training data
“Bedside” prediction models
"Bedside" prediction models

Numerical example: 10 year death due to cardiovascular disease
Why bother?

“As of today, we have deployed the system in 16 hospitals, and it is performing over 1,300 screenings per day”
MedRxiv pre-print only, 23 March 2020, doi.org/10.1101/2020.03.19.20039354
Covid: Extra 1.7m vulnerable added to shielding list

By TILDA TRAYL
Health correspondent

01 day ago | 0 Comments

Coronavirus pandemic

There is to be a large expansion of the number of people being asked to shield in England.

An extra 1.7 million people are expected to be added to the 2.3 million already on the list.

Half of the group have not yet been vaccinated so will now be prioritised urgently by their local GPs.

It comes after a new model was developed that takes into account extra factors rather than just health.

This calculation includes things such as ethnicity, deprivation (by postcode) and weight to work out a person’s risk of becoming seriously ill if they were to catch Covid.

It also looks at age, underlying health issues and prescribed medications.

Prof Andrew Hayward, a member of the Nasa and Emerging Respiratory Virus Threats Advisory Group (Nerva), which has been involved in the modelling, said it was considered a combination of factors such as age, ethnicity and chronic illness and put them together to reach a score.
Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal

Laure Wynants,1,2 Ben Van Calster,2,3 Marc M J Bonten,4,5 Gary S Collins,6,7 Thomas P A Debray,4,8 Maarten De Vos,2,9 Maria C Haller,10,11 Georg Heinze,10 Karel G M Moons,4,8 Richard D Riley,12 Ewoud Schuit,4,8 Luc J M Smits,1 Kym I E Snell,12 Ewout W Steyerberg,3 Christine Wallisch,10,13,14 Maarten van Smeden4

ABSTRACT
OBJECTIVE
To review and critically appraise published and preprint reports of prediction models for diagnosing coronavirus disease 2019 (covid-19) in patients with suspected infection, for prognosis of patients with covid-19, and for detecting people in the general population at risk of being admitted to hospital for covid-19 pneumonia.

were identified for predicting hospital admission from pneumonia and other events (as proxy outcomes for covid-19 pneumonia) in the general population; 18 diagnostic models for detecting covid-19 infection (13 were machine learning based on computed tomography scans); and 10 prognostic models for predicting mortality risk, progression to severe disease, or length of hospital stay. Only one study used patient data from outside of China. The most
• Published on 7 April 2020

• **18 days** between idea and article acceptance (**sprint**)

• Invited by BMJ as the first ever **living review** (**marathon**)

OBJECTIVE

To review and critically appraise published and preprint reports of prediction models for diagnosing coronavirus disease 2019 (covid-19) in patients with suspected infection, for prognosis of patients with covid-19, and for detecting people in the general population at risk of being admitted to hospital for covid-19 pneumonia.
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 Marc M J Bonten,8,10 Darren L Dahly,11,12 Johanna A Damen,8,9 Thomas P A Debray,8,9 Valentijn M T de Jong,8,9 Maarten De Vos,9,13 Paula Dhiman,4,5 Maria C Haller,7,14 Michael O Harhay,15,16 Liesbet Henckaerts,17,18 Pauline Heus,8,9 Michael Kammer,7,19 Nina Kreuzberger,20 Anna Lohmann,21 Kim Luijken,21 Jie Ma,5 Glen P Martin,22 David J McLernon,23 Constanza L Andaur Navarro,8,9 Johannes B Reitsma,8,9 Jamie C Sergeant,24,25 Chunhu Shi,26 Nicole Skoetz,19 Luc J M Smits,1 Kym I E Snell,6 Matthew Sperrin,27 René Spijker,8,9,28 Ewout W Steyerberg,3 Toshihiko Takada,8 Ioanna Tzoulaki,29,30 Sander M J van Kuijk,31 Bas C T van Bussel,1,32 Iwan C C van der Horst,32 Florien S van Royen,8 Jan Y Verbakel,33,34 Christine Wallisch,7,35,36 Jack Wilkinson,22 Robert Wolff,37 Lotty Hooft,8,9 Karel G M Moons,8,9 Maarten van Smeden8

ABSTRACT

OBJECTIVE
To review and appraise the validity and usefulness of published and preprint reports of prediction models for diagnosing coronavirus disease 2019 (covid-19) infection and predicting disease outcome.

DATA SOURCES
PubMed and Embase through Ovid, up to 1 July 2020, supplemented with arXiv, medRxiv, and bioRxiv up to 5 May 2020.

STUDY SELECTION

Wynants et al. BMJ, 2020
Prediction model related to prognosis and diagnosis

3 main types models
1. Patient X is infected / COVID-19 pneumonia diagnostic
2. Patient X will die from COVID-19 / need respiratory support prognostic
3. Currently healthy person X will get severe COVID-19 general population
Data extraction

• 43 researchers, duplicate reviews

• Extraction form based on CHARMS checklist & PROBAST (127 questions)

• Assessment of each prediction model separately if more than one was developed/validated
UPDATE 3

• **232 models** reviewed

• Peer reviewed articles in *Pubmed* and *Embase*

• Pre-prints only until update 2 from *bioRxiv, medRxiv, and arXiv*

• Search: up to 1 July 2020
Some key numbers

Origin of data

• Single country: N = 174, 75% (of which from China: 56%)
• Multiple countries: 18%
• Unknown origin: 7%

Target setting

• Patients admitted to hospital: N = 119, 51%
• Patient at triage centre or fever clinic: 5%
• Patients in general practice: 1%
• Other/unclear: 42%
Some key numbers

