STRATOS Topic Group 6

Evaluating diagnostic tests and prediction models: A focus on calibration

Gary Collins
Professor of Medical Statistics

Centre for Statistics in Medicine
University of Oxford

ISCB
25th August 2016

TG6 Members

Chairs

- Gary Collins
- Carl Moons
- Ewout Steyerberg

Members

- Patrick Bossuyt
- Ben Van Calster
- Petra Macaskill
- Andrew Vickers

Aims of TG6

We aim to provide guidance on how to evaluate the performance of single tests and (multivariable) prediction models, with extensions to issues in the development, validation and impact assessment of prediction models

Outline

- Why we need STRATOS TG6 - why guidance is needed
  - summary of existing systematic reviews
- A focus on calibration
  - brief summary of existing approaches and guidance
  - some examples from the medical literature
- Next steps

Setting the scene: waste in research

- 800 CVD models (Wessler 2015)
- 360 models for predicting CVD (Damen BMJ 2016)
- 263 models in obstetrics (Kleinrouweler 2016)
- 111 prostate cancer models (Shariat 2008)
- 43 type 2 diabetes models (Collins 2011)
- ...many more

- the predictive performance of most models is not evaluated
- ...and when it is, it's often done badly (Collins 2014)

Waste

- incomplete (key details often not reported)
- most papers are never or rarely cited
- most models are never or rarely used (fortunately)
- incorrect methods copied to subsequent studies
  - cycle is never broken
- financial cost
  - peer review
  - publication
  - reading
- ...

- Diabetes
  - Collins (BMC Med 2011); Van Dieren (Heart 2011)
- Cancer
  - Mallet (BMC Med 2010); Altman (Cancer Invest 2009)
- Kidney disease
  - Collins (J Clin Epidemiol 2012)
- Leading medical journals
  - Bouwmeester (PLoS Med 2012)
- Missing data in prognosis studies
  - Burton (Br J Cancer 2004); Masconi (EMPA J 2015)
- External validation studies
  - Collins (BMC Med Res Methodol 2014)
- **many more...**

why is guidance needed? (summary from many reviews)

- Studies tend to be small → overfitting (link to TG5; Design)
- Continuous predictors frequently categorised → leads to poorly performing model (link to TG2)
- Missing data (link to TG1)
- Internal validation rarely done appropriately
  - (random) split-sample regularly done → inappropriate
  - bootstrapping rarely done - often done incorrectly
  - despite existing guidance - evaluating model performance (optimism) in model development studies is poor
- Many models are often not even reported
- → **Developing prediction models often done poorly**

why is guidance needed? (Collins et al. 2014)

- Studies tend to be small (link to TG5; Design)
- Missing data rarely mentioned (link to TG1)
- Discrimination (c-statistic) usually evaluated (not always)
  - 'blank' ROC curves often presented
- Calibration infrequently assessed (often incorrectly/inefficiently)
- Clinical utility (e.g., decision curve analysis) rarely done
  - and often mainly in urology (recent BMJ guidance paper)
- Comparing against other models rarely done
- **Evaluating model performance often done poorly**

TRIPOD Reporting Statement

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement

Gary S. Collins, PhD; Johannes B. Reitsma, MD, PhD; Douglas G. Altman, DSc; and Karel G.M. Moons, PhD

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

TRIPOD Reporting Statement

"Two key aspects characterize the performance of a prediction model: calibration and discrimination. They should be reported in all prediction model papers"

What is calibration?

