Overview of Multivariable Prediction Modelling

Methodological Conduct & Reporting: Introducing TRIPOD guidelines

Gary Collins

Centre for Statistics in Medicine
www.csm-oxford.org.uk

University of Oxford
16-July-2013
Why prediction (prognosis or diagnosis)?

- Provide information to doctors (and patients) to assist therapeutic decisions based on their risk of having a particular disease (diagnosis)
  - Additional testing

- Provide information to patients and their family on their risk of developing a particular disease (prognosis)
  - Lifestyle changes

- Target particular treatments to pre-defined patient strata based on their predicted risk

- Comparing institution performance (provides an adjustment for case-mix)
Prediction models

- **Tools to estimate an individual’s probability or risk**
  - Often referred to as prognostic models, prediction models, predictive scores, risk scores, clinical prediction rules, scoring systems, risk stratification tools...

- **Typically they are**
  - Mathematical formula
  - Simple scores
    - e.g. regression coefficients rounded to the nearest integer – sum the integers
  - Nomograms
    - Graphical representation of a prognostic model
  - Score charts
  - Increasingly available as web calculators or as apps for devices such as iPads/iPhones (Gray et al, Diab Med 2013)
**Table 2. Regression Coefficients and Hazard Ratios**

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\hat{\beta}$</th>
<th>P</th>
<th>Hazard Ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Women [So(10) = 0.95012]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log of age</td>
<td>2.32888</td>
<td>&lt;0.0001</td>
<td>10.27</td>
<td>(5.65–18.64)</td>
</tr>
<tr>
<td>Log of total cholesterol</td>
<td>1.20904</td>
<td>&lt;0.0001</td>
<td>3.35</td>
<td>(2.00–5.62)</td>
</tr>
<tr>
<td>Log of HDL cholesterol</td>
<td>-0.70833</td>
<td>&lt;0.0001</td>
<td>0.49</td>
<td>(0.35–0.69)</td>
</tr>
<tr>
<td>Log of SBP if not treated</td>
<td>2.76157</td>
<td>&lt;0.0001</td>
<td>15.82</td>
<td>(7.86–31.87)</td>
</tr>
<tr>
<td>Log of SBP if treated</td>
<td>2.82263</td>
<td>&lt;0.0001</td>
<td>16.82</td>
<td>(8.46–33.46)</td>
</tr>
<tr>
<td>Smoking</td>
<td>0.52873</td>
<td>&lt;0.0001</td>
<td>1.70</td>
<td>(1.40–2.06)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.69154</td>
<td>&lt;0.0001</td>
<td>2.00</td>
<td>(1.49–2.67)</td>
</tr>
<tr>
<td><strong>Men [So(10) = 0.88936]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log of age</td>
<td>3.06117</td>
<td>&lt;0.0001</td>
<td>21.35</td>
<td>(14.03–32.48)</td>
</tr>
<tr>
<td>Log of total cholesterol</td>
<td>1.12370</td>
<td>&lt;0.0001</td>
<td>3.08</td>
<td>(2.05–4.62)</td>
</tr>
<tr>
<td>Log of HDL cholesterol</td>
<td>-0.93263</td>
<td>&lt;0.0001</td>
<td>0.39</td>
<td>(0.30–0.52)</td>
</tr>
<tr>
<td>Log of SBP if not treated</td>
<td>1.93303</td>
<td>&lt;0.0001</td>
<td>6.91</td>
<td>(3.91–12.20)</td>
</tr>
<tr>
<td>Log of SBP if treated</td>
<td>1.99881</td>
<td>&lt;0.0001</td>
<td>7.38</td>
<td>(4.22–12.92)</td>
</tr>
<tr>
<td>Smoking</td>
<td>0.65451</td>
<td>&lt;0.0001</td>
<td>1.92</td>
<td>(1.65–2.24)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.57367</td>
<td>&lt;0.0001</td>
<td>1.78</td>
<td>(1.43–2.20)</td>
</tr>
</tbody>
</table>

So(10) indicates 10-year baseline survival; SBP, systolic blood pressure.

