

Background Clinical prediction models are used for different purposes, but purpose-specific validation is not usually carried out. The ability of a model to discriminate between true positives and false positives has applications in clinical decision making, screening, and service evaluation. The calibration (goodness-of-fit) of a model is a key indicator of how well a model's predicted outcomes reflect those actually observed. Initial validation of models usually includes assessment of these features but re-evaluation over time might not be performed.

EuroSCORE is an adult cardiac surgery risk model which has been in use since 1998. It predicts in-hospital mortality and is used for clinical decision making and service evaluation. It is widely acknowledged to have demonstrated 'calibration drift', but this has not been formally evaluated in the UK population.

Methods We assessed the performance of EuroSCORE in the Central Cardiac Audit Database (CCAD), covering all NHS cardiac procedures in the UK. Discrimination was tested using the area under the Receiver Operator Characteristic (ROC) curve (AUC). Calibration was assessed with the Hosmer-Lemeshow goodness of fit test. In addition, we developed new models with longer-term outcomes using the data, and tested year-on-year model performance.

Results A total of 399,314 eligible procedures from 1st April 1998 to 31st March 2011 were included in the analysis. Assessing the discrimination of EuroSCORE by financial year showed consistency across the period (AUC values ranging from 0.788 to 0.818). Model calibration, however, drifted considerably with a cumulative mortality over-estimate of 10,801 deaths by the end of the period (increasing from 147 over-estimated deaths in 1998 to 1,500 in 2010). This represented a predicted overall mortality rate of 6.0% compared with the observed rate of 3.4%. We will also present findings relating to year-on-year performance of a panel of models tailored to longer-term outcomes in specific procedures.

Conclusion Models that retain accurate discrimination while undergoing calibration drift may be implemented in settings for longer than is appropriate. A model that maintains good discrimination may be useful in a subset of scenarios, but for most purposes good calibration is also crucial. For models developed for multiple applications, purpose-specific validation and recalibration should be considered. Model performance should be appraised in context and not by indicators in isolation.

OP72 PROSPECTIVE EXTERNAL VALIDATION OF RISK PREDICTION MODELS FOR ACUTE TRAUMATIC BRAIN INJURY IN UK CRITICAL CARE UNITS: THE RAIN STUDY

doi:10.1136/jech-2012-201753.072

¹DA Harrison, ¹KA Griggs, ²M Gomes, ³DK Menon, ¹KM Rowan. ¹*Clinical Trials Unit, Intensive Care National Audit & Research Centre (ICNARC), London, UK;* ²*Department of Public Health & Policy, LSHTM, London, UK;* ³*School of Clinical Medicine, University of Cambridge, Cambridge, UK*

Background Acute traumatic brain injury (TBI) is the leading cause of death and disability in adults aged under 40 years. Statistical models have been developed to predict the risk of mortality or unfavourable outcome (death or severe disability) at six months following acute TBI but to date these risk prediction models have only been validated using existing data sources. The Risk Adjustment In Neurocritical care (RAIN) Study aimed to validate these risk prediction models among adults with acute TBI admitted to UK critical care units.

Methods Ten risk prediction models were identified: four for mortality at six months (the Hukkelhoven model and IMPACT Core, Extended and Lab models); and six for unfavourable outcome at six months (as mortality plus CRASH Basic and CT models). Risk factor data were collected from 67 UK critical care units (including 90% of regional neuroscience centres) from August 2009 to March 2011. Patients were followed up to six months for mortality by

linkage with death registration and unfavourable outcome using the Glasgow Outcome Scale (Extended) administered by postal or telephone questionnaire.

The risk prediction models were validated for calibration (c index), discrimination (Hosmer-Lemeshow test and Cox calibration regression) and overall fit (Brier score). Missing data were handled with multiple imputation.

Results Data were collected for 2,975 eligible patients admitted to critical care following acute TBI. 97% of patients were followed-up for mortality and 81% for unfavourable outcome at six months. Following multiple imputation, mortality and unfavourable outcome at six months were 26% and 57%, respectively. Risk prediction models for mortality at six months had good discrimination (c index 0.75–0.78) and the Hukkelhoven and IMPACT Lab models were well calibrated, although the IMPACT Core and Extended models over-predicted mortality. The models for unfavourable outcome at six months had worse discrimination (c index 0.69–0.71) and all models substantially under-predicted risk of unfavourable outcome. The best performance overall was found for the IMPACT Lab model, which was the most complex model, incorporating laboratory measurements. Models of the next level of complexity (Hukkelhoven, CRASH CT, IMPACT Extended) all performed similarly.

Conclusion Risk prediction models for acute TBI had acceptable discrimination among a large, representative sample of patients admitted to UK critical care units. Calibration was good for mortality but poor for unfavourable outcome, and these models therefore require recalibration for use in this setting.

Friday 14 September 2012

Parallel Session D

Research Methods: Surveys and Use of Routine Data

OP73 EXPLORING IMPACTS OF SURVEY NON-RESPONSE USING RECORD-LINKAGE OF SCOTTISH HEALTH SURVEY DATA (2003 TO 2008)

doi:10.1136/jech-2012-201753.073

¹L Gray, ²G McCartney, ³IR White, ⁴L Given, ¹SV Katikireddi, ¹AH Leyland. ¹*Social and Public Health Sciences Unit, MRC/CSO, Glasgow, UK;* ²*Health Scotland, NHS, Glasgow, UK;* ³*Biostatistics Unit, MRC, Cambridge, UK;* ⁴*Scottish Centre for Social Research, Nat-Cen, Edinburgh, UK*

Background Inference on population health is commonly derived from health survey data, based on the assumption that they are representative of the target communities. Departure from representativeness may weaken external validity leading to biased estimates, with important implications for public health evidence, particularly in relation to prevalence and quantity estimates such as population alcohol consumption. A key aspect determining the extent to which surveys are representative is the level of participant response. Inverse probability weights based on a limited range of demographic variables are usually applied in an attempt to correct for non-response. We aimed to investigate whether weighted estimates of all-cause mortality and mortality from alcohol-related conditions derived from the 2003 Scottish Health Survey (SHeS) – with household response of 67% – reflect those in the population of Scotland.

Methods Baseline observations from SHeS were individually record-linked to mortality data for the 91% of respondents consenting to linkage, and directly age-standardised survey-weighted mortality rates were calculated for the 3117 men and 3980 women aged 20 years and older at interview. Equivalent mortality rates were calculated for the whole of Scotland in the same age group using