Overcoming missing data in a sequential, multiple assignment, randomized clinical trial of patients with schizophrenia

Susan M. Shortreed[1,2] Eric Laber[3] T. Scott Stroup[4] Joelle Pineau[1] Susan A. Murphy[3]

[1]McGill University [2]Group Health Research Institute [3]University of Michigan [4]Columbia University
[1]{susan.shortreed,jpineau}@cs.mcgill.ca [2]shortreed.s@ghc.org [3]{laber, samurphy}@umich.edu
[4]stroups@pi.cpmc.columbia.edu

1 Background

- Sequential, multiple assignment, randomized trials are an effective way for gathering data to learn personalized treatment strategies [1]
- Many analytic techniques have been developed to learn the best individualized treatment strategies from data, including iterative minimization of regrets [2], G-estimation [3], reinforcement learning methods [4, 5, 6], regret-regression [7] and more parametric approaches [8, 9, 10].
- Many practical challenges arise when applying these methods to data collected from clinical trials [11].
- 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

2 The CATIE study: a sequentially randomized trial

- The Clinical Antipsychotic Trials of Intervention and Effectiveness (CATIE) study was an 18 month sequential, multiple assignment, randomized trial of 1460 patients with schizophrenia.
- CATIE was a practical clinical trial, thus had a broad entry criteria and a protocol designed to mimic real life.
  Specifically, participants choose *w*hen to switch treatment
- Two major treatment phases, with monthly follow-ups from baseline
- Two measures of symptoms collected during CATIE
  1. Positive and Negative Syndrome Scale (PANSS) total score
  2. Quality of Life (QOL) scale
- We distinguish between two different types of variables in CATIE:
  Scheduled collected on everyone at pre-specified visits
  end-of-phase collected when a patient entered a new treatment phase
- Table 1 contains a list of variables collected during CATIE

3 Missing information in CATIE

- Dropout 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 treatment
  - Complete information would indicate which patients stayed on phase 1 treatment and who switched into the next treatment phase
  - Missing end-of-phase variables for those who would have switched
- Of the 543 patients who entered phase 2 during CATIE, 42 are missing PANSS and 78 are missing QOLs

Table 1: A list of the variables collected during CATIE, and the times at which each variable was scheduled to be collected. The type of the variable follows in parenthesis.

Variables with no missing information.
Time independent variables:
Age (continuous), Sex (binary), Race (categorical), Tardive dyskinesia status at baseline (binary), Marital status (binary), Patient education (categorical), Hospitalization history in 3 months prior to CATIE (binary), Clinical setting at which patient received CATIE treatment (categorical), Treatment prior to CATIE enrollment (categorical), Phase 1 treatment assignment (categorical), Time in study on phase 1 treatment assignment (continuous).
Variables with missing information.
Time independent variables:
Employment status (categorical), Years since first prescribed anti-psychotic medication at baseline (continuous), Neurocognitive composite score at baseline (continuous), Phase 2 treatment assignment (categorical), Phase 2 randomization arm (binary), Reason for discontinuing phase 1 and 2 (categorical), Reason for discontinuing the CATIE study early (categorical), Total time spent in the CATIE study (continuous).
Variables collected at months 1-18 and at end-of-phase:
Treatment adherence, the proportion of capsules taken since last visit (continuous)
Variables collected at months 0, 1, 3, 6, 9, 12, 15, 18 and at end-of-phase:
Body mass index (continuous), Clinical drug use scale (ordinal), Clinical alcohol use scale (ordinal), Clinical Global Impressions Severity of illness score (ordinal), Positive and Negative Syndrome Scale total score (continuous), Calgary Depression total Score (continuous), Simpson-Angus EP mean scale (continuous), Barnes Akathisia scale (continuous), Total movement severity score (continuous)
Variables collected at months 0, 6, 12, 18 and at end-of-phase:
Quality of Life total score (continuous), SF-12 Mental health summary (continuous), SF-12 Physical health summary (continuous), Illicit drug use (binary)

4 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 observed data, but not with unmeasured information

NMAR not missing at random: probability of being observed depends on unmeasured information

- CATIE trial: a large amount of patient information was collected, including symptoms, side effects and adherence. It is a reasonable assumption that, given this rich source of data, the missing information in CATIE is MAR.
- Most missing data due to drop out, thus we a assume monotone missing data pattern

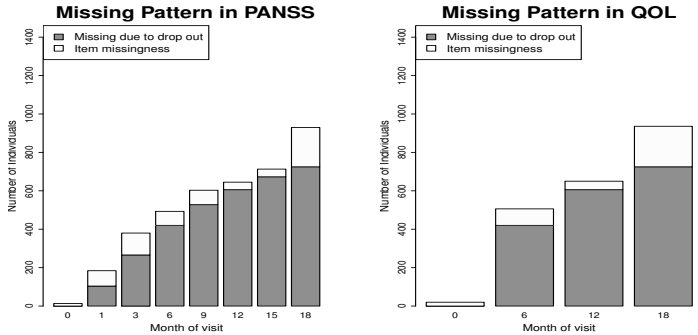

Figure 1: Barplot of missing data in the scheduled PANSS and QOL scores collected during the CATIE study. The total height of the bar shows the absolute number of people who have missing scores at the specified month. The dark grey area represents individuals who have missing scores due to earlier drop out and the unshaded area is the amount of item missingness. The missing data pattern for other scheduled variables collected during the CATIE study is similar to the pattern shown here.

