countries were the following: (1) keep the same date, event, and consequences whenever possible (e.g., Tuesday May 4, a 3 alarm fire, destruction of two hotels and one restaurant); and (2) substitute the place where the event is located (i.e., a city (Cleveland, Ohio)] with a place familiar to the subjects living in the target countries. RESULTS: The event (fire) could be kept in all countries. The date had to be adapted in the Netherlands because it corresponds to a commemoration (Remembrance of the Dead) and would have introduced a bias if kept. The verb “a 3 alarm fire” was impossible to translate literally since no equivalent fire-classification system is used in most target countries (except in Canada). It was decided to use synonyms of “big” to qualify “fire.” Syntax was also an issue especially in Korea, Japan, Romance and Germanic languages where the order of some segments had to be inverted. CONCLUSIONS: Although simple in its structure, the KBANS story grammar proved to be challenging to translate into 24 languages and required a rigorous methodology to preserve the intent of the original.

PM110 STANDARDIZATION OF MENTAL HEALTH ASSESSMENT – USING ITEM RESPONSE THEORY (IRT) TO CROSS-CALIBRATE TWO SELF-REPORTED MENTAL HEALTH TOOLS: THE PATIENT HEALTH QUESTIONNAIRE (PHQ-9) AND THE SF-36V2 MENTAL HEALTH (MH) SCALE

Bjørner JB, White MK, Ylavis AS
Optum, Lincoln, RI, USA

OBJECTIVES: Mental health can be measured by numerous instruments, but scores are usually not directly comparable. The heterogeneity of scale specific metrics seriously impairs comparability across study results and the communication among researchers and clinicians. We aimed to develop and evaluate methods for cross-calibrating two popular mental health tools: the PHQ-9 and the SF-36v2 MH scales.

METHODS: We analyzed data from the United States and the UK including a general population sample (US: 216, UK: 355) and a sample with suspected depression (US: 113, UK: 183). An ensemble of confirmatory bifactor models tested whether the two instruments measured the same construct. Differential item function (DIF) between general population and depression samples was tested using logistic regression. Results estimated PHQ-9 and SF-36v2 MH generalization partial credit model and developed cross-calibration algorithms using the summed score cross-calibration approach. The measurement properties of the instruments were evaluated by test information functions. RESULTS: In the bifactor models, measurement invariance was strongly supported as the two scales measure the same general mental health construct. We found no indication of DIF supporting that the same item parameters apply to the general population and the depression samples. The cross-calibration algorithm revealed a 1:1 linear relation between PHQ-9 score and MH score in the PHQ-9 score range of 0-10 (moderate to severe depression), but a non-linear relation at more extreme scores. The PHQ-9 provided most information for persons with scores in the interval from the general population. The PHQ-9 was invariant to two standard deviations of the MH scale provided more information at the lower and upper extremes. CONCLUSIONS: We successfully developed a procedure for cross-calibrating the PHQ-9 and MH scales. These results can be used to compare scores between the two instruments.

PMI11 INTERNAL VALIDATION OF MAPPING ANALYSES FOR HEALTH TECHNOLOGY ASSESSMENT

Tranman D, Trehanne C
Abacus International, Bicester, UK

OBJECTIVES: Mapping between health status measures is common practice within health economic evaluations. The objective of this analysis was to evaluate the suitability of hold-out validation, whereby models are fitted to a subset of data and then tested on the remaining observations, compared to other methodologies utilising full sample approaches in small to medium sized samples. METHODS: Four models predicting EQ-5D from the SF-12 were estimated using the Medical Expenditure Panel Survey models were estimated using three hypothetical sample sizes of 500, 1,000, and 4,000 observations. For each model and sample size, two hold-out validation specifications were compared against alternative estimators of error: the naïve resubstitution error; repeated 10-fold cross validation; the optimism-corrected bootstrap; the 0.632 bootstrap. The results from these estimators were compared against asymptotic estimates of the true error indices in the remaining observations (n=15,675). Estimators were evaluated by assessment of bias and variance. The exercise was repeated 500 times. RESULTS: Hold-out methods were subject to the largest variance across all estimators and sample sizes. Variance was lower and similar in the full sample estimators (bootstrap and cross-validation methods). The extent of bias in any sample size was associated with the degree to which the algorithms were adaptive to the training sample data. For the two mapping algorithms which were not adaptive to the training sample data, bias was low for all estimators. In the two algorithms which were more adaptive to the training sample data, the naïve resubstitution error was associated with a downward bias, hold-out methods exhibited an upward bias, and all full sample methods exhibited a low degree of bias. CONCLUSIONS: Hold-out validation exhibited the highest variance of all methods in all scenarios. Full-sample designs offer lower variance and are preferable to continued use of hold-out validation with small to medium sized datasets.

