Multiple Imputation strategies for evaluating long term efficacy based on categorical outcomes in longitudinal clinical trials

Ilya Lipkovich, QuintilesIMS
Bohdana Ratitch, QuintilesIMS

24th Annual Biopharmaceutical Applied Statistics Symposium
October 23, 2017

Outline

- Motivating example
  - Case study in assessing long term safety and efficacy of patients with a chronic immune-mediated disease
- Estimands
- Methods
  - Overview of MI
  - Options for imputing categorical outcomes derived from continuous scales
- Simulation study
- Summary

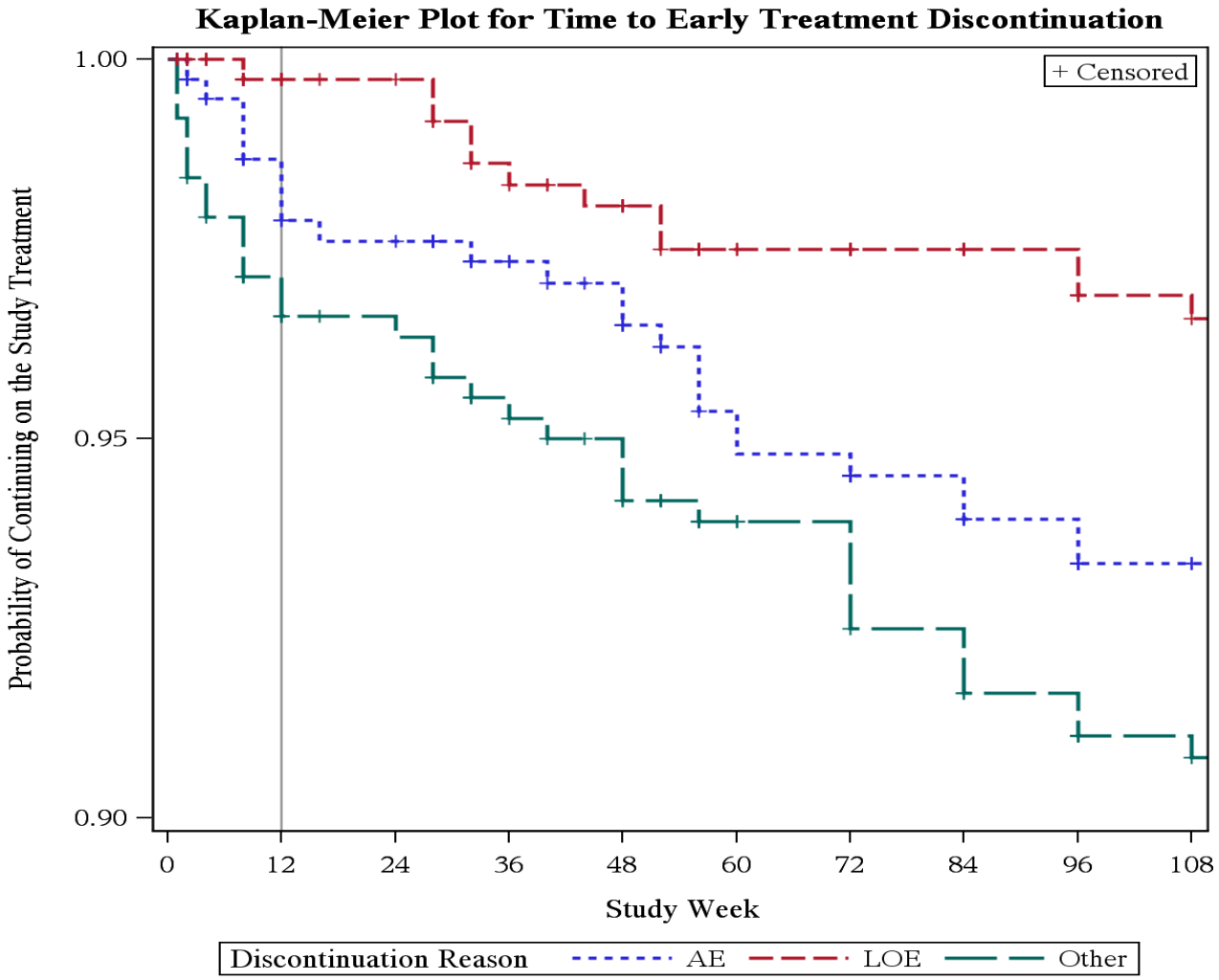
Motivating example

- A randomized, double-blind, multicenter, phase-3 clinical trial to evaluate the long-term efficacy and safety of ixekizumab for the treatment of moderate-to-severe plaque psoriasis
  - For the first 12 weeks: a parallel group design comparing 2 dosage regimens of 80 mg ixekizumab (Q2W or Q4W) versus placebo and active comparator
  - For assessment of long-term safety and efficacy,-80 mg ixekizumab Q4W is evaluated for up to a total of 5 years in patients who participate through the entire study

- The objective of analysis: evaluating long term efficacy of 80 mg ixekizumab Q2/Q4W in presence of missing data up to week 108