Fully Conditional Specification (FCS) [12, 13]

- 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 $v_0, v_1, \ldots, v_{j-1}$

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

- Used a modified version of `R` package `mice`[1] [14]

Bayesian Mixed Effects Method (BMEM) [15, 16]

- Used to enforce smoothness over time in mean of PANSS
- PANSS score measured for individual i at month m modeled by a mixed effect model

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

- $\tilde{v}_{\mathsf{m},i}$ denotes the predictors used for PANSS at month m, and the η_ξ's are the coefficients for the spline constrained so that Eq. (1) is continuous in m
- Used `R` package `pan` [17]

5 Building the 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. This comes from the assumed monotone missing data pattern.
- End-of-phase variables are specific to sequential multiple assignment randomized trials and pose a specific challenge in the imputation procedure
- Reason for discontinuation recorded as one of: clinical determination of inadequate therapeutic effect (lack of efficacy), unacceptable side effects (lack of tolerability), patient inability or refusal to take the assigned antipsychotic (adherence/compliance), or administrative reasons [18]
- Nest all time-varying predictors of end-of-phase variables within reason for discontinuation

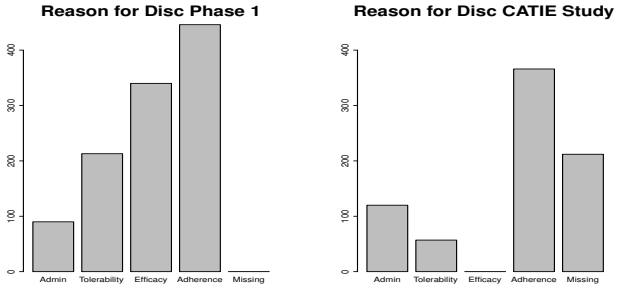

Figure 2: Reason given for discontinuing phase 1 treatment and for discontinuing the CATIE study early.

[1] We modified the `mice` package to correctly allow for complex imputation models requiring interaction terms and to incorporate definitional bounds in the imputations process.

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

Algorithm 1 Algorithm for imputing missing data in CATIE.

Require: Incomplete CATIE data set with all variables listed in Table 1.

Estimate model and use FCS to impute baseline variables with missing data using as predictors all variables in table 1 with no missing information.

for each set of months {1}, {2,3},{4,5,6},{7,8,9}, {10,11,12},{13,14,15},{16,17} **do**

Estimate model using FCS and impute from conditional imputation models for each end-of-phase variable. We estimate one imputation model pooled over the current set of months to increase the stability of the estimates.

1. Estimate model and impute reason for discontinuing treatment.
2. Estimate model and impute end-of-phase variables (except PANSS) for those participants who switched treatments during one of these months. The model for end-of-phase variables includes as predictors the most recently collected scheduled variables and month nested within reason for discontinuation and all baseline variables averaged over all reasons for discontinuation.
3. Impute current treatment for all individuals who switched treatment using the treatment randomization probabilities specified in the CATIE protocol.

Estimate model using FCS and impute scheduled variables (except PANSS) for current month set. Use all baseline variables, previously measured scheduled variables, current phase, and treatment as predictors.

Estimate model using a Bayesian mixed effects model and impute both scheduled and end-of-phase PANSS scores for the current set of months. The model includes all time-varying predictors measured at the same month as PANSS was measured nested within reason for discontinuation (with a separate category for people who did not switch treatment) and all baseline variables and time variables averaged over all reasons for discontinuation.

end for

Estimate model using FCS and impute scheduled variables (except PANSS) for month 18. Use all baseline and previously measured scheduled variables and current phase and treatment as predictors.

Using a Bayesian mixed effects model impute 18-month PANSS score using observed and imputed PANSS from months 1-17 and observed PANSS scores at month 18. Use as predictors all baseline variables and time-varying predictors measured at the month of PANSS observation, and current phase and treatment.

return A complete CATIE data set with all missing values replaced with draws from imputation models.

6 Imputation Diagnostics

- Multiple imputation methodology rests on the untestable assumption that missing data values can be generated from imputation models estimated from the observed data.
- Diagnostics often performed to compare the imputed values with the observed [19, 20]