Target population
- Confirmed COVID-19: N = 108, 47%
- Suspected COVID-19: 36%
- Other/unclear: 18%

Sample size
- Sample size development, median: 338 (IQR: 134—707)
  - number of events, median: 69 (37—160)
- Sample size external validation, median: 189 (76—312)
  - Number of events, median: 40 (24—122)
Models

- Logistic regression (including regularized): 35%
- Neural net (including deep learning): 33%
- Tree-based (including random forest): 7%
- Survival models (including Cox ph): 6%
- SVM: 4%
- Other/unclear: 15%
Reported performance – AUC range

- General population models: 0.71 to 0.99
- Diagnostic models: 0.65 to 0.99
- Diagnostic severity models: 0.80 to 0.99
- Diagnostic imaging models: 0.70 to 0.99
- Prognosis models: 0.54 to 0.99
  prediction horizon varied from 1 to 37 days
Participants
• Inappropriate or unclear in/exclusion or study design

Predictors
• Scored “unknown” in imaging studies

Outcome
• Subjective or proxy outcomes

Analysis
• Small sample size
• Inappropriate or incomplete evaluation of performance
Conclusion update 3 living review

“…models are all at high or unclear risk of bias”

We do “not recommend any of the current prediction models to be used in practice, but one diagnostic and one prognostic model originated from higher quality studies and should be (independently) validated in other datasets”
There have been serious replication issues w/ #AI but now it has become prominent for #COVID19 applications. A recent mortality predictive model w/ 3 biomarkers (LDH, CRP, lymphocyte count) was shown to be irreproducible by 3 groups @NatMachIntell
Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans


Machine learning methods offer great promise for fast and accurate detection and prognostication of coronavirus disease 2019 (COVID-19) from standard-of-care chest radiographs (CXR) and chest computed tomography (CT) images. Many articles have been published in 2020 describing new machine learning-based models for both of these tasks, but it is unclear which are of potential clinical utility. In this systematic review, we consider all published papers and preprints, for the period from 1 January 2020 to 3 October 2020, which describe new machine learning models for the diagnosis or prognosis of COVID-19 from CXR or CT images. All manuscripts uploaded to bioRxiv, medRxiv and arXiv along with all entries in EMBASE and MEDLINE in this timeframe are considered. Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, 62 studies were included in this systematic review. Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. This is a major weakness, given the urgency with which validated COVID-19 models are needed. To address this, we give many recommendations which, if followed, will solve these issues and lead to higher-quality model development and well-documented manuscripts.

Roberts et al. Nature ML, 2021
Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study

Rishi K. Gupta 1,2, Michael Marks 2,3, Thomas H.A. Samuels 2, Akish Luintel 2, Tommy Rampling 2, Humayra Chowdhury 2, Matteo Quartagno 4, Arjun Nair 2, Marc Lipman 5, Ibrahim Abubakar 1, Maarten van Smeden 6, Wai Keong Wong 2, Bryan Williams 7,8 and Mahdad Noursadeghi 2,9, on behalf of The UCLH COVID-19 Reporting Group 10

@ERSpublications
Oxygen saturation on room air and patient age are strong predictors of deterioration and mortality, respectively, among hospitalised adults with COVID-19. None of the 22 prognostic models evaluated in this study adds incremental value to these univariable predictors. https://bit.ly/2Hg24TO

Gupta, ERJ, 2020
b) Deterioration models vs SpO2 on air alone

Gupta, ERJ, 2020

---

d) Mortality models vs age alone

Gupta, ERJ, 2020
## Landscape of clinical prediction models

- 408 models for COPD prognosis (Bellou, 2019)
- 363 models for cardiovascular disease general population (Damen, 2016)
- 263 prognosis models in obstetrics (Kleinrouweler, 2016)
- 258 models mortality after general trauma (Munter, 2017)
- 232 models related to COVID-19 (Wynants, 2020)
- 160 female-specific models for cardiovascular disease (Baart, 2019)
- 119 models for critical care prognosis in LMIC (Haniffa, 2018)
- 101 models for primary gastric cancer prognosis (Feng, 2019)
- 81 models for sudden cardiac arrest (Carrick, 2020)
- 74 models for contrast-induced acute kidney injury (Allen, 2017)
- 73 models for 28/30 day hospital readmission (Zhou, 2016)
- 68 models for preeclampsia (De Kat, 2019)
- 67 models for traumatic brain injury prognosis (Dijkland, 2019)
- 64 models for suicide / suicide attempt (Belsher, 2019)
- 61 models for dementia (Hou, 2019)
- 58 models for breast cancer prognosis (Phung, 2019)
- 52 models for pre-eclampsia (Townsend, 2019)
- 52 models for colorectal cancer risk (Usher-Smith, 2016)
- 48 models for incident hypertension (Sun, 2017)
- 46 models for melanoma (Kaiser, 2020)
- 46 models for prognosis after carotid revascularisation (Volkers, 2017)
- 43 models for mortality in critically ill (Keuning, 2019)
- 42 models for kidney failure in chronic kidney disease (Ramspek, 2019)
- 40 models for incident heart failure (Sahle, 2017)
- 37 models for treatment response in pulmonary TB (Peetluk, 2021)
- 35 models for in vitro fertilisation (Ratna, 2020)
- 34 models for stroke in type-2 diabetes (Chowdhury, 2019)
- 34 models for graft failure in kidney transplantation (Kabore, 2017)
- 31 models for length of stay in ICU (Verburg, 2016)
- 27 models for pediatric early warning systems (Trubey, 2019)
- 27 models for malaria prognosis (Njim, 2019)
- 26 models for postoperative outcomes colorectal cancer (Souwer, 2020)
- 26 models for childhood asthma (Kothalawa, 2020)
- 25 models for lung cancer risk (Gray, 2016)
- 25 models for re-admission after admitted for heart failure (Mahajan, 2018)
- 23 models for recovery after ischemic stroke (Jampathong, 2018)
- 23 models for delirium in older adults (Lindroth, 2018)
- 21 models for atrial fibrillation detection in community (Himmelreich, 2020)
- 19 models for survival after resectable pancreatic cancer (Stijker, 2019)
- 18 models for recurrence hep. carcinoma after liver transplantation (Al-Ameri, 2020)
- 18 models for future hypertension in children (Hamoen, 2018)
- 18 models for risk of falls after stroke (Walsh, 2016)
- 18 models for mortality in acute pancreatitis (Di, 2016)
- 17 models for bacterial meningitis (van Zeggeren, 2019)
- 17 models for cardiovascular disease in hypertensive population (Cai, 2020)
- 14 models for ICU delirium risk (Chen, 2020)
- 14 models for diabetic retinopathy progression (Haider, 2019)
PREDICTION MODELS