PM112 A SYSTEMATIC REVIEW OF METHODOLOGICAL FRAMEWORKS FOR EVALUATION OF ETHICAL CONSIDERATIONS IN HEALTH TECHNOLOGY ASSESSMENT

Anaya N, Schwartz I, Tarinde JE, Campbell K, Goeree R
McMaster University, Hamilton, ON, Canada

OBJECTIVES: While advances have been made in development of ethical frameworks for health technology assessment (HTA), there is no clear agreement on the most useful and practical approach to address ethical aspects in HTA. Moreover, uncertainty remains about appropriate scope and level of details of ethical framework works for HTA. This study seeks to systematically review the literature to identify ethical frameworks and to provide an overview of their methodological features and to gain a better understanding of the areas of commonality and divergence between different frameworks. METHODS: We conducted a systematic search of literature, without limits of time and language, to identify the guidance documents, ethical frameworks published up to October 1st 2015. RESULTS: The review identified 22 frameworks, varying in their philosophical approach, structure, and comprehensiveness. They were designed for different purposes throughout the HTA process, ranging from being used to identify ethical quandaries to a production of analysis of ethical data to supporting decision-makers in making better informed value-sensitive decisions. They frequently promoted analytical methods that combined normative reflection with descriptive approaches to the analysis of values of stakeholders and other societal or technical actors. CONCLUSIONS: The choice of a method for collection and analysis of ethical data seems to depend on the context in which technology is being assessed, the purpose of analysis, and availability of required resources.

PM113 COMPARE PROPENSITY SCORE, PROPENSITY SCORE WITH COVARIATES AND GENETIC ALGORITHM METHODS FOR COVARIATE MATCHING IN OBSERVATIONAL STUDIES

Clancy C, Bakken DG, Wasserman D, Spilman J
Health Partnerships BC, Vancouver, BC, Canada

OBJECTIVES: When the number of individuals as providing informal (unpaid) care for an aging relative. We compare three different methods of covariate matching to determine the effect of caregiving on the mental health status of caregivers. METHODS: Three matching methods were used: 1. Propensity score matching with covariates within calipers; 2. Propensity score matching with calipers; and 3. Genetic algorithm matching. RESULTS: All three methods provide adequate balance on the covariates used for matching. Methods 2 and 3 produce the closest covariate balance, with absolute mean covariate differences less than 0.0008 for all covariates and less than 0.0001 on the core set of covariates. Because methods that censor observations (i.e. matching within calipers) may artificially improve covariate balance, we take the ATT estimate from genetic matching to be the least biased estimate of the true effect. Using a standard 5 point self report measure of mental health, caregivers, on average, report a mental health state that is 5.4% worse than non-caregivers (roughly one-fourth “less healthy” within any given scale range (e.g. 2–3, 3–4). CONCLUSIONS: As all three methods provide adequate balance, our consideration turns to bias reduction and the fact that the genetic matching does not require that we estimate the propensity score prior to matching. We consider the drivers or caregiver MH and implications for health care policy.

PM114 ARE INDUSTRY FUNDED NETWORK META-ANALYSES LOWER QUALITY? A SYSTEMATIC SEARCH OF OVID-MEDLINE FOR NETWORK META-ANALYSES INCLUDING AT LEAST ONE PHARMACEUTICAL

Chambers JD, Gunjal ES, White L, Hendry F, Hop MCG, Pyne WD
Thuffs Medical Center, Boston, MA, USA, 2University of Houston, Houston, TX, USA, 3The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 4Daisy Hill Hospital, Neury, 5Downe Hospital, UK

OBJECTIVES: To compare the quality and transparency of industry supported network meta-analyses with those with non-profit support or no support. METHODS: We systematically searched OVID-Medline for network meta-analyses including at least one pharmaceutical. We reviewed each network meta-analysis and evaluated key general study characteristics, methodology, and transparency using a checklist of objective criteria derived from the ISPOR Taskforce’s recommendations for study conduct and reporting. We reported source of study funding when available. When source of funding was unclear or not reported we contacted the corresponding author. We compared the quality and transparency of industry supported network meta-analyses with those with non-profit support or no support. RESULTS: Two hundred and fourteen studies met our inclusion criteria and were included in our dataset. Source of funding was identified for 211 studies (98.6%). Industry supported studies tended to be published in lower quality medical journals (p<0.01), and typically included fewer studies (p<0.01) and a smaller total number of patients (p<0.05). In terms of study transparency, industry supported studies less often reported the search terms (p<0.01) and, for analyses conducted using a Bayesian framework, presented the model code (p<0.01). Regarding study methodology, industry supported network meta-analyses less often reported a quality assessment of clinical studies included in the network meta-analysis (p<0.01), and less often compared the findings of traditional meta-analysis and network meta-analysis (p<0.01). With respect to presentation of findings, industry supported studies less often reported the full matrix of head-to-head comparisons (p<0.01), or provided a ranking of treatments (p<0.01). CONCLUSIONS: We found that studies with non-support or no support funded tended to be more transparent and rigorous than industry in their funded studies. Studies funded by industry pay particular attention to interpret the findings of network meta-analyses.

