Possible Statistical Methods for Test Conversions

Andrew Zhou, Ph.D
Professor, Department of Biostatistics, University of Washington
Investigator, NACC, University of Washington
Danping Liu, Ph.D
Post-doc fellow, Department of Biostatistics and NACC, University of Washignton

UDS version 2.0 to 3.0

- Many neuropsychological tests are expected to be changed to the new battery due to lower cost and less restrictions to the patients.
- However, researchers may wish to use both the old and new test scores interchangeably, so they need to be equated or made comparable.
- We use one test to illustrate the migration, say, from Mini-Mental State Examination (MMSE) to Montreal Cognitive Assessment (MOCA). The other tests could be treated similarly.

Data pattern

- The data:

	Y_1 (old)	Y_2 (new)	X
Past data (UDS 2.0)	✓		✓
Transition period (training data)	✓	✓	✓
Future data (UDS 3.0)		✓	✓

- The old test is missing in future data and the new test is missing in the past data.
- The relationship between Y_1 and Y_2 would be derived from the training data only.

Scientific questions

- Our ultimate goal is not only to explore the relationship between the old and new test, but also to assist other researchers to address their scientific questions.
- Analysis sample could be a subset of past and/or future data.
- Examples of scientific questions:
  1. A patient get 20 points in MMSE and 23 points in MoCA one year later. Has he improved?
  2. Among a group of MCI patients, is there a significant cognitive decline in their one-year follow-up visit compared to the baseline? Only MoCA is available in the follow-up visit.
  3. How is the decline of MMSE score affected by patient characteristics during a four-year period? Only MoCA is available in the last time point.
  4. Is MoCA a stronger predictor of Alzheimer's disease than MMSE after adjusting for other confounders?

Randomization

- Who should get into the training sample? Ideally, the training sample should be a **random sample** of the NACC patients population.
- Block randomization within each ADC would be an optimal design, but may not be cost effective.
- Randomize ADC is more feasible: randomly select several ADCs; all the patients in these ADCs enter the training sample.
- Possible sampling bias?

Algorithm

- The method was proposed by Hui *et al.* (1997) to establish a standardized score out of different measurement instruments.
- The algorithm is summarized in the following steps:
  1. Start with regression analysis for Y_1 versus Y_2 and Y_2 versus Y_1. Add higher order terms to test if linear relationship holds.
  2. Subtract sample mean from the individual tests: $S = Y_1 - \bar{Y}_1$ and $T = Y_2 - \bar{Y}_2$.
  3. Minimize
  $$\sum_i (aS_i - bT_i)^2$$
  over the entire sample with the constraint
  $$a^2 + b^2 = L,$$
  a normalizing constant.
  4. The standardized score is aS and bT for the old and new test, respectively.

Comments

- The direct standardization approach is easy to implement in practice, and does not use any item level data.
- However, the method assumes a strong linear relationship between the old and new tests, which may not be realistic.
- Useful for marginal comparison, but not for more complicated scientific questions.

Item response theory (IRT)

- Strengths of IRT:
  - The test characteristic curve and calibrated score table provide a nice visualization of the equivalence between the old and new scores.
  - Allows nonlinear relationship.
  - Useful to answer scientific questions on individual patient (Scientific question 1).
  - Helpful to understand the internal structure of a test.
- Weaknesses of IRT:
  - Requires item-level data.
  - Difficult to adjust for the covariates.
  - Do not account for variability while equating the old and new scores.
  - May not take into account subsequent inferences on the population (Scientific questions 2-4) after conversions.

Scientific question 3 -revisit

- How is the decline of MMSE score affected by patient characteristics during a four-year period?
  - A longitudinal study.
  - In the last time point, only MoCA is available but not the MMSE score.
- Options:
  - Direct use MoCA as if it were MMSE - biased results.
  - Convert MoCA to MMSE using the calibrated score table or test characteristic curve - may underestimate the standard error.
  - Multiple imputation - a solution to answer scientific questions 2-4.