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)

situation:
there are 14 competing standards.

14?! ridiculous!
we need to develop
one universal standard
that covers everyone's use cases.
yeah!

soon:
situation:
there are 15 competing standards.

Original: https://xkcd.com/927/
Agenda

• PART I (11:00 – 12:40)
  • State of the medical prediction modeling art
  • Just another prediction model
  • Methods against overfitting
  • Methods for deciding on appropriate sample size

• PART II (13:30 – 15:10)
  • Model performance and validation
  • Heterogeneity over time and place: there is no such thing as a validated model
  • Applied example: ADNEX model
  • Future perspective: machine learning and AI
OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.

STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH ALGORITHMS!

SIX MONTHS LATER: WOW, THIS PROBLEM IS REALLY HARD.
YOU DON'T SAY.
Diagnosis: “starting point of all medical actions”
"Usually doctors are right, but conservatively about 15 percent of all people are misdiagnosed. Some experts think it's as high as 20 to 25 percent"
“Patient trust was essential in the healing process. It could be won by a punctilious bedside manner, by meticulous explanation, and by mastery of prognosis, an art of demanding experience, observation and logic”

Galen, 2nd century AD
The PROGRESS (PROGnosis RESearch Strategy) framework classifies prognosis research into four main types of study:

1. Overall prognosis research
2. Prognostic factor research
3. Prognostic model research
4. Predictors of treatment effect research

More details: https://www.prognosisresearch.com/progress-framework
Overall prognosis research

Figure 1. Cumulative Incidence of Sudden Cardiac Death and All-Cause Mortality After Myocardial Infarction Among Residents of Olmsted County, Minnesota

Adabag et al. JAMA, 2008
Prognostic factor research

Letellier et al. BJC, 2017
Predictors of treatment effect research

Biomarkers as Predictors of Response to Treatment with Motesanib in Patients with Progressive Advanced Thyroid Cancer

Michael B. Bass, Steven I. Sherman, Martin J. Schlumberger, Michael T. Davis, Lisa Kivman, Huan-Mei Khoo, Kimberly H. Notari, Matthew Peach, Yong-jiang Hei, and Scott D. Patterson

Conclusions: Changes in PIGF and soluble VEGF receptor 2 levels after initiation of therapy predicted response to motesanib in patients with advanced differentiated thyroid cancer or metastatic medullary thyroid cancer. Lower baseline VEGF levels were associated with longer PFS. (J Clin Endocrinol Metab 95: 5018–5027, 2010)
PROGRESS framework

The PROGRESS (PROGnosis RESearch Strategy) framework classifies prognosis research into four main types of study:

1. Overall prognosis research Prevalence studies
2. Prognostic factor research Diagnostic test (accuracy) studies
3. Prognostic model research Diagnostic model research
4. Predictors of treatment effect research

More details: https://www.prognosisresearch.com/progress-framework
Focus of this lecture

The PROGRESS (PROGnosis RESearch Strategy) framework classifies prognosis research into four main types of study:

1. Overall prognosis research  Prevalence studies
2. Prognostic factor research  Diagnostic test (accuracy) studies
3. **Prognostic model research**  **Diagnostic model research**
4. Predictors of treatment effect research

More details: https://www.prognosisresearch.com/progress-framework
The phases of prognostic/diagnostic model research

1. Development (including internal validation)
2. External validation(s) and updating
3. Software development and regulations
4. Impact assessment
5. Clinical implementation and scalability

• Most research focusses on development
• Few models are validated
• Fewer models reach the implementation
Not fit for purpose
- Wrong target population
- Expensive predictors
- Discrepancies between development and use
- Complexity/transparency

No validation/impact
- Poor development
- Insufficient reporting
- Incentives
- If done, usually small studies

Regulation/implementation
- MDR
- Soft- and hardware
- Model updating
- Quality control

Not adopted
- User time
- No prediction needed

Picture courtesy: Laure Wynants
Books (focus on development/validation)
Typical regression models

**Binary logistic model**

\[
\Pr(Y = 1) = \expit(\beta_0 + \beta_1 X_1 + \ldots + \beta_P X_P) = \exp(X\beta)/[1 + \exp(X\beta)]
\]

**Multinomial logistic model**

\[
\Pr(Y = j) = \exp(X\beta_j)/[1 + \sum_{h=1}^{J-1} \exp(X\beta_h)]
\]

