Sensible Approaches to Analyses of Incomplete Longitudinal Clinical Trial Data

Mallinckrodt CH[1]

1. Lilly Research Labs. Eli Lilly and Co. Indianapolis, IN. 46285.

Abstract

Recent research has fostered new guidance on the analyses of incomplete data. Common elements from recent guidance are distilled and means for putting the guidance into action are proposed. Whether or not follow-up data after discontinuation of the originally randomized medication and / or initiation of rescue medication contribute to the primary estimand depends on the context. In outcomes trials (intervention thought to influence disease process) follow-up data is often included in the primary estimand, whereas in symptomatic trials (intervention alters symptom severity but does not change underlying disease) follow-up data are often not included. Regardless of scenario, the confounding influence of rescue medications can render follow-up data of little use in understanding the causal effects of the randomized interventions. A sensible primary analysis can often be formulated in the missing at random (MAR) framework. Sensitivity analyses assessing robustness to departures from MAR are crucial. Plausible sensitivity analyses can be pre-specified using controlled imputation approaches to either implement a plausibly conservative analysis or to stress test the primary result, and used in combination with other model-based MNAR approaches such as selection, shared parameter, and pattern-mixture models. The example data set and analyses used in this paper are freely available for public use at www.missingdata.org.uk.

Key Words: Missing Data, Clinical Trials, Sensitivity Analyses

Introduction

Missing data is an ever present problem in clinical trials that can bias treatment group comparisons and inflate rates of false negative and false positive results. However, missing data has been an active area of investigation with many advances in statistical theory and in our ability to implement that theory. These research findings set the stage for new or updated guidance for the handling of missing data in clinical trials. For example, a pharmaceutical industry group published a consensus paper (Mallinckrodt et al, 2008) and an entire text book was devoted to the topic of missing data in clinical trials (Molenberghs and Kenward, 2007). New guidance was released by the EMEA (CHMP, 2010), an expert panel commissioned by FDA issued an extensive set of recommendations (NRC, 2010), and two senior leaders at FDA published their thoughts on the NRC recommendations (O'Neill and Temple, 2012). The intent of this paper is to distill common elements from recent recommendations, guidance documents, and texts to propose means and provide tools for implementing the guidance.

Estimands

An estimand is simply what is being estimated. Components of estimands for longitudinal trials may include the parameter (e.g., difference between treatments in mean change), time point or duration of exposure (e.g., at Week 8), outcome measure (e.g., diastolic blood pressure), population (e.g., in patients diagnosed with hypertension), and inclusion / exclusion of follow-up data after discontinuation of the originally assigned study medication and/or initiation of rescue medication.

Much of the debate on appropriate estimands, and by extension whether or not follow-up data are included in an analysis, centers on whether the focus is on efficacy or effectiveness. Efficacy may be viewed as the effects of the drug if taken as directed: for example, the benefit of the drug expected at the endpoint of the trial, assuming patients took the drug as directed. This has also been referred to as a per-protocol estimand. Effectiveness may be viewed as the effects of the drug as actually taken, and has also been referred to as an ITT estimand (Mallinckrodt et al, 2008).

Referring to estimands in the efficacy vs. effectiveness context ignores the fact that many safety parameters need to be analyzed. It does not make sense to test an efficacy estimand for a safety outcome. A more general terminology for hypotheses about efficacy and effectiveness is *de-jure* (if taken as directed, per protocol) and *de-facto* (as actually taken, ITT), respectively.
The NRC guidance (NRC, 2010) lists the following five estimands:

1. ***Difference in outcome improvement at the planned endpoint for all randomized participants.*** This estimand compares the mean outcomes for treatment vs. control regardless of what treatment participants actually received. Follow-up data (after withdrawal of initially randomized medication and/or initiation of rescue medication) are included in the analysis. Estimand 1 tests *de-facto* hypotheses regarding the effectiveness of treatment policies.

2. ***Difference in outcome improvement in tolerators.*** This estimand compares the mean outcomes for treatment vs. control in the subset of the population who initially tolerated the treatment. This randomized withdrawal design has also been used to evaluate long term or maintenance of acute efficacy. An open label run-in phase is used to identify patients that meet criteria to continue. Patients that continue are randomized (usually double-blind) to either continue on the investigational drug or switch to control.

3. ***Difference in outcome improvement if all patients adhered.*** This estimand addresses the expected change if all patients remained in the study. Estimand 3 addresses *de-jure* hypotheses about the causal effects of the initially randomized drug if taken as directed – an efficacy estimand.