*Estimated regression coefficient

\[ \hat{p} = 1 - S_0(t)^\exp(\sum_{i=1}^{P} \beta_i \bar{X}_i - \sum_{i=1}^{P} \beta_i \bar{X}_i), \]
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 Slow (&lt;100)</td>
</tr>
<tr>
<td>Muscle tone Limp</td>
<td>2 &gt;100</td>
</tr>
<tr>
<td>Reflex irritability No response</td>
<td>1 Weak cry;</td>
</tr>
<tr>
<td>(response of skin</td>
<td>2 Good; strong</td>
</tr>
<tr>
<td>stimulation to feet)</td>
<td>hypoventilation</td>
</tr>
<tr>
<td>Color Blue; pale</td>
<td>1 Some motion</td>
</tr>
<tr>
<td></td>
<td>2 Body pink;</td>
</tr>
<tr>
<td></td>
<td>extremities</td>
</tr>
<tr>
<td></td>
<td>blue</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.
Prognostic model for patients with traumatic bleeding

This prognostic model may be used as an aid to estimate mortality at 28 days in patients with traumatic bleeding.

Country: United Kingdom
Age, years: 18
Hours since injury: 6
Glasgow coma score: 8
Systolic blood pressure: 54
Respiratory rate: 8
Heart rate: 56
Type of injury: Blunt

Risk of 28 day mortality: 20%

(6 red sad faces) death, (89 green smiley faces) alive
20 out of 100 patients with these characteristics will die according to this prognostic model

Online calculator by: Sealed Envelope

gary.collins@cs.m.ox.ac.uk
Framingham Risk Score (iPhone)

Free
Category: Medical
Updated: Jul 15, 2011
Version: 1.5
Size: 2.6 MB
Language: English
Seller: Austin Physician Productivity, LLC
© STATCODER.COM
Rated 9+ for the following:
Infrequent/Mild
Mature/Suggestive Themes

Requirements: Compatible with iPhone, iPod touch, and iPad.
Requires iOS 3.0 or later

Customer Ratings
Current Version:
★★★★ 5 Ratings
All Versions:
★★★★ 103 Ratings

More iPhone Apps by Austin Physician Productivity, LLC

iPhone Screenshots

Framingham 10-year Global CVD Risk

10-year General CVD Risk
coronary heart disease, stroke, peripheral artery disease, or heart failure
Heart Age / Vascular Age

18.4%

40 - 44 | Male
200-239 | 35-39
140-149 | Untreated
Clear Reference
ON Smoker OFF DM

10-year General CVD Risk
coronary heart disease, stroke, peripheral artery disease, or heart failure
Heart Age / Vascular Age

10.0%

40 - 44 | Female
200-239 | 40-44
140-149 | Untreated
Clear Reference
ON Smoker OFF DM

An individual's heart age is calculated as the age of a person with the same predicted risk but with all other risk factor levels in normal ranges. Although called heart age for simplicity of risk communication in primary care, the heart age really reflects vascular age.

An individual's heart age is calculated as the age of a person with the same predicted risk but with all other risk factor levels in normal ranges. Although called heart age for simplicity of risk communication in primary care, the heart age really reflects vascular age.

gary.collins@csm.ox.ac.uk
Increasing interest in prediction models

(("prognostic model") OR ("prediction model") OR ("risk score") OR ("clinical prediction rule") OR ("decision rule") OR ("prognostic index") OR ("prognostic indices") OR ("prediction index") OR ("risk algorithm") OR ("risk stratification") OR ("multivariable prediction"))
Examples of prediction models in UK clinical & public health guidelines

- **Framingham Risk Score & QRISK2 (NICE CG67)**
  - 10-year CVD risk

- **Nottingham Prognostic Index (NICE CG80)**
  - Recurrence & survival in breast cancer patients

- **FRAX & QFracture (NICE CG146)**
  - 10-year osteoporotic and hip fracture risk

- **GRACE/PURSUIT/PREDICT/TIMI (NICE CG94)**
  - Adverse CV outcomes in patients with UA/NSTEMI

- **APGAR (NICE CG132/2)**
  - Newborn prognosis

- **SAPS & APACHE (NICE CG50)**
  - ICU scoring systems

- **Leicester Diabetes Risk Score, QDSCORE, Cambridge Risk score (NICE PH38)**
  - Type 2 diabetes
Phases in prediction model research