Calibration reflects the agreement between predictions from the model and observed outcomes

Available approaches

Statistical tests

- Hosmer-Lemeshow test (logistic regression)
- Nam & D'Agostino test (survival data)
- Grønnesby & Borgan test (survival data)

Problems

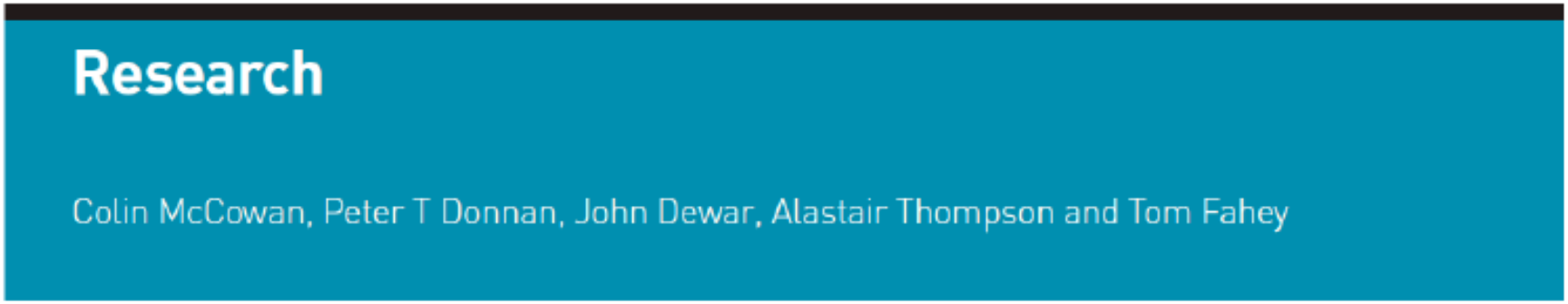
- Conclusions based on a single p-value (>0.05)
- No magnitude of direction of (mis)calibration
- Influenced by sample size and grouping (Kramer 2007)

Available approaches

Graphical approaches

- Plot mean predicted risk vs. observed proportion of events
  - typically (though not always) by tenth of predicted risk
  - influenced by sample size and number of groups
- Flexible nonlinear calibration curve (Austin 2014; Van Calster 2016)
  - $\text{logit}(Y) = a + f(L)$
- Calibration belts (Finazzi 2011; Nattino 2014)
- Extensions for survival data
  - val.surv function in Harrell's rms package in R
  - Royston's psuedo-observations approach (Royston 2014)
- Uncertainty captured by confidence intervals