- The outcome variable was binary based on an underlying rating scale
  - The rating scale (Y): “static Physician Global Assessment” (sPGA): Psoriasis lesions assessed as clear (0), minimal (1), mild (2), moderate (3), severe (4), or very severe (5)
  - The binary outcome (Z) is responder status defined as sPGA =0 or sPGA =1

Motivating example: Dropout patterns



Motivating example: Estimands

- Before analyzing the data it is a good idea to set the estimand (s) of interest = estimation targets
  - A side note: surprisingly, not many protocols/SAP's even mention "estimands": typically state study objectives at a high level, then proceed to describing analytic procedures (e.g., we will do MMRM to evaluate mean changes from baseline in variable X at week 108)

- The ambiguity about the estimation target arises mostly because of post-randomization events: dropouts, rescue arm, other flexible treatment options

Motivating example: Estimands (cont.)

- This ambiguity is often “resolved” by making an implicit assumption that missing data/switching mechanism is ignorable and estimating an estimand involving “counterfactual outcomes”
  - e.g. MMRM implies that patients who dropped out would have had similar outcomes to patients with similar baseline/outcome history prior to dropouts who continued (MAR)
  - this analysis is often combined with sensitivity analyses assuming patients who dropped/switched would do worse than predicted by MAR, if remained on treatment (MNAR)
- On a deeper level, and more directly, the ambiguity should be handled by making explicit selection among different estimands
  - Are we interested to estimate what would have happened, had patents who dropped out remained on treatment (**counterfactuals**)?
  - Or we want to estimate **“de-facto” outcomes** that they may have after dropping out?
- The analyses can be further detailed using different assumptions about potential outcomes (*de facto* or counterfactuals) after discontinuing, perhaps depending on reasons for discontinuation: AE, lack of efficacy, etc.
- As we will see, imputation techniques come in handy due to their flexibility
  - Different assumptions about potential outcomes can be incorporated in the imputation step

Motivating example: Estimands (cont.)

Estimand	Outcome	Time point/ period	Treatment	Conditions under which treatment effect is evaluated	Population
E1	sPGA(0,1)	at visit K	Due to study treatment	if taken as directed	In all subjects
E2	sPGA(0,1)	at visit K	Due to study treatment	if taken as directed until contraindicated	In all subjects
E4	sPGA(0,1)	at visit K	Due to study treatment	if taken as directed	In those who actually took

- E1 can be estimated by imputation of counterfactual outcome
- E2 can be estimated by imputing non-response for those who discontinued (or using some other assumptions for non-completers)
- E4 can be estimated by observed outcome ("completers") at visit K

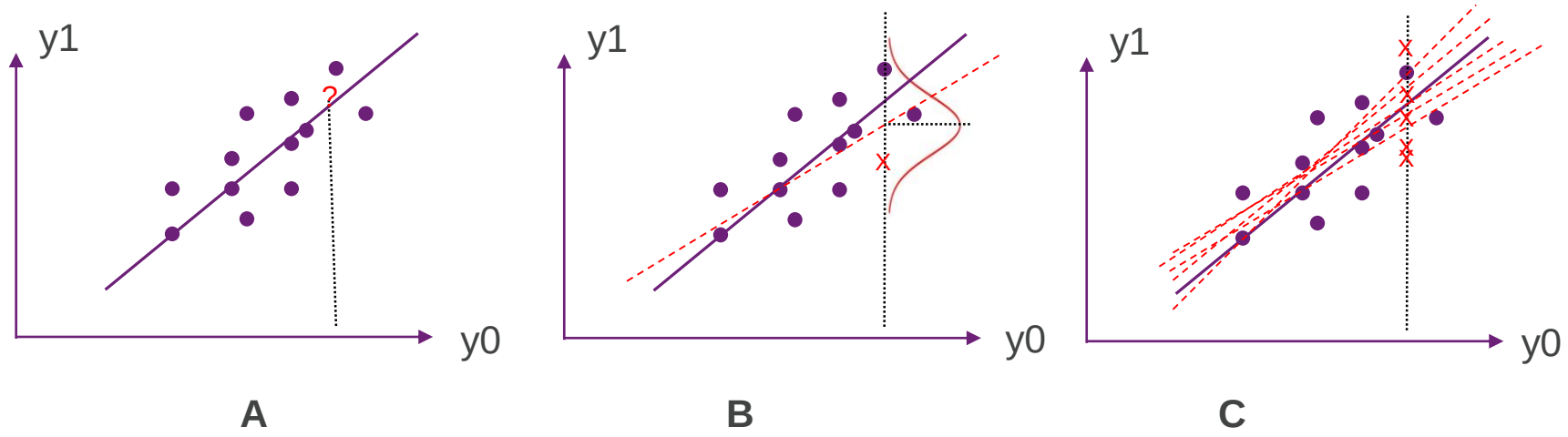
Overview of Multiple Imputation (MI)

- Imputation = fill the holes in the data
- Advantage: allows for standard complete-data analysis methods after imputation
- Single imputation followed by standard analysis underestimates the variability of estimates (SE)
- Multiple Imputation (MI) (Rubin, 1978)
  - Original goal was to impute m completed data sets for public usage in the context of public surveys; assuming imputer is an expert in statistical methods and would be able to access larger set of variables than available for data analysts
  - In clinical trials it may also be desirable to use a simpler model for primary analysis applied to data set(s) completed using a more complex imputation model

- MI Steps
  - Imputation – Generate m plausible values for each missing value with imputation model. The *m* completed datasets reflect uncertainty about the unobserved values.
  - Analysis – Analyze each data set using the desired standard complete-data methods with analysis model
  - Combined Inference – Use standard formulae to combine the *m* estimates found above.

