International initiative

Guidance for key issues of design and analysis of observational studies

TG 6: Evaluating diagnostic tests and prediction models

Petra Macaskill (Sydney, Australia)
Ewout Steyerberg (Rotterdam, the Netherlands)
On behalf of TG 6

Members

- Chairpersons:
  - Ewout Steyerberg (Rotterdam, the Netherlands)
  - Petra Macaskill (Sydney, Australia)
  - Andrew Vickers (New York, USA)

- Confirmed additional members so far:
  - Gary Collins (Oxford, UK)

Scope

Both diagnosis and prognosis will be covered.

- Diagnosis: the focus is on current status
- Prognosis: the aim is to predict risk of a future event

The scope and priorities for the topics to be addressed is open for discussion.

Measures for evaluating the performance of a diagnostic test

Measures for assessing the performance of a diagnostic test are well established.

- Sensitivity and specificity (at a given cut-point for the test).
- Receiver Operating Characteristic (ROC) curve that describes the trade-off in sensitivity and specificity as the cut-point varies.
- Area under the ROC curve (AUC)
- Positive and negative predictive values (depend on pre-test probability of disease)
- Likelihood ratios

All measures have pros and cons!

Measures for evaluating the performance of a diagnostic test

- The performance characteristics of a test are likely to vary according to the context in which the test is to be used (e.g. clinical pathway needs to be considered).
- Test performance should fit with the intended (or potential) use of the test (e.g. triage test)
- An "appropriate" threshold should be used
  - "optimal" threshold
  - Test errors (false positive and false negative results) are unlikely to have equivalent consequences.
  - The balance between benefits and harms is crucial.
- These above issues are also relevant to the development and evaluation of prediction models for diagnosis (decision support models).

Main issues for the start

Guidelines for the evaluation of a prediction model (for diagnosis or prognosis) is seen as a priority.

- Prediction models are being used increasingly to assist with clinical decision making.
  - Statistical guidelines will complement reporting guidelines that are currently being finalised.
  - A framework for evaluation has been published (Steyerberg et al 2010)

- Guidance on approaches to the design, analysis and interpretation are needed to evaluate:
  - Performance of a prediction model.
  - The incremental gain of adding a new test/marker.

Important restrictions

Initially, we will assume:

- a binary outcome
- no error in the measurement of the outcome
- no missing data for the outcome or covariates.

Input from other groups (e.g. TG1, TG2) may be needed to deal with the above issues.

Evaluating model performance
Which methods/measures to use?

Methods and measures are evolving, particularly for prognostic models that take account of time to an event and censoring.

Models are generally assessed in terms of:

- Overall predictive performance
- Discrimination (particularly relevant in diagnosis)
- Calibration

Guidance on the use of existing methods and measures to assess these is needed.

Evaluating model performance
Which methods/measures to use?

Overall predictive performance:

- Various R^2 measures are used for binary outcomes, some specifically designed for survival models. (e.g. Nagelkerke R^2, Pearson R^2, scaled Brier Score).

Discrimination (ability to classify correctly into two outcome categories)

- c-statistic is used routinely for diagnostic (logistic regression) models. Several variants of the c-statistic are available for survival models.

Calibration (agreement between predicted probabilities & observed outcomes)

- difference between overall observed event rate and average predicted probability (calibration-in-the large)
- Hosmer and Lemeshow test (grouped in deciles)
- Testing goodness of fit based on grouping by risk categories
- Calibration plots.

Evaluating model performance
Taking account of harms and benefits

Utility based measures have also been proposed.

Net Benefit provides an assessment of clinical usefulness by taking account of potential harm associated with false positive results when a (treatment) threshold is applied.

A Decision Curve Analysis can be undertaken by displaying NB across the range of thresholds.

Assessing incremental gain of adding a test/marker to a prediction model.