Example 1: McCowan et al. Br J Gen Pract 2011

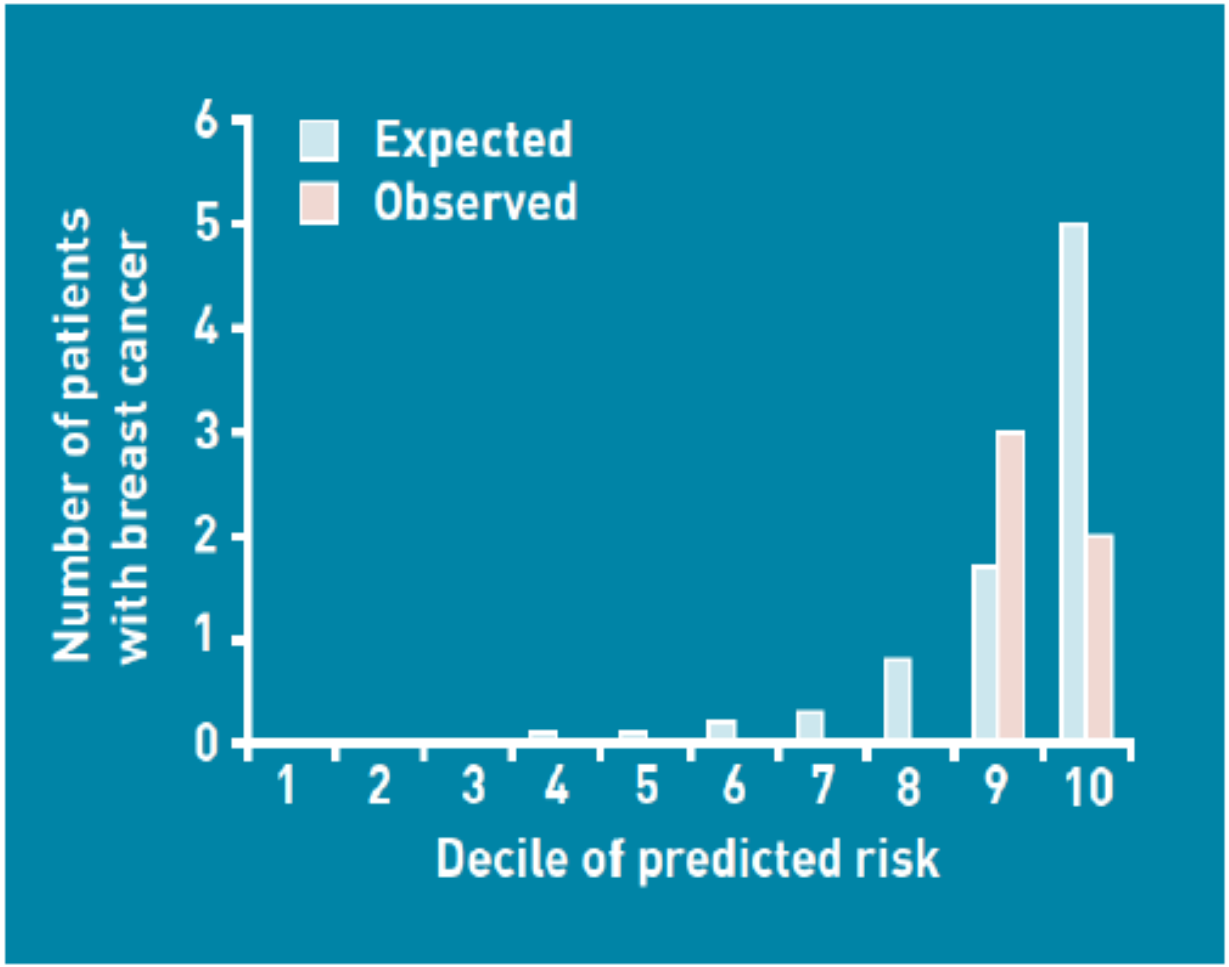


Example 1: McCowan et al. Br J Gen Pract 2011

Validation cohort

There were 202 patients identified by 11 general practices as presenting with symptoms suggestive of breast cancer. However, telephone contact details for 59 patients could not be traced, six patients had the wrong phone contact, and 19 could not be