A simple illustration: the ABC of MI

- y_0 is baseline value with no missing and y_1 is post-baseline with some values missing.
  - Fit regression y_1 vs. y_0 based on observed data – Figure A
  - Step 1: Take account of variability of regression line by sampling from posterior distribution of the model parameters (intercept and slope) – red regression line in Figure B
  - Step 2: To the sampled regression line add random error from a normal distribution centered at the model-predicted value for y_1 and use as the imputed value for the missing value in a fresh imputed data set (shown as X in Figure B)
  - Repeat above steps 1-2 *m* times to create *m* imputed data sets (Figure C)



Combined Inference in MI

$$\hat{\theta}_i = \text{standard analysis estimators (e.g. ML) for } i = 1,\ldots,m$$

$$U_i = \text{estimated variance of } \hat{\theta}_i$$

$$\bar{\theta}_{MI} = m^{-1}\sum_{i=1}^{m}\hat{\theta}_i$$

$$B = (m-1)^{-1}\sum_{i=1}^{m}(\hat{\theta}_i - \bar{\theta}_{MI})^2$$ ← Between-imputation variance

$$\bar{U} = m^{-1}\sum_{i=1}^{m}\hat{U}_i$$ ← Within-imputation variance

$$V = \bar{U} + (1+m^{-1})B$$ ← Total variance

$$(1-\alpha)100\%\, CI : \bar{\theta}_{MI} \pm t_{\nu,1-\alpha/2}\sqrt{V},\ \nu = (m-1)\left(1+\frac{\bar{U}}{(1+m^{-1})B}\right)^2$$

Notice using a Student *t*, rather than normal distribution, in combined inference.

This is needed to account for extra variability in *V* estimator which for finite $m < \infty$ is an inconsistent estimator of the true variance $Var(\bar{\theta}_{MI})$; as $n \to \infty$, $nV_{n,m}$ does not converge to a constant, but has a limiting (Chi-square) distribution.

Combined Inference in MI (cont.)

- The relative increase in variance due to missingness is given by

$$r = \frac{(1+m^{-1})B}{\bar{U}}$$

- The fraction of missing information about θ is given by

$$\hat{\lambda} = \frac{r + 2/(\nu+3)}{r+1}$$

- Accounting for "small sample" variability by further "deflating" degrees of freedom for the t-distribution
  - You can specify the complete data degrees of freedom, ν_0 with option **EDF=** in SAS proc MIANALYZE
  - The adjusted degrees of freedom, ν^* (Barnard and Rubin, 1999):

$$\nu^* = \left(\frac{1}{\nu} + \frac{1}{\hat{\nu}_{obs}}\right)^{-1},$$

$$\hat{\nu}_{obs} = (1-\gamma)\nu_0 \frac{(\nu_0+1)}{(\nu_0+3)},$$

$$\gamma = \frac{(1+m^{-1})B}{V}$$

MI for normal longitudinal data

- Imputation. Two routs
  - For arbitrary missingness patterns use Data Augmentation algorithm based on Markov Chain Monte Carlo (MCMC), draw samples (jointly across time) from $Y_{mis}|Y_{obs}$
  - For monotone missingness pattern, use sequential Bayesian regression: draw a sample from predictive distribution of $Y_{mis}|Y_{obs}$ estimated from regressing the observed Y_2 on Y_1 (Y_1 are outcomes before a subject's dropout, Y_2 - after)
- Both with MCMC and sequential Bayesian regression, we impute missing data for all time points from the same target predictive posterior distribution (approximately with MCMC, exactly with regression)

- Analysis of *m* sets of complete data using "standard methods"

- Combined Inference.
  - Rubin's rules is used to combine the *m* estimated treatment differences and associated standard errors into final estimates and associated CI

MI for normal data with arbitrary missingness. Data Augmentation

- Data Augmentation: parameters + missing values are unobserved quantities
- Iteratively sampling $Y_{mis}^{(i+1)}$ from distribution $p(Y_{mis}|\theta^{(i)}, Y_{obs})$ and $\theta^{(i+1)}$ from full conditional $p(\theta|Y_{obs,}Y_{mis}^{(i+1)})$
- As a result, this sampling process converges - in distribution - to posterior distribution $p(\theta|Y_{obs})$ and predictive distribution of $p(Y_{mis}|Y_{obs})$, effectively integrating out parameters

