Statistics
in Medicine

Research Article

Received 5 May 2015, Accepted 4 January 2016 Published online 16 February 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6890

Analyzing clinical trial outcomes based on incomplete daily diary reports

Neal Thomas,[a*†] Ofer Harel[b] and Roderick J.A. Little[c]

A case study is presented assessing the impact of missing data on the analysis of daily diary data from a study evaluating the effect of a drug for the treatment of insomnia. The primary analysis averaged daily diary values for each patient into a weekly variable. Following the commonly used approach, missing daily values within a week were ignored provided there was a minimum number of diary reports (i.e., at least 4). A longitudinal model was then fit with treatment, time, and patient-specific effects. A treatment effect at a pre-specified landmark time was obtained from the model. Weekly values following dropout were regarded as missing, but intermittent daily missing values were obscured. Graphical summaries and tables are presented to characterize the complex missing data patterns. We use multiple imputation for daily diary data to create completed data sets so that exactly 7 daily diary values contribute to each weekly patient average. Standard analysis methods are then applied for landmark analysis of the completed data sets, and the resulting estimates are combined using the standard multiple imputation approach. The observed data are subject to digit heaping and patterned responses (e.g., identical values for several consecutive days), which makes accurate modeling of the response data difficult. Sensitivity analyses under different modeling assumptions for the data were performed, along with pattern mixture models assessing the sensitivity to the missing at random assumption. The emphasis is on graphical displays and computational methods that can be implemented with general-purpose software. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: multiple imputation; pattern mixture models; clinical trials; incomplete data

1. Introduction

Missing data is a common problem in longitudinal clinical trials, with the potential to lead to loss of statistical efficiency and biased inferences. The recent National Research Council report on the topic [1, 2] highlights the need to limit missing data in trial design and conduct, as well as to use defensible statistical methods in the analysis. Concerning the latter, the report criticized simple methods like analysis of the complete cases and last observation carried forward (LOCF) imputation, and recommended scientifically defensible approaches and sensitivity analyses to assess potential violations of the assumptions of the primary analysis method. We describe methods for handling missing data in a trial assessing a treatment for insomnia, which involved nightly sleep measurements and a substantial amount of missing data from item nonresponse and dropout. Methods described in the protocol were based on the analysis of available data, revealing clear limitations because they rely on the assumption that the missing daily reports are missing completely at random (MCAR, [3]). The analyses make no distinction, for example, between weekly averages of daily values based on 4 instead of the intended 7 values, or 3 actual values instead of 0 values. This practice is so pervasive in therapeutic areas where daily diary data are collected (e.g., insomnia, pain, post-menopausal symptoms) that it is not typically noted when the results are reported. Methods based on multiple imputation (MI, [4]) are described here to address these issues.

Existing MI-based software can produce valid results under the weaker missing at random (MAR) assumption. The MI-based methods form weekly average values for each patient based on exactly seven daily values, some of which may be imputed, thereby addressing potential problems that can occur even

[a]*Pfizer, Inc, Groton, CT, U.S.A.*
[b]*University of Connecticut, Mansfield, CT, U.S.A.*
[c]*University of Michigan, Ann Arbor, MI, U.S.A.*
[*]*Correspondence to: Neal Thomas, Pfizer, Inc., Groton, CT, U.S.A.*
[†]*E-mail: snthomas99@yahoo.com*

under MCAR when the standard errors (SEs) are computed from weekly averages of unequal numbers of available daily reports. Sensitivity analyses are also described to assess deviations from the MAR assumption when values are missing not at random (MNAR, [3]). Graphical displays of the missing data patterns and the multiply-imputed values are emphasized.

For patients who discontinue treatment, it is important to be clear what treatment is being assumed for imputations after discontinuation [5]. Our methods assume the same treatment after discontinuation, because the MAR assumption is not reasonable for imputations where the treatment has changed (for example, from active drug to placebo). In the trial, we describe, it is reasonable to assume that most of the patients who stopped treatment early could have continued to comply with the treatment assigned, except for a small number of patients reporting termination due to adverse events or death, as displayed in the Supporting Information Table A.1. Alternative analysis methods, such as jump-to-placebo [6] or principal stratification [7], were not implemented because there were few of the latter patients.

The clinical trial design and data collection are described in Section 2, which includes a detailed description of the protocol-specified primary analysis and the data used in it. Patterns of missing data with complex dependence between intermittent and dropout missingness and adverse events and temporal proximity of diary reporting to weekly clinic visits are summarized in Section 3. The results of the protocol-specified primary analysis are described in Section 4, along with some other analyses with different missing data approaches specified in the original analysis plan. As previously noted, the common protocol-specified methods depend on the MCAR assumption, and they can be deficient even assuming MCAR, because they do not account for the differing number of measurements included in each 'weekly' average. The first of the MI analyses, which addresses the problems in the protocol-specified method by explicitly imputing all missing daily measurements, is in Section 5. It was specified before examination of the data utilizing a model with a compound-symmetric variance-covariance matrix, and mean daily-diary outcomes that changed only weekly. The model was selected for computational convenience and because we were confident it could be successfully implemented. Once data were available, the impact of the restrictive model specification was assessed using models fitted separately for each treatment group with mean values that changed daily and with models allowing more complex variance–covariance structures selected based on exploratory data analyses. Graphical displays in Section 6.1 compare observed responses amongst patients with different dropout patterns. They show that the observed efficacy measurements are not predictive of missingness, which supports the plausibility of the MAR assumption. Because the MAR assumption cannot be unequivocally established from observed data, sensitivity to violation of the MAR assumption was assessed using pattern-mixture models in Section 6.2. A tipping point analysis was created by modifying an MI analysis to impute missing values that were increasingly unfavorable. When these MNAR imputations were applied to each treatment group, there was minimal impact on estimates of treatment effect, aside from increased variability due to some of the MNAR-imputed values. The unfavorable imputations were also applied to the active treatment groups only, which showed that a shift of approximately one standard deviation (SD) in the missing responses substantively changed the conclusions of the analyses.

