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 altered 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 KBANS story memory task 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
Bjornet JB, White MK, Yarias 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 the two mental health instruments: PHQ-9 and 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: 181, UK: 156). The intergroup 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 models, depression symptoms loaded 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 common 2-parameter IRT model, with a common cut-score 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 level to 2 standard deviations above the average score on 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, Trebarne 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 data, in cross-calibration of the PHQ-9 and SF-36V2 MH scales.

RESULTS: Four models predicting EQ-5D from the SF-12 were estimated using the Medical Expenditure Panel Survey data. 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 naive 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 whenever hold-out was applied to the largest variance across all estimators and sample sizes. Variance was lower whenever hold-out was applied to the data in this model, and was lower in the full sample estimators (bootstrap and cross-validation methods).

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
Ansell K, Schwartz L, Tarrant JE, Campbell K, Geore R
McMaster University, Hamilton, ON, Canada
OBJECTIVES: While advances have been made in development of ethical framework tools 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 tools for HTA. This study seeks to systematically review the literature to identify existing frameworks, their structure, and 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 and technical frameworks published up to 31st December 2014. RESULTS: The review identified 22 frameworks, varying in their philosophical approach, structure, comprehensiveness. They were designed for different purposes throughout the HTA process, ranging from helping producers in identifying for a rigorous methodology to preserve the intent of the original.

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
Cleary C, Bakken DG, Wasserman D, Spilman J
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 4Daisy Hill Hospital, Newry, UK, 5Downe Hospital, UK
OBJECTIVES: When a study involves an increasing number of individuals is providing (uninformal) 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 (unpaid) caregivers within 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. As multiple subsets of a single sample were matched, support that the same item parameters apply to the general population and the depression samples. We considered the genetic matching does not require that we estimate the propensity score prior to matching.

CONCLUSIONS: We consider the drivers or caregiver MH and implications for health care policy.

PRM114
Thufs 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, Newry, UK, 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-profit support or no support funded tended to be more transparent and rigorous than industry supported studies. Studies with industry funding may underestimate the effectiveness of network meta-analyses.

PRM115
AUTOMATIC DEVELOPMENT OF CLINICAL PREDICTION MODELS WITH GENETIC PROGRAMMING: A CASE STUDY IN CARDIOVASCULAR DISEASE
Rangarajan CA, Currie C, Friesche A, Spacie 1
Gairdner University, Cardiff, UK
OBJECTIVES: We aimed to develop and evaluate methods for cross-calibration of the PHQ-9 and SF-36V2 MH scales. These results can be used to compare scores between the two instruments.

RESULTS: The exercise was repeated 500 times.

CONCLUSIONS: We consider the drivers or caregiver MH and implications for health care policy.