**Cox model**

\[
h(X, t) = h_0(t)\exp(\beta_1 X_1 + \ldots + \beta_P X_P) = h_0(t)\exp(X\beta)
\]

**Remarks**
- $X\beta$ is the linear predictor
- Most predictive performance metrics do not directly generalize between outcomes (discrimination/calibration)
- Linearity/additivity assumptions can be relaxed
- Tree based models (e.g. random forest), SVM and neural networks seem to be upcoming
Model users often unaware of models and equations

### Prediction models for 6 month outcome after TBI

<table>
<thead>
<tr>
<th>Admission Characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (14-99 years)</td>
<td>34</td>
</tr>
<tr>
<td>Motor Score</td>
<td>Normal Flexion</td>
</tr>
<tr>
<td>Pupils</td>
<td>Both reading</td>
</tr>
<tr>
<td>Core+CT</td>
<td></td>
</tr>
<tr>
<td>Hypoxia</td>
<td>No</td>
</tr>
<tr>
<td>Hypotension</td>
<td>No</td>
</tr>
<tr>
<td>CT Classification</td>
<td>Diffuse Injury III</td>
</tr>
<tr>
<td>ISAH on CT</td>
<td>Yes</td>
</tr>
<tr>
<td>Epidural mass on CT</td>
<td>Yes</td>
</tr>
<tr>
<td>Core+CT+Lab</td>
<td></td>
</tr>
<tr>
<td>Glucose (3-20 mmol/L)</td>
<td>10 mmol/L</td>
</tr>
<tr>
<td>Hb (6-17 g/dL)</td>
<td>10 g/dL</td>
</tr>
</tbody>
</table>

This model predicts outcome in the following patients:

- Adults with head injury, Glasgow Coma Scale 12 or less.

**Prognostic Results:**

- Predicted probability of 6 month mortality: **Core model: 17%**
- Predicted probability of 6 month unfavourable outcome: **Core model: 32%**

- Predicted probability of 6 month mortality: **Core+CT model: 20%**
- Predicted probability of 6 month unfavourable outcome: **Core+CT model: 31%**

http://www.tbi-impact.org/?p=impact/calc
## Table 3: Prognostic factors for survival of patients with metastatic breast cancer (multivariate analysis)

<table>
<thead>
<tr>
<th>Prognostic factors/index</th>
<th>$p$</th>
<th>Hazard ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>KPS</td>
<td>0.030</td>
<td>0.563</td>
<td>0.336–0.945</td>
</tr>
<tr>
<td>Nottingham prognostic groups</td>
<td>&lt;0.0001</td>
<td>5.147</td>
<td>3.272–8.099</td>
</tr>
<tr>
<td>Liver metastasis</td>
<td>0.275</td>
<td>1.367</td>
<td>0.780–2.394</td>
</tr>
<tr>
<td>Number of metastasis</td>
<td>0.190</td>
<td>1.418</td>
<td>0.841–2.391</td>
</tr>
<tr>
<td>First-/second-/third-line chemotherapy for recurrence or metastasis</td>
<td>0.002</td>
<td>0.503</td>
<td>0.324–0.780</td>
</tr>
</tbody>
</table>

Baseline hazard?
Example: Systematic review by Bellou et al, 2019

Prognostic models for COPD: 228 articles
(408 prognostic models, 38 external validations)

Not reported:
• 12%: **modelling method** (e.g. type of regression model)
• 24%: evaluation of **discrimination** (e.g. area under the ROC curve)
• 64%: **missing data** handling (e.g. multiple imputation)
• 70%: **full model** (e.g. no regression coefficients, intercept / baseline hazard)
• 78%: no evaluation of **calibration** (e.g. calibration plot)

Bellou et al. BMJ, 2019
Reporting guidelines & risk of bias tools

• **TRIPOD**: reporting development/validation prediction models
• **PROBAST**: risk of bias development/validation prediction models

• **STARD**: reporting diagnostic accuracy studies
• **QUADAS-2**: risk of bias diagnostic accuracy studies

• **REMARK**: reporting (tumour marker) prognostic factor studies
• **QUIPS**: risk of bias in prognostic factor studies

• Currently in development: TRIPOD-AI, TRIPOD-cluster, PROBAST-AI, STARD-AI, QUADAS-AI, DECIDE-AI etc.

More info: equator-network.org
Dichotomania

“Dichotomania is an obsessive compulsive disorder to which medical advisors in particular are prone… Show a medical advisor some continuous measurements and he or she immediately wonders. ‘Hmm, how can I make these clinically meaningful? Where can I cut them in two? What ludicrous side conditions can I impose on this?’”

Stephen Senn
It’s all in the title

1. Problems in dichotomizing continuous variables (Altman 1994)
2. Dangers of using "optimal" cutoff points in the evaluation of prognostic factors. (Altman et al 1994)
3. How bad is categorization? (Weinberg; 1995)
4. Seven reasons why you should NOT categorize continuous data (Dinero; 1996)
5. Breaking Up is Hard to Do: The Heartbreak of Dichotomizing Continuous Data (Streiner; 2002)
6. Negative consequences of dichotomizing continuous predictor variables (Irwin & McClelland; 2003)
7. Why carve up your continuous data? (Owen 2005)
9. Categorizing continuous variables resulted in different predictors in a prognostic model for nonspecific neck pain (Schellingerhout et al 2006)
10. Dichotomizing continuous predictors in multiple regression: a bad idea (Royston et al 2006)
11. The cost of dichotomising continuous variables (Altman & Royston; 2006)
12. Leave 'em alone - why continuous variables should be analyzed as such (van Walraven & Hart; 2008)
13. Dichotomization of continuous data--a pitfall in prognostic factor studies (Metze; 2008)
14. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms (Nagagari et al 2011)
15. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents (Bennette & Vickers; 2012)
16. Dichotomizing continuous variables in statistical analysis: a practice to avoid (Dawson & Weiss; 2012)
17. The danger of dichotomizing continuous variables: A visualization (Kuss 2013)
18. The “anathema” of arbitrary categorization of continuous predictors (Vintzileos et al; 2014)
19. Ophthalmic statistics note: the perils of dichotomising continuous variables (Cumberlaid et al 2014)