2. Clinical trial design and data

2.1. Design

The case study is based on a randomized double-blind parallel-group placebo-controlled study of a compound for chronic insomnia sponsored by a large pharmaceutical company during phase 2 of development. There was a 1-week blinded placebo run-in period before randomization. The randomization visit is defined as day 1 of the study. The daily diary collected during the morning of day 1 is regarded as part of the pre-randomization baseline period. There were five treatment groups: (PBO) and 15, 30, 45, and 60 mg of the active compound. There were approximately 135 randomized patients per group. It was planned that each patient would receive their assigned treatment for 4 weeks. There were 10 randomized patients who did not start dosing; they are excluded throughout. Weekly visits were scheduled (0,1,2,3,4) for data collection. Data from a post-dosing safety visit are not included in our analysis. The data for each patient are longitudinal, with repeated measurements based on patient-reported outcomes and one clinician assessment of severity. Baseline age, sex, race, and clinical site are included in the data set. In addition to the weekly visit, patients called a data collection system each morning from their first screening visit until their week 4 final-dosing visit and responded to questions about their sleep the previous night.

Table I. Daily diaries contributing to subjective time awake after initial sleep onset at week 4.

Days between visits 3 and 4	Row total	Days contributing to the weekly average											
		1	2	3	4	5	6	7	8	9	10	11	14
2	1	1	0	—	—	—	—	—	—	—	—	—	—
3	1	0	0	1	—	—	—	—	—	—	—	—	—
4	4	0	0	1	3	—	—	—	—	—	—	—	—
5	31	0	3	2	11	15	—	—	—	—	—	—	—
6	81	1	0	1	12	24	43	—	—	—	—	—	—
7	342	1	0	5	7	32	81	216	—	—	—	—	—
8	64	0	1	0	2	4	9	19	29	—	—	—	—
9	21	0	0	0	0	1	1	4	4	11	—	—	—
10	5	0	0	0	1	0	1	1	0	1	1	—	—
11	1	0	0	0	1	0	0	0	0	0	0	0	—
12	2	0	0	1	0	0	0	0	0	0	0	1	—
13	1	1	0	0	0	0	0	0	0	0	0	0	—
16	1	0	0	0	0	0	0	0	0	0	0	0	1

Counts are the number of patients amongst those with week 3 and 4 visits. Row totals are patients with the specified days between visits. Impossible combinations are marked with a '—'.

Weekly summaries were not determined by the common practice of setting time windows around the scheduled date, with visits outside the window excluded. Instead, the visit designation was based on the reported visit number on the case report form. Most subjects followed the visit schedule closely, but there were patients with substantial deviations. The first two columns of Table I summarize the distribution of patients by their days between the weeks 3 and 4 visits.

2.2. *Data*

There were five variables collected from the daily phone calls: subjective time awake after initial sleep onset (SWASO, minutes), subjective latency to sleep onset (minutes), subjective number of awakenings after sleep onset (SNAASO), subjective total sleep time (minutes), and sleep quality (0–100, higher is better). The primary endpoint, SWASO, was derived from the daily phone diary data by averaging the daily values between each weekly visit, as is commonly carried out with daily diary data (e.g., [8, 9]). The number of days between visits varied, and measurements from all days between visits were averaged. If there were <4 diary reports between visits, the statistical analysis plan specified that the weekly average was missing. Weekly values for each of the other sleep measures were formed using the same approach.

The SWASO endpoint requires definition when a patient reports no awakenings during the night (SNAASO = 0). In this case, the SWASO value was coded as missing in the database and treated as missing in the original study analyses. Combined with the use of available cases when computing the weekly averages, this approach creates an MNAR condition that causes underestimation of the effect of the drug when patients successfully sleep through the night. In all of the analyses reported here, when a patient reports SNAASO = 0, the corresponding SWASO value will be assigned 0 awake time, and it will not be regarded as missing.

The 11 variables collected at the weekly visits, which record the patients' recall of day-time function, drowsiness, and other sleep-related conditions, are described in the Supporting Information Table A.1. A limited set of commonly occurring adverse events (e.g., headaches and dizziness) were included in our data. A final status was also obtained from the case report form for each patient at their final visit, which indicates whether the patient finished treatment as planned, withdrew consent, stopped because of pregnancy and so on.

2.3. *Pre-specified analyses*

The primary statistical analysis plan pre-specified a mixed-model repeated-measures analysis for the change from baseline in weekly average SWASO, with site, treatment, visit, baseline SWASO, treatment-by-visit interaction, and baseline-by-visit interaction as fixed predictors, and an unstructured covariance structure. Each dose was compared with placebo at week 4 to measure persistent effect. A similar analysis was planned for each secondary endpoint. The primary pre-specified intention-to-treat analysis excluded

randomized patients who did not receive a dose of study drug. We followed this convention and excluded these patients throughout. Other randomized patients without any post-randomization endpoints were included in our analyses, as documented in Section 3.

3. Patterns of missing data

3.1. Missingness in daily telephone diaries

Plots of missingness rates for the SWASO variable during the first 28 days after randomization are displayed in the Supporting Information Figure A.1. The rate of missingness increases over time with a pattern of lower missing rates on the days of scheduled clinic visits and bigger increases in missingness rates the day following clinic visits due to dropouts that occur at the visit. Additional analyses showed there was much less missing diary data on the days of clinic visits. These trends are attenuated in Figure A.1 because not all patient visits occur on the planned schedule. The highest-dose group has consistently higher missingness rates than the placebo rate.

Most of the 'weekly' averages were computed based on collection time intervals that spanned 4 to 9 days, as displayed in Table I, which summarizes the collection period before the final visit. The table also shows that many patients had unplanned missing diary entries during this collection period. The proportion of the weekly averaged endpoints computed with at least one missing daily value ranged from 0.3 to 0.5 across the four weekly visits.

3.2. Missingness in weekly averages

The missing data rates for the SWASO averages, as defined in the protocol, at week 4 for the 0-, 15-, 30-, 45-, and 60-mg-dose groups are 0.18, 0.25, 0.16, 0.16, and 0.25, respectively. These rates are sums of the dropout and intermittent missing (weekly) rates in Table II. The only notable dose-related pattern is a higher rate of missing data at the early visits for the highest dose. To explore potential reasons for the elevated missing rate in the highest-dose group, the frequencies of the reasons for the end of dosing were examined (displayed in the Supporting Information Table A.2). Dropouts due to adverse events were also more frequent in the highest-dose group. The excess adverse events were not concentrated in a small number of related categories.