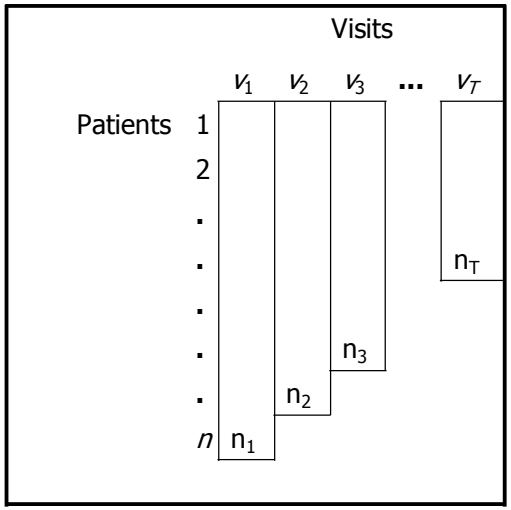
```
proc MI data = mydata out = miout  nimpute =50 seed =123;
   mcmc chain=single nbiter=200 niter=100 initial=em (itprint maxiter=200);
   var x y1 y2 y3 y4 … yT;
 run;
```

MI for normal longitudinal data with monotone missingness. Bayesian regression

- Idea: Use the fact that multivariate normal factors into a product of conditional distributions that are also normal
  - $p(Y_1, Y_2, Y_3, \dots, Y_T) = p(Y_T|Y_2, Y_3, \dots, Y_{T-1})p(Y_{T-1}|Y_2, Y_3, \dots, Y_{T-2})p(Y_2|Y_1)p(Y_1)$
- SAS proc MI with monotone option can be used to fill-in missing values (Y). Impute data using Bayesian regression (Schafer, 1997)

```
proc MI data = mydata out = miout nimpute =50 seed =123;
   monotone method=reg;
   var y1 y2 y3 y4 .. yT;
 run;
```

$$y_{it}^{*} = \beta_0^{*} + \beta_1^{*} y_{i1} + \beta_2^{*} y_{i2} \dots + \beta_{t-1}^{*} y_{i,t-1} + z_i \sigma_t^{*}$$



Bayesian Regression Method: Stage one

- Fit a linear Least Squares regression Y_t on $Y_1, \ldots, Y_{t-1}$ model to the n_t complete cases, obtain $\hat{\beta}_t, \hat{\sigma}_t^2$
- Draw values of $\boldsymbol{\beta}^*$ and σ^{*2} from their posterior distributions under Jeffreys prior

$$\sigma_t^{*2} \Big| Y \sim \text{Inv}\chi^2(n_t - t, \hat{\sigma}_t^2)$$

$$\boldsymbol{\beta}_t^* \Big| Y, \sigma_t^2 \sim N_p(\hat{\boldsymbol{\beta}}, \mathbf{V}_t \sigma_t^{*2})$$

$$\mathbf{V}_t = (\mathbf{X}'\mathbf{X})^{-1}$$

with $\mathbf{X}$ combining intercept and $Y_1, .., Y_{t-1}$

Bayesian Regression Method. Stage two

- Using available measurements for the dropouts from the previous visits ($\boldsymbol{X}$) the missing values are imputed by random draws from the conditional predictive distribution
- This approach uses the predictions made from the linear regression equations as imputations for each missing value

$$y_t^* \sim N(\mathbf{X}\boldsymbol{\beta}_t^*, \sigma_t^{*2}).$$

Combining MCMC and regression for monotone patterns

- Use MCMC for arbitrary pattern to complete data to monotone and produce m such data sets
- Apply Bayesian regression method for monotone data to each completed set
- Example of SAS code:

```
proc MI data=mydata out=miout nimpute=50 seed =123;
   mcmc impute = monotone;
   var x y1 y2 y3 y4 .. yT;
 run;
```

```
 proc MI data = miout out = miout nimpute =1 seed =123;
   by _Imputation_;
   monotone method=reg;
   var x y1 y2 y3 y4 .. yT;
  run;
```

Sequential imputation for binary outcomes using logistic regression for monotone data

- Imputation of binary repeated measures
  - Consider binary outcome (Z=0,1) the assessment of patient clinical response measured at post baseline visits 2,3,5 (variables z2 z3 z5)
  - We can use logistic regression with monotone option (add baseline covariates (X) with no missing values), as shown below:

```
proc mi data=datain out=datain_mi seed=4566765 nimpute=50;
  var x z2 z3 z5;
  class z2 z3 z5;
  monotone logistic;
run;
```

Sequential imputation from conditional distributions for mixed type data

- What if there are no responders yet at visit 2?
- We can use some other outcome (e.g. continuous variable Y) at visit 2 in the sequential imputation process:

```
proc mi data=datain out=datain_mi seed=4566765 nimpute=50;
   var  x y2 z3 z5;
   class z3 z5;
   monotone reg (y2=x);  /*impute y2 using baseline covariate x*/
   monotone logistic (z3=x y2); /*impute z3 using x and y2*/
   monotone logistic (z5=x y2 z3); /*impute z5 using x,y2,z3*/
run;
```