contacted. Of the 118 patients contacted, 16 declined participation and five failed to return written consent; this gave a total of 97 patients providing data for the validation study. Of these, 73 (75%) were referred to the symptomatic breast clinic; five (5%) were subsequently diagnosed as having breast cancer.

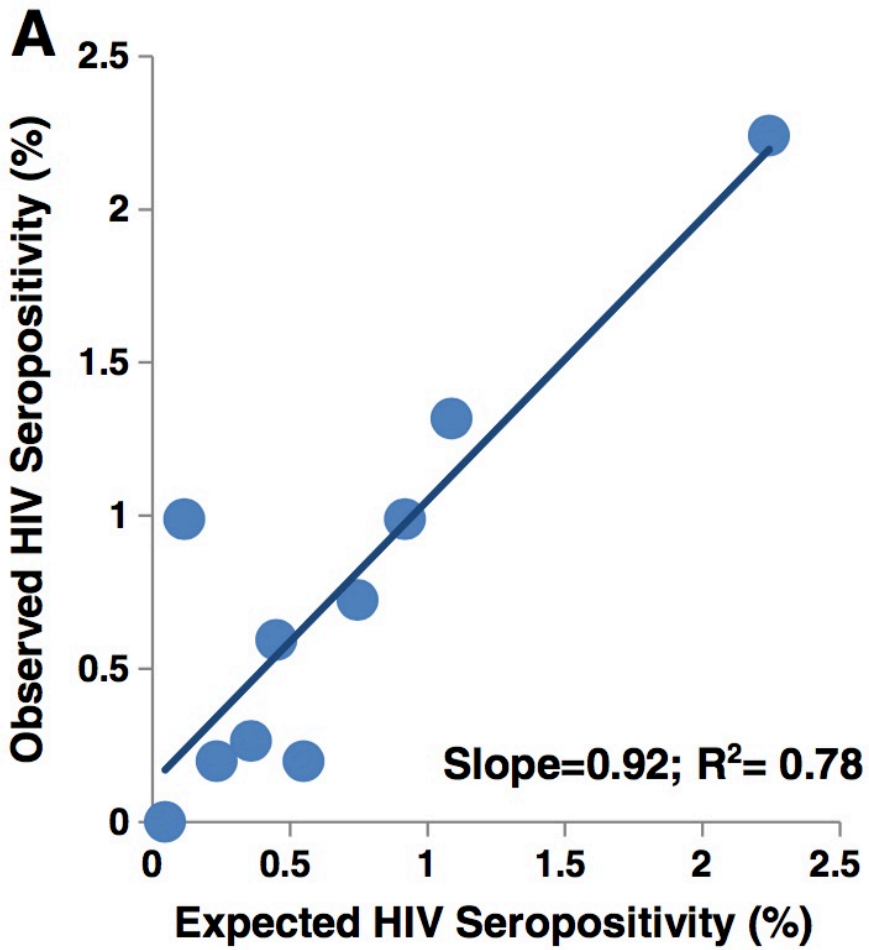
Example 1: McCowan et al. Br J Gen Pract 2011