Further examination of the patients who dropped out of the study early showed there were 36 randomized patients missing all weekly diary data. It is not apparent why they do not have baseline diary data. All of them dropped out of the study without any post-dosing endpoints. Of these patients, 13 reported stopping because of an adverse event, out of 26 such patients in the entire study. The numbers of these patients in the 0-, 15-, 30-, 45-, and 60-mg-dose groups are 2, 2, 0, 1, and 8, respectively. These patients largely account for the higher rates of missingness and adverse events observed in the high-dose group. Another potential reason for more dropouts in the highest-dose group is described in Section 4.1.

Figure 1 (produced by the R package, VIM, [10]) summarizes the frequency of missing data patterns. The weekly averaged endpoints based on the daily diary data display many different missing data patterns, but the patterns associated with monotone dropout are the most common. The missing data patterns are sorted from completely missing to completely observed by the order of the first occurrence of a missing

Table II. Dropout and intermittent missing rates for the subjective time awake after initial sleep onset endpoint.

Dose	Cumulative dropout rates Week				Intermittent missing rates Week				
	1	2	3	4	0	1	2	3	4
0	0.06	0.08	0.10	0.14	0.06	0.02	0.02	0.03	0.04
15	0.04	0.11	0.15	0.20	0.04	0.01	0.01	0.04	0.05
30	0.04	0.09	0.13	0.14	0.04	0.00	0.00	0.02	0.02
45	0.04	0.10	0.13	0.16	0.05	0.02	0.01	0.02	0.00
60	0.09	0.15	0.18	0.22	0.09	0.04	0.04	0.00	0.03

The intermittent missing rate refers to the weekly means of the daily values and applies to patients remaining in the study at a visit. Note that a patient can return for their week 4 visit and still have missing daily diary data for the week.

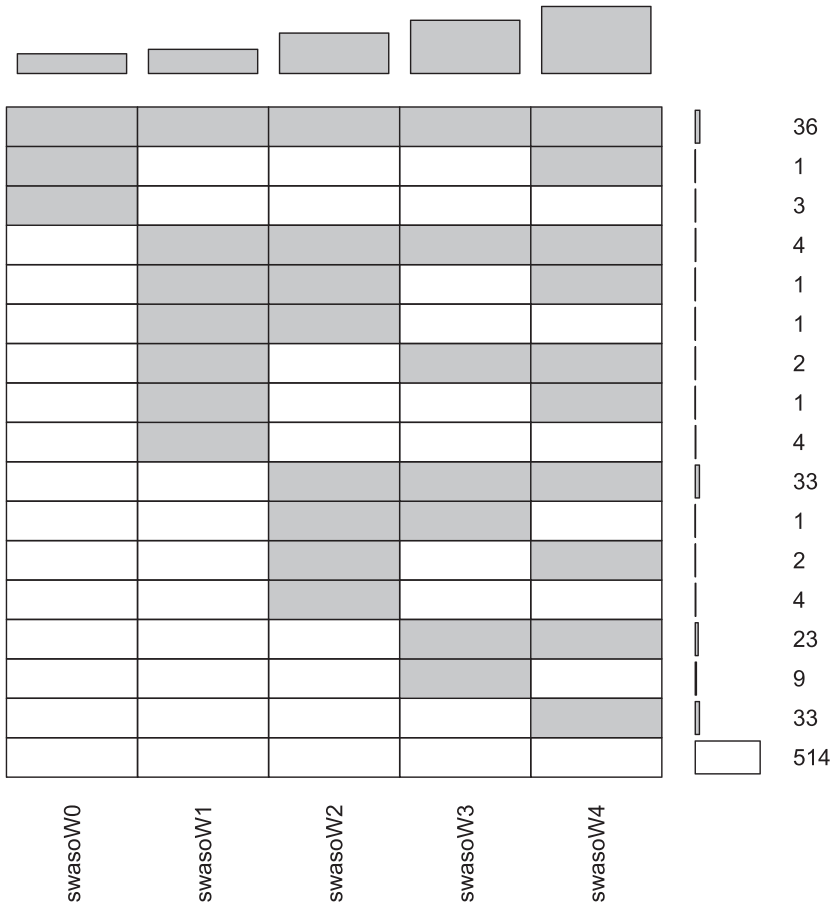

Figure 1. Missing data patterns for the weekly averaged subjective time awake after initial sleep onset variable. The grey shaded regions represent missing data. The patterns are ordered by the first occurrence of missing data. The right-hand numbers are the frequencies of each pattern.

value. The large number of patients without diary data was first revealed by this display. The histogram at the top of the figure displays the increase in missing data by visit.

4. Efficacy based on common methods applied to weekly averaged values

4.1. Treatment effect estimates

The mean dose–response curve for the primary endpoint at week 4 is displayed in the Supporting Information Figure A.2. The drug effect increases with dose to a plateau consistent with an Emax model [11], which has been observed for most effective compounds [12]. The other endpoints based on the daily diaries have similar response trends, which are also displayed in Figure A.2. The endpoints collected at the weekly visits are displayed in the Supporting Information Figure A.3. The patient and clinician global assessment endpoints have trends similar to the endpoints based on the daily diaries. All of the other secondary endpoints, which measure daytime function, display non-monotone dose–response. This is likely due to residual drug effect during the day because some drug remains in the body, as has been observed with other sleep medications [e.g., [13]]. The residual drug might also affect compliance with the dosing regimen and dropout status.

Treatment effect estimates (and SEs) for the primary endpoint at week 4 are in the first three rows of Table III. Estimators include maximum likelihood for the longitudinal mixed model (MLLM), which was described in Section 2.3, a corresponding linear model estimator applied to complete data at baseline and week 4 created by LOCF and a linear model estimator using available cases (ACs) with baseline and week 4 data. The methods are applied to the weekly averaged values, which were computed as described in Section 2.2. There were no pronounced differences between estimators, when assessed across all of the endpoints (displayed in the Supporting Information Tables A.3 and A.4). The estimated SEs differed

Table III. Estimates and standard errors for the week 4 subjective time awake after initial sleep onset endpoint from different missing data methods.