4. ***Difference in areas under the outcome curve during adherence to treatment, and,***
5. ***Difference in outcome improvement during adherence to treatment.***
Estimands 4 and 5 assess *de-facto* hypotheses regarding the initially randomized drug. These estimands are based on all patients and simultaneously quantify treatment effects on the outcome measure and the duration of adherence. As such, there is no missing data due to patient dropout.

Each estimand has strengths and limitations. Estimand 1 tests hypotheses about treatment policies. However, the most relevant research questions are often about the causal effects of the investigational drugs, not treatment policies. This is also relevant for product labels where patients hope to learn what they may expect if they take the product as prescribed. In the intention-to-treat (ITT) framework where inference is drawn based on the originally assigned treatment, the inclusion of follow-up data when rescue medications are allowed can mask or exaggerate both the efficacy and safety effects of the initially assigned treatments, thereby invalidating causal inferences for the originally assigned medication (Mallinckrodt and Kenward, 2009).

O'Neill and Temple (2012) noted that including follow-up data in the primary estimand may be more useful in outcomes trials (where the presence / absence of a major health event is the endpoint and/or the intervention is intended to modify the disease process), whereas in symptomatic trials (symptom severity is the endpoint) the complications of follow-up data are usually avoided in the primary estimand.

Estimand 2 focuses on a patient subset and would not be applicable when inference to all patients was desired. Relevance of this estimand is further complicated because in most situations it is not known who will tolerate, and thus all patients must be exposed to the safety risks of the drug, whereas efficacy inferences apply only to the tolerators.

Although knowing what happens if a drug is taken as directed, as is done for estimand 3, is important, it is also hypothetical because in actual clinical settings there will always be some patients who do not adhere (NRC, 2010).

For estimands 4 and 5, assessing drug effectiveness during adherence ignores that in many instances benefit disappears when patients stop taking the medication (Permutt and Pinheiro, 2009; Kim, 2011). In such situations, estimands 4 and 5 overestimate effectiveness at the planned endpoint of the trial.

None of the estimands proposed in the NRC guidance (NRC, 2010) address *de-facto* (effectiveness) hypotheses for the initially randomized medication at the planned endpoint of the trial. The estimands in the NRC guidance were not intended to be an exhaustive list. Therefore, a 6^{th} estimand is proposed that may be particularly relevant in the early evaluations and initial regulatory approvals of new medications.

6. ***Difference in outcome improvement in all randomized patients at the planned endpoint of the trial attributable to the initially randomized medication.*** The key attributes of estimand 6 are also summarized in Table 1 (Mallinckrodt et al, 2012). Estimand 6 assesses effectiveness at the planned endpoint, focusing on the causal effects attributable to the initially randomized medications. Conceptually, estimand 1 and estimand 6 require follow-up data. Unlike estimand 1, the intent with estimand 6 is to avoid the confounding effects of rescue medications. However, ethical considerations often mandate that rescue medications be allowed after patients discontinue randomized study medication.

Estimand 3 and estimand 6 focus on causal effects of the initially randomized medications, in all randomized patients, at the planned endpoint of the trial. Estimand 3 focuses on what would have happened if patients adhered to treatment and estimand 6

focuses on what was actually observed. Estimand 3 addresses *de-jure* (efficacy) hypotheses and estimand 6 addresses *de-facto* (effectiveness) hypotheses. Estimand 3 and estimand 6 can be used in combination as the primary and secondary estimands, an approach that would be particularly useful in trials assessing symptomatic treatments. For example, in a proof-of-concept study focus may be primarily on efficacy, but as development progresses focus may shift towards effectiveness if the conditions under which the drug is studied are naturalistic enough to be generalized to clinical practice.

Given the confounding effects of rescue medications and the ethical need to allow them, one approach to testing *de-facto* hypotheses is to impute the data after discontinuation of the initially randomized study medication under the assumption that initially randomized active medications have no effect after they are discontinued. This assumption is often reasonable in trials of symptomatic interventions (O'Neill and Temple, 2012).

Estimation of this estimand has most commonly been done by imputing values using baseline observation carried forward (BOCF). However, using baseline values as the measure of no benefit ignores the improvements that are often seen in trials due to non-pharmacologic reasons and would be valid only in those situations where there was no change in a placebo group over time. Alternative means to test *de-facto* hypotheses have come into the literature recently and these alternatives are described in a subsequent section.

Several approaches may also be taken in estimation of *de-jure* estimands. For example, although endpoint contrasts are often the focus, regression parameters (e.g., linear or linear plus quadratic slopes) for treatment vs. control can be compared.

