DIVISION OF INTEGRATED CARE & PUBLIC HEALTH
UNIVERSITY of WASHINGTON
Department of Psychiatry & Behavioral Sciences

IMPACT Research Methods

Updated 3/6/2014

These articles focus on research methods used in the IMPACT trial, including screeners and scales used to measure.

1. Callahan CM, Unverzagt FW, Hui SL, Perkins AJ, Hendrie HC. **Six-item screener to identify cognitive impairment among potential subjects for clinical research.** *Med Care.* 2002;40(9):771-81.

   OBJECTIVE: To design a brief cognitive screener with acceptable sensitivity and specificity for identifying subjects with cognitive impairment. DESIGN: Cohort one is assembled from a community-based survey coupled with a second-stage diagnostic evaluation using formal diagnostic criteria for dementia. Cohort two is assembled from referrals to a specialty clinic for dementing disorders that completed the same diagnostic evaluation. SETTING: Urban neighborhoods in Indianapolis, Indiana and the Indiana Alzheimer Disease Center. PATIENTS: Cohort one consists of 344 community-dwelling black persons identified from a random sample of 2212 black persons aged 65 and older residing in Indianapolis; cohort two consists of 651 subject referrals to the Alzheimer Disease Center. MEASUREMENTS: Formal diagnostic clinical assessments for dementia including scores on the Mini-mental state examination (MMSE), a six-item screener derived from the MMSE, the Blessed Dementia Rating Scale (BDRS), and the Word List Recall. Based on clinical evaluations, subjects were categorized as no cognitive impairment, cognitive impairment-not demented, or demented. RESULTS: The mean age of the community-based sample was 74.4 years, 59.4% of the sample were women, and the mean years of education was 10.1. The prevalence of dementia in this sample was 4.3% and the prevalence of cognitive impairment was 24.6%. Using a cut-off of three or more errors, the sensitivity and specificity of the six-item screener for a diagnosis of dementia was 88.7 and 88.0, respectively. In the same sample, the corresponding sensitivity and specificity for the MMSE using a cut-off score of 23 was 95.2 and 86.7. The performance of the two scales was comparable across the two populations studied and using either cognitive impairment or dementia as the gold standard. An increasing number of errors on the six-item screener is highly correlated with poorer scores on longer measures of cognitive impairment. CONCLUSIONS: The six-item screener is a brief and reliable instrument for identifying subjects with cognitive impairment and its diagnostic properties are comparable to the full MMSE. It can be administered by telephone or face-to-face interview and is easily scored by a simple summation of errors.

2. Unutzer J, Choi Y, Cook IA, Oishi S. **A web-based data management system to improve care for depression in a multicenter clinical trial.** *Psychiatr Serv.* 2002;53(6):671-3, 8.

3. Lowe B, Unutzer J, Callahan CM, Perkins AJ, Kroenke K. **Monitoring depression treatment outcomes with the patient health questionnaire-9.** *Med Care.* 2004;42(12):1194-201.

BACKGROUND: Although effective treatment of depressed patients requires regular follow-up contacts and symptom monitoring, an efficient method for assessing treatment outcome is lacking. We investigated responsiveness to treatment, reproducibility, and minimal clinically important difference of the Patient Health Questionnaire-9 (PHQ-9), a standard instrument for diagnosing depression in primary care. METHODS: This study included 434 intervention subjects from the IMPACT study, a multisite treatment trial of late-life depression (63% female, mean age 71 years). Changes in PHQ-9 scores over the course of time were evaluated with respect to change scores of the SCL-20 depression scale as well as 2 independent structured diagnostic interviews for depression during a 6-month period. Test-retest reliability and minimal clinically important difference were assessed in 2 subgroups of patients who completed the PHQ-9 twice exactly 7 days apart. RESULTS: The PHQ-9 responsiveness as measured by effect size was significantly greater than the SCL-20 at 3 months (-1.3 versus -0.9) and equivalent at 6 months (-1.3 versus -1.2). With respect to structured diagnostic interviews, both the PHQ-9 and the SCL-20 change scores accurately discriminated patients with persistent major depression, partial remission, and full remission. Test-retest reliability of the PHQ-9 was excellent, and its minimal clinically important difference for individual change, estimated as 2 standard errors of measurement, was 5 points on the 0 to 27 point PHQ-9 scale. CONCLUSIONS: Well-validated as a diagnostic measure, the PHQ-9 has now proven to be a responsive and reliable measure of depression treatment outcomes. Its responsiveness to treatment coupled with its brevity makes the PHQ-9 an attractive tool for gauging response to treatment in individual patient care as well as in clinical research.

4. Perkins AJ, Kroenke K, Unützer J, Katon W, Williams JW, Hope C, et al. **Common comorbidity scales were similar in their ability to predict health care costs and mortality.** *J Clin Epidemiol.* 2004;57(10):1040-8.

OBJECTIVE: To compare the ability of commonly used measures of medical comorbidity (ambulatory care groups [ACGs], Charlson comorbidity index, chronic disease score, number of prescribed medications, and number of chronic diseases) to predict mortality and health care costs over 1 year. STUDY DESIGN AND SETTING: A prospective cohort study of community-dwelling older adults (n=3,496) attending a large primary care practice. RESULTS: For predicting health care charges, the number of medications had the highest predictive validity (R(2)=13.6%) after adjusting for demographics. ACGs (R(2)=16.4%) and the number of medications (15.0%) had the highest predictive validity for predicting ambulatory visits. ACGs

and the Charlson comorbidity index (area under the receiver operator characteristic [ROC] curve=0.695-0.767) performed better than medication-based measures (area under the ROC curve=0.662-0.679) for predicting mortality. There is relatively little difference, however, in the predictive validity across these scales. CONCLUSION: In an outpatient setting, a simple count of medications may be the most efficient comorbidity measure for predicting utilization and health-care charges over the ensuing year. In contrast, diagnosis-based measures have greater predictive validity for 1-year mortality. Current comorbidity measures, however, have only poor to moderate predictive validity for costs or mortality over 1 year.

5. Tang L, Song J, Belin TR, Unutzer J. **A comparison of imputation methods in a longitudinal randomized clinical trial.** *Stat Med.* 2005;24(14):2111-28.

It is common for longitudinal clinical trials to face problems of item non-response, unit non-response, and drop-out. In this paper, we compare two alternative methods of handling multivariate incomplete data across a baseline assessment and three follow-up time points in a multi-centre randomized controlled trial of a disease management programme for late-life depression. One approach combines hot-deck (HD) multiple imputation using a predictive mean matching method for item non-response and the approximate Bayesian bootstrap for unit non-response. A second method is based on a multivariate normal (MVN) model using PROC MI in SAS software V8.2. These two methods are contrasted with a last observation carried forward (LOCF) technique and available-case (AC) analysis in a simulation study where replicate analyses are performed on subsets of the originally complete cases. Missing-data patterns were simulated to be consistent with missing-data patterns found in the originally incomplete cases, and observed complete data means were taken to be the targets of estimation. Not surprisingly, the LOCF and AC methods had poor coverage properties for many of the variables evaluated. Multiple imputation under the MVN model performed well for most variables but produced less than nominal coverage for variables with highly skewed distributions. The HD method consistently produced close to nominal coverage, with interval widths that were roughly 7 per cent larger on average than those produced from the MVN model.