- **Development**
  - Predictor selection, model building
  - Internal validation (evaluating optimism)
    - Split sample (random) – inefficient - not very useful
    - Cross-validation
    - Bootstrapping

- **Evaluate performance (external validation)**
  - Temporal & geographical validation
  - Independent validation (i.e. independent investigators)

- **Impact study**
  - Does the prognostic model improve patient outcomes?
  - Does the prognostic model change clinician behaviour?
  - Is the prognostic cost effective?
Model development

• **Select important candidate predictors**
  – Trying to avoid selection based on statistically significant univariable associations with outcome

• ** Appropriately handle (acknowledge) missing data**

• **Fit a multivariable model**

• **Estimate the predictive performance**
  – Calibration and discrimination
  – Quantify any optimism from overfitting
    • Use bootstrapping (avoid randomly splitting a dataset)
External validation

• **Separate dataset (not a random split)**
  – Different centres (geographical validation)
  – Different time period (temporal validation)
  – Different case-mix
  – Possibly with different definitions of predictors and outcome

• **Ideally conducted by independent researchers**

• **Validation is not...**
  – repeating all the steps in the development study
  – refitting the model and evaluating the re-fitted model on a new dataset
  – refitting the model and ‘comparing’ the regression coefficients on a new dataset
identified and fewer than 3 cases to go untested and undetected.

When our final prediction model was refitted to ARIC/CHS, we obtained consistent results: All of the risk factors were significant ($P \leq 0.001$), and the magnitude of the associations was similar, with an AUC of 0.74 (Appendix Table 3, available at www.annals.org).

The Appendix Figure (available at www.annals.org) shows the prevalence of undiagnosed diabetes for individ...
What does it validation entail?

• Access to the underlying ‘static’ mathematical model (simplified model), source code etc...

  – Models often not published or only partially
    • Model intercepts, baseline survival often missing

  – An unpublished model is limited (useless) – lack of transparency and reproducibility (e.g. FRAX)

  – Evaluate performance
    • Discrimination, calibration, $R^2$, Brier score...
    • Model comparison
    • Clinical consequences (net benefit)
Evaluating the model performance

• Prediction models are characterised by

  – **Calibration**
    • agreement between observed outcomes and predictions
      – Often ignored
      – Preferably assessed graphically

  – **Discrimination**
    • ability to distinguish between patients who do and do not experience the event of interest
      – Usually reported (c-index)
Impact studies

• Does the prediction model offer an improvement over current standard practice?

• Quantify effect of using the prognostic model on
  – Decision making, patient outcomes, cost effectiveness

• Typically a comparative study
  – Cluster RCT, interrupted time series, step wedge, before & after study etc...
  – Intervention could be assistive or directive (Moons, Heart 2012)

• However, there a very few impact studies
  – Good example is the STarT Back trial (Lancet 2011)
  – Treatment based on patient’s prognosis
    • Improved patient outcomes and cost-effective
Increasing interest in prediction models

(("prognostic model") OR ("prediction model") OR ("risk score") OR ("clinical prediction rule") OR ("decision rule") OR ("prognostic index") OR ("prognostic indices") OR ("prediction index") OR ("risk algorithm") OR ("risk stratification") OR ("multivariable prediction"))
Published prediction models

- 111 models for prostate cancer (Shariat 2008)
- 102 models for TBI (Perel 2006)
- 83 models for stroke (Counsell 2001)
- 54 models for breast cancer (Altman 2009)
- 43 models for type 2 diabetes (Collins 2011)
  - 20+ more models have since been published!
- 31 models for osteoporotic fracture (Steurer 2011)
  - Omitted FRAX due to insufficient information
- 29 models in reproductive medicine (Leushuis 2009)
- 26 models for hospital readmission (Kansagara 2011)
- 13 models for tooth decay (Ritter 2010)