Analyses

In order to choose an appropriate analysis, the mechanism(s) leading to the missingness must be considered. In longitudinal clinical trials MCAR is not likely; MAR is often plausible but never provable; and, going beyond MAR to MNAR requires assumptions that are not testable. Hence, no single MNAR analysis can be definitive (Verbeke and Molenberghs, 2000).

Consensus is emerging that a primary analysis based on MAR is often reasonable, whereas complete case and single imputation methods that require MCAR and / or other restrictive assumptions are not reasonable (Molenberghs and Kenward, 2007; Mallinckrodt et al, 2008; NRC, 2010). Primary analyses based on MAR may be especially reasonable when combined with rigorous efforts to maximize retention on the initially randomized medications. Methods common in the statistical literature based on MAR include likelihood-based analyses, multiple imputation (MI) and weighted generalized estimating equations (wGEE) (Molenberghs and Kenward, 2007). The specific attributes of each method can be used to best tailor an analysis to the situation at hand.

With an MAR primary analysis, assessing robustness of conclusions to departures from MAR via sensitivity analyses is essential. Although there may be need for additional

sensitivity analyses inspired by trial results, a parsimonious set of plausible sensitivity analyses should be pre-specified and reported.

Three common families of MNAR analyses are shared-parameter models, pattern-mixture models, and selection models (Molenberghs and Kenward, 2007). As typically implemented, these approaches assess *de-jure* estimands. Selection models are conceptually multivariate models for repeated measures, where one variable is the efficacy outcome from the primary analysis and the second is the repeated binary outcome for dropout that is modeled via logistic regression. Pattern-mixture models fit a response model for each pattern of missing values weighted by their respective probabilities. Patterns are often defined by time of dropout, but could be defined by other means, such as reason for discontinuation. In shared-parameter models a set of latent variables, latent classes, and/or random effects is assumed to drive both the measurement and missingness processes. Shared-parameter models can be thought of as multivariate models, where one variable is the continuous efficacy outcome from the primary analysis and the second is (typically) a time to event analysis for dropout.

Recently, another family of methods referred to as controlled imputation has seen increasing discussion in the literature and use in practice. Controlled imputation approaches such as those discussed by Little and Yao, 1996; Carpenter and Kenward, 2007; Ratitch and O'Kelly, 2011) can be thought of as specific versions of pattern-mixture models. The basic idea is to construct a principled set of imputations that exhibit a specific statistical behavior, often a departure from MAR, in order to assess either sensitivity of *de-jure* estimands or as a primary means to assess *de-facto* estimands (Teshome et al; 2012).

In the MAR setting, MI uses separate imputation models for the drug and placebo (control) arms (in a two-arm study). For MNAR analyses, one sub-family of approaches within controlled imputation, referred to as reference-based imputation, uses one imputation model (or in some manner borrows information) from the reference (e.g., placebo, or standard of care) group but then applies that model to both the drug and placebo arms. Alternatively, a single imputation model can be developed from all the data and applied to both arms.

Using one imputation model for both treatment arms diminishes the difference between the arms compared with MAR approaches that use separate imputation models for each arm. The intent is to generate a plausibly conservative efficacy estimate that can be used to define the lower bound of values for the set of sensitivity analyses; or, to generate an estimate of effectiveness that reflects a change in or discontinuation of treatment.

Controlled imputation can also be used to assess sensitivity by repeatedly adjusting the imputations to provide a progressively more severe stress test to assess how extreme departures from MAR must be to overturn the primary result. For example, the analysis can assume that patients who discontinued had outcomes that were worse than otherwise similar patients that remained in the study (NRC, 2010; Carpenter and Kenward, 2007). The difference (adjustment) in outcomes between dropouts and those who remain can be a shift in location (mean) or slope, and is referred to as delta.

Typically, only the experimental arm is delta-adjusted while the control arm is handled using an MAR-based approach. Delta-adjustment can be applied to only the first visit with missing data or to all visits with missing data; and, delta adjustment can be applied as part of a visit-by-visit imputation or after completion of all imputations.

Delta adjustment after imputation simply subtracts a constant from the imputed values and the adjustment at a visit does not influence imputed values at other visits. With delta-adjustment in visit-by-visit imputation, missing values are imputed as a function of both actually observed and previously imputed delta-adjusted values. In this setting, delta-adjustment influences imputed values at the visit to which it is applied and also influences imputed values at subsequent visits through the imputation model. Delta adjustment applied to every visit in a visit-by-visit imputation results in an accumulation of adjustments and thus implies a greater departure from MAR than delta-adjustment at a single visit.