	15 mg vs. PBO			60 mg vs. PBO			Pooled
Method	Est	SE	%Mis Info	Est	SE	%Mis Info	Res SD
MLLM-week	−2.93	5.97		−26.61	6		44.7
LOCF-week	−2.7	5.84		−24.42	5.92		46.6
AC-week	−1.59	6.46		−23.22	6.46		45.9
MVNMI1-day	−4.43	5.15	11	−26.57	5.15	12	39.4
MVNMI2-day	−4.23	5.17	11	−26.22	5.09	9	39.5
MVNMI3-day	−5.2	5.16	12	−26.17	5.07	9	39.3
MVNMI4-day	−5.09	5.25	11	−26.54	5.25	12	40.1
MVNMI5-day	−4.19	5.23	10	−26.33	5.27	12	40.1

The 'week' in methods indicates it was applied to weekly data. The 'day' indicates it was applied to daily data, which was subsequently averaged.
MLLM, mixed longitudinal model; LOCF, linear model applied to LOCF data at Week 4; AC, available cases.

by less than 10% and, as anticipated, increased with methods in the order from LOCF, MLLM, to AC. The standardized differences between the estimates (difference divided by the *SE* of an estimate) were <0.5, except for a few differences as large as 1.0 for the LOCF-based and AC-based estimates.

4.2. Distribution of the subjective time awake after initial sleep onset endpoint

The distribution of the primary SWASO endpoint was examined to assess the appropriateness of the planned analyses and to support selection of imputation models. Boxplots of SWASO by dose group for individual study days displayed right-skewed distributions. The transformed values are closer to normally distributed, but a floor at 0 remains. Boxplots of the distributions for selected study days after applying the square root transform are displayed in the Supporting Information Figure A.4.

Boxplots for the weekly averages of the daily values, with and without transformation, display distributions similar to the corresponding daily values, but the averaging produces closer agreement to the normal distribution. As a consequence, models for daily values will be applied to the square root-transformed values. Because the original scale is more interpretable and the skewness after weekly averaging is not severe, daily values will be back transformed before applying the primary analysis methods.

5. Multiple imputation methods applied to daily values

By basing the weekly averages on available cases, we in effect impute the missing nights using the mean outcome for the nights reported that week. This assumes that the reported nights are representative of all the nights in that week; sleep patterns are assumed to be no different for the nights where no report was provided. Also, weekly averages are treated as having the same precision, regardless of how many measurements are included in the average. The analysis in this section still assumes the missing data are MAR, but imputes missing nights based on a regression of the missing on the recorded nights and uses multiple imputation to reflect the imputation uncertainty.

Model-based multiple imputation of missing daily diary values was performed for the primary SWASO endpoint. The completed daily values were averaged to produce weekly values, and the primary analysis was applied to these data. Completed data sets were created for study days −5 to 28 for each patient. The weekly averages corresponding to baseline and four post-randomization visits were computed from exactly seven daily values determined by the planned visit schedule. This differs from the protocol-specified weekly averages, which could include more than seven values, because there were no planned windows around the weekly visits. The imputations were performed for the square root-transformed values. The daily values were back transformed before weekly averages were computed. Any negative imputed values were set to 0 before back transformation. One-hundred imputed data sets were generated for each multiple imputation method.

The imputation models assume a multivariate normal (MN) distribution for the transformed daily SWASO values (including the baseline values) with means determined by a multiple linear regression. The transformed SWASO values are denoted by Y_{ij}, where patients are indexed by $i = 1, \ldots, N$ and

study days are indexed by $j = -5, \ldots, 29$. The patient-specific mean models include the following fully observed predictors denoted by X_i: age (continuous), sex, race, and one post-randomization variable, the reason for terminating dosing (planned end of study, AE or death, other). The regression parameters associated with X are denoted by β. The dose group is denoted by T_i, with values $(0, 15, 30, 45, 60)$. A potentially different mean value for each study day for each dose is denoted by δ_j^T. For some of the models, the daily means are assumed to be the same within a study week:

$$\delta_j^T = \Delta^T_{\lceil (j-1)/7 \rceil} ,$$

where $\lceil\ \rceil$ denotes the integer ceiling and the Δ_k^T, $k = 0, \ldots, 4$ represent the weekly means. Because of the randomization, $\delta_j^T \equiv \delta_j$ when $j \leq 1$ and $\Delta_0^T \equiv \Delta_0$. For the initial models, equi-correlated variance matrices were specified through random (normal) patient-specific terms denoted by $\theta_i, i = 1, \ldots, N$, with mean of 0 and variance ψ^2. The residuals about the daily mean values are denoted by ϵ_{ij}, with variance σ^2.

The first model assumes that the daily means change weekly:

$$Y_{ij} = X_i'\beta + \Delta^{T_i}_{\lceil (j-1)/7 \rceil} + \theta_i + \epsilon_{ij} . \tag{1}$$

The second imputation model is the same as the first model except that it was fit separately for each dose group, and thus implicitly included interaction terms between dose and all of the main effects in the model. The second model also included separate within and between variance parameters for each dose group. The first two models were fit using the R package PAN [14, 15]. Results from these models are denoted by MVNMI1 and MVNMI2.

A third imputation model, denoted by MVNMI3, was similar to model (1), except that the mean values were allowed to change daily rather than weekly:

$$Y_{ij} = X_i'\beta + \delta_j^{T_i} + \theta_i + \epsilon_{ij} . \tag{2}$$

The model was fit, and the imputations were generated using the general-purpose Bayesian Markov chain Monte Carlo program STAN [16]. Results from the same model fit using the PAN software (not shown) were similar. The same diffuse prior distributions (i.e., diffuse normal prior distributions for fixed effects and diffuse gamma distributions for random effects) were utilized in both programs. The flexibility of the general-purpose software can be used to impute from many alternative models for the mean and variance structures, but it is somewhat slower to execute. The imputation approaches described here took from 30 min to a full day to create 100 imputed data sets on a mid-range desktop computer.

