Overcoming missing data in a SMAR clinical trial of patients with schizophrenia

Susan M. Shortreed
McGill University and Group Health Research Institute
shortreed.s@ghc.org

Eric Laber
University of Michigan
laber@umich.edu

T. Scott Stroup
Columbia University
stroups@pi.cpmc.columbia.edu

Joelle Pineau
McGill University
jpineau@cs.mcgill.ca

Susan A. Murphy
University of Michigan
samurphy@umich.edu

Background

- Sequential, multiple assignment, randomized (SMAR) trials are an effective way for gathering data to learn personalized treatment strategies
- Practical challenges remain before learning treatment strategies from clinical data
- Missing data: drop out and missed exams can lead to missing information in clinical trials
- Imputation replaces missing data with predicted values to obtain complete data sets
- Having complete data is especially important when outcome is a function of the whole trial

The CATIE study

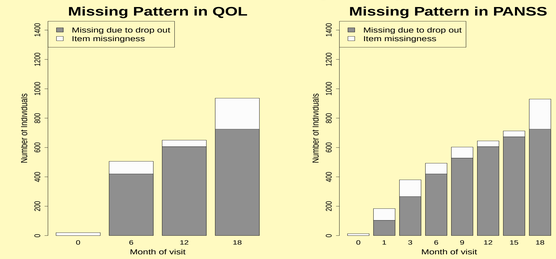
- Clinical Antipsychotic Trials of Intervention and Effectiveness (CATIE) study was an 18 month SMAR clinical trial of 1460 patients with schizophrenia
- CATIE had a broad entry criteria and a protocol designed to mimic real life
  - Participants chose *when* to switch treatment
- 2 major treatment phases, monthly follow-ups
- 2 symptom measures collected during CATIE:
  1. Positive and Negative Syndrome Scale (PANSS) total score
  2. Quality of Life (QOL) scale
- We distinguish between two types of variables:
Scheduled: collected at pre-specified months
End-of-phase: collected when a patient entered a new treatment phase

Missing information in CATIE

- Drop out in studies of patients with schizophrenia is often high
- 755 (52%) participants dropped out of CATIE
- Only 312 patients have PANSS and QOL scores measured at every scheduled visit
- 509 patients drop out while still on phase 1 trt
- Complete information would indicate which patients stayed on phase 1 treatment and who switched into the next treatment phase
  - Missing end-of-phase 1 variables for those who would have switched
- Of the 543 patients who entered phase 2 during the CATIE trial, 42 are missing their end-of-phase PANSS score and 78 are missing their QOL score

Overcoming missing data

MCAR, missing completely at random: probability of being observed is *independent* of the data
MAR, missing at random: probability of being observed can be associated with the observed data, but not with unmeasured information
NMAR, not missing at random: probability of being observed depends on unmeasured information
- CATIE: large amount of patient information was collected, reasonable to assume that given this rich source of data, we have MAR
- Most missing data due to drop out, thus we assume a monotone missing data pattern



Fully Conditional Specification (FCS) [**1,2**]
- Scales well in number of variables; allows flexibility of different models for each type of variable
- Denote set of variables by $v_0, v_1, v_2, ..., v_J$, ordered by time, with v_1 collected first and v_J last; v_o denotes variables with no missing information
- Model for complete data formed via conditional models for each variable v_j given $v_0, v_1, \ldots, v_{j-1}$

$$P(v_{j_{\text{miss}}} \mid v_0, v_1, v_2, ..., v_{j-1}, v_{j_{\text{obs}}}) = \prod_{\ell=1}^{j} P(v_{\ell_{\text{miss}}} | v_0, v_1, v_2, ..., v_{\ell-1}, v_{\ell_{\text{obs}}})$$

- Use a modified version of `R` package `mice` [3]

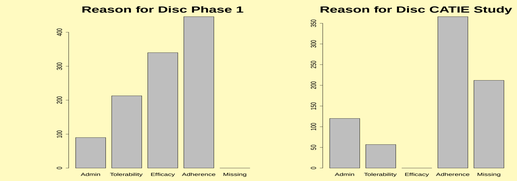
Bayesian Mixed Effects Method (BMEM) [**4,5**]
- Used to enforce smoothness over time in mean of PANSS
- PANSS score measured for person i at month m modeled by a mixed effects model

$$(\gamma_0 + g_i) + \underset{\sim}{\gamma}^{\mathsf{T}} \tilde{v}_{m,i} + \sum_{\xi=1}^{17} \eta_\xi (m - \xi)_+ + \epsilon_{m,i},$$