Sequential imputation from conditional distributions for mixed type data

- Another situation may be that Z is binary outcome derived from continuous Y and we want to impute z5 using x, y2, y3

```
proc mi data=datain out=datain_mi seed=4566765 nimpute=50;
   var  x y2 y3 z5;
   class z5;
   monotone reg (y2=x);  /*impute y2 using baseline covariate x*/
   monotone reg (y3=x y2); /*impute y3 using x and y2*/
   monotone logistic (z5=x y2 y3); /*impute z5 using x,y2,y3*/
run;
```

The case of arbitrary missingness. Fully conditional specification

- MI is done by repeatedly sampling from full conditional distributions (rather than sampling from a joint multivariate distribution)
  - Independently developed under different names by several authors, e.g. “Fully conditional specification” by van Buuren, 2007
  - It is possible to specify models for which no known joint distribution exits.The conditionally specified model may be incompatible in the sense that the joint distribution does not exist
- Software implementation:
  - R package MICE (Multivariate Imputation by Chained Equations), van Buuren Groothuis-Oudshoorn:
    - http://cran.r-project.org/web/packages/mice/index.html
  - IVEware (Imputation and Variance Estimation Software, version 2.0), Raghunathan et al., 2001., can be called from SAS
    - http://www.isr.umich.edu/src/smp/ive/
  - SAS v 9.3 and higher. statement FCS in PROC MI

The case of arbitrary missingness. Fully conditional specification (Cont.)

- The general algorithm
  - Associate with each outcome variable $Y_j, j = 1, \ldots, p$ a parameter vector θ_j that governs conditional distribution of Y_j given the rest of variables combined in matrix $\boldsymbol{Y}_{-j}$.
  - At initial stage we fill-in missing values $Y_j^{*(0)}$ (e.g. with mean imputation or a sequence of simple univariate models) and form $Y_j^{(0)} = \left(Y_j^{obs}, Y_j^{*(0)}\right)$
  - After that we iteratively update the imputed values and parameters θ_j by sampling from conditional distributions (the order of variables should be pre-specified for both initial fill-in and subsequent imputation updates):

$\theta_j^{(i+1)} \sim p(\theta_j^{(i+1)} | Y_j^{obs}, \mathbf{Y}_{-j}^{(i)})$, where $\mathbf{Y}_{-j}^{(i)} = (\mathbf{Y}_{-j}^{obs}, \mathbf{Y}_{-j}^{*(i)})$

$Y_j^{*(i+1)} \sim p(Y_j | \theta_j^{(i+1)}, Y_j^{obs}, \mathbf{Y}_{-j}^{(i)})$, j=1,..,p

```
proc mi;
   class z5;
   fcs reg(y2= x) reg(y3= x y2) logistic(z5= x y2 y3);
   var x y2 y3 z5; /*order specified by var statement*/
run;
```

Predictive Mean Matching (PMM)

- PMM is a “hot-deck” type imputation for monotone missing data (Schenker and Taylor, 1996), that is this procedure guarantees that the imputed values for a rating scale will be “in range.”
- The imputed value for a given observation with missing outcome y_{it} is randomly drawn from a set of k_0 observed values whose predicted values are closest to the predicted value for the missing observation
- Predicted values are obtained using a Bayesian regression model for Y_t, given previous outcomes $Y_{t-1}, \ldots, Y_0$
- The default value of $k_0 = 5$
- In SAS use proc MI (REGPMM option for MONOTONE statement)

When MI is most useful?

- When direct multivariate likelihood analysis is not available, and multivariate distribution may be hard to specify, e.g.
  - Repeated binary outcomes
  - Outcomes of mixed types (continuous and categorical), MI can be done by repeatedly sampling from full conditional distributions
- Imputation of underlying continuous data when primary analysis model is for derived outcome
  - e.g. analysis is for clinical response defined as Z=(Y1 > c1) and (Y2 > c2) at last scheduled evaluation
- Implementing “custom” imputation rules for patients with certain conditions
  - impute QOL score = 0 after dropout if caused by a stroke/death
  - Impute non-response for patients who discontinued because of AE or LOF
- Implementing “Inclusive” imputation strategy
  - MI can utilize auxiliary information in imputer’s model, which may not be available or undesirable to use in analyst’s model (Meng, 1994; Collins et al., 2001).