The estimated magnitudes of the within (σ^2) and between patient (ψ^2) variability were roughly equal for all of the fitted models. The variance–covariance matrix of the square root-transformed daily SWASO values was examined by pooling the residuals across dose groups after applying dose group analysis of variance to the available SWASO values for each day. The pairwise correlations displayed a weak trend toward increased correlation for nearby days. A more pronounced difference was much higher correlations between daily values collected during the post-randomization period compared with correlations involving days from the baseline period. The empirical (5th, 95th) percentiles for correlations including at least one baseline value are (0.23, 0.45), while they are (0.46, 0.67) for post-randomization values. The residual SD also displayed a marked increase following the end of the baseline period that continued to increase more gradually during the post-randomization period. An empirical fit of the trend was obtained by a least squares fit of the daily SDs on study day, which is given by $f(day) = 3.72 + 0.203 \log\{0.5 + (day)I(day > 1)\}$. The daily SDs and the empirical curve, $f(day)$, are displayed in the Supporting Information Figure A.5.

Two additional imputation models, denoted by MVNMI4 and MVNMI5, were fit to better represent the changing variances and correlations. In model MVNMI4, both the within and between patient random terms were multiplied by f:

$$Y_{ij} = X_i'\beta + \delta_j^{T_i} + f(j)\theta_i + f(j)\epsilon_{ij} . \tag{3}$$

This model has an increased variance over time while maintaining a common correlation. The increasing multiplier over time was applied to the between patient random term only in model MVNMI5:

$$Y_{ij} = X_i'\beta + \delta_j^{T_i} + f(j)\theta_i + \epsilon_{ij} . \tag{4}$$

This model yields increasing variance with time and higher correlations between values from later study days compared with early days, as was observed in the complete case analysis.

Results based on the multiply-imputed data are in the lower portion of Table III; the reported results include only comparisons of the lowest and highest doses to placebo at week 4. All of the imputation models yielded estimates of treatment effect that were substantively similar to the MLLM approach based on the protocol-specified weekly averages, but the estimates for the 15-mg dose trended toward larger effects. The maximum absolute difference between the MLLM and various MI estimates divided by the SE of the MLLM estimate for the 15- and 60-mg effects versus placebo were 0.38 and 0.07, respectively. A difference of 0.4, even if it were replicated across repeated data sets, would only reduce the coverage of a nominal 95% interval to 93% (p. 14, [17]).

Plots of observed and imputed daily values for individual patients displayed agreement in location and trend over time. The proportions of imputed values across the models that were originally negative and subsequently truncated ranged from 0.01 to 0.039. The proportion of zero values in the observed data was 0.115, and the corresponding proportion after imputation ranged from 0.099 to 0.101. Figure 2 displays the observed values and the first five imputed daily values from model MVNMI3 for three patients treated with the 60-mg dose who have common missing data patterns. Patients '239' and '270' were selected for display because they had a pronounced tendency to repeatedly report rounded times (e.g., $\sqrt{60}$ and $\sqrt{120}$), which the normal-based imputation models cannot accurately reproduce. Aside from this common situation, the imputed values appeared in good visual agreement with the observed values. Boxplots in Figure 3 summarize all of the imputed values for the three patients. They display more clearly the distributions of the missing values implied by the imputation model.

The SEs from the MVN imputation models were smaller than those produced by the other methods. Most of the difference between the MVN-based SEs and those from the MLLM and LOCF methods is due to the smaller residual SDs estimated in the primary analysis model, which are displayed in the final column of Table III. It is not apparent why the MVN-based imputations yielded less residual variation in

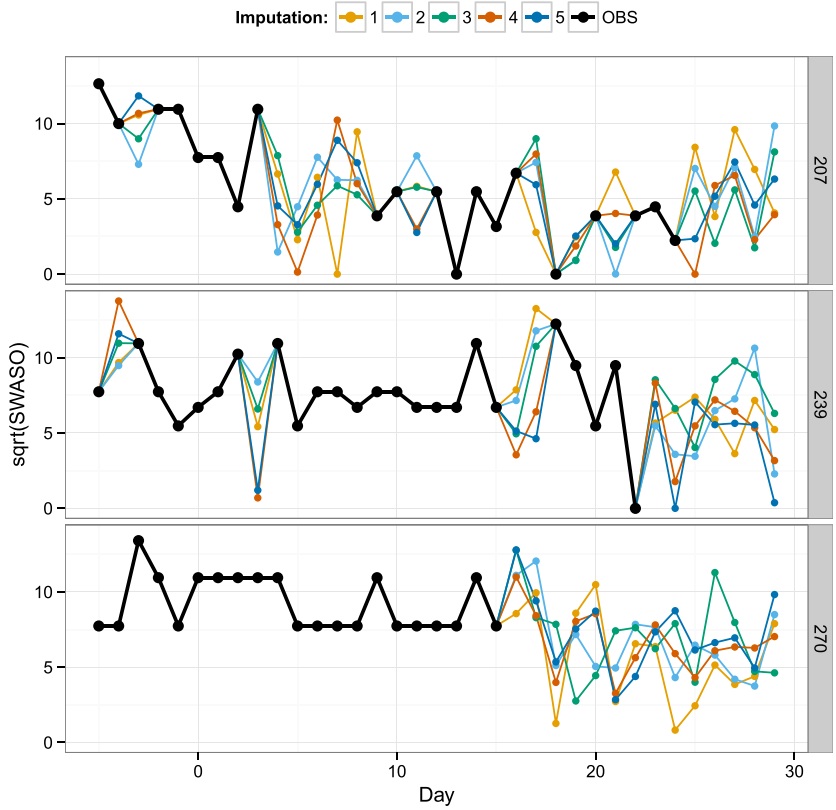

Figure 2. Longitudinal plot of $\sqrt{SWASO}$ for three patients treated with the 60-mg dose. The observed data are displayed using bold black lines and points. The first five sets of imputed values from model MVNMI3 are displayed using smaller, lighter lines and symbols. Software to produce this graphic is included in the Appendix. SWASO, subjective time awake after initial sleep onset.