The flexibility and transparent assumptions of controlled imputations allows the methods to be tailored to the clinical setting and the analytic goals. However, these are comparatively new approaches and their attributes in various scenarios have not been fully characterized.

Example

The data set used in this example was somewhat contrived to avoid implications for marketed drugs. Nevertheless, the key features of the original data were preserved. The original data were from an antidepressant clinical trial reported by Goldstein et al (2004). The trial contained four treatment arms, with patients randomized in a 1:1:1:1 ratio to two doses of an experimental medication (subsequently granted marketing authorizations in most major jurisdictions), an approved medication, and placebo. Postbaseline assessments on the Hamilton 17-item rating scale for depression ($HAMD_{17}$) (Hamilton, 1960) were taken at baseline and weeks 1, 2, 4, 6, and 8. In this re-analysis the Week-8 observations were not included. All patients from the original placebo arm were included along with a contrived drug arm that was created by randomly selecting patients from the three non-placebo arms.

Completion rates were 76% (64/84) for drug and 74% (65/88) for placebo. Visitwise mean changes for patients that completed the trial versus those who discontinued early are summarized in Figure 1. Patients who discontinued early had less favorable outcomes than completers, suggesting that missing data did not arise from an MCAR mechanism.

The analysis plan focused on estimand 3, a *de-jure* (if taken as directed) efficacy hypothesis. The key assumption of the direct-likelihood primary analysis, and the focus of sensitivity analyses, was that missing data arose from an MAR mechanism. Other assumptions not tested here which can be objectively evaluated include assumptions regarding time trends, correlation structure, and error distribution. A secondary goal of the analysis was to assess a *de-facto* (effectiveness) hypothesis, estimand 6. Sensitivity analyses included: 1) inclusive models in the MAR framework via MI and wGEE; 2)

model-based MNAR methods, including selection, pattern mixture, and shared parameter models; and, 3) reference-based and delta-adjustment controlled imputations.

The primary analysis used a restrictive model. Mean changes from baseline were analyzed using a restricted maximum likelihood (REML)-based repeated measures approach. The analysis included the fixed, categorical effects of treatment, investigative site, visit, and treatment-by-visit interaction, as well as the continuous, fixed covariates of baseline score and baseline score-by-visit-interaction. An unstructured (co)variance structure shared across treatment groups was used to model the within-patient errors. The Kenward-Roger approximation was used to estimate denominator degrees of freedom and adjust standard errors. Analyses were implemented using SAS PROC MIXED (SAS, 2008). The primary comparison was the contrast between treatments at the last Visit (Week-6). Results from the primary analysis are summarized in Table 1.

Within group LSMEAN changes at Week 6 were -7.05 for drug vs. -4.41 for placebo. Negative values indicated improvement. Therefore, the advantage of drug over placebo was -2.64 (SE=1.02, P= 0.010) – a difference that was statistically significant, but not so robust that the high rate of missing data could be disregarded.

A parametric selection model was implemented using SAS PROC MIXED for starting values and for certain analyses and PROC IML was used to build and solve necessary equations. In the measurement model the primary outcome was modeled using the repeated measures model as in the primary analysis. The drop out model was a logistic regression that fit the log odds of dropout as a function of separate intercepts (Ψ_1,Ψ_2) for each treatment group, along with separate linear regression coefficients for previous (Ψ_3,Ψ_4) and current (possibly unobserved) efficacy outcomes (Ψ_5,Ψ_6). Hence the measurement and drop out models were linked as the dependent variable from the measurement model was an independent variable in the dropout model. The parameters Ψ_5 and Ψ_6 were of interest because they were the "MNAR" part of the model. Fitting separate models for each treatment allowed for different departures from MAR for drug and placebo groups. In addition to estimating parameters from the data, a wide range of values for Ψ_5 and Ψ_6 were input for illustration purposes. Whenever possible, sensitivity analysis should be based on a pre-defined, plausible range of values for Ψ_5 and Ψ_6.

Results from selection model analyses are summarized in Table 1. As expected, assuming MAR by inputting Ψ_5 and $\Psi_6 = 0.0$ (first row of Table 2) yielded a treatment contrast of -2.64, matching the primary direct likelihood analysis. When all parameters were estimated (second row of Table 5) the treatment contrast was -2.48, with SE = 1.09, and P = 0.023. Therefore, compared with the MAR primary analysis the MNAR selection model yielded a slightly smaller treatment contrast, a slightly larger standard error, and a slightly larger but still significant p value.