Simulation study. Data Structure

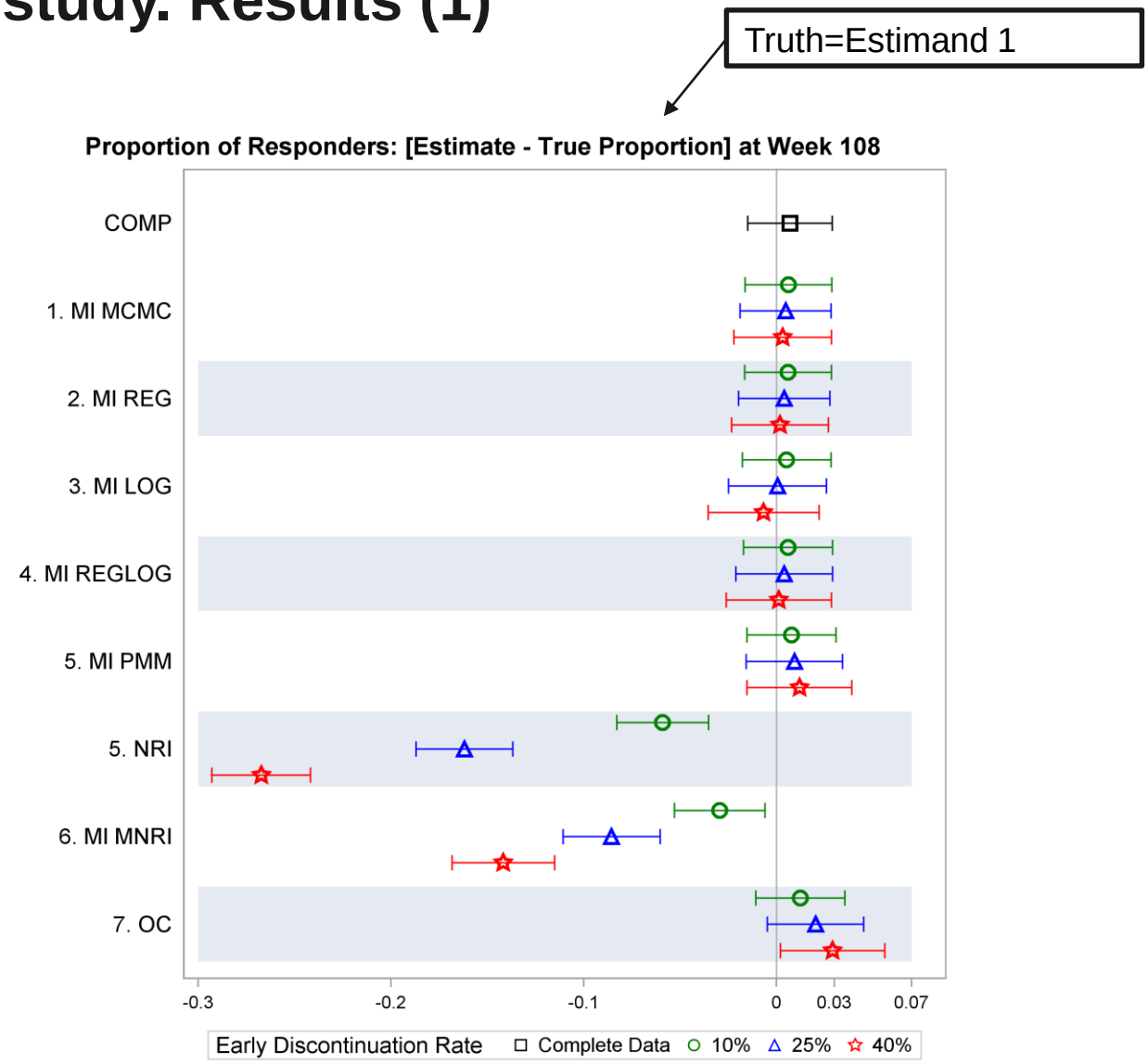
- Outcome data – repeated measures simulated as multivariate normal
  - $y_{s,k}^{comp}$=$\mu_k + \varepsilon_{s,k}$, where $\varepsilon_{s,k} \sim MN(0, \Sigma)$, *k*=1,..,*K* (*K*= 22 visits including baseline)
  - The means μ_k and covariance matrix Σ were chosen to mimic the distribution of sPGA scores in the real dataset
  - The parameters (means and covariance structure) were borrowed from the real data set, the resulting response rate for sPGA(0,1) was about 0.74 at week 108

- Missingness mechanism (MAR)
  - The simulation model was a multinomial logistic regression for a four-level dependent variable representing treatment continuation (*r*=0) and three reasons for treatment discontinuation: lack of efficacy (*r*=1), adverse event (*r*=2), and other reasons (*r*=3)
  - $\log\left(\frac{p_{s,k}^r}{p_{s,k}^0}\right) = \alpha_0^r + \alpha_1^r * (Y_k - Y_{k-1}) + \alpha_2^r * (Y_k + Y_{k-1}) + \alpha_2^r * k + interaction\ terms$
  - Coefficients of logistic model were calibrated so as to achieve a desired overall dropout rate (*d*): Scenarios with *d* = 10%, 25%, and 40% were generated