Unnecessary dichotomizations of predictors

- **Biological implausible** step-functions in predicted risk
- **Source of overfitting** when cut-off is chosen based on maximizing predictive performance
- **Loss of information** (e.g. Ensor et al. show dichotomizing BMI was equivalent to throwing away 1/3 of the data)

Dichotomization/categorization remains **very prevalent**
- Wynants et al. 2020 (COVID prediction models): 48%
- Collins et al. 2011 (Diabetes T2 prediction models): 63%
- Mallett et al. 2010 (Prognostic models in cancer): 70%
Unnecessary dichotomizations of predictions

OUR ACCURATE PREDICTION MODEL SAYS THAT YOU WILL DIE WITHIN 5 YEARS OF CARDIOVASCULAR DISEASE

ACTUALLY, YOUR RISK IS ONLY 5%, BUT A 4% RISK THRESHOLD MAXIMIZED PREDICTIVE PERFORMANCE
# Predictimands

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Estimand</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ignore treatment</td>
<td>Risk of the event, regardless of treatment</td>
<td>Risk of cardiovascular events where some patients will initiate statins according to routine-care prescriptions</td>
</tr>
<tr>
<td>Composite</td>
<td>Risk of the event or treatment initiation</td>
<td>Risk of a composite of cardiovascular death, myocardial infarction and treatment with revascularisation (PCI or CABG)</td>
</tr>
<tr>
<td>While untreated</td>
<td>Risk of the event occurring before treatment is started</td>
<td>Risk of dying while on the waiting list for a liver transplant</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>Risk of the event if treatment were never started</td>
<td>Risk of a natural pregnancy without IVF treatment</td>
</tr>
</tbody>
</table>

Van Geloven et al. EJE, 2020
Predictimands

Van Geloven et al. EJE, 2020
Agenda

• PART I (11:00 – 12:40)
  • State of the medical prediction modeling art
  • Just another prediction model
  • Methods against overfitting
  • Methods for deciding on appropriate sample size

• PART II (13:30 – 15:10)
  • Model performance and validation
  • Heterogeneity over time and place: there is no such thing as a validated model
  • Applied example: ADNEX model
  • Future perspective: machine learning and AI
Overfitting =

image source: https://bit.ly/3eENofJ
THE BEST WAY TO EXPLAIN OVERFITTING
Classical consequence of overfitting

Bell et al. BMJ, 2015
Shortest routes to overfitting

Study Design and Setting: A previously published model predicted the likelihood of having a mutation in germline DNA mismatch repair genes at the time of diagnosis of colorectal cancer. This model was based on a cohort where 38 mutations were found among 870 participants, with validation in an independent cohort with 35 mutations. The modeling strategy included stepwise selection of predictors from a pool of over 37 candidate predictors and dichotomization of continuous predictors. We simulated this strategy in small subsets of a large contemporary cohort (2,051 mutations among 19,866 participants) and made comparisons to other modeling approaches. All models were evaluated according to bias and discriminative ability (concordance index, c) in independent data.
Study Design and Setting:  
repair genes at the time of diagnosis, with validation in at least 10% from a pool of over 37 candidate large contemporary cohort (2,051 were evaluated according to bias
Routes to preventing overfitting

• **Sample size**: gather data on a sufficient* number of individuals
  - Use it all on model development (avoid training-test splitting)

• Careful candidate **predictors preselection**
  - Especially if data are small

• **Avoid** and be aware of the **winner’s curse**
  - Data driven variable selection, model selection

• **Avoid** unnecessary **dichotomizations**

• Regression **shrinkage** / penalization / regularization

* Topic of 4th topic of this talk
calibration plot

- ideal
- model

observed

predicted
calibration plot

observed

predicted

ideal

model
Shrinkage

Post-estimation shrinkage factor estimation
• Van Houwelingen & Le Cessie, 1990: uniform shrinkage factor, $S_{VH}$
• Sauerbrei 1999: parameterwise shrinkage factors

Regularized regression (shrinkage during estimation)
• Ridge regression: L2-penalty on regression coefficient
• Lasso: L1 penalty
• Elastic net: L1 and L2 penalty
Shrinkage

Post-estimation shrinkage factor estimation
\[
\Pr(Y = 1) = \expit[\beta_0^* + S_{VH}(\beta_1 X_1 + \ldots + \beta_p X_p)]
\]

Regularized regression (shrinkage during estimation)
\[
\ln L_p = \ln L_{ml} - \lambda \left[ (1 - \alpha) \sum_{p=1}^{P} \beta_p^2 + a \sum_{p=1}^{P} |\beta_p| \right]
\]
Ridge regression: \( a = 0 \), Lasso: \( a = 1 \), Elastic net \( 0 < a < 1 \)
Uniform shrinkage factor

\[ S_{VH} = 1 + \frac{p}{n \ln \left( 1 - R_{CS\_app}^2 \right)} \]