$\tilde{v}_{m,i}$ denotes the predictors collected both prior to, and at month m, and η_ξ's are the coefficients for the spline on month
- Use `R` package `pan` [6]

Building CATIE imputation models

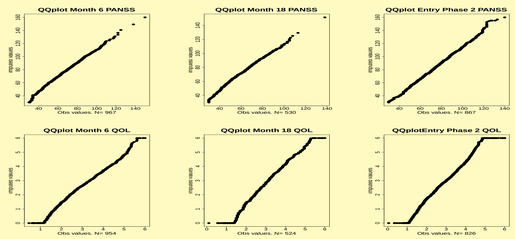
- **Goal**: learn imputation models for the missing data from the observed data
- Use as predictors: all *previously* measured scheduled variables
- Imputation of scheduled variables: use FCS for all but PANSS and BMEM for PANSS
- End-of-phase variables are specific to SMAR trials and pose a specific challenge in the imputation procedure
- Reason for discontinuation recorded as one of: lack of efficacy, lack of tolerability, adherence/compliance, or administrative reasons [7]
- Nest all time-varying predictors of end-of-phase variables within reason for discontinuation



Assumptions for end-of-phase variable models:
1. Had a participant who dropped out of CATIE on phase 1 trt remained in the study, they would have decided to switch off of their phase 1 trt.
   - Consider dropping out of the study switching into next treatment phase in an imputed data set
   - Use recorded reason for dropping out of the study as imputed reason for switching treatment phases and use all variables collected at study drop out as corresponding end-of-phase variables
2. We can pool over a set of months to learn end-of-phase imputation models
   - A small number of individuals choose to switch treatment each month
   - Increases stability in the estimates

Imputation Diagnostics

- Cannot test if data truly are MAR
- Compare imputations with obs. values [8,9]



Discussion

- Treatment of schizophrenia is notoriously difficult and requires personalized adaption of treatment due to lack of efficacy of treatment, poor adherence and intolerable side effects
- Sequential, multiple assignment, randomized trials implemented in a practical clinical trial setting provide excellent opportunities for learning personalized treatment strategies
- Possibilities for informing clinical practice are great, so are the challenges
- Imputation methods can be used to overcome missing data, a difficulty associated with almost all clinical trials
- If QOL was a variable of interest, QOL imputation models would need additional work
- Sensitivity analysis can evaluate the impact of any violations to the MAR assumption on a particular analysis [10]
- Using a rich set of predictors and prior knowledge to pick intelligent models helps to ensure validity of MAR assumption

References

1. S van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *SMMR*, 1(3):219-242, 2007.

2. S van Buuren, JPL Brand, CGM Grootheuis-Oudshoorn, and DB Rubin. Fully conditional specification in multivariate imputation. *J Stat Comp and Sim*, 76(12):1049-1064, 2006.

3. S van Buuren and K Groothuis-Oudshoorn. MICE Multivariate imputation by chained equations in R. *J Stat Software*, forthcoming.

4. P Diggle, P Heagerty, K-Y Liang, and S Zeger. *Analysis of longitudinal data*. Oxford Univ Press, 2002

5. JL Schafer. *Analysis of incomplete multivariate data*. C & H, 1997.

6. JL Schafer. *Multiple imputation for multivariate panel of clustered data*, 2009. R package version 0.2-6.

7. TS Stroup, et al. The National Institute of Mental Health clinical antipsychotic trials of intervention effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophrenia Bull*, 29(1):15-31, 2003.

8. K Abayomi, A Gelman and M Levy. Diagnostics for multivariate imputations. *JRSS, Series C*, 2008.

9. EA Stuart, M Azur, C Frangakis, and P leaf. Multiple imputation with large data sets: A case study of the children's mental health initiative. *AJE*, 169(9):1133-1139, 2009.

10. JR Carpenter, MG Kenward, IR White. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *SMMR*, 16(3):259-75, 2007.