Results from inputting values for Ψ_5 and Ψ_6 are summarized at the bottom of Table 5. Negative (positive) values for Ψ_5 and Ψ_6 led to within group mean changes that were greater (less) than from the MAR results. This result makes sense in that if better (worse) outcomes were more likely to be missing, had they been observed means would have showed greater (smaller) improvement.

When Ψ_5 and Ψ_6 differed, treatment contrasts followed a consistent pattern. Whenever Ψ_5 (the regression coefficient for the drug group) was less than Ψ_6 (the regression coefficient for the placebo group) the treatment contrast was greater than from the MAR primary analysis; when Ψ_5 was greater than Ψ_6 the treatment contrast was smaller than in MAR. Across the range of input values the endpoint treatment contrast ranged from -1.40 to -3.78. Given the lack of a pre-specified plausible range of input values no inference is drawn from this range of results.

Several caveats apply to the selection model results above and to MNAR models generally. These models inherently rely on untestable assumptions. They are highly sensitive to influential observations and distributional assumptions. Different models with similar maximized likelihoods (*i.e.*,, with similar plausibility with respect to the observed data) can have completely different implications for the dropout process. And, an alternate parameterization of the selection model that fits the increment from the penultimate to the final visit rather than the final outcome itself can lead to meaningfully different interpretations of the dropout process.

Table 1. Results from selection model analyses

Model[1]	Ψ_5[2]	Ψ_6[2]	Endpoint LSmean Change Drug	Placebo	Endpoint Contrast	Standard Error	P value
MAR	0.0	0.0	7.05	4.41	-2.64	0.98	0.007
Estimate	-0.13	-0.16	7.48	5.00	-2.48	1.09	0.023
Input	0.0	0.2	7.07	3.67	-3.40	1.02	< .001
Input	0.0	-0.2	7.06	5.13	-1.93	0.97	0.047
Input	0.0	-0.4	7.07	5.67	-1.40	0.97	0.150
Input	0.2	0.0	6.39	4.43	-1.96	1.02	0.054
Input	-0.2	0.0	7.69	4.41	-3.28	0.97	< .001
Input	-0.4	0.0	8.18	4.40	-3.78	0.97	< .001

1. Estimate indicates all model parameters were estimated and the values in the Ψ_5 and Ψ_6 columns are estimates of those parameters; input indicates values for Ψ_5 and Ψ_6 were input and the values in the Ψ_5 and Ψ_6 columns are the input values.
2. Ψ_5 and Ψ_6 are the regression coefficients (drug and placebo, respectively) for the association between the current, possibly missing efficacy scores and the logit for probability of dropout

Pattern-mixture models were implemented by imputing missing values using the non-future dependent type of complete case and neighboring case missing value restrictions (CCMV and NCMV, respectively). See Molenberghs and Kenward (2007) for detailed descriptions of the restrictions. Dropout patterns were defined by the visit where the last observation for the primary analysis was obtained. Imputations were implemented using SAS PROC MI and PROC MI ANALYZE (SAS, 2008). Completed data sets were

analyzed using the same repeated measures model as for the primary analysis with the addition of terms for dropout group and its interactions with treatment and time.

Results from pattern-mixture model analyses are summarized in Table 2. Endpoint treatment contrasts using NCMV and CCMV restrictions were -2.95 and -2.67, respectively, with similar standard errors and p values < 0.01. Given both approaches yielded similar results, no attempt is made to justify one as more relevant than the other and final inference from the sensitivity analyses is based on results from NCMV.

Table 2. Results from pattern-mixture model analyses

Identifying Restriction[1]	Endpoint Contrast	Standard Error	P value
CCMV	-2.68	0.99	0.007
NCMV	-2.95	0.99	0.003

1. CCMV = non future dependent complete case missing value
2. NCNV = non future dependent neighboring case missing value

The intent for the shared parameter model was to model efficacy outcomes using he same repeated measures model as for the primary analysis, to use the efficacy outcomes in the time to dropout part of the model. However, due to convergence problems a parametric model for time (dependent variable linearly related to square root of time) was used rather than modeling time as unstructured. Two shared-parameter models were implemented using SAS PROC NLMIXED. The first model had no linkage between the measurement and dropout models. A second model linked the dropout and measurement models via separate random intercepts and slopes by treatment group.

Results from shared-parameter model analyses are summarized in Table 3. Using no linkage between the dropout and measurement models yielded an endpoint contrast of –2.92. Using the separate intercept and slope linkages by treatment group yielded a slightly larger endpoint contrast of -3.00, with a standard error that was also larger, resulting in a small increase in the p value for the MNAR model.