PM115 AUTOMATIC DEVELOPMENT OF CLINICAL PREDICTION MODELS WITH GENETIC PROGRAMMING: A CASE STUDY IN CARDIOVASCULAR DISEASE

Nashizge CA, Currie CJ, Freese A, Spasic I
Cardiff University, Cardiff, UK
OBJECTIVES: Genetic programming seems a promising technique for the automated development of clinical prediction models using cardiovascular disease as a case study. METHODS: We compared genetic programming and the commonly used Cox regression technique in the development of a cardiovascular risk score using data from the SMART study, a prospective cohort study designed to identify predictors of future cardiovascular events in patients with symptomatic cardiovascular disease. The primary outcome was any cardiovascular event, comprising cardiovascular death, non-fatal stroke and myocardial infarction. The discrimination of the model was assessed in terms of discrimination and calibration. RESULTS: 3,873 patients were enrolled in the study 1996–2006, aged 19–82 years and with 460 cardiovascular events during the 19–7 months follow-up period. The discrimination of both models was also comparable, indicating similar disagreement between observed and predicted risks. The C-index of the genetic programming model being smaller (0.65; 95% CI: 0.63–0.66) but not significantly different from that of the Cox regression model (0.71; 0.67–0.75). The calibration of both models was also comparable, indicating similar disagreement between observed and predicted risks. The wider CI reflected the additional data uncertainty from the small number of events.

OBJECTIVES: To develop and test a novel approach to estimate the ICER via a new bootstrap approach based upon the sample coefficient of variance and optimized via computational intelligence. METHODS: A novel bootstrap ICER estimation approach was developed and that was compared to the standard univariate Cox regression. The predictive power of the model was assessed in terms of discrimination and calibration. RESULTS: The proposed approach yielded less statistical estimation error than the typical bootstrap approach across all cases, with the relative efficiency of point estimates ranging from +0.06% to +113.35%. An equal or improved coverage error for the CI was also consistently achieved, deviating from the population value by zero (i.e., perfect coverage) to 0.0200 versus from 0.0060 to 0.0210. Subsequently, an improved shortening of the CI length was noted. The relative bias suggested slightly increased over the bias estimated by the confidence interval of the computed and the comparator. Applying the equations for two hypothetical points, (7.6 QALY, US$100,000) and (8.6 QALY, US$150,000), we found that the ICER of 50 bounds with the lower and the upper limits, respectively, 40.6 and 60.8 US$(x1000)/QALY for exponential model, and (8.6 QALY, US$150,000), we found that the ICER of 50 bounds with the lower and the upper limits, respectively, 40.6 and 60.8 US$(x1000)/QALY for exponential model, and as well for 49.3 and 53.1 for quadratic model. Those estimates were not so much different as the limits of 43.9 and 65.8 for the ICER of 51.4, obtained by the regression analysis presented in the ISPOR 2015 conference. CONCLUSIONS: Our approach can be used in multiple and science-based methods to estimate the ICER. It would be useful for negotiations and decisions in value-based pricing in which a range of ICER must be considered beyond a single threshold ratio.

OBJECTIVES: To develop predictive models to identify CHF and COPD patients at high risk of hospitalization using predictive analytics for patient outreach. METHODS: We conducted 1000 Monte Carlo simulations for each of the outcome models where covariates were related only to the exposure, but substantial bias tended to increase with more covariates. Simulation results showed that a pair of cost and benefit is given on the C(cost)-E(benefit) plane. In order to find a cost-benefit function connecting the two points, an analytical method was developed by means of curve-fitting technique with exponential and quadratic modeling. The resultant cost-benefit function was further analytically expanded to the derivative, dC/dE, where we call it "tangent limit". Example calculations of the tangent limits were conducted for each model. RESULTS: The analytical development resulted in the following equations of the cost-benefit function and the derivative for each model: C = (1/q) * Exp((E - p)/q) for exponential model, and C = (1/q) * exp(-p/q) for quadratic model, where p and q are parameters selected by the curve-fitting technique. The model was applied to a control group of 10,000 patients and to the outcome model. The results showed that the model was able to identify patients at high risk of admission have been developed, which were used to generate a list of patients with high probability of admission for case management outreach.

