countries were the following: (1) keep the same date, event, and consequences whenever possible (i.e., 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 changed in The Netherlands because it corresponds to a commemoration (Remembrance of the Dead) and would have introduced a bias if kept. The verbatim "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 RBANS story memory-test proved to be challenging to translate into 24 languages and required a rigorous methodology to preserve the intent of the original.

PRM110

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

Bjorner JB, White MK, Yarlas 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-calibration of two popular mental health tools: the PHQ-9 and the SF-36v2 MH scale. **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: 169, UK: 153). Multigroup 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 DIF tests. We estimated IRT item parameters using a multigroup generalized 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 model, all items loaded strongly on a common factor, supporting that 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 fairly linear relation between PHQ-9 score and MH score in the PHQ-9 score range of 10-20 (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 average down to two standard deviations below average, but 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.

PRM111

INTERNAL VALIDATION OF MAPPING ANALYSES FOR HEALTH TECHNOLOGY ASSESSMENT

Trueman D, Treharne 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 in the remaining observations, compared to other methods of internal validation 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.

PRM112

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

Assasi N, Schwartz L, Tarride 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 frameworks for HTA. This study seeks to systematically review the literature to identify existing ethics frameworks for HTA in order 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 or practical frameworks published up to October 1st 2013. **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 helping HTA producers in identification, appraisal and 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.

RESEARCH ON METHODS – Statistical Methods

PRM113

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

Claeys C, Bakken DG, Wasserman D, Spilman J

KJT Group, Inc., Honeoye Falls, NY, USA

OBJECTIVES: As the population ages an increasing number of individuals are 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 states of informal caregivers. Covariate matching methods pair observations from different treatment groups by matching the members of each pair on a set or vector of covariates that would be randomly distributed across the groups in a randomized trial. **METHODS:** Multiple waves of an online survey conducted among a representative sample of U.S. adults yielded 740 informal caregivers and 2260 non-caregivers. We applied three different methods for covariate matching to determine the "average effect of treatment on the treated" (ATT) of caregiving on mental health states (MH): 1. Propensity score within calipers; 2. Propensity score and covariates within 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 best covariate balance, with absolute mean covariate differences less 0.0008 on all covariates and less than 0.00001 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 matching, 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.

PRM114

ARE INDUSTRY FUNDED NETWORK META-ANALYSES LOWER QUALITY?

Chambers JD[1], Gunjal SS[2], Winn A[3], Kennedy IR[4], Hoey MG[5], Pyo J[1]

[1]*Tufts Medical Center, Boston, MA, USA,* [2]*University of Houston, Houston, TX, USA,* [3]*The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA,* [4]*Daisy Hill Hospital, Newry, UK,* [5]*Downe 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 ($p<0.05$) 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-profit support or no support funded tended to be more transparent and rigorous than industry supported studies. Study findings emphasize that users of network meta-analyses should take great care to account for study quality when interpreting the findings of network meta-analyses.

PRM115

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

Bannister CA, Currie CJ, Preece A, Spasic I

Cardiff University, Cardiff, UK