Table 3. Results from shared parameter model analyses

Model	Endpoint Contrast	Standard Error	P value
Naïve model (MAR)	-2.92	0.93	0.002
Int + slope by trt linkage	-3.00	1.03	0.004

Three methods of reference-based imputations were implemented, each based on MI using a multivariate repeated measures model where the means for the drug arm were altered using information from the placebo arm. Although it is commonly stated that 5

rounds of imputation is sufficient to yield a high degree of efficiency, stability of results is important in this setting. Our experience suggests that many more rounds of imputation are needed to stabilize results. These analyses used 5000 imputations.

The jump to reference (J2R) method used the active means up to withdrawal and then jumps to the means in the placebo arm after withdrawal. That is, immediately upon withdrawal, all benefit from the treatment was gone, thereby modeling the effectiveness of a symptomatic treatment with a short duration of effect. The J2R approach usually results in the largest decrease in the difference between the experimental and reference group of the three methods used in this example.

In the copy reference (CR) method, previous outcomes were included in the imputation model to imput values as if drug treated patients who dropped out had been on placebo throughout the study. Therefore, if a patient had good outcomes while on drug, those favourable outcomes contributed to the predictions of the missing values based on placebo imputation model. This approach generally results in a more gradual decay of the treatment effect compared with J2R This approach is useful for modeling the effectiveness of a symptomatic treatment with a longer duration of action, conditions matching those in the example data. Although some minor differences in implementation exist, the CR method is conceptually similar to the approach termed placebo multiple imputation (pMI) in Teshome et al (2012) and detailed in Ratitch and O'Kelly (2011).

The copy increment from reference (CIR) method used the active means up to withdrawal but then increments using the changes in the mean from visit to visit seen in the placebo arm. Therefore, improvement prior to withdrawal is maintained, but after withdrawal the trajectory is parallel to that for the placebo. This approach models effectiveness of a disease modifying treatment and usually has the least impact of these three sensitivity analyses.

Results from the reference-based imputations are summarized in Table 4. The endpoint contrast from CR was -2.20, with a standard error somewhat greater than in the primary analysis, and p = 0.028. Therefore, when interpreted as an MNAR sensitivity analysis this result supports robustness of the primary analysis. When interpreted in the effectiveness context, the CR result suggested that 83% (-2.20 / -2.64) of the effect if taken as directed (efficacy) was maintained as actually taken in this study.

Table 4. Results from reference-based multiple imputation

	LSMEAN changes		Endpoint	Standard	
Method	Drug	Placebo	Contrast	Error	P value
J2R	-6.28	-4.30	-1.98	1.01	0.051
CR	-6.46	-4.26	-2.20	0.99	0.028
CIR	-6.52	-4.25	-2.28	0.99	0.022

In addition, progressive stress tests were implemented via MI with delta-adjustment. Two forms of adjustment were applied by either delta-adjusting only the first visit with

missing data or by adjusting all visits with missing data. In both cases imputations were performed visit-by-visit, with patients' delta-adjusted imputed data contributing to imputed values at subsequent visits. The "tipping point" was identified by repeating the imputation process with progressively larger deltas. Analyses were implemented as previously described for other MI based approaches.

Delta-adjustment stress test results are summarized in Table 6. When applying the delta adjustment to only the first missing visit the delta had to be a worsening of 4 points on the $HAMD_{17}$ in order to overturn the primary result. When applying the delta adjustment to all visits the magnitude of the adjustment had to be a worsening of 2 points on the $HAMD_{17}$ in order to overturn the primary result.

Table 5. Results from delta-adjustment multiple imputation

Value of Delta Adjustment	Adjustment method	Endpoint Contrast	P value
0	First missing visit only	-2.74	.008
1.0	First missing visit only	-2.56	.013
2.0	First missing visit only	-2.38	.022
3.0	First missing visit only	-2.20	.035
4.0	First missing visit only	-2.02	.055
0	All visits	-2.74	.008
1.0	All visits	-2.38	.021
2.0	All visits	-2.02	.054

Results from sensitivity analyses are summarized in Table 6. Across the various model-based sensitivity analyses the advantages of drug over placebo at endpoint were generally close to or greater than the primary result. Therefore, the model-based sensitivity analyses support the robustness of the primary analysis to departures from MAR.