Standard MI approach

1. Convert from Y_1 to Y_2 (the similar method can also be applied to converting from Y_2 to Y_1)
2. Estimate the imputation model from the training data, i.e., the conditional distribution of $Y_2 | Y_1, X$.
3. Generate Y_2 for every subject in the analysis data except those in the training set to obtain an imputed data set.
4. Analyze the imputed data set to address the scientific question of interest.
5. Repeat step 2 and 3 for several times (usually 10 or 20 times), and obtain the estimated parameters from each imputed data set.
6. Combine the estimators using Rubin's rules, yielding a single pooled estimate and its standard error.

Advantages of MI approach

1. The appropriately done MI method will yield a statistically justifiable standard error for the final estimate.
2. More broadly, the main analysis for any of the hypothetical studies mentioned earlier can be done m times ($m = 10$ or 20, say), and the relevant parameter estimates (which might be a regression coefficient or other statistic) would be obtained by Rubin's rules.

Bias of the Rubin's rule

- The application of Rubin's rule requires the following conditions (Wang and Robins, 1998)
  - The imputation model and the analysis model are derived from the same data set.
  - Proper imputation – imputed values are drawn from the posterior predictive distribution.
  - Parametric imputation model.
- A simulation result:

	Bias (%)	SD	SE	Coverage (%)
Single imputation	-2.5	0.044	0.033	80.4
Rubin's rule	-2.4	0.022	0.056	99.6

Unique features of the problem

- The imputation model is estimated from “exogenous” data (training sample).
- Completely missing in the analysis sample: for example, we need to impute MMSE for every subject in the future data.
- Rubin’s combination rile for MI methods was developed for missing data without exogenous data.
- Rubin’s combination rule tends to overestimate the standard errors, so new MI combination is need for the analysis model.

Modified multiple imputation procedures

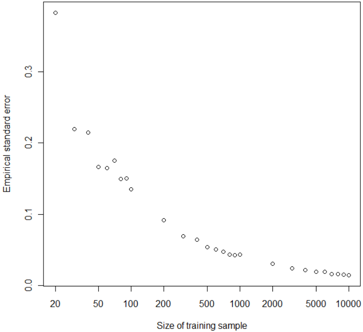
- All the imputed data sets are combined and analyzed together as if they are from i.i.d. observations.
- The “sandwich” variance formula needs to be derived to capture the variabilities due to (a) estimating the imputation model and (b) complete data analysis.
- Robins and Wang (2000) proposed similar techniques, but they did not use exogenous training data for the imputation.
- This procedure does not limit the method used for analysis, but variance estimates may differ for different analysis model - a potential obstacle for practical use.

Outstanding issues

- Shall the conversion to UDS 3.0 be done once and for all at NACC, without regard to the topics of future analytic studies, or, should the conversion to UDS 3.0 be tied to future longitudinal data projects that need the method to handle missing test data in longitudinal data?
- Hot to minimize bias in sampling of the training set? Sample size?
- Variables to use in the imputation model?

Sample size

- A simulation based on scientific question 2: multiple imputation to estimate the population mean. Analysis sample size: 1,000; training sample size: 20 - 10,000.
- In this example, a training sample size of 1,000 - 2,000 may be reasonable. More explorations needed for other scenarios.



Summary

- IRT and direct standardization methods provide marginal comparison between old and new tests.
- Imputation techniques are helpful to future analysis of UDS data, especially for studies utilizing past and future data together.
- Because of several unique features of the imputation, we need to modify the multiple imputation procedures and derive the "correct" standard error estimates.

References

1. Crane PK *et al.* (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology*, **61**, 1018-1027.
2. Gibbons LE *et al.* (2011). Migrating from a legacy fixed-format measure to CAT administration: calibrating the PHQ-9 to the PROMIS depression measures. *Qual Life Res*, DOI 10.1007/s11136-011-9882-y.
3. Hui SL *et al.* (1997). Universal standardization of bone density measurements: a method with optimal properties for calibration among several instruments. *Journal of Bone Mineral and Research*, **12**, 1463-1470.
4. Robins JM and Wang N (2000). Inference for imputation models. *Biometrika*, **87**, 113-124.
5. Wang N and Robins JM (1998). Large sample inference in parametric multiple imputation. *Biometrika*, **85**, 935-948.