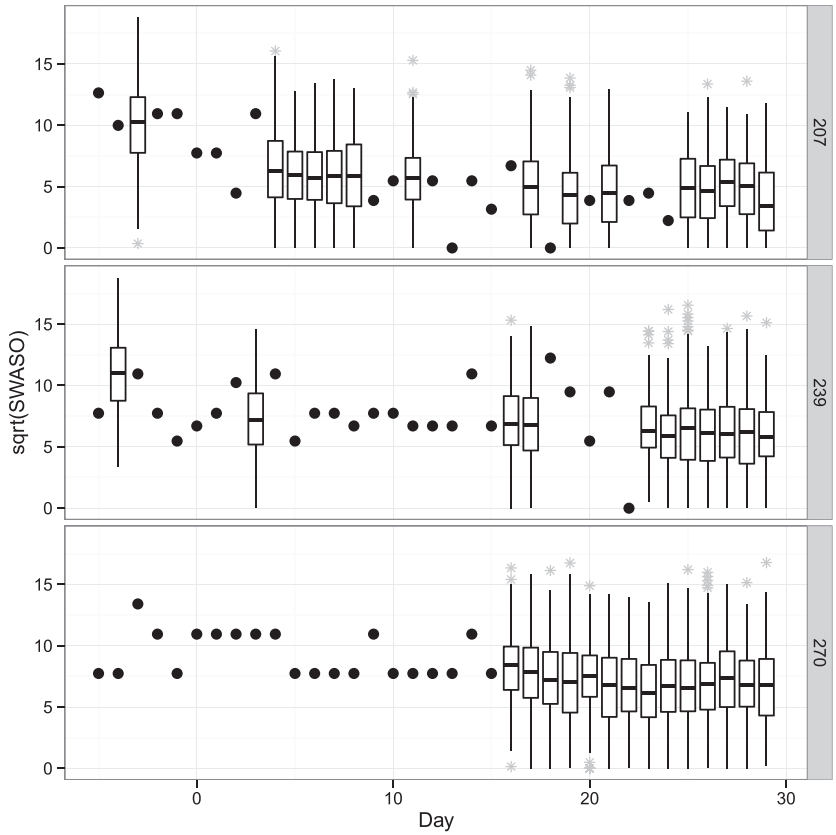

Figure 3. Boxplots summarizing all of the imputed values from model MVNMI3 for the three patients treated with the 60-mg dose in Figure 2. The observed values are represented by the larger black dots. The patient ID's are displayed in the right panels. Software to produce this graphic is included in the Appendix. SWASO, subjective time awake after initial sleep onset.

the weekly averaged values. The larger SEs for the AC method are due primarily to its smaller sample sizes. The differences in the estimates and SEs did not change the substantive conclusions of this trial, but in a trial with treatment differences near boundaries for statistical significance, changes of the magnitude observed would likely yield p-values below and above the boundary.

6. Pattern mixture models and MNAR analyses

6.1. Comparing data distributions from different missing data patterns

Most of the weekly missing data are monotone because of dropouts. Data from weeks common to different missing data patterns were compared with check potential dependence of dropout status on observed efficacy data. In Figure 4, boxplots of the weekly SWASO values from the completers are compared with corresponding boxplots of patients who dropped out after 1, 2, or 3 weeks, respectively. The weekly averaged SWASO values computed per protocol are displayed. The observed data from the different dropout patterns are very similar. Other weekly endpoint plots within dose groups were also assessed and displayed close agreement. The observed responses up to any visit do not predict who will subsequently drop out or complete the study. Because the observed data cannot unequivocally demonstrate MAR or related conditions, sensitivity analyses are described in the next section under MNAR models.

6.2. Multiple imputation -based assessment of missing not at random pattern mixture models

The pattern mixture models [18, 19] in this section are constructed from the MVN model for complete data, MVNMI3, which has different means for each study day within dose group, and a common equicorrelated variance–covariance matrix. The posterior means for the within (σ^2) and between (ψ^2) patient variances from the fitted model are $\sigma^2 = 8.9$ and $\psi^2 = 8$. An MNAR model similar to the one in Guisti

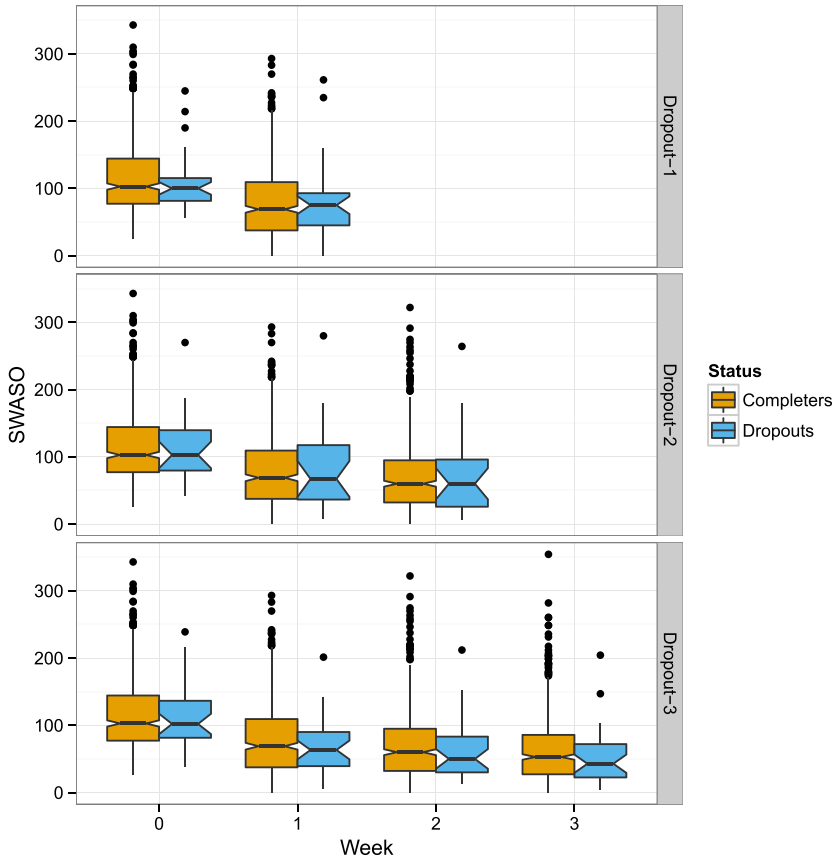

Figure 4. Response for patients with monotone missing (dropout) patterns compared with completers for weekly averaged subjective time awake after initial sleep onset (SWASO). The number of completers is 514. The numbers of dropouts at weeks 2, 3, and 4 are (33,23,33). Software to produce this graphic is included in the Appendix.