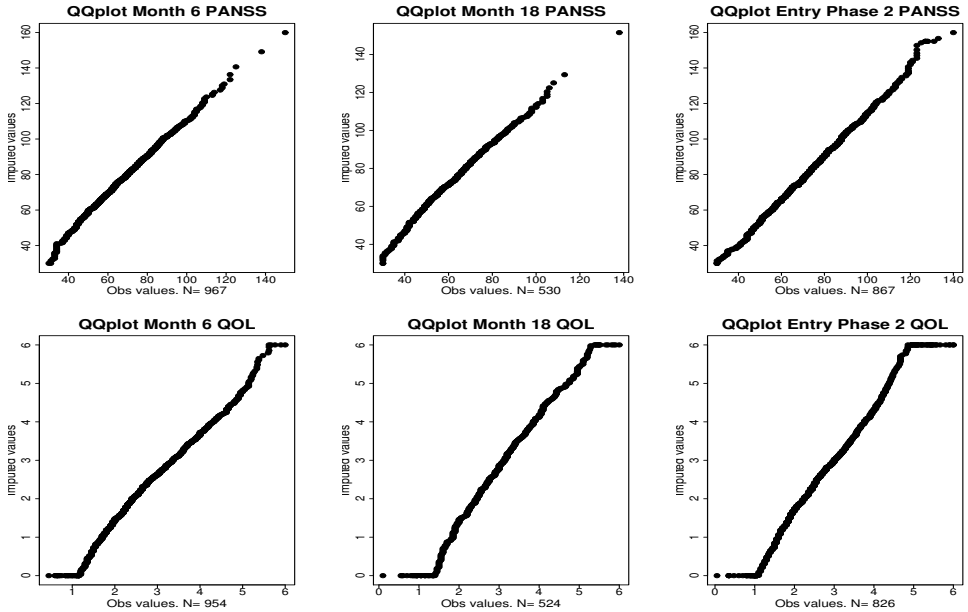

Figure 3: QQ-plots of imputed versus observed PANSS and QOL scores measured at months 6, 18 and end-of-phase 1. The missing data distribution contains the imputed values from twenty-five imputations (and none of the observed values).

7 Discussion

- Treatment of schizophrenia is notoriously difficult and requires personalized adaption of treatment due to lack of efficacy of treatment, poor adherence or intolerable side effects.
- Sequential, multiple assignment, randomized trials implemented in a practical clinical trial setting, such as CATIE, provide excellent opportunities to apply methods for learning personalized treatment strategies to a diverse population in a randomized setting.
- While the possibilities for informing treatment are great, so are the challenges.
- Imputation methods can be used to overcome missing data, a difficulty associated with almost all clinical trials
- In CATIE, models for imputing QOL score need additional work if QOL a variable of interest in an analysis
- Sensitivity analysis should be performed to evaluate the impact of any violations to the MAR assumption on a particular analysis [21].
- Using a rich set of predictors and prior knowledge to pick intelligent models, helps to ensure validity of MAR assumption

References

[1] Susan A Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24:1455–1481, 2005.

[2] Susan M Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, 65(2):331–366, 2003.

[3] J M Robins. Optimal structural nested models for optimal sequential decisions. In D Y Lin and P Heagerty, editors, *Proceedings of the Second Seattle Symposium on Biostatistics*, pages 189–326, New York, NY, USA, 2004. Springer.

[4] B Chakraborty, S A Murphy, and V Strecher. Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19:317–343, 2010.

[5] S A Murphy. A generalization error for Q-learning. *Journal of Machine Learning Research*, 6:1073–1097, 2005.

[6] J Pineau, M G Bellemare, A J Rush, A Ghizaru, and S A Murphy. Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence*, pages S52–S60, 2007.

[7] R Henderson, P Ansell, and D Alshibani. Regret-regression for optimal dynamic treatment regimes. *Biometrics*, 2010.

[8] E Arjas and O Saarela. Optimal dynamic regimes: Presenting a case for predictive inference. *The International Journal of Biostatistics*, 6, 2010.

[9] R Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1957.

[10] D P Bertsekas and J N Tsitsiklis. *Neuro-dynamic Programming*. Athena Scientific, Belmont, NH, USA, 1996.

[11] S M Shortreed, E Laber, D J Lizotte, J Pineau, and S A Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine Learning*, (to appear).

[12] Stef van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242, 2007.

[13] S van Buuren, J P L Brand, C G M Groothuis-Oudshoorn, and D B Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064, 2006.

[14] Stef van Buuren and Karin Groothuis-Oudshoorn. MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, forthcoming*, 0:00–00, 2010.

[15] Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002.

[16] J L Schafer and R M Yucel. Computational strategies for multivariate linear mixed models with missing values. *Journal of Computational and Graphical Statistics*, 11:421–442, 2002.

[17] Joseph L Schafer. *Multiple imputation for multivariate panel or clustered data*, 2009. R package version 0.2-6.

[18] T Scott Stroup, Joseph P McEvoy, Marvin S Swartz, Matthew J Byerly, Ira D Glick, Jose M Canive, Mark McGee, George M Simpson, Michael D Stevens, and Jeffrey A Lieberman. The National Institute of Mental Health clinical antipsychotic trials of intervention effectiveness (CATIE) project: schizophrenia trial design and protocol deveplopment. *Schizophrenia Bulletin*, 29(1):15–31, 2003.

[19] Kobi Abayomi, Andrew Gelman, and Marc Levy. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society, Series C*, pages 273–291, 2008.

[20] Elizabeth A Stuart, Melissa Azur, Constantine Frangakis, and Philip Leaf. Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology*, 169(9):1133–1139, 2009.

[21] J R Carpenter, M G Kenward, and I R White. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medidcal Research*, 16(3):259–75, 2007.