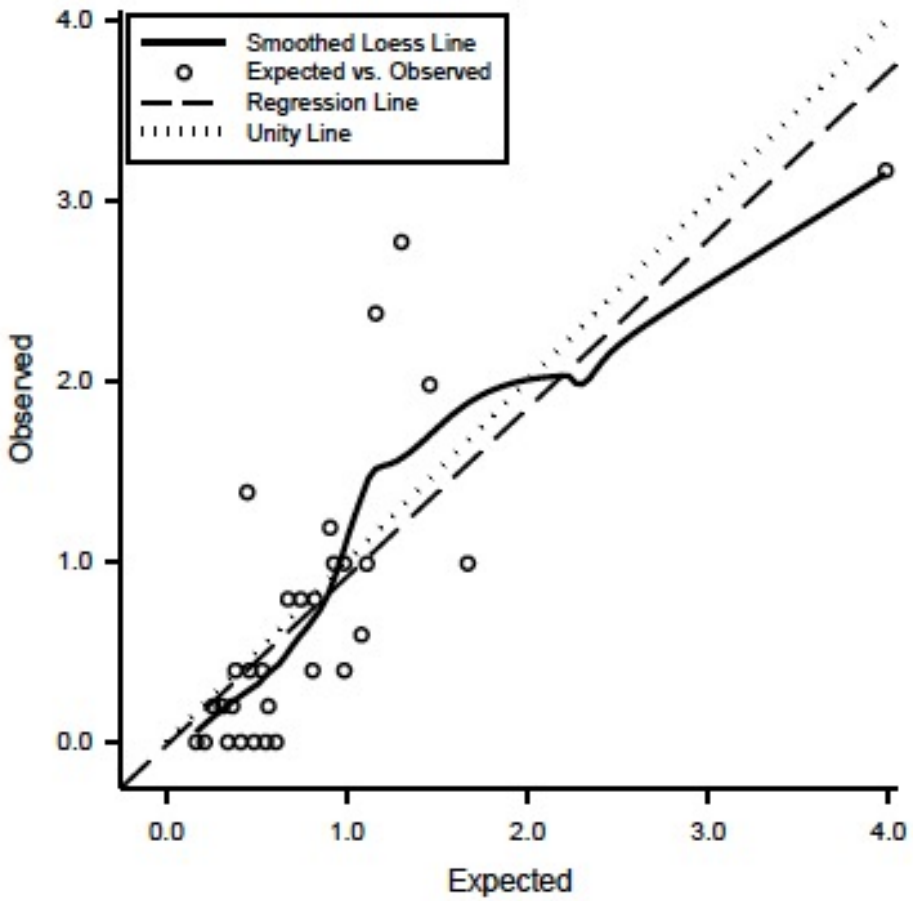
Example 1: McCowan et al. Br J Gen Pract 2011

the derivation model and the expected and observed cancers recorded (Figure 1). All observed breast cancers occurred in the top two deciles (top quintile) of expected risk. A Hosmer-Lemeshow goodness-of-fit test for the calibration of the model (HLGOFCS) shows no significant difference between expected and observed breast cancers (HLGOFCS = 7.02, $P = 0.73$), but the plot suggests the number of cancers was overestimated for those at highest risk (top decile).