and Little [20] is fit with the data distribution given by (2), except that each missing SWASO value is increased (worsened) by the amount $c * \sigma$. If Y_{ij} is missing with imputed values $Y_{ij,imp}$, $imp = 1, \ldots, 100$, the imputed values under the MNAR model are $Y_{ij,imp} + c\sigma_{imp}$, where σ_{imp} are drawn from the posterior distribution of σ as part of the generation of the imputations. Following the recommendation in Guisti and Little [20], low, medium, and high values of the sensitivity multiplier are $c = 0.8, 1.2, 1.6$. A second MNAR model was fit using the same approach, but missing data following dropout were assigned a larger offset derived from the total variance observed in the SWASO variable after accounting for dose group and baseline characteristics, $c\sqrt{\sigma^2 + \psi^2}$. The offset for intermittent missing values remained $c * \sigma$.

Results for the models with differing sensitivity parameters are in Table IV. The mean difference for the 15-mg dose versus placebo decreased, while the mean difference for the 60-mg dose increased. The changes do not substantively impact the interpretation of the results, but the largest changes for the 15-mg dose were >1.5 SE. The change in the 15-mg dose was not anticipated because this group did not have more missing data. It occurred because there were more imputations in the upper tail of the distribution for this group. The addition of the offsets combined with the back transformation from the square root scale produced a more skewed distribution and the difference in the regression-based primary complete data analyses. Many of the imputations under the MNAR models appear as outliers when plotted with the observed data in plots like those in Figure 2. This is reflected in the increases in the estimated SD and SE in Table IV. Similar to the results from different models under the MAR assumption in Section 5, there was no high sensitivity of the results to the different MNAR models, but the combination of changes in the estimates and increased SEs could affect the interpretation of a study with less robust treatment differences.

The two MNAR models described here are not dependent on the treatment group. MNAR models that use different imputation models for different treatment groups [6, 19, 21, 22] were also explored by modifying the second MNAR model. The offsets were only applied to patients who dropout from the active treatment groups, with no offsets applied to intermittent missing values. The results for these

Table IV. Estimates and standard errors for the week 4 subjective time awake after initial sleep onset endpoint from different missing not at random models.

	15 mg vs. PBO			60 mg vs. PBO			Pooled
Method	Est	SE	%Mis Info	Est	SE	%Mis Info	Res SD
MVNMI3-day	−5.2	5.16	12	−26.17	5.07	9	39.3
AllLow-day	−2.39	5.74	16	−27.48	5.63	13	42.7
AllMed-day	−0.58	6.23	17	−27.9	6.12	14	46.1
AllHigh-day	1.49	6.86	17	−28.16	6.75	14	50.8
DropLow-day	−1.6	5.99	16	−26.92	5.89	14	44.5
DropMed-day	0.77	6.78	16	−26.94	6.68	14	50.4
DropHigh-day	3.49	7.83	15	−26.75	7.74	13	58.6
DiffLow-day	5.48	5.77	12	−16.22	5.69	11	43.7
DiffMed-day	12.28	6.51	11	−9.64	6.45	10	49.8
DiffHigh-day	20.03	7.6	9	−1.99	7.55	9	58.6

The *All* in the method denotes the same offset for all missing measurements, *Drop* denotes higher offsets following dropout, and *Diff* denotes higher offsets applied differentially to the active treatment groups. Low, Med, and High are offset multiples of 0.8, 1.2, and 1.6.
placebo; SE, standard error; SD, standard deviation.

models are in the lower portion of Table IV. As anticipated, these models substantially discount the results from the active treatment groups. An offset with $c > 1$ eliminates the clinical and statistical significance of even the large effect observed in the high-dose group. These sensitivity results are dubious, however, as they predict a substantial increase in SWASO for the lowest-dose group and a small decrease for the high-dose group. There is no apparent mechanism to explain a large differential in potential responses for dropouts in the active groups, and the results are in sharp contrast with the observed effects that are consistent with dose–response across numerous measures of sleep and the previously observed effects of related compounds. The effect of the high dose versus placebo on square root SWASO is approximately $0.4\sqrt{\sigma^2 + \psi^2}$, so even the 'low'-sensitivity setting implies a much larger shift than would occur under jump-to-placebo models [6, 21, 22]. An overall summary of the MNAR results could be obtained by specifying a distribution for the offset parameter c, which would assign higher probabilities to smaller values, and then applying the methods in [23].

7. Conclusions

Missing data rates for the weekly endpoints of approximately 15–25% at week 4 are within expectations based on past experience with trials of similar duration. There is evidence that dropout might be related to dose, but less so for intermittent missing values. The higher number of patients reporting adverse events on the highest dose support this conjecture, but the number of dropouts spread across five treatment groups is too low for definitive conclusions. No simple model using a small number of measurements was found that could predict which measurements would be missing.

The estimated proportions of missing information computed from the multiple imputations in Table III show that recovery of some of the missing information is possible because of the correlation between the numerous diary and the baseline values. The differences in estimates and SEs for the treatment effects from the different models under the MAR assumption and the MNAR models without differential behavior by treatment group were not large enough to change the substantive interpretation of the results. The differences would be large enough, however, to create ambiguity in the results from a trial with smaller treatment effects that achieved borderline statistical significance. Models under MNAR with differential behavior by treatment group attenuated the effects for all active treatment groups when dropouts have increased SWASO of approximately one SD. The SWASO values for dropouts implied by a shift of one SD yield patients with response patterns that were not observed in any patients, including those treated with placebo, so the relevance of such models to any estimate is dubious.