Shrinkage < 1 means overfitting
lower values = more overfitting

Formula from Riley et al.
Shrinkage/tuning/penalization hard to estimate

\[ R_{app}^2 \] is 0.56 for model B when using the maximum of 253 participants

“shrinkage works on the average but may fail in the particular unique problem on which the statistician is working.”
Shrinkage works on the average

Simulation results, averaged over 4k scenarios and 20M datasets

van Smeden et al. SMMR, 2019
Shrinkage may not work on a particular dataset

Maximum likelihood, Ridge, Lasso

RMSD log(slope): root mean squared distance of log calibration slope to the ideal slope (value = log(1))

Van Calster et al. SMMR, 2020
Shrinkage may not work on a particular dataset

“We conclude that, despite improved performance on average, shrinkage often worked poorly in individual datasets, in particular when it was most needed. The results imply that shrinkage methods do not solve problems associated with small sample size or low number of events per variable.”
What about the variance in estimated risk?

van Smeden et al. SMMR, 2019
Sources of prediction error

\[ Y = f(x) + \varepsilon \]

For a model \( k \) the **expected test prediction error** is:

\[
\sigma^2 + \text{bias}^2 \left( \hat{f}_k(x) \right) + \text{var} \left( \hat{f}_k(x) \right)
\]

Irreducible error \( \approx \) Mean squared prediction error

What we don’t model \( \approx \) How we model

(with \( \text{E}(\varepsilon) = 0, \text{var}(\varepsilon) = \sigma^2 \), values in \( x \) are not random)

See equation 2.46 in Hastie et al., the elements of statistical learning, https://stanford.io/2voWjra
Agenda

• PART I (11:00 – 12:40)
  • State of the medical prediction modeling art
  • Just another prediction model
  • Methods against overfitting
  • Methods for deciding on appropriate sample size

• PART II (13:30 – 15:10)
  • Model performance and validation
  • Heterogeneity over time and place: there is no such thing as a validated model
  • Applied example: ADNEX model
  • Future perspective: machine learning and AI
You get what you pay for
Sample size considerations

• Sample **size matters** when designing a **study of any kind**

• When designing a prediction study precedes data collection (**prospective**)
  • **How many** units/individuals/events do I need to collect data on?

• When designing a prediction study on existing data (**retrospective**)
  • Is my dataset **large enough** to build a model?
  • How much **model complexity** (e.g. number predictors considered) can I afford?
One in ten rule

From Wikipedia, the free encyclopedia

In statistics, the one in ten rule is a rule of thumb for how many predictors can be derived from data when doing regression analysis (in particular proportional hazards models and logistic regression) without risk of overfitting. The rule states that one predictive variable can be studied for every ten events. It is not applicable to ordinary least squares linear regression, where it is suggested that as few as two events per predictor are sufficient.
"For EPV values of 10 or greater, no major problems occurred. For **EPV values less than 10**, however, the regression coefficients were **biased in both positive and negative directions**."
Peduzzi et al. JCE, 1996
More simulation studies

A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis
Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alan R. Feinstein

For EPV values of 10 or greater, no major problems

citations: 5,736

“a minimum of 10 EPV [...] may be too conservative”
citations: 2,438

Performance of logistic regression modeling: beyond the number
of events per variable, the role of data structure
Delphine S. Courvoisier, Christophe Combescure, Thomas Agoritsas,
Angèle Gayet-Ageron, Thomas V. Perneger

“substantial problems even if the number of EPV exceeds 10”
citations: 216

Citations based on Google Scholar, Oct 30 2020
More simulation studies

For EPV values of 10 or greater, no major problems

"substantial p"
Log(odds) is consistent but finite sample biased

On the bias of various estimators of the logit and its variance with application to quantal bioassay

BY JOHN J. GART AND JAMES R. ZWEIFEL
National Cancer Institute, Bethesda, Md

Small-Sample Bias of Point Estimators of the Odds Ratio from Matched Sets

Nicholas P. Jewell
Department of Biomedical and Environmental Health Sciences, School of Public Health, University of California, Berkeley, California 94720, U.S.A.

A comparative investigation of methods for logistic regression with separated or nearly separated data

Georg Heinze*†

Logistic Discrimination and Bias Correction in Maximum Likelihood Estimation

J. A. Anderson and S. C. Richardson

Bias Correction in Generalized Linear Models

By GAUSS M. CORDEIRO† and PETER McCULLAGH
Universidade Federal de Pernambuco, Brazil
University of Chicago, USA

BMC Medical Research Methodology

Research article
Bias in odds ratios by logistic regression modelling and sample size
Szilard Nemes*1, Junmei Mao Jonasson1, Anna Genell1 and Gunnar Steineck1,2

M.vanSmeden@umcutrecht.nl | Twitter: @MaartenvSmeden
BIAS CORRECTION IN MAXIMUM LIKELIHOOD LOGISTIC REGRESSION

ROBERT L. SCHAEFER

Department of Mathematics and Statistics, Miami University, Oxford, Ohio, U.S.A.

\[
\text{bias } (\hat{\beta}) \approx \left( -\frac{1}{2} \right) (X^T V X)^{-1} X^T \left\{ \text{trace} \left( X (X^T V X)^{-1} E^k (X^T V X)^{-1} X^T \text{Var} (\varepsilon) \right) \right\} \\
\left( -\frac{1}{2} \right) (X^T V X)^{-1} X^T V \left\{ (1 - 2\pi_k) x_k^T (X^T V X)^{-1} x_k \right\}
\]

(6)

No rationale for 1 variable per 10 events criterion for binary logistic regression analysis


• Examine the reasons for substantial differences between the earlier EPV simulation studies
• Evaluate a possible solution to reduce the finite sample bias

van Smeden et al. BMC MRM, 2016
No rationale for 1 variable per 10 events criterion for binary logistic regression analysis

Maarten van Smeden¹**, Joris A. H. de Groot¹, Karel G. M. Moons¹, Gary S. Collins², Douglas G. Altman², Marinus J. C. Eijkemans¹ and Johannes B. Reitsma¹

- Examine the reasons for substantial differences between the earlier EPV simulation studies (simulation technicality: handling of “separation”)
- Evaluate a possible solution to reduce the finite sample bias
• Examine the reasons for substantial differences between the earlier EPV simulation studies (simulation technicality: handling of “separation”)

• Evaluate a possible solution to reduce the finite sample bias
Firth’s "correction" aims to reduce finite sample bias in maximum likelihood estimates, applicable to logistic regression.