Example 2: Hsieh et al. Am J Emerg Med 2014



Example 2: Hsieh et al. Reply (still wrong)



Example 3: Garcia-Valentin et al. Eur J Cardiothorac Surg 2016 (copying method from a previous study)

with a substantial difference between observed and expected values. Regarding EuroSCORE II calibration, we can find statistically significant differences between observed and expected mortalities according to the Hosmer–Lemeshow test, although values for observed and expected mortality are very similar. This test is currently under debate for some problems, although it was chosen for this study as it was the one used in the internal validation of the original paper [14]. This could be interpreted as a calibration failure, but absolute difference between observed and expected values were only 0.8%. On the other hand, if we consider the SDs, we can

Example 3: Garcia-Valentin et al. Reply

the analysis [5].

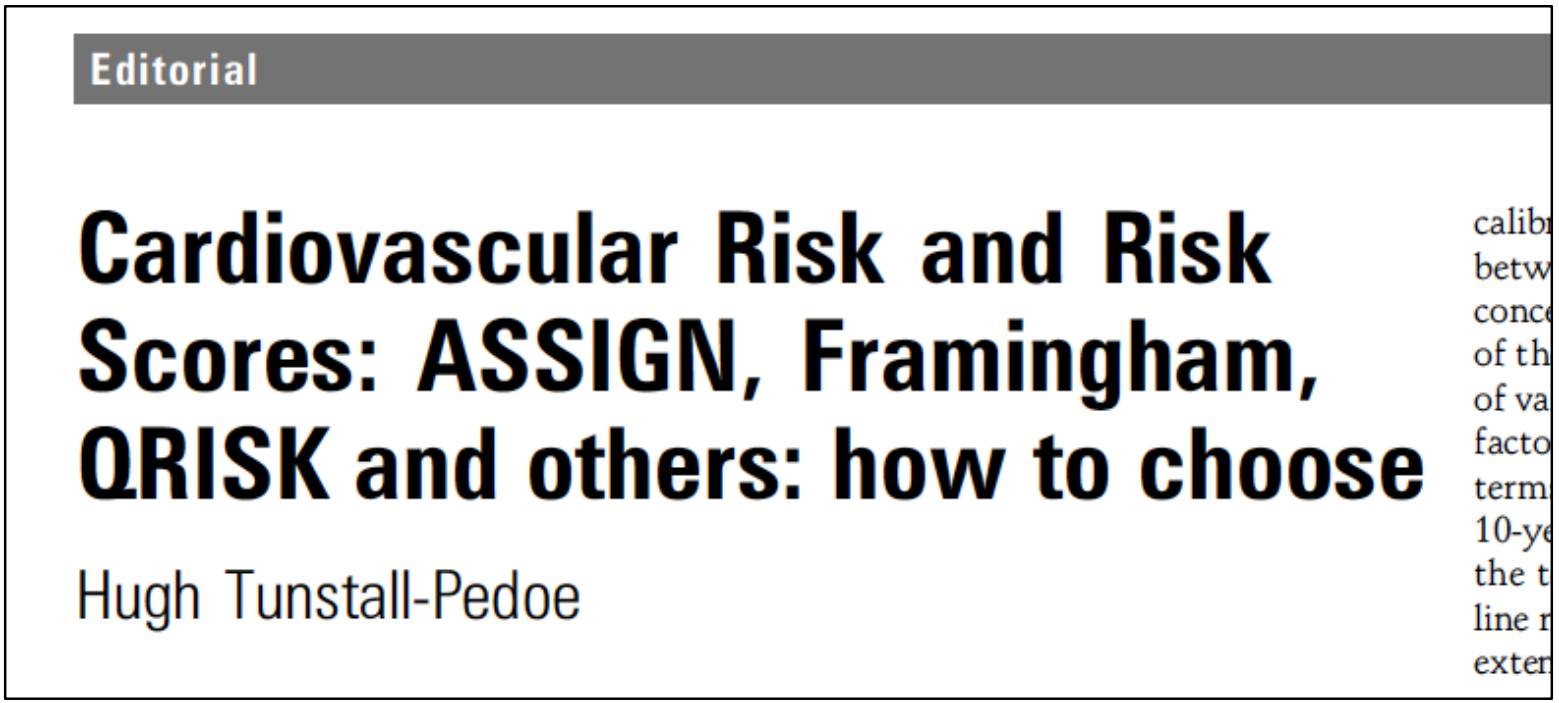
Novel calibration methods were considered during the design of our study although we found some advantages in the Hosmer–Lemeshow test. Readers are used to it by its wide utilization and this makes possible to easily compare results with previous studies in the same terms. Perhaps we failed to adequately describe this in our article and we apologize for this. Although we acknowledge these limitations, we do not share this negative opinion about the Hosmer–Lemeshow test. Consequently, we cannot accept the suggestion of intentional flaw that they