With controlled imputations, previous experience suggested the CR approach provided a clearly conservative, but plausible, estimate of the treatment effect for an efficacy hypothesis and a reasonable assessment of effectiveness. The advantage of drug over placebo from CR was approximately 83% of the magnitude of the primary result, with statistical significance preserved.

Preservation of statistical significance need not be a requirement of sensitivity analyses when assessing robustness of the primary result. However, in those cases where significance is preserved from a clearly conservative analysis this may be sufficient to declare the primary result robust to departures from MAR.

In the delta-adjustment stress testing analyses, deltas required to overturn the primary result ranged from 2 to 4 points on the $HAMD_{17}$ depending on the specific method. Given a residual variance of 36 (see Table 3), the residual standard deviation was 6.0. Therefore, the tipping points correspond to 1/3 and 2/3 of the residual standard deviation.

Table 6. Summary of missing data sensitivity analysis results

Method[1]	Endpoint Contrast	Standard Error	P value
Model-based approaches			
Likelihood (primary analysis)	-2.64	1.01	0.010
MI with inclusive model	-2.54	1.12	0.024
wGEE with inclusive model	-3.03	1.09	0.006
SM	-2.48	1.09	0.023
PMM (NCMV)	-2.67	0.99	0.007
SPM	-3.03	1.03	0.004
Controlled imputation approaches			
CR	-2.20	0.99	0.028

Delta adjustment first missing visit only significance lost when delta ≥ 4.0
Delta adjustment all visits significance lost when delta ≥ 2.0

1. SM = selection model; PMM = pattern mixture model, SPM = shared parameter model

Discussion

Recent research has produced useful guidance and recommendations regarding the prevention and treatment of missing data. The intent of this paper has been to provide a practical guide and tools for applying the recent guidance. The programs and example data used in this paper are freely available at www.missingdata.uk.org. The web site also provides additional details on the programs and how to use them.

Approaches and ideas presented in this paper are not intended as specific prescriptions for all trials. As the clinical contexts vary between studies, so too should the trial design and conduct options to reduce missing data, along with the specific form of the sensitivity analyses. Despite the idiosyncrasies of specific situations, several general points are clear. Most importantly, trials should be designed and conducted to maximize the proportion of patients that adhere to the study prescribed treatments. Given that missing data cannot be eliminated, it is also important to clearly state objectives and estimands, and to pre-specify the primary analysis and its assumptions, along with sensitivity analyses that are based on plausible assumptions.

Some argue that follow-up data collected after discontinuation of the initially randomized study drug and / or initiation of rescue medication should usually be included in the primary analysis (NRC, 2010). Others point to a more nuanced usage wherein follow-up

data are often part of the primary estimand in outcome trials, but not in symptomatic trials (O'Neill and Temple, 2012).

In many cases both efficacy and effectiveness at the planned endpoint of the trial will be of interest because it is important to know what happens when a drug is taken as directed (efficacy) and to know what happens when the drug is taken as in actual practice (effectiveness). The choice between efficacy and effectiveness as the primary estimand should be influenced by whether the trial design and conduct is more consistent with rigorously controlled efficacy assessments or more naturalistic effectiveness assessments. Whether or not follow-up data should be collected and / or included in the primary estimand can be considered on a case-by-case basis. However, given the confounding influences of rescue medications, the role for follow-up data in the analysis of symptomatic treatment trials would usually be secondary.

A primary analysis based on MAR is often reasonable. Likelihood-based methods, MI, and wGEE are all useful MAR approaches whose specific attributes can be considered when tailoring a primary analysis to specific situations. With an MAR-based primary analysis a focal point of sensitivity assessments is the impact of departures from MAR on estimates of the primary treatment contrast. To this end, the model-based family of MNAR methods such as selection models, pattern-mixture models and shared-parameter models can be considered. Prior experience can guide analytic decisions such as plausible ranges of input values for selection models, appropriate linkages between analysis and dropout models in the shared parameter setting, or appropriate identifying restrictions for pattern-mixture models.

Controlled-imputation methods can be especially useful in constructing analyses to assess specific departures from MAR and for assessing effectiveness because the assumptions are transparent. If a plausibly conservative controlled imputation analysis agrees sufficiently with the primary result, as it did in the example data, the primary result can be declared robust to departures from MAR. Alternatively, a tipping point (progressive stress-testing) format can be used to assess how severe departures from MAR must be in order to overturn conclusions from the primary analysis. If, as in the example data, severe departures from MAR are required to negate the primary result, the primary result can be declared robust to departures from MAR.