It makes clever use of the “Jeffries prior” (from Bayesian literature) to penalize the log-likelihood, which shrinks the estimated coefficients.

Nice theoretical justifications.
Averaged over 465 simulation conditions with 10,000 replications each
Averaged over 465 simulation conditions with 10,000 replications each
Firth’s correction and predictive performance

Maximum likelihood, Firth’s correction

RMSD log(slope): root mean squared distance of log calibration slope to the ideal slope (value = log(1))

Van Calster, SMMR, 2020, DOI: 10.1177/0962280220921415
Beyond events per variable

• Sample size has **a big influence** on the performance of prediction models
  • Giving appropriate weights to predictors is not a simple task

• Challenge is **to avoid overfitting and lack of precision** in the predictions
  • Requires adequate sample size

• How much data is sufficient? Depends
  • **Model complexity** (e.g. # predictors, EPV)
  • **Signal:noise ratio** (e.g. R² or C-index)
  • **Performance required** (e.g. high vs low stake medical decisions)
Sample size for binary logistic prediction models: Beyond events per variable criteria

Maarten van Smeden, Karel GM Moons, Joris AH de Groot, Gary S Collins, Douglas G Altman, Marinus JC Eijkemans, and Johannes B Reitsma

Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes

Richard D. Riley, Kym I.E. Snell, Joie Ensor, Danielle L. Burke, Frank E. Harrell Jr, Karel GM Moons, Gary S. Collins

Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes

Richard D Riley, Kym I.E. Snell, Joie Ensor, Danielle L. Burke, Frank E. Harrell Jr, Karel GM Moons, Gary S. Collins
Calculating the sample size required for developing a clinical prediction model

Clinical prediction models aim to predict outcomes in individuals, to inform diagnosis or prognosis in healthcare. Hundreds of prediction models are published in the medical literature each year, yet many are developed using a dataset that is too small for the total number of participants or outcome events. This leads to inaccurate predictions and consequently incorrect healthcare decisions for some individuals. In this article, the authors provide guidance on how to calculate the sample size required to develop a clinical prediction model.

Richard D Riley professor of biostatistics1, Joie Ensor lecturer in biostatistics1, Kym I E Snell lecturer in biostatistics1, Frank E Harrell Jr professor of biostatistics2, Glen P Martin lecturer in health data sciences3, Johannes B Reitsma associate professor4, Karel G M Moons professor of clinical epidemiology4, Gary Collins professor of medical statistics5, Maarten van Smeden assistant professor1 2 4 5 6

Riley et al., BMJ, 2020
Our proposal

- Calculate sample size that is needed to
  - minimise potential overfitting
  - estimate probability (risk) precisely

- Sample size formula’s for
  - Continuous outcomes
  - Time-to-event outcomes
  - Binary outcomes (focus today)

Riley et al., BMJ, 2020
Example

Development and validation of the ISARIC 4C Deterioration model for adults hospitalised with COVID-19: a prospective cohort study


Summary

Background Prognostic models to predict the risk of clinical deterioration in acute COVID-19 cases are urgently required to inform clinical management decisions.

Methods We developed and validated a multivariable logistic regression model for in-hospital clinical deterioration (defined as any requirement of ventilatory support or critical care, or death) among consecutively hospitalised adults with highly suspected or confirmed COVID-19 who were prospectively recruited to the International Severe Acute Respiratory and Emerging Infections Consortium Coronavirus Clinical Characterisation Consortium (ISARIC4C) study across 260 hospitals in England, Scotland, and Wales. Candidate predictors that were specified a priori were considered for inclusion in the model on the basis of previous prognostic scores and emerging literature describing routinely measured biomarkers associated with COVID-19 progression. We used internal–external cross-validation to evaluate discrimination, calibration, and clinical utility across eight National Health Service (NHS) regions in the development cohort. We further validated the final model in held-out data from an additional NHS region (London).
Example

- **COVID-19 prognosis hospitalized patients**

  **Composite outcome:** “deterioration” (in-hospital death, ventilator support, ICU)

  **A priori expectations**
  - Event fraction at least 30%
  - 40 candidate **predictor parameters**
  - **C-statistic** of 0.71 (conservative est)
  - Cox-Snell $R^2$ of 0.24

  **Methods**
  We developed and validated a multivariable logistic regression model for in-hospital clinical deterioration (ventilation, critical care, or death) among consecutively hospitalised adults with highly suspected or confirmed COVID-19 who were prospectively recruited to the International Severe Acute Respiratory and Critical Care Registry (ISARIC4C) study across 260 hospitals in England, Scotland, and Wales. Candidate predictors that were specified a priori were selected based on previous prognostic scores and emerging literature describing COVID-19 prognosis. We used internal–external cross-validation to assess model calibration, and clinical utility across eight National Health Service (NHS) regions in the development cohort. We further validated the final model in held-out data from an additional NHS region (London).
Restricted cubic splines with 4 knots: 3 degrees of freedom