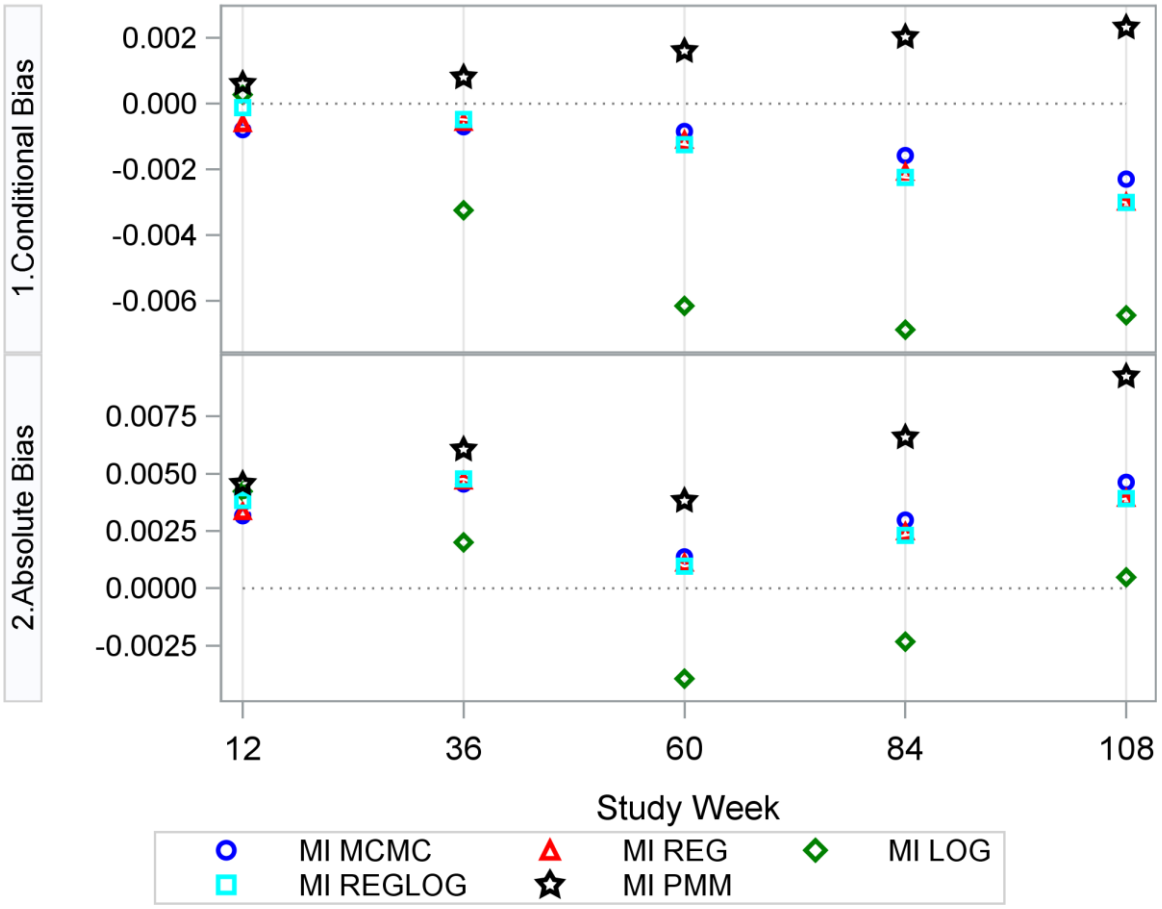
Simulation study. Methods

Label	Description	Estimand
COMP	estimate from complete datasets (no dropout benchmark)	E1
MI MCMC	multiple imputation of numeric sPGA scores using the MCMC method	E1
MI REG	multiple imputation of numeric sPGA scores using sequential regression at all visits	E1
MI LOG	multiple imputation of binary sPGA(0,1) responder status using sequential logistic regression at all visits	E1
MI REGLOG	multiple imputation of numeric sPGA scores using sequential regression at all but the last visits and logistic regression of the binary sPGA(0,1) responder status at the last visit	E1
MI PMM	multiple imputation of numeric sPGA scores using sequential predictive mean matching at all visits	E1
NRI	non-responder imputation for all early discontinuations	E2
MI mNRI	non-responder imputation for early discontinuations due to AEs and lack of efficacy and using MI PMM for subjects discontinued due to other reasons	E2
OC	estimate from observed data in study completers	E4

Simulation study. Results (1)



Simulation study. Results (2)