In addition to the methods illustrated in this paper, macros are also available at www.missingdata.org.uk to conduct influence and residual diagnostics, and other descriptive analyses. Future efforts of the DIASWG on missing data will include making available semi- and non-parametric versions of relevant methods, doubly robust methods, sensitivity analyses for categorical data, and applying sensitivity analyses retrospectively to large data pools in order to gain experience and provide perspective.

References

Carpenter JR, Kenward MG. Missing data in randomised controlled trials – a practical guide. (2007). Available at http://missingdata.lshtm.ac.uk/downloads/rm04_jh17_mk.pdf (accessed 23 January 2012).

Carpenter J, Roger J, and Kenward M. Analysis of Longitudinal Trials with Missing Data: A Framework for Relevant, Accessible Assumptions, and Inference via Multiple Imputation. 2011 (Submitted).

Daniel R, Kenward M. A method for increasing the robustness of multiple imputation, *Computational Statistics and Data Analysis.* (2012); 56, 1624-43

Committee for Medicinal Products for Human Use (CHMP). Guideline on missing data in confirmatory clinical trials. 2010. EMA/CPMP/EWP/1776/99 Rev. 1

Goldstein DJ, Lu Y, Detke MJ, Wiltse C, Mallinckrodt C, Demitrack MA: Duloxetine in the treatment of depression: a double-blind placebo-controlled comparison with paroxetine. J Clin Psychopharmacol 2004;24: 389-399.

Hamilton M: **A rating scale for depression.** *J Neurol Neurosurg Psychiatry* 1960, **23:** 56-61.

Khan A, Schwartz K, Redding N, Kolts R, Brown WA. Psychiatric Diagnosis and Clinical Trial Completion Rates: Analysis of the FDA SBA Reports. Neuropsychopharmacology (2007) 32, 2422–2430

Kim Y. Missing Data Handling in Chronic Pain Trials. Journal of Biopharmaceutical Statistics. 2011; 21: 2, 311 — 325

Lane PW. Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. Pharm Stat. 2008; 7:93-106.

Little, R.J.A., and Rubin, D.B. (2002). Statistical Analysis with Missing Data, 2nd edition. New York: Wiley.

Little R and Yau L, 1996, Intention-to-treat analysis for longitudinal studies with drop-outs, *Biometrics*, 1324-1333

Mallinckrodt CH, Raskin J, Wohlreich MM, Watkin JG, Detke MJ, The efficacy of duloxetine: A comprehensive summary of results from MMRM and LOCF in eight clinical trials. BMC Psychiatry. 2004; 4:26.

Mallinckrodt CH, Lane PW, Schnell D, Peng Y, and Mancuso JP. Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials. *Drug Information Journal. 2008;* **42**:305-319.

Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, **Mallinckrodt** CH, Carroll RJ. Analyzing incomplete longitudinal clinical trial data. Biostatistics. 2004; 5:445-464.

Molenberghs G, Kenward MG. (2007), *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons.

National Research Council (2010). The prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

O'Neill RT and Temple R. (2012). The Prevention and Treatment of issing Data in Clinical Trials: An FDA Perspective on the Importance of Dealing With It. Clinical Pharmacology and Therapeutics. doi:10.1038/clpt.2011.340

Permutt T and Pinheiro J. (2009). Dealing with the missing data challenge in clinical trials. *Drug Information Journal*. **43(**, 403-408.

Ratitch B, O'Kelly M. Implementation of Pattern-Mixture Models Using Standard SAS/STAT Procedures. *PharmaSUG* 2011. Available at http://pharmasug.org/proceedings/2011/SP/PharmaSUG-2011-SP04.pdf (accessed October 4, 2011)

Roger J, Ritchie S, Donovan C, Carpenter J. Sensitivity Analysis for Longitudinal Studies with Withdrawal. *PSI Conference*, May 2008. Abstract at http://www.psiweb.org/docs/2008finalprogramme.pdf (accessed 23 January 2012)

Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley: New York, 1987.

SAS Institute Inc. 2008. SAS/STAT® 9.2. User's Guide. Cary, NC: SAS Institute Inc.

Siddiqui, O, Hung, H.M., O'Neill, R.O. MMRM vs. LOCF: A Comprehensive Comparison Based on Simulation Study and 25 NDA Datasets. J. Biopharmaceutical Statistics, **19(2)**, 227-246.

Teshome B, Lipkovich I, Molenberghs G, Mallinckrodt CH. Placebo Multiple Imputation: A new approach to sensitivity analyses for incomplete longitudinal clinical trial data. Submitted.

Verbeke G, Molenberghs G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.