BIOMED workshop

Performance of clinical prediction models

Thursday, May 16 2013, from 14h till 17h

PLACE:

KU Leuven, Department of Electrical Engineering, Kasteelpark Arenberg 10, 3001 Leuven, room 00.62

Programme:

14h00-14h30: Arnaud Installé, ESAT-SCD (SISTA)

14h30-14h40: Questions and Discussion

14h40-15h10: Ben Van Calster, Department of Development & Regeneration, KU Leuven

15h10-15h20: Questions and Discussion

15h20-15h40: Coffee Break

15h40-16h10: Kirsten Van Hoorde, ESAT-SCD (SISTA)

16h10-16h20: Questions and Discussion

16h20-16h50: Laure Wynants, ESAT-SCD (SISTA)

16h50-17h00: Questions and Discussion

ABSTRACTS:

14h00-14h30: Arnaud Installé, ESAT-SCD (SISTA)

Influence of sample size on predictive performance of ovarian tumour models

Machine-learning can play a valuable role in improving the quality of clinical diagnostic models. In the context of the International Ovarian Tumour Analysis (IOTA) studies, several such models have been created, based on different machine-learning algorithms, such as logistic regression and Least-Squares Support Vector Machines (LS-SVM).

The predictive performance of these models depends on various factors. Two of the most important ones are sample size and machine-learning algorithm used.

In this talk, we will analyze predictive performance of models based on the IOTA data set. We will use both logistic regression and LS-SVM, and compare their performance as well.

14h50-15h10: Ben Van Calster, Department of Development & Regeneration, KU Leuven

Comparing clinical prediction models: the interplay of sensitivity and specificity and the impact of model misspecification

When comparing clinical prediction models, it is essential to estimate the magnitude of change in performance rather than rely solely on statistical significance. The Net Reclassification Improvement (NRI) is widely used to assess improved classification by adding markers to risk prediction models or by comparing non-nested models. NRI does not consider misclassification costs in contrast with decision-analytic alternatives. We aimed to investigate measures that estimate change in classification performance when adding a new biomarker to a risk prediction model, assuming two-group classification based on a single risk threshold. Using simulated data, we investigate the change in sensitivity and specificity (ΔSe and ΔSp) and study the influence of ΔSe and ΔSp on the NRI (i.e. sum of ΔSe and ΔSp) and decision-analytic measures (Net Benefit or Relative Utility). We observed that even when a strong marker is added and/or the extended model has a dominating receiver operating characteristic curve, it is possible that either ΔSe (for thresholds below the event rate) or ΔSp (for thresholds above the event rate) is negative. In these cases decision-analytic measures provide more modest support for improved classification than NRI, but in general all measures confirm that adding the marker improved classification accuracy.

Exceptions may occur when important interaction effects are omitted, models are miscalibrated or continuous predictors are not modeled appropriately. Our results underscore the necessity of reporting ΔSe and ΔSp separately. When a single summary measure is desired, decision-analytic measures allow for a simple incorporation of the misclassification costs.

15h40-16h10: Kirsten Van Hoorde, ESAT-SCD (SISTA)

Calibration tools for polytomous risk prediction models

Risk prediction models are used to assist clinicians in making optimal treatment decisions. Therefore the estimated risks should be calibrated, i.e. correspond to observed risks. For binary prediction models several calibration tools exist to assess different aspects of model calibration, e.g. calibration-in-the-large, calibration slope, logistic calibration and (non-)parametric calibration plots. We extend these tools to risk prediction models for nominal outcomes developed using baseline-category multinomial logistic regression.

The multinomial logistic re-calibration model is a baseline-category fit of outcome *Y* with *k* categories, in which each category *i* (*i*=1,…,*k*-1) is compared to reference category *j* using $\log\left[P\left(Y=i\right)/P\left(Y=j\right)\right]=a_i+b_i\cdot \mathrm{lp}_i$, with lp_i the linear predictor for category *i* versus *j*. Thus, each equation contains only the linear predictor of the category involved. Calibration-in-the large $\left(a_i \mid b_i=1\right)$ assesses whether or not risks are too high or low, and calibration slopes $\left(b_i\right)$ whether or not risks are overfitted or underfitted. A multinomial parametric calibration plot uses the logistic re-calibration model, thus assuming linearity of lp_i. A non-parametric alternative estimates $a_i'+b_i'\cdot s\left(\mathrm{lp}_i\right)$, with $s(\ )$ a vector spline, which is a natural extension of the cubic smoothing spline to vector responses.

An illustrative case study is presented on the prediction of ovarian tumors as benign, borderline or invasive. The prediction model is based on 2037 patients from oncology referral centers, and is evaluated on 1107 patients from public hospitals.

16h20-16h50: Laure Wynants, ESAT-SCD (SISTA)

Performance measures in clustered data

Increasingly, multicenter studies are being conducted. An important reason to collect data in multiple centers is that this enhances the generalization of the results. The recruitment of subjects from a wider population and a broader range of clinical settings results in a sample that is more typical of the target population. Multicentre data have a clustered nature: patients from the same centre are likely to be more similar than patients from different centres. Additionally, performance measures may primarily reflect the performance in the largest centers.

Performance measures for clustered data exist but are not commonly used. We discuss cluster-corrected discrimination indices such as cluster-corrected sensitivity and cluster-corrected specificity. We also discuss two options for correcting the AUC for clustering: the within-cluster AUC proposed by Van Oirbeek and the AUC determined by a random effects meta-analysis of cluster-specific AUC's. We further discus cluster-corrected calibration. All methods will be illustrated on the IOTA 3 dataset, collected with the purpose of classifying ovarian tumors as benign or malignant.