OBJECTIVES: Using empirical data we demonstrated that a prediction model developed by the novel technique of genetic programming has a comparable predictive ability to that of Cox regression. We conducted 1000 Monte Carlo simulations for each of the outcome models where covariates were related only to the exposure, but substantial bias tended to increase with more covariates. Simulation results showed that a pair of cost and benefit is given on the C(cost)-E(benefit) plane. In order to find a cost-benefit function connecting the two points, an analytical method was developed by means of curve-fitting technique with exponential and quadratic modeling. The resultant cost-benefit function was further analytically expanded to the derivative, dC/dE, where we call it "tangent limit". Example calculations of the tangent limits were conducted for each model. RESULTS: The analytical development resulted in the following equations of the cost-benefit function and the derivative for each model: C = (1/q) * Exp((E - p)/q) for exponential model, and C = (1/q) * exp(-p/q) for quadratic model, where p and q are parameters selected by the curve-fitting technique. The model was applied to a control group of 10,000 patients and to the outcome model. The results showed that the model was able to identify patients at high risk of admission have been developed, which were used to generate a list of patients with high probability of admission for case management outreach.

OBJECTIVES: Genetic programming seems a promising technique for the automated development of clinical prediction models using cardiovascular disease as a case study. METHODS: We compared genetic programming and the commonly used Cox regression technique in the development of a cardiovascular risk score using data from the SMART study, a prospective cohort study designed to identify predictors of future cardiovascular events in patients with symptomatic cardiovascular disease. The primary outcome was any cardiovascular event, comprising cardiovascular death, non-fatal stroke and myocardial infarction. The discrimination of the model was assessed in terms of discrimination and calibration. RESULTS: 3,873 patients were enrolled in the study 1996–2006, aged 19–82 years and with 460 cardiovascular events during the 19–7 months follow-up period. The discrimination of both models was also comparable, indicating similar disagreement between observed and predicted risks. The C-index of the genetic programming model being smaller (0.65; 95% CI: 0.63–0.66) but not significantly different from that of the Cox regression model (0.71; 0.67–0.75). The calibration of both models was also comparable, indicating similar disagreement between observed and predicted risks. The wider CI reflected the additional data uncertainty from the small number of events.

OBJECTIVES: To develop and test a novel approach to estimate the ICER via a new bootstrap approach based upon the sample coefficient of variance and optimized via computational intelligence. METHODS: A novel bootstrap ICER estimation approach was developed and that was compared to the standard univariate Cox regression. The predictive power of the model was assessed in terms of discrimination and calibration. RESULTS: The proposed approach yielded less statistical estimation error than the typical bootstrap approach across all cases, with the relative efficiency of point estimates ranging from +0.06% to +113.35%. An equal or improved coverage error for the CI was also consistently achieved, deviating from the population value by zero (i.e., perfect coverage) to 0.0200 versus from 0.0060 to 0.0210. Subsequently, an improved shortening of the CI length was noted. The relative bias suggested slightly increased over the bias estimated by the confidence interval of the computed and the comparator. Applying the equations for two hypothetical points, (7.6 QALY, US$100,000) and (8.6 QALY, US$150,000), we found that the ICER of 50 bounds with the lower and the upper limits, respectively, 40.6 and 60.8 US$(x1000)/QALY for exponential model, and as well for 49.3 and 53.1 for quadratic model. Those estimates were not so much different as the limits of 43.9 and 65.8 for the ICER of 51.4, obtained by the regression analysis presented in the ISPOR 2015 conference. CONCLUSIONS: Our approach can be used in multiple and science-based methods to estimate the ICER. It would be useful for negotiations and decisions in value-based pricing in which a range of ICER must be considered beyond a single threshold ratio.

OBJECTIVES: Using empirical data we demonstrated that a prediction model developed by the novel technique of genetic programming has a comparable predictive ability to that of Cox regression. We conducted 1000 Monte Carlo simulations for each of the outcome models where covariates were related only to the exposure, but substantial bias tended to increase with more covariates. Simulation results showed that a pair of cost and benefit is given on the C(cost)-E(benefit) plane. In order to find a cost-benefit function connecting the two points, an analytical method was developed by means of curve-fitting technique with exponential and quadratic modeling. The resultant cost-benefit function was further analytically expanded to the derivative, dC/dE, where we call it "tangent limit". Example calculations of the tangent limits were conducted for each model. RESULTS: The analytical development resulted in the following equations of the cost-benefit function and the derivative for each model: C = (1/q) * Exp((E - p)/q) for exponential model, and C = (1/q) * exp(-p/q) for quadratic model, where p and q are parameters selected by the curve-fitting technique. The model was applied to a control group of 10,000 patients and to the outcome model. The results showed that the model was able to identify patients at high risk of admission have been developed, which were used to generate a list of patients with high probability of admission for case management outreach.