Approaches include:

- Change in c-statistic (criticised as insensitive)

- Reclassification of individuals across risk thresholds.
  - Net Reclassification Improvement (NRI)
    - all "movements" are treated as equivalent in terms of weight
    - concerns about the properties of this measure are growing
  - Integrated Discrimination Improvement (IDI is a continuous extension of NRI)

- Change in Net Benefit

Evaluating model performance

Characteristics of some traditional and novel performance measures

from Steyerberg et al, 2010

Aspect	Measure	Visualization	Characteristics
Overall performance	R^2 Brier	Validation graph	Better with lower distance between Y and $\hat{Y}$. Captures calibration and discrimination aspects.
Discrimination	C statistic	ROC curve	Rank order statistic; Interpretation for a pair of patients with and without the outcome
	Discrimination slope	Box plot	Difference in mean of predictions between outcomes; Easy visualization
Calibration	Calibration-in-the-large	Calibration or validation graph	Compare mean(y) versus mean($\hat{y}$); essential aspect for external validation
	Calibration slope		Regression slope of linear predictor; essential aspect for internal and external validation related to 'shrinkage' of regression coefficients
	Hosmer-Lemeshow test		Compares observed to predicted by decile of predicted probability
Reclassification	Reclassification table	Cross-table or scatter plot	Compare classifications from 2 models (one with, one without a marker) for changes
	Reclassification calibration		Compare observed and predicted within cross-classified categories
	Net Reclassification Index (NRI)		Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right correction
	Integrated Discrimination Index (IDI)	Box plots for 2 models (one with, one without a marker)	Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes
Clinical usefulness	Net Benefit (NB)	Cross-table	Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)
	Decision curve analysis (DCA)	Decision curve	

Assessing incremental gain of adding a test/marker to a prediction model.

The debate continues, but a recent paper Pepe (2013) helps to provide a way forward.

Research Article

Statistics in Medicine

Received 19 December 2011, Accepted 11 December 2012 Published online 7 January 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5727

Testing for improvement in prediction model performance

Margaret Sullivan Pepe,[a*†] Kathleen F. Kerr,[b] Gary Longton[a] and Zheyu Wang[b]

Authors have proposed new methodology in recent years for evaluating the improvement in prediction performance gained by adding a new predictor, Y, to a risk model containing a set of baseline predictors, X, for a binary outcome D. We prove theoretically that null hypotheses concerning no improvement in performance are equivalent to the simple null hypothesis that Y is not a risk factor when controlling for X, $H_0 : P(D = 1|X, Y) = P(D = 1|X)$. Therefore, testing for improvement in prediction performance is redundant if Y has already been shown to be a risk factor. We also investigate properties of tests through simulation studies, focusing on the change in the area under the ROC curve (AUC). An unexpected finding is that standard testing procedures that do not adjust for variability in estimated regression coefficients are extremely conservative. This may explain why the AUC is widely considered insensitive to improvements in prediction performance and suggests that the problem of insensitivity has to do with use of invalid procedures for inference rather than with the measure itself. To avoid redundant testing and use of potentially problematic methods for inference, we recommend that hypothesis testing for no improvement be limited to evaluation of Y as a risk factor, for which methods are well developed and widely available. Analyses of measures of prediction performance should focus on estimation rather than on testing for no improvement in performance.

Assessing Impact

- Good model performance does not imply that the test, marker or predicted risk score will have a beneficial impact.

- Cost effectiveness analysis can address this issue:
  - Reliable/accurate inputs for costs etc may be difficult to obtain
  - Estimates (and conclusions) are likely to vary by setting (e.g. country).

How to start

- Identify sub-topics that are seen as important and where sensible advice can be provided
- Review and summarize relevant literature
- Recruit people with relevant expertise to take the lead and/or collaborate on identified sub-topics
- Ensure that feedback/input is sought from a key people working in the area who may hold a range of views

Relevant Literature

- Extensive reference lists are provided in existing papers.
- A key objective is to update these references with more recent articles.
- Of particular interest are papers that evaluate the properties of existing methods and measures.