Very few of these models have been ‘validated’ in new data and compared. Models are easy to create, no intention of being used, extra line on a CV => rubbish!
Vickers & Cronin. *Everything you wanted to know about evaluating prediction models (but were too afraid to ask).* J Urol, 2010.
vs radiotherapy). The model has high discrimination (AUC of 0.78) and good calibration (see Fig. 2). In other words, the model is terrific in all ways other than that it is completely useless. So why did we create it? In short, because we could: we have a dataset, and a statistical package, and add the former to the latter, hit a few buttons and voila, we have another paper. It is tempting to speculate that the ubiquity of nomograms in the uro-
package, and add the former to the latter, hit a few buttons and voila, we have another paper. It is tempting to speculate that the ubiquity of nomograms in the urological literature is simply because it is particularly easy research to do: you do not need to collect any data or even think of an interesting scientific question. We would argue that a predictive model should only be published if it is has a compelling clinical use, and that is rarely the case.
Poor reporting...and conduct

- Collins et al (BMC Medicine, 2011)  
  - diabetes
- Mallet et al (BMC Medicine, 2010)  
  - cancer
- Collins et al (J Clin Epidemiol, 2012)  
  - kidney disease
- Bouwmeester et al (PLoS Medicine, 2012)  
  - general medical journals
- Altman (Cancer Invest, 2009)  
  - breast cancer
- Burton (BJC, 2004)  
  - missing data in prognosis studies
- Many more....
Conclusions from the systematic reviews

• **Poor reporting**
  - Number of events often difficult to identify
    • Candidate predictors (and number) inadequately defined
  - Insufficient information to report EPV
    • 40% studies (Mallett 2010; 33% Collins 2011)
  - How candidate predictors were selected
    • Unclear in 25% studies (Bouwmeester 2012)
  - How the multivariable model was derived
    • Unclear in 77% studies (Mallett 2010)
Conclusions from the systematic reviews

• **Poor reporting**
  - Missing data rarely mentioned (41% Collins, 2010; 45% Collins 2012)
    - Often an exclusion criteria (often not specified)
    - Complete-case usually carried out
  - Model often not reported in full
    - intercept missing for logistic regression, baseline survival missing for Cox regression models
  - Ranges of continuous predictors rarely reported
Conclusions from the systematic reviews

- **Methodological shortcomings including**
  - Small sample size (number of events) [EPV<10]
  - Large number of candidate predictors
  - Calibration rarely assessed
    - 85% (Altman); 74% not done (Collins); 46% not done (Bouwmeester)
  - Dichotomization of all/some continuous predictors
    - 63% of studies (Collins); 70% of studies (Mallett)
  - Previously published models often ignored
  - Inadequate validation
    - Reliance on random-split (often using an already small dataset) to validate

- **Lack of comparing competing models on the same dataset**
Sample size

Evaluation of mortality following severe burns injury in Hungary: External validation of a prediction model developed on Belgian burn data

N. Brusselaers\textsuperscript{a,b}, I. Juhász\textsuperscript{c}, I. Erdei\textsuperscript{c,d}, S. Monstrey\textsuperscript{a,b}, S. Blot\textsuperscript{b,e,f,\ast}
3.3. **External validation of the Belgian prediction model**

The whole population and several subgroups were also analysed with the Belgian prediction model. In total 43 deaths were predicted, whereas 32 deaths were observed (Table 2). Only six patients had a predicted mortality of 50% or higher and all those patients died (five deaths predicted). No patients were admitted with a total score above 8 points. The prediction was accurate in all categories, only in the ‘low risk’ categories with 1 or 2 points there was an overestimation (Table 2). The observed/expected ratio of the total population was 0.74 (32/43) (Table 2).

The ROC-curve analysis showed an area under the curve of 0.94 (95% CI: 0.89–0.98) (Fig. 2) [24]. The accuracy of the model was also analysed for both genders separately, which showed a slightly higher predictive value for males: AUC females 0.90 (95% CI: 0.78–1.00), AUC males 0.96 (95% CI: 0.93–0.99).

For the paediatric population this model was very accurate, but further analyses are obliged with a higher
number of patients (AUC of 1.00 and 95% CI of 1.00–1.00). The maximum score for children is 7 points (with corresponding predicted mortality of 75%), and in this population only 1 child had a score >3, namely the child who deceased with a score of seven points. The goodness of fit (for the whole population) with the Hosmer-Lemeshow was acceptable (5.8; p = 0.055) [25].
What they concluded

next burn centre is organised. From each of the 6 burn units the next centre is always within reach of less than an hour, in contrast to the rural Debrecen region, where other burn that patients in the lowest risk categories are less severely burned in this Hungarian population, leading to this over-estimation. Nevertheless, this prediction model has a high

discriminative power of prediction for the total adult and paediatric populations separately, and for both genders, as illustrated by ROC curve analysis (Fig. 2).