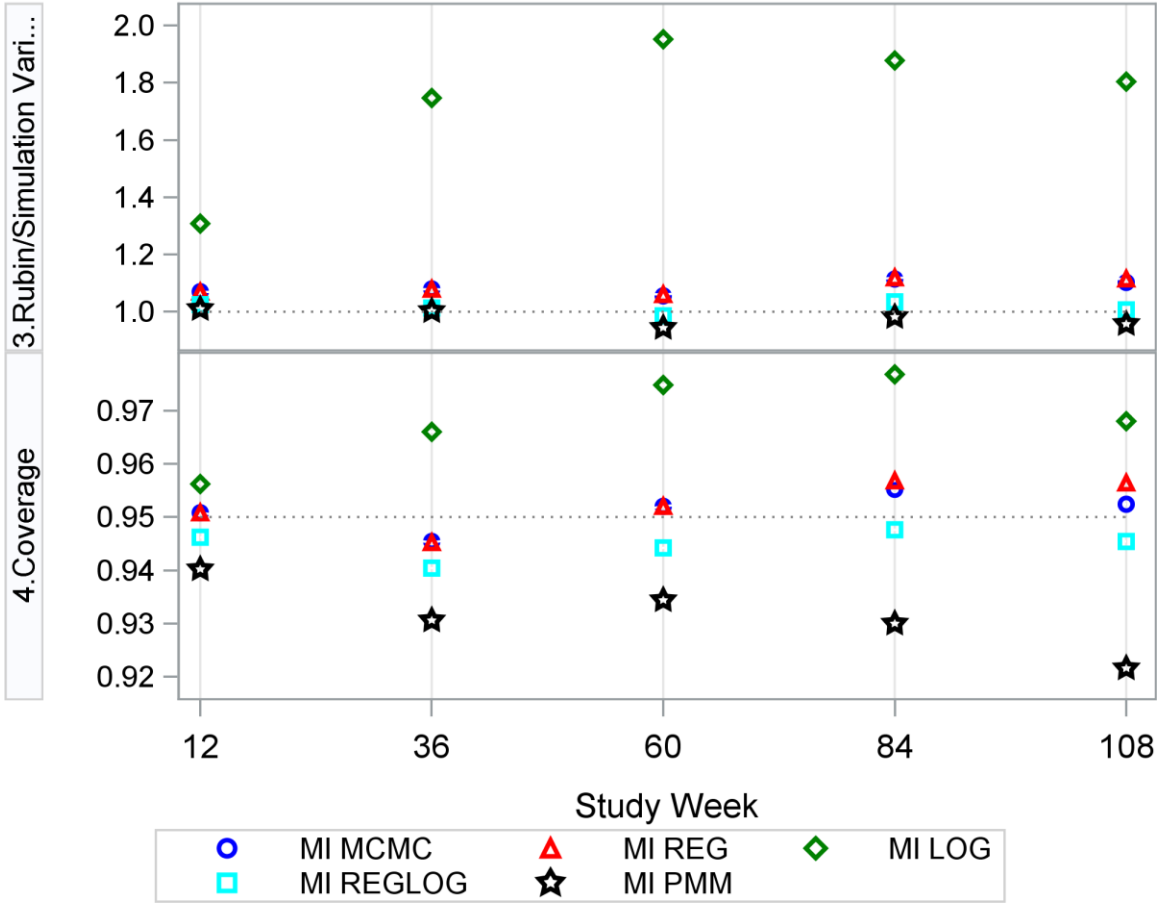


Absolute bias=Simulation Ave($\hat{\pi}_k^j - \pi_k$)

Conditional bias=Simulation Ave($\hat{\pi}_k^j - \hat{\pi}_k^{j,comp}$), $j = 1,..,5000, k = 1,..,22$

QuintilesIMS™

Simulation study. Results (3)



Summary and Discussion

- MI provides flexible framework for evaluating different estimands for categorical outcomes
- Imputation methods utilizing underlying continuous data have an advantage over those imputing categorical outcome directly
  - Lipkovich et al. (2005) reported simulations comparing similar imputation strategies with non-imputation based methods
- In the simulation study, normal MCMC and sequential regression MI (that did not respect the discrete nature of underlying ordinal outcome) had better performance than sequential logistic regression or predictive mean matching
  - see also Lipkovich et al. (2014) for review and simulations comparing imputation methods that "respect" ranges of clinical scale vs. using multivariate normal

- Specific results
  - Simulation study did not reveal drastic differences in bias, but found **dropout dependent** bias for MI PMM (positive) and MI LOG (negative)
  - Bias in MI PMM also drove its low coverage, making inference invalid
  - Despite bias, MI LOG had a higher coverage, as bias was offset by its **large** model-based (Rubin's) standard error => resulting in valid albeit conservative inference

References

- Ratitch B, Lipkovich I, Erickson JS, Zhang L, Mallinckrodt CH. (2017) Points to Consider for Analyzing Long Term Efficacy Outcomes in Clinical Trials, *Pharmaceutical Statistics*, under review
- Rubin D B (1978) Multiple imputations in sample surveys, in *Proc Survey Res Meth Sec Am Statis Assoc*, pp 20-34, Washington DC, American Statistical Association.
- Schenker N, Taylor JMG (1996). Partially parametric techniques for multiple imputation.*Computational Statistics and Data Analysis* 22:425–446.
- Schafer JL (1997) *Analysis of incomplete multivariate data*, Chapman and Hall, London
- Lipkovich I, Duan Y, Ahmed S. (2005) Multiple Imputation Compared with REPL and GEE for Analysis of Binary Repeated Measures in Clinical Studies. *Pharmaceutical Statistics.* 4:267-285.
- van Buuren, S. (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research.* 16, 219–242.
- Collins LM, Schafer JL, Kam C. (2001) A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological methods.* 6:330-351.
- Lipkovich I, Kadziola Z, Xu L, Sugiharad T, Mallinckrodt CH. (2014) Comparison of several multiple imputation strategies for repeated measures analysis of clinical scales: Truncate or not to? *Journal of Biopharmaceutical Statistics.* 24, 924-943.

Questions ?

Ilya.lipkovich@quintiles.com
Bohdana.Ratitch@quintiles.com

Thank You!