Example 3: Garcia-Valentin et al. Reply

about the actual amount of time readers dedicate to the "Discussion" section, but it sounds appropriate that they should produce actual data to support their opinion, considering their strong scientific and methodological background. Our article underwent an exhaustive peer review process that involved 2 editors, 3 reviewers and 2 independent statisticians. Possible misinterpretations were reassessed, and no additional problems in our methodology detected.

We thank Collins and Le Manach for reminding the community their recommendations, which we will consider. It is appropriate that explanations about unclear methods or data should be demanded. We understand the deep disappointment of Collins and Le Manach for what they consider a suboptimal methodology in our contribution. Scientific thinking should also stay away from radical ideas and disqualification, and should be respectful towards other thoughts that differ from one's own.

Example 4: (re)educating



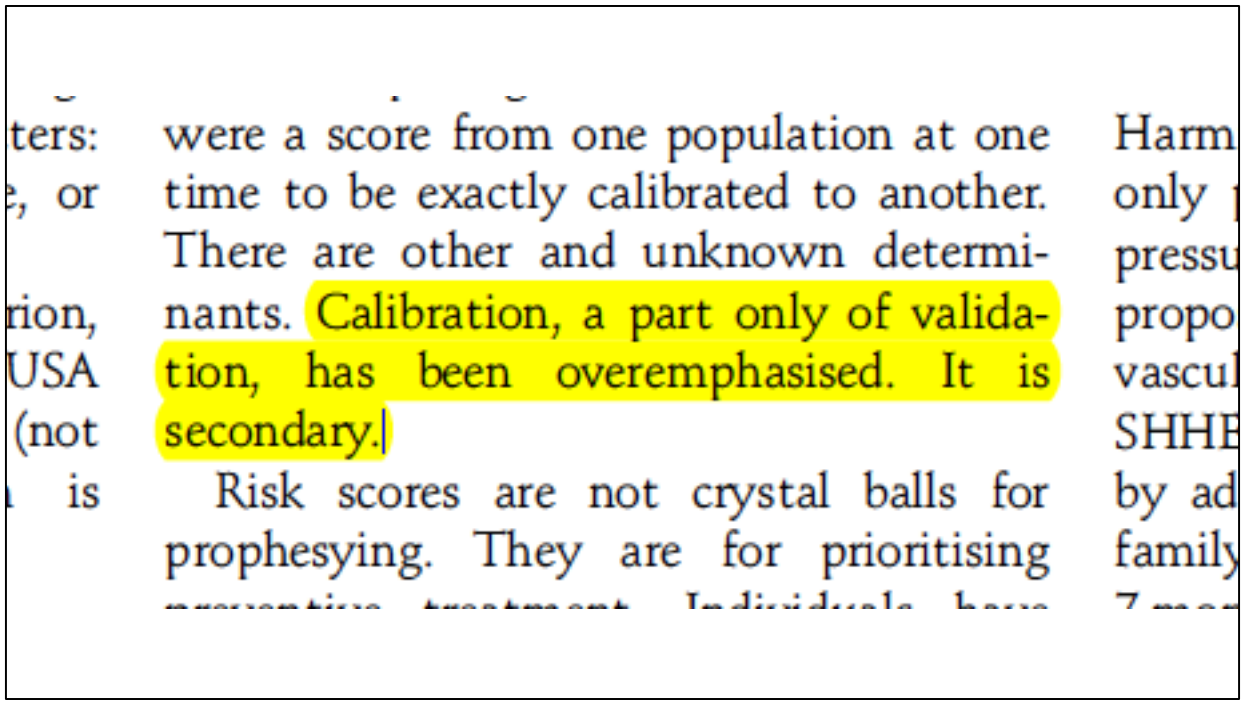
Editorial

Cardiovascular Risk and Risk Scores: ASSIGN, Framingham, QRISK and others: how to choose

Hugh Tunstall-Pedoe

calib
betw
conc
of th
of va
facto
term
10-ye
the t
line r
exter

Example 4: (re)educating



ters:
, or
rion,
USA
(not
is

were a score from one population at one time to be exactly calibrated to another. There are other and unknown determinants. Calibration, a part only of validation, has been overemphasised. It is secondary.

Risk scores are not crystal balls for prophesying. They are for prioritising

Harm
only
pressu
propo
vascul
SHHE
by ad
family

Example 4: (re)educating

:orial score and are ranked according to nan
:his estimated risk to assess whether they but
ustify treatment. More important than sho

calibration is the score's discrimination between future cases and non-cases, by concentrating future cases at the top end of the distribution, the crucial component of validation. Table 1 shows how different factors and two scores discriminate in terms of the percentage of subsequent

- Numerous different ways to assess calibration
  - predominantly in Stat Med; J Clin Epidemiol for statisticians
- Plenty of overview papers discussing general aspects of prediction modelling, including calibration (often vague)
  - some useful like the TRIPOD E&E paper, 2099 BMJ prognosis series, 2013 PROGRESS series, Steyerberg book + many others
  - ...but some less useful/misleading (often in clinical journals)
- Little (no) useful guidance for calibration for non-statisticians
  - advantages and disadvantages of different approaches
  - prediction model studies often largely conducted without a statistician or a methodologist

Planned Activities

1. Identify / summarise existing guidance (including advantages and disadvantages) of different aspects of performance evaluation and measures in prediction modelling
   - for example, focussing on assessing calibration
   - break it down in the the various STRATOS levels of knowledge
2. Initiate new systematic reviews of published studies in the medical literature
   - Identify current practice (what is being done in clinical studies)

www.stratos-initiative.org/contact2

Contact

Currently, the STRATOS initiative is not funded and all members are participating voluntary in the project. Therefore some delays may happen depending on your specific request. Please note that STRATOS is a project to develop guidance documents. Please understand that we do not help with issues concerning the design and analysis of specific studies.

1. **I want to become a member**
2. **I want to contact a topic group**

Please write an E-Mail to contact@stratos-initiative.org stating the Topic Group(s) in the subject heading. We will forward your E-Mail to the chairs of the respective Topic Group(s). Further information about the topic groups can be found here.

#	Topic group
1.	Missing Data
2.	Selection of Variables and Functional Forms in multivariable analysis
3.	Descriptive and initial data analysis
4.	Measurement Error and Misclassification
5.	Study Design
6.	Evaluating diagnostic tests and prediction models
7.	Causal inference
8.	Survival analyses
9.	High-dimensional data