Multivariate normal models were utilized after data transformation to form the multiply-imputed missing SWASO values. With some data-driven adjustments to these models, the mean, variance, and correlation structure of the data could be represented. The clear preference for reporting rounded times (e.g., 30 min and 1 h) and the tendency of patients to repeat the same values for several consecutive days could not be easily reproduced with the normal-based models. Some hot deck-type approaches with sam-

pling of observed SWASO values were explored, such as predictive mean matching [4, 24], but the results are not presented here because they were unstable, depending on methodological features such as the order the variables were imputed. Alternative approaches involving weights based on estimates of the missingness probabilities were not developed here [25, 26]. Such models would also be difficult to specify because of the complex multivariate nature of the data and the added complexities that arise because of the dependence between missingness and the scheduling of weekly clinic visits.

With current desktop computing and general-purpose statistical software, it is feasible to account for missing daily diaries in aggregated endpoints. The most challenging aspect is the specification of models that can adequately represent the missing data. This problem becomes more difficult when pre-specification of the models is required for confirmatory trials. The approach used to assess model sensitivity fitted several models with flexible mean functions and different variance–covariance structures. MVN models, however, are unlikely to reproduce some of the features present in subjectively reported diary data.

R programs to create graphical displays like those in Figures 2–4 are included in the the Supporting Information [27, 28].

Acknowledgements

The authors thank the reviewers for many helpful suggestions. This project was partially supported by award number K01MH087219 from the National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

References

1. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press: Washington, DC, 2010.
2. Little R, D'Agostino R, Cohen M, Dickersin K, Emerson S, Farrar J, Frangakis C, Hogan J, Molenberghs G, Murphy S, Neaton J, Rotnitzky A, Scharfstein D, Shih W, Siegel J, Stern H. Special report: the prevention and treatment of missing data in clinical trials. *New England Journal of Medicine* 2012; **367**(14):1355–1360.
3. Rubin D. Inference and missing data. *Biometrika* 1976; **63**(3):581–592.
4. Rubin D. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
5. Little R, Kang S. Intention-to-treat analysis with treatment discontinuation and missing data in clinical trials. *Statistics in Medicine* 2015; **34**(16):2381–2390.
6. Ratitch B, O'Kelly M, Tosiello R. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics* 2013; **12**(6):337–347.
7. Zhang J, Rubin DB. Estimation of causal effects via principal stratification when some outcomes are truncated by death. *Journal of Educational and Behavioral Statistics* 2003; **28**(4):353–368.
8. Merck Sharp & Dohme Corp. Safety and efficacy study of suvorexant in participants with primary insomnia (mk-4305-028). ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US), 2015. Available from: https://clinicaltrials.gov/ct2/show/NCT01097616NLMIdentifier:NCT01097616 2000-[Access on 5 March 2015].
9. Farrar J, Young J, LaMoreaux L, Werth J, Poole M. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* 2001; **94**:149–158.
10. Templ M, Alfons A, Kowarik A, Prantner B. VIM: Visualization and imputation of missing values, 2013. Available from: http://CRAN.R-project.org/package=VIM, r package version 4.0.0 [Accessed on May 2015].
11. Iliadis A, Macheras P. *Modeling in Biopharmaceutics, and Pharmacodynamics: Homogeneous and Heterogeneous Approaches*. Springer: New York, 2006.
12. Thomas N, Sweeney K, Somayaji V. Meta-analysis of clinical dose–response in a large drug development portfolio. *Statistics in Biopharmaceutical Research* 2014; **6**(4):302–317.
13. Food, Administration D. Suvorexant advisory committee meeting briefing document, 2013. Available from: http://www.fda.gov/downloads/advisorycommittees/committeesmeetingmaterials/drugs/peripheralandcentralnervoussystemdrugsadvisorycommittee/ucm352970.pdf [Accessed on May 2015].
14. Schafer J, Yucel R. Computational strategies for multivariate linear mixed effects models with missing values. *Journal of Computational and Graphical Statistics* 2002; **11**:437–457.
15. Zhao J, Schafer J. pan: multiple imputation for multivariate panel or clustered data, 2013. R package version 0.9.
16. Stan Development Team. *Stan: A C++ Library for Probability and Sampling*, 2013. Available from: http://mc-stan.org/, 1.3 edn. [Accessed on May 2015].
17. Cochran W. *Sampling Techniques* (3rd edn). Wiley: New York, 1977.
18. Little R. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistial Association* 1993; **88**(421):125–134.
19. Little R, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics* 2006; **52**(4):1324–1333.
20. Giusti C, Little R. An analysis of nonignorable nonresponse to income in a survey with a rotating panel design. *Journal of Official Statistics* 2011; **27**(2):211–229.

21. Carpenter J, Roger J, Kenward M. Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics* 2013; **23**(4): 1352–1371.
22. Ayele B, Lipkovich I, Molenberghs G, Mallinckrodt C. A multiple-imputation-based approach to sensitivity analyses and effectiveness assessments in longitudinal clinical trials. *Journal of Biopharmaceutical Statistics* 2014; **24**(2):211–228.
23. Siddique J, Harel O, Crespi C. Addressing missing data mechanism uncertainty using multiple-model multiple imputation: application to a longitudinal clinical trial. *The Annals of Applied Statistics* 2012; **6**(4):1814–1837.
24. Little R. Missing data adjustments in large surveys (with discussion). *Journal of Business Economics and Statistics* 1988; **6**:287–301.
25. Robins J, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of American Statistical Association* 1995; **90**:122–129.
26. Preisser JS, Lohman KK, Rathouz P. Performance of weighted estimating equations for longitudinal winary data with drop-outs missing at random. *Statistics in Medicine* 2002; **21**:3035–3054.
27. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2011.
28. Wickam H. *ggplot2: Elegant Graphics for Data Analysis*. Springer: New York, 2009. Available from: http://had.co.nz/ggplot2/book [Accessed on May 2015].

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.