A final interesting remark is the higher male mortality.

What do authors do…?

- **Numerous systematic reviews have shown studies developing risk prediction models are poorly conducted and poorly reported**
  - Little or no intention of the model being used

- **Unlikely that external validation are well conducted and sufficiently reported**
  - Anecdotally we know they are fraught with numerous flaws and poorly interpreted
  - Makes deciding whether to consider using a prediction model tricky
  - Overwhelming focus only on discrimination

- **Ultimately a lot of rubbish is published**
  - Clear need for educational articles (for clinical readers/authors/reviewers), not tucked away in specialist / methods journals – including ‘how to read such papers’
  - (also) clear need for educational articles for statisticians not familiar with risk prediction
• Consensus-based guidelines for improving the quality of reporting of multivariable prediction modelling studies

• Focus on reporting (but aspects of methodological conduct also touched upon)

• 3-day meeting held in Oxford
  – 24 participants (statisticians, epidemiologists, clinicians, journal editors)

• Funded by Cancer Research UK, ZonMW and the Medical Research Council
TRI POD Members

Gary Collins (Oxford)
Karel Moons (UMC Utrecht)
Doug Altman (Oxford)
Hans Reitsma (UMC Utrecht)

Ginny Barbour (PLoS Med)
Nancy Cook (Harvard)
Joris de Groot (UMC Utrecht)
Trish Groves (BMJ)
Frank Harrell (Vanderbilt)
Harry Hemingway (UCL)
Michael Kattan (Cleveland Clinic)
Andre Knottnerus (JCE)
Susan Mallett (Oxford)
Cindy Mulrow (Ann Intern Med)
Richard Riley (Birmingham)
Peter Rothwell (Oxford)
Patrick Royston (MRC)
Willi Sauerbrei (Freiburg)
Ian Stiell (OHRI, Canada)
Overview of TRI POD (1/5)

• **Title & Abstract**
  – Identify study as a model developing or validating a prediction model
  – Clear summary of objectives, study design etc...

• **Background**
  – Rationale for developing/validating model
  • Reference to existing models (important)
Overview of TRIPOD (2/5)

- **Methods**
  - Source of data (e.g. RCT, registry)
  - Study dates
  - Study setting (e.g. primary care, #centres, location)
  - Eligibility criteria
  - Details of any treatment given
  - Clear definition of outcome (how, when, blinding)
    - Different in the validation study?
  - Clear definition of all predictors (how, when, blinding)
    - Different in the validation study?
  - Rationale of sample size
• **Methods (continued)**
  - Handling of missing data
  - How were continuous predictors handled?
  - Type of model, model building strategy (including predictor selection)
  - How were the predictions calculated (validation)?
  - Performance measures (c-index, calibration)
  - Model-updating (e.g. recalibration)
  - Creation of risk groups
• **Results**
  
  – Summary of participants (demographics, clinical features, missing data), number with and without outcome + summary of follow-up
  
  – For validation studies, a comparison of case-mix with the development dataset
  
  – Describe the final model so that predictions on new patients can be made
    
    • Including any simplified models (with predicted risks)
  
  – Performance measures (e.g. calibration, discrimination, net benefit, model comparison)
• **Discussion**
  – Study limitations (overfitting, few event per predictor, omission of known predictors)
  – Discuss validation results with reference to development data or other validation data
  – Potential clinical use of the model, implications for future research

• **Other Information**
  – Availability of data, prediction model (web calculator), protocol etc...
  – Funding (and role of)
• Key minimal information deemed important to report

• Help authors, peer reviewers, editors, readers and potential users

• Educational – providing guidance, cautioning against particular approaches

• To be submitted for publication in January 2014

• Improve evaluating risk of bias (PROBAST) if more information is reported
Internal validation strategies

![Boxplot showing the relationship between c-statistic and events per variable. The plot illustrates the performance of different validation strategies, including apparent, split-sample validation, cross-validation, and optimism-corrected methods.]