Note: EPV rule also calculates degrees of freedom of candidate predictors, not variables!
Calculate required sample size

**Criterion 1.** Shrinkage: expected heuristic shrinkage factor, \( S \geq 0.9 \) (calibration slope, target < 10% **overfitting**)

**Criterion 2.** Optimism: Cox-Snell \( R^2 \) apparent - Cox-Snell \( R^2 \) validation < 0.05 (**overfitting**)

**Criterion 3:** A small margin of error in overall risk estimate < 0.05 absolute error (precision estimated baseline **risk**)

**(Criterion 4: a small margin of absolute error in the estimated **risks**)**
Calculation

R code:
> require(pmsampsize)
> pmsampsize(type="b",rsquared=0.24,parameters=40,prevalence=0.3)

```
NB: Assuming 0.05 acceptable difference in apparent & adjusted R-squared
NB: Assuming 0.05 margin of error in estimation of intercept
NB: Events per Predictor Parameter (EPP) assumes prevalence = 0.3

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Samp_size</th>
<th>Shrinkage Parameter</th>
<th>Rsq</th>
<th>Max_Rsq</th>
<th>EPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criteria 1</td>
<td>1290</td>
<td>0.900</td>
<td>40</td>
<td>0.24</td>
<td>0.71</td>
</tr>
<tr>
<td>Criteria 2</td>
<td>962</td>
<td>0.871</td>
<td>40</td>
<td>0.24</td>
<td>0.71</td>
</tr>
<tr>
<td>Criteria 3</td>
<td>323</td>
<td>0.900</td>
<td>40</td>
<td>0.24</td>
<td>0.71</td>
</tr>
<tr>
<td>Final</td>
<td>1290</td>
<td>0.900</td>
<td>40</td>
<td>0.24</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Minimum sample size required for new model development based on user inputs = 1290, with 387 events (assuming an outcome prevalence = 0.3) and an EPP = 9.68
```
A few alternative scenarios

- $r^2 = 0.24$, parameters = 40, prevalence = 0.3 -> $\text{EPV} \geq 9.7$
- $r^2 = 0.12$, parameters = 40, prevalence = 0.3 -> $\text{EPV} \geq 21.0$
- $r^2 = 0.12$, parameters = 40, prevalence = 0.5 -> $\text{EPV} \geq 35.0$
- $r^2 = 0.36$, parameters = 40, prevalence = 0.2 -> $\text{EPV} \geq 5$
The sample size that meets all criteria is the MINIMUM required

• **Why minimum?** Other criteria may be important. e.g. missing data, clustering, variable selection

Future directions
• More evidence/guidance to choose values for the sample size criteria
• Sample size for validation (for continuous outcome just published)
• Sample size for different outcomes (e.g. multi-category)
• Sample size taking into account variable selection
• Simulation based approaches
• High-dimensional data?
Should a risk prediction model be developed?

Is prediction needed?  
- yes  
- no  
- don’t know  
- step

off for whom?  
- patients with characteristics X  
- no sort of  
- step

data on such patients?  
- yes  
- no  
- step

existing prediction model?  
- yes  
- no  
- step

data with 100+ events?  
- yes  
- no  
- step

check again  
- checked, nothing there  
- no more...  
- step

large dataset?  
- yes  
- no  
- step

predictions?  
- unique, found nowhere else  
- yes, commonly available in envisioned setting of application  
- step

Proceed with disentangling, penalties where possible, rigorous internal and/or external validations, study model calibration, think hard about dealing with missing data and imperfect outcome measurements, don’t forget to report everything including your intercept (last, follow TANDEM guidelines)

Fig. 24.6 Maarten van Smeden’s flowchart for the question: should a new prediction model be developed? [633]

Steyerberg, 2019
Work in collaboration with:

• Carl Moons
• Hans Reitsma
• Ben Van Calster (Leuven)
• Laure Wynants (Maastricht)
• Richard Riley (Keele, materials for this presentation)
• Gary Collins (Oxford, materials for this presentation)
• Ewout Steyerberg (Leiden)
• Many others

Contact: M.vanSmeden@umcutrecht.nl
Reference list

Bass et al., JCEM, 2010, doi: 10.1210/jc.2010-0947
Bell et al. BMJ, 2015, doi: 10.1136/bmj.h5639
Bellou et al. BMJ, 2019, doi: 10.1136/bmj.f3358
Gupta et al. ERJ, doi: 10.1183/13993003.03498-2020
Letellier et al. BJU, 2017. doi: 10.1038/bjc.2017.352
Mallett et al. 2010, doi: 10.1186/1741-7015-8-21
Peduzzi, JCE, 1996, doi: 10.1016/s0895-4356(96)00236-3
Steyerberg, 2019, doi: /10.1007/978-3-030-16399-0
TRIPOD statement, 2015; Collins et al., Ann Intern Med, doi: 10.7326/M14-0697
van Houwelingen, Statistica Neerlandica, 2001
van Geloven et al. EJE, 2020, doi: 0.1007/s10654-020-00636-1
van Smeden et al. Clinical prediction models: diagnosis versus prognosis, JCE, in press
van Smeden et al., SMMR, 2019, doi: 10.1177/0962280218784726
Riley et al., BMJ, 2020, doi: 10.1136/bmj.m4441
van Smeden et al. SMMR, 2019, doi: 10.1177/0962280218784726
Wynants et al. 2020, BMJ, doi: 10.1136/bmj.m1328